Speech Production and Perception

Volume 6

Speech production and perception: Learning and memory

Edited by
Susanne Fuchs
Joanne Cleland
Amélie Rochet-Capellan

PETER LANG

Learning and memory processes are basic features of human existence. They allow us to (un)consciously adapt to changes in our social and physical environment in a variety of ways and may have been a precursor for survival in human evolution. Through several reviews and original work the book focuses on three key topics that enhanced our understanding of the topic in the last twenty years: first, the role of real-time auditory feedback in learning, second, the role of motor aspects for learning and memory, and third, representations in memory and the role of sleep on memory consolidation.

Susanne Fuchs is a phonetician and speech motor control researcher at the Leibniz-Zentrum Allgemeine Sprachwissenschaft in Berlin. She investigates the biological grounding of spoken language, iconicity and its origin in sensorimotor properties as well as the effect of motion on cognitive processes.

Joanne Cleland is a researcher and Speech and Language Therapist at the University of Strathclyde in Glasgow. She studies clinical phonetics and articulatory analysis of Speech Sound Disorders in children.

Amélie Rochet-Capellan is a researcher at French CNRS. In the framework of embodied cognition, she studies the links between orofacial and limb sensorimotor control and language in typical speakers and speakers with intellectual deficiencies.

www.peterlang.com

Speech production and perception: Learning and memory

SPEECH PRODUCTION AND PERCEPTION

Edited by Susanne Fuchs and Pascal Perrier

VOLUME 6

Notes on the quality assurance and peer review of this publication:
Prior to publication, the quality of the work published in this series is double blind reviewed by external referees appointed by the editorship. The referee is not aware of the author's name when performing the review; the referees' names are not disclosed.

Susanne Fuchs / Joanne Cleland /
Amélie Rochet-Capellan (eds.)

Speech production and perception: Learning and memory

PETER LANG

Bibliographic Information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the internet at http://dnb.d-nb.de.

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

An electronic version of this book is freely available, thanks to the support of libraries working with Knowledge Unlatched. KU is a collaborative initiative designed to make high quality books Open Access for the public good. More information about the initiative and links to the Open Access version can be found at www.knowledgeunlatched.org

This work was supported by a grant from the French-German University (UFA) Saarbrücken and by a grant from the ANR-DFG to the Salammbo project (FU791/8-1).

Printed by CPI books GmbH, Leck.

ISSN 2191-8651
ISBN 978-3-631-72691-4 (Print)
E-ISBN 978-3-631-79786-0 (E-PDF)
E-ISBN 978-3-631-79787-7 (EPUB)
E-ISBN 978-3-631-79788-4 (MOBI)
DOI 10.3726/b15982

Peter Lang GmbH Internationaler Verlag der Wissenschaften Berlin

Peter Lang – Berlin · Bern · Bruxelles · New York · Oxford · Warszawa · Wien

This publication has been peer reviewed.

www.peterlang.com

Contents

List of Contributors

Louis ten Bosch
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

Jana Brunner
Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Germany

Tiphaine Caudrelier
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Joanne Cleland
University of Strathclyde, Glasgow, United Kingdom

Jonathan Delafield-Butt
University of Strathclyde, Glasgow, United Kingdom

Marion Dohen
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Mirjam Ernestus
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands;
Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

Susanne Fuchs
Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS), Berlin, Germany,

Pamela Fuhrmeister
Department of Speech, Language, and Hearing Sciences, University of Connecticut, United States

Phil Hoole
Institut für Phonetik und Sprachverarbeitung, Ludwig-Maximilians-Universität München, Germany

Eugen Klein
Institut für Deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Germany

Louise McKeever
University of Strathclyde, Glasgow, United Kingdom

Lisa Morano
Centre for Language Studies, Radboud University, Nijmegen, the Netherlands

Amélie Rochet-Capellan
Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

James M. Scobbie
Queen Margaret University, United Kingdom

Susanne Fuchs, Joanne Cleland, Amélie Rochet-Capellan

Preface

Learning and memory processes are basic features of human existence and are also reported in other species (Clayton & Dickinson, 1998). They allow us to (un)consciously adapt to changes in our social and physical environment in a variety of ways and may have been a precursor for survival in human evolution. Although learning and memory processes have been at the center of psychological, linguistic and philosophical research, and discussed from the earliest existence of these disciplines, there is still much to learn.

In the domain of speech production and perception, the focus of the present book, there has been a renaissance in terms of the subject's matter. Three major topics will be addressed in this book through reviewing previous work; discovering research gaps and summarizing potential future research directions; or with original work. These three major topics are: 1.) the role of real-time sensory (auditory) feedback for learning, 2.) the role of motor aspects for learning and memory (including recent technological developments which may support learning in people with specific needs) and 3.) representations in memory and the role of sleep on memory consolidation with a specific focus on second language learning.

Regarding the first topic, computational and technological developments in recent years have made it possible to alter sensory feedback of a speaker in *real-time*. That is, a speaker's spoken language can be recorded, manipulated, and played back with such a short delay that the speaker considers it as his/her own speech. These developments made it possible to investigate the role of auditory feedback in speech production and learning and determine how and when speakers adapt to changes in auditory feedback. The first two chapters of the book focus on this topic. Tiphaine Caudrelier and Amélie Rochet-Capellan provide a review of two decades of research initiated by Houde & Jordan's (1998) pioneering study on auditory-motor learning in response to formant perturbations. The chapter starts with an overview of the impact of Houde and Jordan's work across different research fields. Then, based on 77 studies using formant perturbations,

the authors present the systems and procedures associated with this paradigm. They also provide a comprehensive review of the research topics addressed by these studies and their main results. The chapter concludes with suggestions for future research, including using sensorimotor learning to further explore the nature of speech production representations.

The second chapter presents a recent study on real-time feedback perturbation by Eugen Klein, Jana Brunner, and Phil Hoole. They work on inter- and intra-individual variability of adaptation processes during auditory perturbation of vowel formants. Specifically, the authors investigate the influence of experimental task demands – such as the alternating perturbation of the second formant and the consonantal context of the perturbed vowel – on speakers' compensatory adjustments. Examining the adaptation process with due regard to its temporal dimension, the authors show that its variability is strongly associated with speakers' exploratory behavior and cannot be exclusively ascribed to the characteristics of speakers' internal models of speech motor control.

The next three chapters deal with the particular role of motor aspects in learning and memory consolidation. Different populations are investigated. In a review of literature on speech sound errors in people with autism, Louise McKeever, Joanne Cleland and Jonathan Delafield-Butt begin to explore the underlying causes of speech production differences in people with autism. Two major theoretical accounts for the prevalence of speech sound errors are highlighted: the speech attunement framework, and deficits in speech motor control. Both theories provide explanations for how children with autism may come to have difficulty learning to produce speech which is in line with their typically developing peers. The chapter concludes by suggesting that both the speech attunement framework and the theory of impaired speech motor control may be complimentary, rather than competing, theories and suggests further empirical work to test this assertion.

In the following chapter Joanne Cleland and James Scobbie focus on learning new speech motor plans in children with speech sound disorders. They first describe the concept of categorising persistent speech sound disorder in children as a disorder characterised by erroneous motor plans. They then go on to explain how various different forms of articulatory visual biofeedback (namely, electropalatography, electromagnetic

articulography and ultrasound tongue imaging) can be used to allow children to view their articulators moving in real time and use this information to establish more accurate motor plans. A novel theoretical account of how these articulatory biofeedback techniques might lead to establishment of new motor plans is given. The chapter concludes with an illustrative case study of a child with persistent velar fronting who acquired a new motor plan for velar stops using ultrasound visual biofeedback.

The role of motor aspects is then discussed by Marion Dohen, who provides a comprehensive review of the role of manual gestures in word learning and memorizing. Typically developing children as well as children with specific needs are the focus of this review. The findings from a selection of empirical studies serves for answering general questions about potential advantages, efficiency and types of manual gestures in learning novel words. Motor aspects of manual gestures, i.e. producing an additional gesture during learning is compared with findings where manual gestures are perceived only. Finally, Marion Dohen discusses three potential explanations as to why manual gestures might enhance learning novel words.

The last two chapters are dedicated to learning new languages and how this information is consolidated in memory. Starting in the early 90s of the last century there have been several theoretical and empirical attempts to justify that it is not abstract linguistic representations, but rather episodic traces (exemplars) that are stored in memory (see Smith 2015 for review). These may include fine phonetic detail, for example detail about the speaker's voice, the communicative situation and so on. More recently, these approaches have been unified to hybrid models, since neither the one nor the other can solely account for learning a language. Lisa Morano, Louis ten Bosch, and Mirjam Ernestus follow then with their work on different mental representations stored in memory. They place themselves in the Complementary Learning Systems framework which assumes the use of abstract and exemplar mental representations in speech comprehension. Specifically, their work focuses on second language (L2) listeners' exemplar representations of words. Their particular novel finding is that L2 exemplars are faithful representations of the speech signal. Thus, unlike abstract representations, exemplars have not been altered by listeners' L1 phonological filter. The authors also found significant evidence that L2

listeners used abstract representations in an experiment that had been specifically designed to trigger exemplar effects.

In the last twenty years, evidence has been accumulated that sleep may play a major role for memory. Pamela Fuhrmeister reviews some recent work suggesting that memory consolidation during sleep is important for non-native speech sound learning. While factors that influence learning of difficult speech sound contrasts have received a lot of attention in the literature, less is known about how what happens after learning can affect consolidation and retention of newly learned phonetic information. Studies from other domains, such as motor learning, are reviewed, and these suggest that certain tasks that follow training can interfere with consolidation of new information. Hints of these effects can already be found in studies in the speech domain, and the author argues that the sleep and memory consolidation literature should inform future speech research and that future research should consider not only how speech sounds are best learned, but also how they are most optimally consolidated and retained.

Besides the intellectual merit of the authors and reviewers, this book was only possible thanks to the financial support the French-German University (UFA) in Saarbrücken. They supported the publication of the book and the international winter school on "Speech production and perception: Learning and memory" that took place 2017 in Chorin, Germany. Furthermore, parts of the presented work were supported by a grant from the ANR-DFG to the Salammbo project (FU791/8-1). Finally, we would like to acknowledge all reviewers who helped improving clarity and structure of the chapters. We are very delighted that the Crowdfunding Initiative "Knowledge Unlatched" selected our book and allowed it to be open access. Thanks to all the unknown contributors!

References

Caudrelier, T., & Rochet-Capellan, A. (2019). Changes in speech production in response to formant perturbations: An overview of two decades of research. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Clayton, N.S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. Nature 395 (6699): 272–274.

Cleland, J. & Scobbie, J. M. (2019). Acquisition of new speech motor plans via articulatory visual biofeedback. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Dohen, M. (2019). Do manual gestures help the learning of new words? A review of experimental studies. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Fuhrmeister, P. (2019) Interference in memory consolidation of non-native speech sounds. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Houde, J.F., & Jordan, M.I. (1998). Sensorimotor adaptation in speech production. *Science*, 279(5354), 1213–1216.

Klein, E., Brunner, J. & Hoole, P. (2019). Spatial and temporal variability of corrective speech movements as revealed by vowel formants during sensorimotor learning. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Louise McKeever, Joanne Cleland, Jonathan Delafield-Butt (2019). Aetiology of speech sound errors in autism. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Morano, L., ten Bosch, L. & Ernestus, M. (2019). Looking for exemplar effects: testing the comprehension and memory representations of r'duced words in Dutch learners of French. In S. Fuchs, J. Cleland, & A. Rochet-Capellan (eds.) *Speech production and perception: Learning and memory*. Frankfurt/M.: Peter Lang GmbH. Internationaler Verlag der Wissenschaften.

Smith, R. (2015). Perception of speaker-specific phonetic detail. In S. Fuchs, D. Pape, C. Petrone, P. Perrier (eds.). Individual Differences in Speech Production and Perception. Peter Lang GmbH. Internationaler Verlag der Wissenschaften, 12–38.

Tiphaine Caudrelier, Amélie Rochet-Capellan

Changes in speech production in response to formant perturbations: An overview of two decades of research

Abstract: One way to investigate speech motor learning is to create artificial adaptation situations by perturbing speakers' auditory feedback in real time. Formant perturbations were introduced by Houde and Jordan (1998), providing the first evidence that speakers adapt their pronunciation to compensate for these perturbations. Twenty years later, this chapter provides an overview of the general impact of Houde and Jordan's work in speech research and beyond, as well as a more detailed review of studies that involve formant perturbations. The impact of Houde and Jordan's work appears to be cross-disciplinary. Although mainly related to speech production and perception, it has also been cited in the limb movement and even animal research, mainly as evidence of adaptive sensorimotor control. Formant perturbations research has expanded rapidly since 2006, spreading across the world and many research teams. We identified 77 experimental studies focused on formant perturbations which we then analyzed with regard to technical and theoretical issues. This analysis showed that various apparatuses and procedures were used to address important topics of speech research. A primary interest has been in feedback and feedforward control mechanisms in speech. These mechanisms were addressed in different populations, including adults and children with typical vs. atypical development, with behavioral or neurophysiological approaches, or both. Some formant perturbations studies more specifically focused on the integration of auditory and somatosensory feedback in speech production, while others explored the interaction between speech production and perception of phonemic contrasts. Some research questioned the processes and the nature of speech representations by investigating generalization of adaptation to formant perturbations. Finally, a few studies were interested in the effect of extraneous variables such as surface effects or speakers' general cognitive abilities. Altogether, these studies provide insights into speech motor control in general and into the understanding of sensorimotor interactions in particular. The field has developed recently and may still expand in the future, as it allows us to address fundamental topics in speech research such as perception-production links or abstract vs. exemplar representations. Future

research with formant perturbations may also further connect sensorimotor adaptation to linguistic and cognitive factors and in particular to working and long-term memory.

Keywords: perturbation, real-time auditory feedback, formants, speech units, learning

1. Introduction

As an "extraordinary feat of motor control" (Kelso, Tuller, Vatikiotis-Bateson, & Fowler, 1984, p. 812), speech production is a challenging research topic, highly influenced by movement sciences (Grimme, Fuchs, Perrier, & Schöner, 2011; Maas et al., 2008). Speech motor control indeed shares numerous features with other sensorimotor systems and in particular with limb motor control. Among these features, sensorimotor adaptability of speech is of particular interest to speech science as the basis of speech rehabilitation (Maas et al., 2008), and since it is ubiquitous in daily life. Common examples include, among others, changes in the way we speak according to our interlocutor or to the surroundings, such as speaking louder when talking with someone with a hearing impairment or in a noisy environment (Garnier, Henrich, & Dubois, 2010); or spontaneously imitating our interlocutor's speech sounds (Pardo, 2006). Speech motor control also adapts throughout the lifespan to natural or accidental alterations of our sensory systems or vocal tract geometry, temporarily or more permanently (Jones & Munhall, 2003; Lane et al., 2007). These adaptations allow maintenance of some level of intelligibility despite vocal tract growth, hearing loss, orofacial surgery, or when wearing a dental apparatus, losing teeth, speaking while eating etc. Being essential to speech production, sensorimotor adaptation of speech is the topic of numerous studies. For the purpose of this chapter, we will focus on studies that involved specific perturbation of formants. Formants are frequencies corresponding to peaks of acoustical energy, the relative values of which characterize vowels. Research in this field, and especially Houde and Jordan's work, was inspired by the study of visuomotor adaptation in the limb movement literature (Houde & Jordan, 1998).

Pioneering work on adaptation of different visuomotor activities appeared at the end of the 19th century (Held, 1965; Stratton, 1897). This

work introduced a now common approach to assessing visuomotor adaptation that consists of investigating changes in movement in response to a systematic distortion of visual feedback, such as prism adaptation. As an illustration, Stratton (1897) reported his own and extreme everyday life experience while wearing an apparatus for eight days that reversed the retinal image upside down and left to right. On the first day, "the entire scene appeared upside down". He felt nauseous. His movements were "laborious", "embarrassed", "inappropriate" (p. 344), required a lot of attention and were "extremely fatiguing" (p. 344). By the start of the third day things were much better, with no sign of "nervous distress" (p. 349). At the end of the fourth day, he "preferred to keep the glasses on rather than sit blindfolded" (p. 351/352). When the apparatus was removed on day eight, it took him some time to go back to normal feelings and motions.

Later work on visuomotor adaptation focused on more specific activities, less dramatic and more local and short-term changes, with a focus on reaching movements performed with rotations of the visual field. In this context, it has been repetitively demonstrated that when movements are achieved while the visual field is shifted by a specific angle (α), participants first miss the target by the same angle α. However, with repetition, they progressively learn to adapt their movements to the new feedback and reach the target accurately again. When they return to normal vision, after-effects and transfer effects are observed: participants miss the training target (*after-effects*) and/or a new target (*transfer*) by an angle more or less close to $-\alpha$. These effects vary as a function of the angular distance between the training and the testing targets (Krakauer, Pine, Ghilardi, & Ghez, 2000; Shadmehr & Mussa-Ivaldi, 1994). Sensorimotor adaptation has been attributed early on to feedforward control (i.e. predictive control based on learnt sensorimotor mappings) in contrast to forward closed-loop control (i.e. online processing of sensory inputs), visible in correction to unexpected perturbations (Golfinopoulos, Tourville, & Guenther, 2010; Houde & Chang, 2015). These notions are defined later in this chapter.

Twenty years ago, Houde and Jordan (1998) introduced an analogous procedure of visuomotor rotation adaptation to question feedforward control in speech, which used real-time alterations of formant frequencies in vowels. By altering the frequencies of the first and/or second formants (F1 and F2 respectively) it is possible to make a vowel sound like another

vowel. For example, by decreasing F1 and increasing F2, the vowel /ɛ/ would sound closer to the vowel /ɪ/, as illustrated in Figure 1. This alteration *displaces* the auditory feedback, in the same way as prism vision displaces the visual position of the target. For example, the speaker says "head", speaking into a microphone and wearing headphones (Figure 1.A). The signal is processed in real time so that F1 and F2 formants are moved towards "hid" (Figure 1.B), and played-back into the headphones. The consequence for the speaker is a discrepancy between the auditory target expected from the planned movements ("head") and the auditory target they actually got (~"hid"). In other words, similar to visuomotor adaptation, the speaker first *misses* the auditory target (Figure 1.C, "Training start"). With practice – repetition of shifted utterance(s) with the same perturbation – the speaker adapts to the perturbation (Figure 1.C, "Training end"): To *reach* the auditory target "head" again in the presence of the perturbation, they produce formants in the opposite direction to the perturbation. In our example, this corresponds to the production of an utterance closer to "had". When the feedback is returned to normal or masked with a noise, for the same vs. different utterance(s) than the training one(s), after-effects vs. transfer effects are observed (Figure 1.C, column "After-effect" and "Transfer"). This suggests that the compensation is not only an online feedback control change but also affects auditory-motor mappings supporting feedforward control, in a more or less utterance or segment-specific way. The procedure was later adapted to address feedback control by investigating online compensation to unexpected perturbations (Purcell & Munhall, 2006b).

Adaptation to formant perturbations has been investigated per se, or used as a paradigm to address more general issues in speech science. The current chapter reviews research in formant perturbations by analyzing Houde and Jordan's seminal study (Houde & Jordan, 1998, 2002) and the scientific literature that has referred to it. Using this approach (detailed in the first section of the chapter) we can see the cross-disciplinary impact of Houde and Jordan's work and in particular, identify the main topics of the scientific literature that have cited this work (reported in the second part of the chapter). Among the collected papers, only a subsection corresponded to empirical studies involving formant perturbations. Based on the analysis of these studies, including review of their reference lists, the latter parts

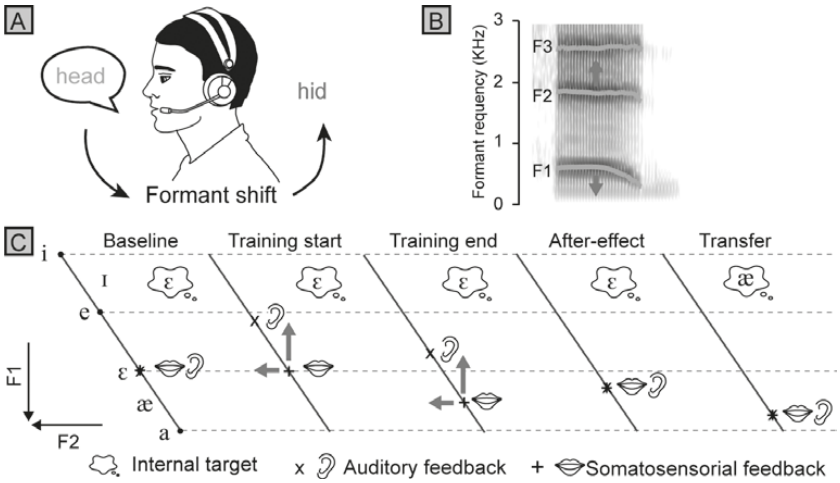

Figure 1: The auditory prism adaptation. (A) The speaker speaks into a microphone; his feedback is altered such as when he produces "head" he is hearing a signal closer to "hid"; (B) To do so, F1 and F2 are changed in real time; (C) Before the introduction of the perturbation (Baseline) the auditory feedback is consistent with the target. The first exposure to the perturbation (Training start) induces a discrepancy (or an error) between the auditory feedback and the planed target. With repetitive exposure to the perturbation, the talker changes his production to compensate for the perturbation (Training end). When the perturbation is removed after-effects and/or transfer effects are observed.

of the chapter provide: (1) a description of the main apparatuses and paradigms used in formant perturbations studies; (2) an overview of the research topics addressed using these perturbations and the main reported results; and (3) some perspectives for future research.

2. Paper collection and analysis

As we were interested in the impact of Houde and Jordan's work and also wanted to provide an analytical review of formant perturbations studies, we first analyzed the published work that referred to Houde and Jordan (1998 and/or 2002) from 1999 to 2018 (last update on July 6th 2018). This was performed using the "Cited by" function in Google Scholar. We choose this approach rather than keyword research, as we wanted to collect various sorts of publications, and because it appeared to be the

Table 1: Number of references in each category of the first level of selection (see text for details)

Formant shift		No formant shift		Not in English	Error ref.	Total
Rejected	Kept	Rejected	Kept			
26	72 (+ 2, Houde & Jordan 1998 and 2002)	140	287	35	22	584

most systematic way to collect publications in the field. To compensate for potential errors and omissions by Google Scholar, the results were then analyzed very closely.

An analysis by year of Google Scholar output resulted in a total of 584 references (including the two papers by Houde and Jordan, see Table 1). As a first step, we excluded documents that were not written in English or that corresponded to reference errors (57 in total, see Table 1). Among the 527 remaining references, we distinguished between those without vs. with an empirical study that included formant perturbations. In the former category (n=427, without formant perturbation), we kept only journal papers for a thematic analysis of Houde and Jordan's broad impact (n=287). In the latter category (n=100, with formant perturbations), we first kept all the documents except PhD or Master theses, posters or abstracts to conferences (74 references kept, 26 rejected). Note that there were 11 PhD theses; most of them were associated with journal publications. For consistency in criteria, we did not include Frank (2011)'s PhD thesis, even though it is often cited by studies investigating linguistic effects on formants adaptation. Its results were never published in peer-reviewed papers.

Three more papers were added that included formant perturbations. One paper that did not cite Houde and Jordan was found in the reference list of the selected papers (Niziolek & Guenther, 2013); and two papers in course of publication at the time of writing that we were aware of (Caudrelier, Perrier, Schwartz, & Rochet-Capellan, 2018; Klein, Brunner, & Hoole, in this book). The general characteristics of the documents including formants perturbations are described in Table 2. Technical

Table 2: Number of papers considered for the analysis of formant perturbations according to source and type. Houde & Jordan (1998, 2002) are included.

	Journal papers	Proceedings papers	Reports/ chapters	Total
Google Scholar	55	17	2	74
Other sources	1	1	1	3

papers as well as papers investigating compensation to unexpected formant perturbations were included.

The full list of analyzed papers related to formant perturbation is available in Table 4, with their main related research topic indicated. As the paper collection is based mainly on the "cited by" function of Google Scholar some papers may be missing despite our careful attention. However, we believe our analysis provides an accurate picture of the field at the time it was run.

3. Overall impact of Houde and Jordan's seminal work

The overall impact of Houde and Jordan (1998, 2002) is illustrated in Figure 2. We distinguished seven broad categories of research: (1) formant perturbations studies (n=77); (2) studies that investigated speech compensation and/or adaptation to other auditory perturbations or equivalent situations (n=91) or (3) to an alteration of the vocal tract (n=16); (4) empirical or theoretical papers on speech production (n=61) or (5) on speech perception (n=46); (6) studies involving non-speech actions (n=25); and (7) experimental or theoretical papers involving animals (n=43). Five papers were not considered, as they were difficult to classify in these categories. We first analyzed the journal papers that did not empirically test formant perturbations. As described above, this involved 286 articles. Broad research topics were identified mainly from abstract reading. A subset of papers was selected and read in more detail to illustrate the different topics. The articles on formant perturbations will be reviewed in detail in the next sections. We will now briefly overview the research topics in the six other categories. References in the following section are illustrative.

3.1. Compensation/adaptation of speech production to various auditory perturbations

Speech compensation and adaptation were investigated prior to the development of formant perturbation studies and used various methods. These methods continued to be used in some of the later work that cited Houde and Jordan. About half of the papers in this first category investigated speech modifications in reaction to either an unexpected or a predictable modification of F0 in different populations and conditions. A number of papers in this topic were published by Jones et al. (Jones & Munhall, 2000); Larson et al. (Burnett & Larson, 2002); or Hanjun et al. (Li et al., 2016). The other half of the studies investigated speech modifications in reaction to other types of auditory perturbations such as delayed auditory feedback (Chon, Kraft, Zhang, Loucks, & Ambrose, 2013); changes in intensity or noise level (Maas, Mailend, & Guenther, 2015); hearing loss (Palethorpe, Watson, & Barker, 2003); real or simulated use of cochlear implants (Casserly, 2015; Lane et al., 2007); or replacement of the auditory feedback by a stranger's voice (Hubl et al., 2014). Other work modified consonant features such as frication (Shiller, Sato, Gracco, & Baum, 2009) or voicing (Mitsuya, MacDonald, & Munhall, 2014). Self-regulation in adaptation to formant perturbations was also linked with interpersonal auditory-motor regularizations in speech such as phonetic convergence (Pardo, 2006).

3.2. Compensation/adaptation of speech production to perturbations of the vocal tract dynamics or geometry

Research on compensation and adaptation to perturbations affecting the somatosensory feedback is another field closely connected to adaptation to formant perturbations. Houde and Jordan's work was thus cited by studies involving an alteration of the vocal tract geometry or dynamics. This includes dental prostheses (Jones & Munhall, 2003); lip tubes in children and adults (Ménard, Perrier, & Aubin, 2016); false palates (Thibeault, Ménard, Baum, Richard, & McFarland, 2011); mechanical forces applied to the jaw with a robot (Tremblay, Shiller, & Ostry, 2003); or more permanent changes such as those induced by oropharyngeal cancer treatments (de Bruijn et al. 2012).

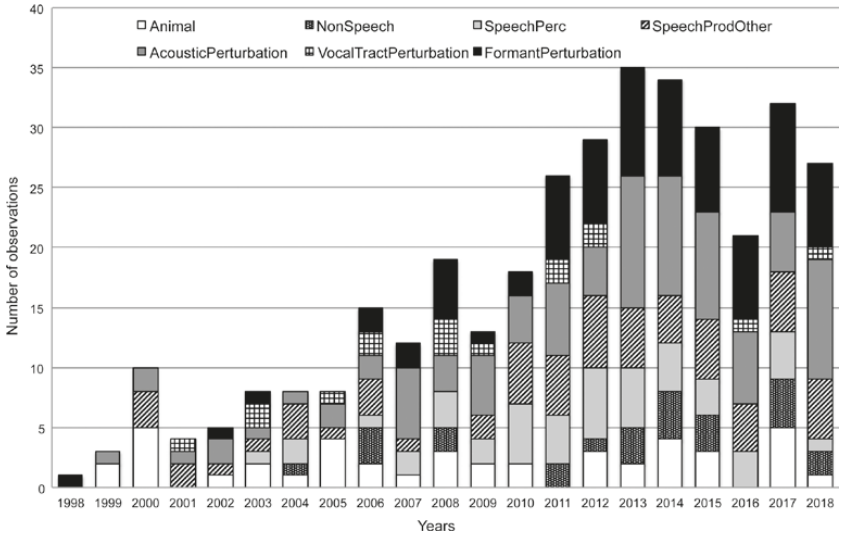

Figure 2: Overall impact: number of analyzed papers by year and categories.

3.3. Empirical or theoretical papers on speech production

Houde and Jordan's work is cited by empirical and theoretical research on speech production. For example, adaptation to formant perturbations is mentioned by studies providing further evidence of the role of auditory feedback in speech motor control, such as work linking auditory acuity to the production of speech contrasts (Perkell et al., 2004); auditory perceptual learning with improvement in production (Shiller, Rvachew, & Brosseau-Lapré, 2010); comparing overt and covert speech (Brumberg et al., 2016) or analyzing the neurophysiological activities of the auditory cortex during speech production (Curio, Neuloh, Numminen, Jousmäki, & Hari, 2000). Adaptation to formant perturbations provides support for neurocomputational models of speech production such as the Directions Into Velocity of Articulators model (DIVA, Golfinopoulos et al., 2010) or the State Feedback Control model (SFC, Houde & Chang, 2015), both models assuming a feedback and a feedforward control mechanism. Further information about these control mechanisms will be provided in the section describing formant perturbation studies related to this topic.

3.4. Empirical or theoretical papers on speech perception

Adaptation to formant perturbations is also taken as evidence of sensorimotor integration in speech. As such, it is relevant for papers probing or discussing the role of the motor system in speech perception (Sato, Troille, Ménard, Cathiard, & Gracco, 2013) or in theoretical papers related to the dual-stream model of language processing. Basically, this model proposes a cortical ventral stream that maps speech sounds to concepts, and a dorsal stream for auditory-motor mapping. Adaptation to formant perturbations is then cited as an evidence that a dorsal auditory-motor integration path is still functional in adulthood (Hickok & Poeppel, 2004).

3.5. Non-speech movement studies

Various non-speech studies cited Houde and Jordan's work to illustrate sensorimotor adaptation in humans. These studies focused on activities involving auditory feedback such as piano playing (Pfordresher & Palmer, 2006); or the learning of artificial auditory-arm movement maps (van Vugt & Ostry, 2018). Some papers were also interested in other kinds of sensorimotor adaptations such as swallowing (Wong, Domangue, Fels, & Ludlow, 2017), or visuomotor adaptation of limb movements (Wei et al., 2014). Note that as formant perturbations studies were inspired by visuomotor adaptation, they often referred to limb movement literature. The converse seems not necessarily true as our research suggests that few works on limb adaptation have cited Houde and Jordan's work. This result should be taken cautiously as limb movement research could cite other studies using formant perturbations to illustrate the adaptability of speech motor control, and we only collected papers that reference Houde and Jordan using "cited by" functionality of Google Scholar.

3.6. Animal studies

Finally, animal studies have early, and regularly, cited Houde and Jordan's work (Figure 2), with a main focus on the role of auditory feedback in action control. Over half of these papers were dedicated to birdsong and published by Brainard et al. and/or Doupe et al. and/or Sober et al. Many of these papers include studies of birdsong production or learning using auditory perturbations with behavioral and/or neurophysiologic recordings, as

well as interspecies comparative reviews about the processing of auditory feedback of self-produced sounds (Brainard & Doupe, 2000; Doupe & Kuhl, 1999; Sober & Brainard, 2009). Analogous works were done in bats (Smotherman, Zhang, & Metzner, 2003) and primates (Eliades & Miller, 2017).

To summarize, this non-exhaustive analysis of the overall impact of Houde and Jordan's seminal work suggests that it is (as expected) cited by papers investigating speech compensation and adaptation to other types of sensory perturbations. Most of the scientific questions in this first set of papers overlap with the research topics we will review based on the more detailed analysis of formant perturbations studies in the related section of this chapter. In a broad context, adaptation to formant perturbations is often interpreted as evidence for sensorimotor integration and sensorimotor plasticity in speech production and perception. It is cited to illustrate auditory feedback and feedforward control mechanisms in speech production, as explained below, and taken as an example of such mechanisms (and their plasticity) in studies investigating animal vocalizations, singing, music playing, but also inter-personal convergence or coordination of movements.

Note that more research topics related to formant perturbation studies may be found by including "2nd order" connections to Houde and Jordan's work (i.e. references that cite any of the studies on formant perturbations).

4. Methods in formant perturbation studies

In this section, we provide an overview of the apparatuses used to apply real-time formant perturbation and a description of the main procedures identified in the collected papers.

4.1. Real-time formant perturbation

The systems used to shift formants in the collected papers are summarized in Table 3. Paper details can be found in Table 4. With regards to formant perturbation, it is important to emphasize that in order to preserve the best quality of self-perception, the real-time modification of formants in speakers' auditory feedback should meet some requirements, specifically:

(1) The signal should be processed and played back fast enough for the speaker not to perceive any delay (less than 30ms, see Yates, 1963). Specific digital signal processing boards (DSP), including systems from the music industry were used, especially in earlier work. Nowadays, this can be achieved at a software level, on a PC with appropriate sound card and software to analyze and change formants. For the same code, the achieved delay can vary depending on the operating system and hardware.
(2) The parameters of the signal processor should be adapted to the speaker and/or to the vowel. This parameterization improves the formant detection and the reliability of the perturbation.
(3) Perception of unperturbed feedback (bone conduction and air conduction outside the headphones) should be reduced as much as possible. Different approaches were used to achieve this aim, such as:
    - Using whispered speech (Houde & Jordan, 1998, 2002) although subsequent studies were run with normal speech;
    - Using closed headphones or insert earphones to reduce the perception of the air-conducted signal. The occlusion effect of the headphones on adaptation was recently investigated with no significant difference in the magnitude of F1 adaptation between the use of the closed Sennheiser "HD 265" and the insert Etymotic Research ER2 (Mitsuya & Purcell, 2016);
    - Increasing the level of the feedback in the headphones, up to 87dB SPL (Villacorta et al., 2007);
    - And/or using a masking noise mixed with the played back signal to mask bone-conducted speech.
(4) The shifted vowel should have clearly distinguishable F1 and/or F2 values, and the shift should be consistent with these values. For this reason, the vowel /ɛ/ is chosen in most of the studies as shifting more extreme front or back vowels could be limited by overlap in F1–F2 or F0–F1 frequencies (Mitsuya, MacDonald, Munhall, & Purcell, 2015), and this vowel allows upward and downward perturbations.

Different research groups have developed their own formant perturbation systems (Table 3) with four main categories: (1) The two systems developed by Houde described with more details in Houde's PhD (Houde, 1997) for whispered speech (1.a), and then in Katseff, Houde, & Johnson

(2012) for voiced speech (1.b); (2) The system developed and used by Munhall, Purcell and collaborators that used a specific hardware; (3) The system used by Perkell and Guenther's teams that first included specific hardware (Villacorta et al., 2007) and was then adapted as a free software for Matlab. It supports various auditory perturbations, including changes in F1 and/or F2, but also more complex ones such as formant trajectory perturbations (Cai, Boucek, Ghosh, Guenther, & Perkell, 2008; Tourville, Cai, & Guenther, 2013). The last version is called "Audapter" and can be download on github.com (https://github.com/shanqing-cai/audapter_matlab, this link was retrieved July, 6, 2018); (4) The last system was developed in parallel by three teams: Max et al., Ostry et al., and Shiller et al. It uses a device from the music industry (VoiceOne, TC Helicon) that by default allows shifting of all the formants while preserving F0. This system was used as a way to alter all formants in the same direction (Max & Maffett, 2015) or, with supplementary signal processing steps, including filtering and mixing, as a way to perturb F1 only (Rochet-Capellan & Ostry, 2011). A few papers were dedicated to the presentation and first

Table 3: Main signal processing systems used in the literature to perturb formants in real time (references indicate the publication describing the system) and number of papers using the system.

	System 1	System 2	System 3	System 4	
References	**Houde (1997); Katseff et al. (2012)**	***Purcell & Munhall (2006ab)***	***Villacorta et al. (2007); Cai et al. (2008); Tourville & al.* (2013)**	**Feng et al. (2011); Rochet-Capellan & Ostry (2011); Shum et al. (2011)**	**Others**
Signal processing	1.a. Whispered speech: Analysis-synthesis process, DSP- 96 board, Ariel, Inc. 1.b.Voiced speech: "Feedback Alteration Device" – Sinewave synthesis	National Instruments PXI-8176 embedded controller	Texas Instruments C6701 Evaluation Module DSP board **then** C-extension Mex for Matlab, opened access – Audapter	Electronic speech processor from music industry VoiceOne; TC Helicon + filters	Other software or hardware solutions –
Number of papers	10	23	2 **then** 20	19	3

evaluation of these different perturbation systems. This was the case with Cai et al. (2008) and Tourville et al. (2013) and with the preliminary work by Shih, Suemitsu, & Akagi (2011). Two papers also presented a method to perturb formants in populations in which speech acoustics have deteriorated, by coupling articulatory synthesis with Audapter (Berry, North, & Johnson, 2014; Berry, North, Meyers, & Johnson, 2013).

As displayed in Table 4, most of the studies involved native speakers of English, mainly from North America. Other languages were investigated in a few comparative studies or in relation to other research questions as described in the next section. Potential generalization of these findings to other languages and populations should therefore be taken with caution.

4.2. Main procedures in formant perturbation studies and related concepts

The main procedures identified in the collected papers about formant perturbations are summarized in Figure 3. These procedures will be referred to in relation to the research topics detailed in the next section. Two main approaches can be distinguished:

(1) Unexpected formant perturbation during the production of prolonged utterances: This first approach was used in only a few of the collected papers (n=11, ~14 % of the papers with formant perturbations, see Table 4). The perturbation is only applied to a small proportion of utterances so that talkers cannot anticipate the perturbation. Moreover, the utterances are produced with long vowel duration (steady-state vowels) so that corrective answers result from online processing of the auditory feedback (cf. Figure 3, procedure P4). This correction is called *compensation*.

(2) Systematic and constant perturbation over a number of utterances: This second approach was used in the majority of the papers (n=66, ~86 %, Table 4). The basic procedure is represented in Figure 3, procedure P1. It involves the production of utterances with "natural" duration, in general. After a baseline with unaltered auditory feedback, the perturbation is introduced either gradually or abruptly, and then systematically applied at a constant level. Depending on the research group, changes in formant production

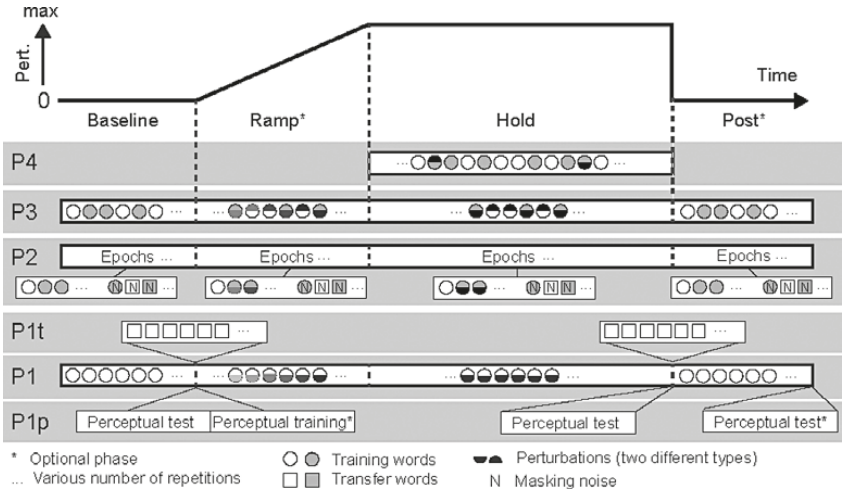

Figure 3: Overview of procedures used in formant perturbations studies. Duration of experimental phases and perturbations were variable across studies. P1 is the basic procedure to study auditory-motor adaptation, used in Munhall et al.'s studies. It was adapted to investigate the transfer of adaptation (P1t) (MacDonald, Pile, Dajani, & Munhall, 2008; Rochet-Capellan, Richer, & Ostry, 2012) and the effect of auditory motor adaptation on perception (P1p) (Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014) or the effect of perceptual training on sensorimotor adaptation (Lametti, Krol, Shiller, & Ostry, 2014). P2 is the procedure used in Houde & Jordan (1998) and then by Perkell et al. (Villacorta, Perkell, & Guenther, 2007). It is structured in epochs with training words produced with feedback followed by training words and generalization words produced with a masking noise. P3 is the multiple perturbation procedure developed in Rochet-Capellan & Ostry (2011), during which words are produced in random order with specific perturbation associated with each word. P4 is the compensation procedure to unpredictable perturbations. In this last case, long steady-state vowels are produced and the perturbation is introduced randomly for a small proportion of utterances to assess online correction (Purcell & Munhall, 2006b). Grey scale gradient in the ramp phase represents the progressive introduction of the shift.

at the end of the training phase are referred to as *compensation* (cf. Houde & Jordan, 1998; Purcell & Munhall, 2006b) or *adaptation* (cf. Rochet-Capellan, Richer & Ostry, 2012, Martin et al., 2018), and residual changes when the feedback is returned to normal after training are referred to as *adaptation* or *after-effect*, respectively.

This procedure was also used to assess *generalization* (or *transfer*) of adaptation to untrained utterances, either in the course of the training phase (Figure 3, procedure P2) or after the training (Figure 3, procedure P1t), as presented in the next section.

Hereafter, *adaptation* will refer to changes observed at the end of the training phase in response to a systematic perturbation. *Compensation* will mainly refer to changes in response to unpredictable perturbations but will also be used to qualify the direction of adaptive responses (by contrast with following responses that go in the same direction as the perturbation).

5. Research topics tackled with formant perturbations

In this section, we provide a thematic review of the collected papers that included an empirical study of formant perturbation. As much as possible, we chose to associate each paper with a main topic but obviously a paper could be related to more than one topic. Table 4 provides a list of all the cited references and their main associated research topics.

5.1. Properties of feedback and feedforward control

Many studies involving formant perturbations are related to the role of auditory feedback in speech motor control and distinguish between feedback and feedforward control mechanisms. Feedback control is a closed-loop system that involves the sensory consequences of the current motion. It is regarded as too slow to account for rapid control and rapid adjustments observed in fast coordinated actions. Rapidity and adaptability of motion were identified early on as evidence of a feedforward control mechanism by researchers in visuomotor adaptation. The core idea is that the brain makes predictions of the sensory consequences of its actions based on an efference copy of the motor command (Houde & Jordan, 2002). These predictions involve mappings between motor and sensory representations also called *internal models* (Purcell & Munhall, 2006a) or sensorimotor memories (see Perrier, 2012, for a discussion of the nature of internal models in speech). The DIVA (Golfinopoulos et al., 2010) or the SFC (Houde & Chang, 2015) neurocomputational models of speech production assume the existence of both feedback and feedforward control networks that involve auditory and somatosensory systems. When

the prediction based on *internal models* does not match the actual sensory input, the internal representations are changed to reduce this prediction "error" so that future movements performed in similar conditions will be accurate. This mechanism is claimed to underlie sensorimotor adaptation.

In this context, a first subset of studies with formant perturbations was designed to "Investigate the nature, level of details, and use of internal models in speech production" (Max, Wallace, & Vincent, 2003, p. 1053) and to "begin to parameterize the formant feedback system" (MacDonald, Goldberg, & Munhall, 2010 p. 1060). The main contribution of these studies is to describe the role of auditory feedback in the control of formant production, and the adaptability of this control. In these papers, adaptability is mainly explained or taken as an evidence for feedforward internal models.

To address the properties of adaptation to formant perturbations, Houde and Jordan (2002) analyzed in more detail the adaptation phenomenon introduced in Houde and Jordan (1998). The results highlight some properties of feedback and feedforward control that were subsequently discussed and investigated in later work, involving various types of formant perturbations and procedures.

The first observation of Houde and Jordan was that the changes in F1 and F2 production in talkers' speech were compensatory responses, in the opposite direction to the perturbation. This result has been reproduced consistently in later work when between-speaker data are aggregated. Individual data suggests that some speakers follow the shift, however. For example, in a meta-analysis of their own studies of adaptation to formant perturbations, MacDonald et al. (2011) found that 26 out of 116 female speakers followed F1 or F2 shifts when their production of "head" was perturbed toward "had". A possible explanation is that non-adapted speakers may not be able to dissociate their own production from the auditory feedback (Vaughn & Nasir, 2015). Following the formant shift rather than compensating for it was actually the most frequent behaviour observed in a preliminary study investigating compensation in Japanese speakers to unexpected perturbations of F1, F2 and F3 (Shih et al., 2011). Aside from this study, all other published work on formant perturbations observed significant compensatory adaptation in acoustic analyses, whereas preliminary analyses of articulatory correlates of adaptation are

Table 4: List of all the studies related to formant perturbation included in the present review. The first column provides the reference of the article. The 2nd column gives the language of participants (Du: Dutch, En: English, Fr: French, Ge: German, Ja: Japanese, Ko: Korean, Ma: Mandarin, Ru: Russian, Sp: Spanish). Column 3 is related to the perturbation systems, which are described in Table 3 (briefly, 1.a: Houde & Jordan (1998), 1.b. Katseff et al. (2012); 2: Purcell & Munhall, (2006a); 3: Audapter and its previous versions; 4: VoiceOne, TC Helicon, 5: Others) and column 4 indicates whether an article is mainly dedicated to the description of a perturbation system. Each study has been classified into either compensation (to unpredictable perturbations, column 5) or adaptation (to sustained perturbations). Columns 7 to 14 show whether the article is related to each of the main research topics presented in the present review. A cross indicates that the article is cited in the corresponding subsection, while a (X) indicates it is not although it is related to the topic.

References	Language	Perturbation System	System description	Compensation	Adaptation	Properties of feedback and feedforward control	Perception acuity and sensory integration	Perceptual & phonological categories	Transfer/Specificity and speech units	Pathology affecting speech production	Neural basis of speech motor learning	Development	Surface effects & speakers' characteristics
Alsius, Mitsuya, Latif, & Munhall, 2017	En	2			X		(X)						**X**
Berry, Jaeger, Wiedenhoeft, Bernal, & Johnson, 2014	En	3			X	**X**			**X**				
Berry, North, & Johnson, 2014	En	3	X										
Berry, North, Meyers, & Johnson, 2013	En	3	X										
Bourguignon, Baum, & Shiller, 2014	En	4			X			**X**					
Bourguignon, Baum, & Shiller, 2015	En	4			X			**X**					
Bourguignon, Baum, & Shiller, 2016	En	4			X			**X**					
Cai, Beal, Ghosh, Tiede, Guenther, & Perkell, 2012	En	3			X		(X)			**X**			

Table 4: Continued

References	Language	Perturbation System	System description	Compensation	Adaptation	Properties of feedback and feedforward control	Perception acuity and sensory integration	Perceptual & phonological categories	Transfer/Specificity and speech units	Pathology affecting speech production	Neural basis of speech motor learning	Development	Surface effects & speakers' characteristics
Cai, Boucek, Ghosh, Guenther, & Perkell, 2008	Ma	3	X		X								
Cai, Ghosh, Guenther, & Perkell, 2010	Ma	3			X	**X**			**X**				
Cai, Ghosh, Guenther, & Perkell, 2011	En	3		X		**X**							
Caudrelier, Perrier, Schwartz, & Rochet-Capellan, 2016	Fr	3			X			(X)	**X**				
Caudrelier, Perrier, Schwartz, & Rochet-Capellan, 2018	Fr	3			X				(X)				**X**
Caudrelier, Schwartz, Perrier, Gerber, & Rochet-Capellan, 2018	Fr	3			X			(X)	**X**				
Daliri, Wieland, Cai, Guenther, & Chang, 2018	En	3			X		(X)			**X**			
Lametti, Krol, Shiller, & Ostry, 2014	En	4			X			**X**					
Lametti, Nasir, & Ostry, 2012	En	4			X		**X**						
Lametti, Smith, Freidin, & Watkins, 2018	En	4			X						**X**		
Demopoulos et al., 2018	En	1b		X	X					**X**		(X)	
Deroche, Nguyen, & Gracco, 2017	En	4			X		(X)				**X**		
Dimov, Katseff, & Johnson, 2012	En	1b			X								**X**
Eckey & MacDonald, 2015	Ge	5		X			**X**						
Feng, Gracco, & Max, 2011	En	4			X		**X**						
Houde & Jordan, 1998	En	1a			X	**X**			**X**				

(*continued on next page*)

Table 4: Continued

References	Language	Perturbation System	System description	Compensation	Adaptation	Properties of feedback and feedforward control	Perception acuity and sensory integration	Perceptual & phonological categories	Transfer/Specificity and speech units	Pathology affecting speech production	Neural basis of speech motor learning	Development	Surface effects & speakers' characteristics
Houde & Jordan, 2002	En	1a			X	X							
Ito, Coppola, & Ostry, 2016	En	4			X		(X)				X		
Katseff & Houde, 2008	En	1b			X		(X)						
Katseff, Houde, & Johnson, 2012	En	1b			X		X						
Klein, Eugen; Brunner, Jana; Hoole, Phil (sous press)	Ru	3			X				X				(X)
Lametti, Rochet-Capellan, Neufeld, Shiller, & Ostry, 2014	En	4			X			X					
MacDonald & Munhall, 2012	En	2			X		X						
MacDonald, Goldberg, & Munhall, 2010	En	2			X	X	X						
MacDonald, Johnson, Forsythe, Plante, & Munhall, 2012	En	2			X							X	
MacDonald, Pile, Dajani, & Munhall, 2008	En	2			X				X				
MacDonald, Purcell, & Munhall, 2011	En	2		X		X							
Martin et al., 2018	Sp	1b			X		X						
Max & Maffett, 2015	En	4			X	X							
Max, Wallace, & Vincent, 2003	En	5			X	X							
Mitsuya & Purcell, 2016	En	2			X	X							
Mitsuya, MacDonald, Munhall, & Purcell, 2015	En	2			X		X						
Mitsuya, MacDonald, Purcell, & Munhall, 2011	En	2			X			X					
Mitsuya, Munhall, & Purcell, 2017	En	2				X							

Table 4: Continued

References	Language	Perturbation System	System description	Compensation	Adaptation	Properties of feedback and feedforward control	Perception acuity and sensory integration	Perceptual & phonological categories	Transfer/Specificity and speech units	Pathology affecting speech production	Neural basis of speech motor learning	Development	Surface effects & speakers' characteristics
Mitsuya, Samson, Ménard, & Munhall, 2013	Fr	2			X			**X**					
Mollaei, Shiller, & Gracco, 2013	En	4			X					**X**			
Mollaei, Shiller, Baum, & Gracco, 2016	En	4		X			(X)			**X**			
Munhall, MacDonald, Byrne, & Johnsrude, 2009	En	2			X	**X**							(X)
Neufeld, Purcell, & Van Lieshout, 2013	Ko	2			X	**X**							
Niziolek & Guenther, 2013	En	3		X				**X**					
Parrell, Agnew, Nagarajan, Houde, & Ivry, 2017	En	1b		X	X					**X**			
Pile, Dajani, Purcell, & Munhall, 2007	En	2			X				**X**				
Purcell & Munhall, 2006a	En	2			X	**X**							
Purcell & Munhall, 2006b	En	2		X		**X**							
Purcell & Munhall, 2008	En	2			X	**X**	**X**						
Reilly & Dougherty, 2013	En	3		X			**X**	(X)					
Reilly & Pettibone, 2017	En	3			X				**X**				
Rochet-Capellan & Ostry, 2011	En	4			X				**X**				
Rochet-Capellan, Richer, & Ostry, 2012	En	4			X				**X**				
Sato & Shiller, 2018	Fr	3			X		(X)				**X**		**X**
Schuerman, Nagarajan, & Houde, 2015	En	1b			X			**X**					
Schuerman, Nagarajan, McQueen, & Houde, 2017	En	1b			X			**X**					

(*continued on next page*)

Table 4: Continued

References	Language	Perturbation System	System description	Compensation	Adaptation	Properties of feedback and feedforward control	Perception acuity and sensory integration	Perceptual & phonological categories	Transfer/Specificity and speech units	Pathology affecting speech production	Neural basis of speech motor learning	Development	Surface effects & speakers' characteristics
Schuerman, Meyer, & McQueen, 2017	Du	3			X			**X**			(X)		
Sengupta & Nasir, 2015	En	2			X						**X**		
Sengupta & Nasir, 2016	En	2			X						**X**		
Sengupta, Shah, Gore, Loucks, & Nasir, 2016	En	2			X					**X**	(X)		
Shih, Suemitsu, & Akagi, 2011	Ja	5		X		**X**							
Shiller & Rochon, 2014	En	4			X			**X**				(X)	
Shiller, Lametti, & Ostry, 2013	En	4			X			**X**					
Shum, Shiller, Baum, & Gracco, 2011	En	4			X						**X**		
Terband & Van Brenk, 2015	Du	3			X							**X**	
Terband, Van Brenk, & van Doornik-van der Zee, 2014	Du	3			X			(X)		**X**		(X)	
Tourville, Cai, & Guenther, 2013		3	**X**										
Tourville, Reilly, & Guenther, 2008	En	3		X							**X**		
Trudeau-Fisette, Tiede, & Ménard, 2017	Fr	2			X		**X**			(X)			
van den Bunt, Groen, Ito, Francisco, Gracco, Pugh, & Verhoeven, 2017	Du	4			X			(X)		**X**			
Vaughn & Nasir, 2015	En	2			X	**X**							
Villacorta, Perkell, & Guenther, 2007	En	3			X	**X**	**X**		**X**				
Zheng, Vicente-Grabovetsky, MacDonald, Munhall, Cusack, & Johnsrude, 2013	En	2		X							**X**		

less clear. Max et al. (2003) analyzed acoustic changes to perturbation of all formants in the same direction in relation to jaw and tongue movement during adaptation. No consistent behaviour were observed in articulatory kinematics. Similar results were obtained in a pilot study in one Korean speaker with an F2 shift (Neufeld, Purcell, & Van Lieshout, 2013), while clearer tongue compensation movements were reported in speakers with blindness (Trudeau-Fisette, Tiede, & Ménard, 2017). On the other hand, while the majority of studies on adaptation to formant perturbations found significant compensatory responses, it was also shown that adaptation vanishes when perturbed feedback is delayed by more than 100ms (Max & Maffett, 2015), or is at least largely reduced (Mitsuya, Munhall, & Purcell, 2017).

Houde and Jordan also reported that maximal changes at the end of training did not fully compensate for the perturbation. This result was systematically reproduced in later studies. As an illustration, in Purcell & Munhall (2006a), the maximal adaptation to a 200Hz upward vs. downward shift of F1 compensated for about 30 % of the perturbation, regardless of the number of repetitions during the hold phase. This also suggests that adaptation is a fast process, in agreement with Max et al. (2003)'s observation that compensatory responses occurred after only a few repetitions. However, a F1 perturbation of at least 60Hz (80Hz on average across conditions) was required in Purcell & Munhall (2006a) to initiate the compensatory response. Similar thresholds were reported in later work, regardless of the delay in the auditory feedback (Mitsuya et al., 2017) and the occlusion of the headphones (Mitsuya & Purcell, 2016). Furthermore, MacDonald et al. (2010) highlighted a linear relationship between the magnitude of the perturbation and the magnitude of changes in speakers' utterances for perturbation magnitudes up to +200Hz in F1 and -250Hz in F2, compensating for 25 % of the perturbation in F1 and 30 % in F2. With larger perturbations, there was no improvement, and a decrease even appeared in response to perturbations larger than 300Hz in F1 and larger than 400Hz in F2. Similar limits were observed by Katseff and colleagues (Katseff & Houde, 2008; Katseff et al., 2012), as discussed in the next section. Comparable adaptations were reported in the meta-analysis provided by MacDonald et al. (2011), with an average of 26.5 % for F1 and 23.2 % for F2. Moreover, in this last analysis, changes in F1 in

speakers' production weakly correlated with changes in F2, suggesting a specific control of the two parameters and the existence of speaker-specific strategies. The magnitude of the response was also found to vary according to the vowel in *pet, bus* and *law* utterances in Max et al. (2003). Further work addressing this last point with regard to more specific research topics is presented in the next section.

Houde and Jordan also noticed that inter-speaker variability was not related to a speaker's awareness of the auditory shift. When interviewed after the study, talkers reported they were unaware of the perturbation or of any change in their production. By contrast, Purcell & Munhall (2006a) reported that 40 % of their participants indicated awareness of "some kind of change in the auditory feedback over the course of the experiment", with only 8 % noticing that the perturbation transformed the vowel into a different one. However, the magnitude of adaptation did not seem to be related to the responses in this interview. This difference to Houde and Jordan might be related to the abrupt suppression of the perturbation after training in Purcell & Munhall (2006a) (Procedure P1, Figure 3) that was probably perceived by the speakers, while Houde and Jordan assessed how adaptation was sustained using catch trials with masking noise (Procedure P2, Figure 3). Munhall, MacDonald, Byrne, & Johnsrude (2009) then confirmed that the awareness of the perturbation does not influence adaptive behavior, as discussed later in the "*Surface effects & speakers' characteristics"* subsection.

Another important result in Houde and Jordan was that changes for perturbed utterances were larger than changes for utterances produced with a masking noise. The authors discussed this result as evidence that "vowel production could be partly under immediate auditory feedback control" (Houde & Jordan, 2002, p. 307). By contrast, in their preliminary study of adaptation to a shift of all formants in the same direction, Max et al., (2003) argued that the modifications in talkers' production should be considered as adaptive responses rather than reactive changes, as they already occur at vowel onset, and have been observed for sustained vowels as well as vowels with shorter duration. The variability of changes in formants according to the vowel's parts were not systematically investigated in adaptation studies as most of the studies used a single steady-state value, often around the middle of the vowel. However, in their preliminary

work, Berry, Jaeger, Wiedenhoeft, Bernal, & Johnson (2014) suggested that this single value might not be the most appropriate, depending on consonant context and coarticulatory effects. Vaughn and Nasir (2015) also provided evidence that full trajectory analysis might better capture adaptation phenomena. The relationship between formant values in consecutive trials (as measured with one-lag cross correlation analyses), in the absence of any perturbation, may also be predictive of adaptation magnitude (Purcell & Munhall, 2006a). Altogether, these results suggest that changes observed over the course of adaptation to a perturbation result probably from a mix of feedback and feedforward control.

Houde and Jordan (2002) suggested investigating compensation to formant perturbations in steady-state vowels to determine the role of online feedback in formant control. Studies focusing on compensation to an unexpected formant perturbation in sustained vowels usually analyzed changes at different points of the vowel. For instance, in Purcell and Munhall (2006b) upward vs. downward perturbations of F1 were applied randomly in five utterances of "head" over 100 utterances of different CVC words. Results show partial compensation, with on average, 16.3 % vs. 10.6 % of the upward vs. downward shifts, but with high variability for the same talker between utterances and between talkers. However, this study was not designed to measure the delay in compensatory response. This delay was found in later studies to be around 160ms, at least when F1 is shifted upward (e.g. Tourville, Reilly, & Guenther, 2008), and when more complex spatial or temporal perturbations of formants trajectories are applied during the production of short sentences (Cai, Ghosh, Guenther, & Perkell, 2011). The smaller compensation of perturbation observed in studies involving unexpected perturbation compared to studies involving systematic perturbation, as well as the delay required to observe a compensatory response, confirm the idea that responses produced in the presence of the perturbation in adaptation studies are at least partially adaptive.

One of the most intriguing outcomes of Houde and Jordan (2002) was that the modification in formants was still present when talkers came back a month later to run a control study evaluating changes in production without perturbation. This long-term effect was attributed by the authors to implicit memory of the task or specific control mechanisms for

whispered speech. Although not reproduced in later work – as there was no study with equivalent long-term assessment in our review at least – Purcell and Munhall, (2006a) showed that 115 repetitions without perturbation after the training phase were not enough to fully return to the baseline state. The explanation introduced by Houde and Jordan echoes the idea that auditory-motor learning could be specific to some situations or ways of speaking, as discussed in generalization studies. The ability to memorize specific ways of speaking according to the situation could be a way to support fast speech adaptability in known situations. This idea could be further investigated by means of transfer of adaptation from one context to another as discussed below.

Finally, large inter-speaker variability was also pointed out in Houde and Jordan (2002) and then observed in all the subsequent studies. MacDonald et al. (2010, 2011) suggested that this variability is not clearly related to the variability in baseline production, nor to the size of the vowel space. Inter-speaker variability, as well as partial compensation, in formant adaptation studies was often discussed in terms of a tradeoff between auditory and somatosensory feedback. For example, Purcell and Munhall (2006a) suggested that "Some [speakers] may rely more on kinesthetic feedback and thus are not influenced as much by acoustic feedback" (p. 975), while Houde and Jordan (2002) suggested "it may be that there are differences across participants as to the degree to which they rely on auditory feedback" (p. 308). The tradeoff between auditory and somatosensory feedback, as well as the role of sensory acuity in adaptation was then explored in several papers, as described in the next section.

5.2. Perception acuity and sensory integration

Formant perturbations' paradigms involve modifying the auditory feedback, i.e. sensory input of speech control system, and measuring the outcome in terms of speech production, or motor control. Hence these paradigms are by nature relevant to the question of the relationship between perception and production. Several aspects of this relationship have been investigated over the past two decades.

First, adaptation to auditory perturbations may be influenced by speakers' sensory acuity. Auditory acuity has been positively correlated with adaptation magnitude in two studies (Martin et al., 2018;

Villacorta et al., 2007) involving 13 and 31 subjects respectively. Auditory acuity measurements were based on discrimination tasks in both cases. Villacorta et al. focused on acuity for F1 while Martin et al. measured acuity based on pitch and loudness, as well as melody discrimination tasks. A possible interpretation of the relation between adaptation magnitude and auditory acuity is that better acuity could lead speakers to have smaller goal regions for their production, resulting in higher adaptation (Villacorta et al., 2007). However, auditory feedback may not be the only feedback used to control speech production. Feng et al. (2011) investigated the relationship between the adaptation magnitude of F1 and the auditory acuity for F1, as well as somatosensory acuity for jaw position. They did not find a reliable correlation. However, fewer subjects were involved in this study than in previously cited ones (8 subjects vs. 13 and 31).

Feng et al. also combined a somatosensory perturbation induced by a robotic device pulling the jaw, with an auditory shift on F1. Using this procedure, they found that speakers mainly compensated for the auditory perturbation. They suggested that auditory feedback may be dominant over somatosensory input, but that their relative weight could evolve with speech experience. Using similar methods, Lametti, Nasir, and Ostry (2012) found that all speakers adapted for at least one of the two perturbations. The group who adapted to the somatosensory perturbation (half of the participants) did not significantly compensate for the auditory perturbation while the group that did not adapt to the jaw perturbation significantly compensated for the F1 shift. This observation suggests a speaker-specific sensory preference for either auditory or somatosensory inputs. In addition, the weights attributed to auditory and somatosensory feedback may vary according to the articulator (i.e. vocal folds, tongue or jaw) to control. Indeed, no correlation has been found in the magnitude of adaptation in F0, F1 and F2 across speakers while altering them simultaneously or separately (Eckey & MacDonald, 2015; MacDonald & Munhall, 2012). Interestingly, Trudeau-Fisette et al. (2017) showed that speakers with blindness adapted more to an F2 shift than control speakers, independently of their auditory acuity, and that they also produced larger articulatory changes in response to the auditory shift. Speakers with blindness may rely more on auditory feedback than control speakers, who

may have more precise somatosensory goals, probably built and supported by visual perception of speech.

However, sensory preference in the control of speech, which can be modeled by different weights attributed to each kind of sensory feedback, may also evolve with experience. Most studies on auditory-motor adaptation report a partial compensation for the auditory perturbation as already mentioned in the previous section. Some studies showed that the percentage of compensation relative to the magnitude of the perturbation decreases when the magnitude of the perturbation increases, reaching an asymptote, and can even tend to decrease for larger perturbations (Katseff & Houde, 2008; Katseff et al., 2012; MacDonald et al., 2010). Katseff et al. (2012) interpreted this phenomenon as evidence that the weights attributed to auditory and somatosensory feedback may vary according to experience: "For small discrepancies between auditory and somatosensory feedback, auditory feedback takes precedence, and for large discrepancies between auditory and somatosensory feedback, somatosensory feedback takes precedence" p. 307. Thus, a high-amplitude shift may lead the speech system to consider auditory feedback as unreliable and therefore give more weight to somatosensory feedback. In addition, the relative importance of sensory input may depend on the specific sounds produced. Several studies observed less compensation in closed vowels than in open vowels (Mitsuya et al., 2015; Purcell & Munhall, 2008; Reilly & Dougherty, 2013). This could be explained by better-specified somatosensory information in the former than in the latter case (Mitsuya et al., 2015). Another possible explanation is that the importance of F1 as an acoustic cue in perception may depend upon the vowel (Reilly & Dougherty, 2013).

5.3. Perceptual and phonological categories

Speech perceptual space is structured by phonological categories, which are delimited by perceptual boundaries. Niziolek and Guenther (2013) showed an effect of perceptual boundaries on the magnitude of compensation to unpredictable auditory perturbations. They observed that if the auditory signal resulting from the perturbation is near a boundary, the compensation, as well as the cortical activation, is higher than when it

is far from a boundary, the magnitude of the shift being equal. In addition, various studies have investigated the relation between perceptual boundary and adaptation to sustained auditory perturbations.

The influence of perceptual boundaries on adaptation can be investigated using perceptual learning on the perceptual contrast that is at stake in the adaptation paradigm. For instance, Shiller, Lametti, & Ostry (2013) manipulated speakers' perceptual boundaries between "head" and "had" through perceptual training preceding auditory-motor adaptation to a perturbation consisting of altering "head" into "had" (see Procedure P1p on Figure 3). The group whose boundary was shifted towards "head" (i.e. who was more likely to classify ambiguous stimuli as "had") adapted more to the auditory perturbation than the group whose boundary was shifted towards "had" by the perceptual training. Similarly, children adapted more to a perturbation transforming /beb/ into /bab/ after a perceptual training manipulating /ɛ/-/æ/ boundary towards /ɛ/ than before training (Shiller & Rochon, 2014). They also adapted more than children having undergone a perceptual training on an unrelated contrast. Furthermore, Lametti, Krol, et al. (2014) observed in adults that the amount of adaptation to auditory-feedback perturbation was correlated with the position of the perceptual boundary obtained through perceptual training.

Instead of using perceptual training, changes in perceptual boundaries were obtained by manipulating the pitch and formant of the carrier phrase "please say what this word is…" (Bourguignon, Baum, & Shiller, 2015, 2016). In this study, the group exposed to high carrier-phrase (high pitch and formants) had the boundary between 'bit' and 'bet' shifted toward 'bet'. They adapted more to an auditory feedback alteration transforming /ɛ/ into /ɪ/ than the speakers exposed to low carrier-phrase (low pitch and formants). This finding suggests that "context-dependent plasticity in speech perception may also transfer to production" (Bourguignon et al., 2016, p. 1040). Interestingly, Bourguignon, Baum, and Shiller (2014) also showed an effect of the lexical status that can be interpreted in terms of perceptual boundaries. In their study, a group of speakers produced pseudo-words that resulted in real word when auditory perturbation was applied (e.g. "kess" changed into "kiss"). Another group produced real words that were transformed into pseudo-words by the same formant shift

(e.g. “less” changed into “liss”). The first group showed greater adaptation than the second group, indicating a lexical effect on auditory-motor adaptation.

The influence of phoneme categories on speech motor adaptation was also highlighted in cross-language studies. Mitsuya, MacDonald, Purcell, and Munhall (2011) contrasted the adaptation to upward and downward shifts in F1 in three groups: English speakers pronouncing “head”, Japanese speakers producing the Japanese word /he/ and Japanese speakers learning English, producing “head”. The magnitude of adaptation was equivalent in all groups in response to the downward shift, but the adaptation was smaller in Japanese than in English speakers in response to the upward shift. This difference is evidence for the influence of the phonological system in adaptation. Mitsuya, Samson, Ménard, and Munhall (2013) also showed differences between English speakers and French speakers in the adaptive response to the same auditory perturbation. In this study, a perception test suggested that this language effect on adaptation was mediated by a difference in perceptual boundaries: larger adaptation in French speakers was related to greater sensitivity to some phonetic contrasts.

Reciprocally, the influence of adaptation on perceptual boundaries has also been investigated. Lametti et al. (2014) incorporated perceptual tests in a classic auditory-motor procedure (Figure 3, procedure P1p), before and after the training phase – during which adaptation occurs – as well as after the after-effect phase, used here as a wash-out of adaptation. They observed that auditory-motor adaptation resulted in a shift of a perceptual boundary in the phonetic range of what speakers produced but not what speakers heard. For instance, speakers who produced “head” and heard an auditory feedback shifted toward “had”, compensated by producing an utterance closer to “hid”. Their perceptual boundary between “head” and “hid” was shifted toward “head”, that is, speakers became more likely to report hearing ‘hid’ in the perceptual test, while there was no effect on the perceived boundary between “head” and “had”. This result suggested that the change in perception was specifically driven by speech motor adaptation and not by the auditory input during learning. The interpretation of these results, together with the results of other studies on the effect of auditory-motor adaptation on categorical perception, was recently specified in a Bayesian modeling framework, suggesting

that speech motor adaptation results both in speech sound remapping and changes in phoneme categories (Patri, Perrier, Schwartz, & Diard, 2018). Yet, using a similar paradigm to that of Lametti et al. (2014), Schuerman, Meyer, and McQueen (2017) did not find significant influence of auditory-motor adaptation on related perceptual boundaries. It should be noted that this experiment had fewer subjects than Lametti et al. (2014); was run with speakers of Dutch as opposed to English; and used a continuum with isolated vowels rather than a continuum between words, during the perceptual test. However, this last study also recorded EEG signals during initial vs. final perception tests. The analysis of ERPs to the stimuli of the /ɛ/-/ɪ/ continuum revealed changes in N1 and P2 components for ambiguous stimuli, which correlated with the magnitude of adaptation as measured by F1. The effect on both N1 and P2 suggest that auditory-motor adaptation influences both early perception and late perceptual decisions. Interestingly, Schuerman, Nagarajan, and Houde (2015) and Schuerman et al. (2017) showed that the adaptation to an auditory perturbation of F2 shifting the front vowel /i/ towards the back-vowel /u/ resulted in a shift in the perceptual boundary between "see" and "she". More specifically, the shift in perceptual boundaries depended on the behavior of speakers during the adaptation task: speakers who followed the auditory perturbation had their perceptual boundary shifted in the opposite direction to that of speakers who compensated for the auditory feedback. This last group was more likely to categorize ambiguous stimuli as "see" than "she", the place of articulation of the consonant /s/ being more anterior than /ʃ/. These findings are in agreement with the idea that some transfer of adaptation may occur between vowels and consonants articulated with a similar tongue position.

While this impact of a change in production on the perception of another contrast is actually a transfer from production to perception, the term *transfer* is typically investigated in speech production itself, from one utterance to another.

5.4. Transfer/specificity and speech units

In the limb movement literature, generalization of motor learning is the "ability to correctly extrapolate to contexts that are different from our limited experience" (Krakauer, Mazzoni, Ghazizadeh, Ravindran, &

Shadmehr, 2006, p. 1798). This extrapolation could be the result of an interpolation of previous experiences (Mattar & Ostry, 2007). Generalization has been extensively investigated in motor learning research, and in speech, in particular to address the specificity of motor adaptation and the underlying representations (Tremblay, Houle, & Ostry, 2008). Transfer of adaptation is usually defined as a positive generalization, as opposed to interference (Krakauer et al., 2006). However, we will use generalization or transfer to designate changes observed in untrained utterances after adaptive training, going in the same direction as adaptation. When no significant transfer is observed, changes related to adaptation are considered to be specific to the training utterance.

The investigation of generalization or transfer of adaptation relied on two different motivations. The first set of work focused on generalization as a way to assess the global vs. specific nature of auditory-motor mapping. This approach is derived from limb movement studies that analyzed generalization of visuomotor adaptation to address the global vs. specific nature of visuomotor mapping. The second set of work, that is sometimes an extension of the first one, considered generalization of auditory-motor learning as a way to assess the nature of speech production units, by questioning the linguistic level of auditory-motor mapping. This second approach was introduced by Houde and Jordan (1998) and is consistent with earlier work on transfer of perceptual learning to assess speech perception units (e.g. Chambers et al., 2010).

Different procedures were used to investigate generalization of auditory-motor adaptation. The first one is structured in "epochs" (Figure 3, P2). Each epoch includes utterances with feedback on and utterances with a masking noise, which can be either the training utterances or different utterances. Transfer is evaluated at the end of the training phase, when the perturbation is maximal by measuring changes in transfer (or test) utterances as compared with their baseline. Using this procedure, Houde and Jordan (1998) found significant transfer from the training words sharing the same vowel /ɛ/ ("pep", "peb", "bep", and "beb"), shifted toward /ɪ/ or /æ/ to the various test words (same vowel as training words – "gep", "peg", "teg", or different vowels – "pip", and "pap"). The amount of transfer was variable depending on the test word, but not statistically different. Consistent results were reported in Villacorta et al. (2007),

where adaptation on the vowel /ɛ/ for nine CVC words to an F1 perturbation, significantly generalized to the same vowel in different CVC or to the vowels in "pit", "pat", and "pot". Results were less consistent for "put" and "pete" and seemed to depend on the direction of the perturbation. Still with a similar procedure, but with perturbation of F1 trajectory in speakers of Mandarin, Cai et al., (2008) and Cai et al. (2010) found gradients of generalization that depended on the similarity in formant trajectory between a training triphthong and the tested utterances. Finally, Reilly and Pettibone (2017) tested generalization from the vowels /i/ vs. /æ/ (embedded in a set of CVC utterances) to /i/, /ɛ/ and /æ/ (also in CVC) produced with a masking noise. In both training conditions, /ɛ/ was the "near" vowel in test utterances, while /i/ and /æ/ were either the same as the training vowel or the "far" vowels, depending on the training condition. Adapted speakers exhibited significant generalization to all vowels, regardless the training vowel. However, correlation between adaptation and generalization were unclear suggesting that generalization may depend on multiple factors and may be sensitive to inter-speaker variability.

Similar procedures, mixing training and transfer trials, were used in limb movement studies. However, the approach was later criticized. In particular, with this procedure, "the patterns of generalization observed are difficult to interpret, as transfer could reflect an averaging that takes places when subjects experience several training conditions simultaneously" (Rochet-Capellan et al., 2012 p. 1711). For this reason, other studies tested transfer after the training phase, when the feedback is turned-back to normal (Procedure P1t on Figure 3). In preliminary work, MacDonald et al. (2008) compared transfer tested in the course of training vs. after training. In both cases, speakers were trained on "head" shifted towards "had" and transfer was tested on the production of "hid" with unaltered feedback. When the transfer utterance "hid" was inserted during training, changes in "hid" were observed at the beginning of the training phase, but then its production came back to baseline. When tested after training, no change was observed at all in "hid". Overall this suggests that adaptation is specific to the trained vowel, although it slightly depends on the training conditions. Pile, Dajani, Purcell, and Munhall (2007) then observed similar adaptation and lack of generalization toward "hid" or "hayed" (i.e. /hed/). Both studies were published in proceedings and were preliminary,

with restricted analyses. Later work by Rochet-Capellan et al. (2012) evaluated how adaptation to a perturbation of F1 in /pen/, /ben/, /ken/, /gen/, /ten/, /den/, /pan/, /pin/ then affect the production of /pen/ produced without perturbation. Results were consistent with previous work that tested generalization with a mixed procedure (Figure 3, P2): generalization was variable according to the training word and seemed to depend on the acoustical proximity between the training and the testing utterance. Another important result of this work was that the after-effect, assessed on the training utterance after the transfer phase, was still significant, suggesting that the production of the transfer utterance with normal feedback did not wash out adaptation. This last result is consistent with the idea that learning is related to the training experience, and at least to some extent specific to this experience.

Another way to assess specificity of adaptation is to evaluate how speakers can specifically compensate for several perturbations in the same training session. This approach is inspired by limb movement studies and in particular Osu, Hirai, Yoshioka, and Kawato (2004). In Rochet-Capellan and Ostry (2011), speakers produced "head" and "had" in random order with F1 shifted downward in "head" and upward in "had" and conversely (Procedure P3 in Figure 3). On average, speakers were able to change F1 frequency in opposite directions for "head" and "had", suggesting that auditory-motor mapping is specific to each vowel. To assess whether auditory-motor mapping could be specific to a word, the authors then evaluated multiple adaptations for "head" and "bed" shifted in opposite directions and "ted" un-shifted. Again, on average, specific adaptation in opposite directions were observed for "head" and "bed" while F1 in "ted" remained unchanged, suggesting that different auditory-motor mappings could be built for a same vowel in different words. Similar results were obtained recently by Klein, Brunner and Hoole (in this book) with a Russian vowel in /d/ vs. /g/ CV syllables and a perturbation of F2. The authors also provided analysis of speakers' data showing symmetrical vs. asymmetrical profiles of adaptation.

Altogether, these results suggest that generalization of auditory motor adaptation occurs in a way that depends on the similarity between the training and the testing utterance and that specific control can be achieved, at least under specific conditions. The results were interpreted as an

indication of global control for vowel production vs. specific control. Furthermore, generalization from a vowel to the same vowel in different contexts suggests that auditory-motor mapping could occur at the level of the phoneme. It is thus a way to question the structure of feedforward mapping, and the nature of its underlying representations (Houde & Jordan, 1998). The fact that transfer is in general smaller than after-effect suggests that word context may play a role. The idea that multiple representations may coexist in auditory-motor mapping of speech was directly assessed in recent papers by Caudrelier et al. (Caudrelier et al., 2016; Caudrelier et al., 2018). In this work, several linguistic levels were contrasted by assessing transfer on test utterances that shared either the same vowel, and/or the same syllable or was the same word as the training utterance. Transfer was smaller (although significant) at the vowel level than transfer to the same syllable, which was lower than after-effect in the same word, suggesting that these three levels – words, syllables, phonemes – could coexist in parallel in the structure of the speech sound map. This conclusion is consistent with multiple traces connectionist models of long-term memory (Ans, Carbonnel, & Valdois, 1998; Carbonnel, Charnallet, & Moreaud, 2010) in the sense that multiple units could emerge as common information of multiple experiences (Goldinger, 1998; Hintzman, 1986). Specific production of the vowel to the syllable or word context also questions the role of coarticulation in adaptation and transfer of adaptation, a topic introduced in a preliminary paper by Berry et al. (2014).

In addition to the theoretical insights mentioned above, a better understanding of generalization in speech may have clinical implications in speech rehabilitation (e.g. after stroke), since transfer from training with a speech therapist to daily life is essential (Aichert & Ziegler, 2013). Other clinical applications are described in the next section.

5.5. Pathology affecting speech production

Auditory feedback perturbation paradigms may be instrumental in the understanding of mechanisms underlying disorders related to or affecting speech production. In particular, low compensation or adaptation observed in patients with a given pathology is regarded as evidence for a lack of sensorimotor integration or as an impairment of feedforward control mechanisms.

Stuttering is suspected to be driven by abnormal integration of sensory input in speech motor control, and has been an early target for auditory perturbation studies, and more recently for studies using formant perturbations. Cai et al. (2012) observed smaller compensation to unpredictable perturbation of formants in persons who stutter compared to control participants. The latency of compensation was however found to be equivalent in both groups. According to the authors, this suggests impairment of the inverse model responsible for translating auditory error detection into proper correction in motor commands. Reduced responses to formant perturbations were also observed in adaptation studies, with systematic perturbations. Sengupta, Shah, Gore, Loucks, & Nasir (2016) found smaller adaptation in adults who stutter as compared with control speakers that was also related to anomalous EEG phase coherence. This hints at a miscommunication between speech sensory and motor areas, which confirms a potential deficit in sensorimotor integration in people who stutter. A recent study, Daliri, Wieland, Cai, Guenther, & Chang (2018) also found reduced adaptation in adults who stutter compared to control speakers. However, the difference was not observed in children who stutter as compared with their aged-match controls. These results suggest that reduced adaptation observed in adults may be a consequence of compensatory strategies induced by the pathology rather than a root cause.

Terband, Van Brenk, and van Doornik-van der Zee (2014) used a similar adaptation paradigm as Daliri et al. (2018) with children with CAS (Childhood Apraxia of Speech). CAS was described as "a disordered development of the functional synergies/coordinative structures that underlie speech motor coordination causing impairment of the forward model leading to poor feedforward control" (Terband et al., 2014, p. 66). In agreement with this description, children with CAS were shown to follow the auditory perturbation on average, while their aged-match controls adapted to the perturbation by compensating for it.

Van den Bunt et al. (2017) used formant adaptation to assess the nature of the phonological deficit observed in dyslexia, known as a "difficulty in acquiring fluent word-decoding skills" (p. 1). Adults with dyslexia showed greater adaptation and after-effects than control speakers to a formant feedback perturbation that doesn't cross a phonemic boundary (i.e. an

allophonic perturbation). Moreover, a negative correlation was observed between reading skills and the magnitude of adaptation: the worse the reading score, the larger the adaptation. This result could be interpreted as a weaker perceptual magnet effect (Kuhl et al., 2008) in speakers with dyslexia and supports theories claiming that dyslexia is associated with a greater distinction between allophones, which may lead phoneme categories to be less prominent. However, a condition with a perturbation crossing the phonetic boundary is required to further support this hypothesis.

Compensation or adaptation to formant perturbations were also investigated in populations with neurogenetic or neurodegenerative diseases. Demopoulos et al. (2018) used adaptation to formant perturbation to address the origin of the speech production deficit observed in young individuals with a subtype of autism (due to a 16p11.2 deletion). The adaptation was reduced in this population as compared with age-matched controls while compensation to unexpected perturbation of F0 was larger. According to the authors, this suggests that feedforward models could be altered in people with 16p11.2 deletion, leading to an over-reliance on feedback control. A comparable profile of larger compensation to unexpected perturbation of F0 was observed in patients with Parkinson Disease (PD). However, both compensation and adaptation to unexpected vs. constant formant perturbation were reduced in speakers with PD as compared with age-matched control speakers (Mollaei, Shiller, Baum, & Gracco, 2016; Mollaei, Shiller, & Gracco, 2013). The authors interpreted the difference in pitch and formant compensation in terms of somatosensory and muscle activation deficits of the larynx and oral cavity. This dissociation between compensation to F0 vs. formant perturbations calls into question the conclusion of Demopoulos et al. (2018): as feedback control was only assessed with F0 in speakers with 16p11.2 deletion, it remains unclear whether they indeed rely more on feedback control in general or if the effect was specific to F0 control. Finally, Parrell, Agnew, Nagarajan, Houde, & Ivry (2017) found that speakers with cerebellum degeneration compensate for unexpected formant perturbations more than their age-matched controls, while they show weaker adaptation to sustained perturbation. This suggests that the cerebellum plays an important role in feedforward control, and probably less in feedback control. The involvement of the cerebellum in feedback control is discussed in the next section.

5.6. Neural basis of speech motor learning

The neural correlates of speech motor control and learning have been investigated through a variety of techniques, including EEG, fMRI, rTMS and tDCS.

fMRI is not suitable to observe changes in the timeframe of adaptation to sustained perturbation because it could be confounded with low-frequency noise observed in fMRI (Zheng et al., (2013). However, it is feasible to investigate the neural networks involved in feedback control using unpredictable perturbations. In Tourville et al. (2008), trials under altered auditory feedback (as opposed to normal feedback) were associated with increased bilateral activation in posterior auditory cortex (including posterior Superior Temporal Gyrus, pSTG, and Planum Temporale, PT). This observation is regarded as evidence for the existence of auditory error cells, dedicated to detect errors in auditory feedback. The increased activation in right pSTG was observed to be enhanced when auditory perturbation outcomes were close to a perceptual boundary. In addition, Tourville et al., (2008) found increased right activation in ventral Motor and Premotor Cortex (vMC and vPMC, respectively) and anterior medial cerebellum (amCB). This suggests that feedback control involves mainly the right hemisphere whereas the left hemisphere, which is known to be dominant in speech production, would be mainly associated with feedforward control. Zheng et al., (2013) conducted further fMRI investigation. Their experimental procedure consisted of production trials with normal feedback, altered feedback (with F1 shift) and feedback with masking noise. Speakers then passively listened to every signal corresponding to their auditory feedback in the production session. Combining fMRI with an analysis of neural pattern similarity analysis enabled differentiation of three functional networks: an error signal network (including right AG, right SMA, and bilateral cerebellum), a passive listening network, and a network responding to both production and passive listening conditions, that may correspond to sensorimotor integration, located in bilateral Inferior Frontal Gyrus (IFG).

The Inferior Parietal Lobe (IPL), which comprises Supramarginal Gyrus (SMG) and Angular Gyrus (AG) may be involved in multisensory integration. An rTMS stimulation applied over the SMG just before the auditory-motor adaptation procedure reduced adaptation responses in

comparison with a sham stimulated group (Shum et al., 2011). Similarly, a tDCS stimulation applied over IPL affected auditory-motor adaptation (Deroche, Nguyen, & Gracco, 2017). More specifically, anodal stimulation aiming at facilitating neuronal excitability resulted in stronger adaptation magnitude whereas cathodal stimulation, which has an inhibitory effect, prevents auditory-motor adaptation to predictable perturbations.

Lametti, Smith, Freidin, and Watkins (2018) investigated the specific role of two areas involved in motor control, the cerebellum and the premotor cortex. In this experiment, anodal tDCS was applied during the baseline phase and the training. The auditory perturbation consists of an F1 shift making the training words "bed", "head" and "dead" sound more like "bad", "had" or "dad", respectively. Stimulations over either motor cortex or cerebellum were both found to lead to higher adaptation and/or after-effect than in the sham-stimulated group. Interestingly, stimulation over the cerebellum increased error compensation on F1, while stimulation of the motor cortex also led to adaptation in F2. Adaptation in F2 when altering F1 only has been reported for the front vowel /ɛ/ with variable size-effects (MacDonald et al., 2011; Rochet-Capellan & Ostry, 2011; Villacorta et al., 2007). Changing F2 in answer to a perturbation of F1 may be a strategy to reach an appropriate phoneme auditory category, as F1 and F2 vary at the same time in the contrast of front vowels. Thus, the cerebellum is suggested to contribute to error correction only, while motor cortex may lead to more general adaptation, possibly related to previously learnt movements.

While rTMS and tDCS can reveal the functional role of a specific brain area, neuronal oscillations as observed in EEG combined with phase coherence analysis may provide insights into the communication between brain areas as proposed by Sengupta and Nasir (2015). Phase coherence over a specific brain area can also represent a measure of this area's engagement. In this study, a redistribution of phase coherence in specific frequency bands (theta and gamma bands) occurred at the end of the training phase and was related to the amount of speakers' adaptation. This phenomenon was interpreted as a sign of the establishment of a new feedforward map (i.e. associating an auditory target to a motor gesture that enables the speaker to reach it) together with increased engagement of sensorimotor areas. Sengupta and Nasir (2016) then found that by late training, power in specific frequency bands during speech planning and

speech production was related to whether speakers were adapting to the auditory perturbation or not. Finally, Sato and Shiller (2018) analyzed event-related potentials (ERPs) during adaptation to an increase of F1. They observed that electro-cortical potentials at certain temporal windows (N1, P2) amplitude mirrors adaptation, as larger adaptation magnitude correlated with smaller N1/P2 amplitude. This larger speaking-induced suppression with learning was interpreted as an indication of auditory prediction during speaking.

5.7. Speech development

Auditory perturbation is an artificial way to generate speech learning, which otherwise occurs in *natural* situations: learning a new language, as well as during the development of speech. Studying adaptation to perturbations in typical adult speakers might help understand potential mechanisms occurring in these natural situations. It also questions the way children learn speech sounds. Daliri et al. (2018) and Terband et al. (2014) studied adaptation in atypical development, as reported in the "Pathology" section. Shiller and Rochon (2014) investigated the relation between adaptation on perceptual boundaries in children, as reported in the "*Perceptual and phonological categories*" section. MacDonald, Johnson, Forsythe, Plante, and Munhall (2012) and Terband and Van Brenk (2015) focused on adaptation in typically developing children at different ages. Terband and Van Brenk (2015) found greater adaptation in 4 to 9-year-old children than in adults, although the magnitude of adaptation did not correlate with age in the group of children, and the proportion of children exhibiting a consistent compensatory response was lower than in adults. MacDonald et al. (2012) showed that 4-year-old children adapted to a sustained perturbation with a similar magnitude of adaptation as adults, whereas 2-year-old toddlers did not adapt at all. This could suggest that toddlers ignore their own auditory feedback to focus on external stimulation or have an immature feedforward control. According to Messum and Howard (2012), this observation contradicts the widely held view that children learn speech sounds by imitation, which would require them to listen to what they produce and try to make it match what they want to imitate. Instead, it supports the idea that a child learns to speak thanks to a tutor: "Mothers reflect (or mirror) what

their children say, but such imitation generally takes the form of reformulation into well-formed sounds of the ambient language, rather than simple mimicry" (Messum & Howard, 2012, p. 160). Thus, plasticity observed in adults in the situation of adaptation to auditory perturbations may be different in nature to what occurs in the early speech development.

5.8. Surface effects and speakers' characteristics

Other effects related to speakers or context, like the characteristics of the prompt during the adaptation procedure, may influence speech adaptation. Alsius, Mitsuya, Latif, and Munhall (2017) investigated the influence of the stimulus used to prompt the training word "head" by contrasting visual and auditory modalities as well as linguistic vs non–linguistic prompts. No effect of the sensory modality was found on the magnitude of adaptation but linguistic prompts ("head" as a spoken or written word) were found to induce more adaptation than non-linguistic prompts (a cross or a tune). Similarly Sato and Shiller, (2018) found no difference in the magnitude of adaptation between visual and auditory modalities. In addition, Caudrelier et al. (2018) investigated whether naming a picture or reading a word aloud would make a difference in adaptation and in transfer. Although no effect was found in the adaptation response, the pattern of generalization was influenced by the prompt used during the transfer phase, regardless of the training prompt, hinting at possible surface effects.

With regards to speakers' abilities, Martin et al. (2018) found no correlation between general executive control and adaptation magnitude. In a preliminary study, Dimov, Katseff, and Johnson, (2012) investigated the influence of speakers' characteristics including some social and personal aspects. In particular, less empowered subjects were found to adapt more than more empowered ones. Finally, Munhall et al. (2009) reported equivalent adaptation in naïve speakers and in speakers who were informed of the shift and who were asked to compensate or not. These results suggest that auditory-motor recalibration is at least in part an automatic process. More work is required to better understand the complexity of adaptive profiles that might be determined by numerous factors, as discussed in the next section.

6. Research outlook on formant perturbations

In this section, we identify some perspectives for future studies in adaptation to formant perturbations, in relation to methodological aspects as well as to some of the reviewed research questions.

6.1. Toward standards to investigate and report adaptation to formant perturbations

Various interests have motivated adaptation to formant perturbations studies in various teams. This induced the use of different methods to alter formants but also different procedures and analyses. These methodological differences often make studies difficult to compare directly. Therefore some standards should be developed, in particular to facilitate meta-analyses of formants perturbations studies, at least with regards to the way to report the methods and the results. Munhall, Purcell and collaborators studies are very interesting in this regard, as they have involved a significant number of speakers and have used similar methods to alter formants to run the adaptation and to analyze the data. A number of questions should be taken into consideration when designing and reporting studies. Some of them may also require further methodological studies, in line with Munhall and collaborators work. For instance:

- Should the participants be only females or males? What is the effect of mixing vs. not mixing gender on adaptation?
- This first question could be crossed with the effect of the type of perturbation: should the perturbation be absolute vs. relative, formant values being clearly different across gender? What is the effect of shifting only F1 vs. F1 and F2 in opposite direction?
- Whether participants are monolingual or multilingual should be controlled and reported, and as far as possible kept available for meta-analysis. Indeed, adaptation seems influenced by perceptual categories, which are related to phonological systems of languages. One of the best ways to address the question would be to be able to compare large datasets recorded around the world in the different research topics.
- What is the real effect of bone conduction on adaptation? This question has not been addressed systematically, although it has been considered in the conception of apparatuses to shift formants. Most studies used

quite high sound intensity of feedback and/or mixed the signal with noise. The effect of the feedback level, the signal to noise ratio as well as the type of noise on adaptation were not systematically reported.

- What is the real effect of the perturbation on the signal heard by speakers? This question is rarely investigated in papers, while the obtained perturbation can be far from the expected one (Mitsuya et al., 2015). In particular, when using existing packages such as Audapter, delay in feedback should be checked, as it could depend on the properties of the OS and computer hardware. The evaluation of formants provided by the tool, especially when applying unusual shifts, should be verified as there is no guarantee that the system will be able to track and shift the formants in the expected way. This is true for all the systems and could be easily verified by comparing the obtained formant values with corresponding spectrograms or with values assessed by an independent formant assessment software. This approach was used in Reilly and Pettibone (2017).
- Due to the high variability in adaptation magnitude between participants, apparent differences on some parameters of adaptation between conditions are often found to be non-significant. Some effects and, in particular, surface effects such as visual vs. audio prompts might exist but may require testing a large number of speakers to reach significance. This could also be the case for effects related to the direction of the perturbation or to the number of trials during the hold phase as well as to the way the perturbation is introduced. At the very least, non-significant results between different groups of speakers should be interpreted carefully, in relation to this large variability.

These examples suggest that methodological aspects should be directly addressed and clearly reported to help teams working in the field share standards and enable the constitution of large databases. Large between-subjects variability suggests that adaptation to formant perturbations is a complex phenomenon, influenced by different factors. Multifactorial analyses such as introduced in Dimov et al. (2012) could be run on large datasets, but this requires – at the very least – recording of systematic information about the participants and reporting clear information about the perturbation and its real effect.

6.2. Topics which will benefit from further investigation

Due to the broad range of research topics addressed by formant perturbation studies, more studies are still required to reproduce or better understand some results. This is particularly the case for the effect of adaptation on categorical perception, as results between studies have been sometimes inconsistent. Only a few studies were published on the effect of adaptation to formant perturbations on categorical perception of speech (Lametti et al., 2014; Schuerman et al., 2015; Schuerman et al., 2017; Schuerman, Nagarajan, et al., 2017) with some inconsistent findings between Lametti et al. (2014) and Schuerman et al. (2017). The two studies were run with speakers of different languages (English vs. Dutch) and with different types of continua for the perceptual test (words vs. vowels). It would be useful to gain more awareness of other attempts with non-significant or inconsistent profiles of perceptual changes following adaptation if any exist. This will avoid a publication bias towards significant-only results that seems to be a sensitive topic for this research question, in particular as the effects of speech production on changes of categorical boundaries may be sensitive to numerous variables, including the number of speakers, their gender, regional accent, languages skills etc. Replication is also required as the involvement of the motor system in perception is an important challenge for speech research more generally.

Investigating the development of feedback and feedforward control systems and their potential interaction in typically developing children is also an important topic to further develop using formant perturbations paradigm. Moreover, using compensation to unpredictable perturbations in conjunction with sustained perturbations in atypical speakers may shed light on the root causes of some pathologies affecting speech production. For instance, van den Bunt et al. (2017) provides a rather convincing explanation about the sensorimotor bases of dyslexia, which could be further investigated in children. As adaptation has been shown to interact with phoneme categories, it allows investigating the development of phonological categories in both typical children and children with phonological disorders.

An important topic also under-investigated so far is the influence of extraneous factors (i.e. not directly related to language or speech) on

auditory-motor adaptation. First results by Munhall et al. (2009) suggested that the magnitude of the compensation is relatively independent of the awareness of the experimental aim and that speakers compensate even when asked not to compensate. This suggests that adaptation is quite independent from higher cognitive functions such as attention. Martin et al. (2018) also found no significant contribution of general executive control skills on adaptation. However, the preliminary work by Dimov et al. (2012) suggests that variables related to speakers' social status may play a role. Further investigations linking working memory abilities, attention levels etc. to formant adaptation will help tackle the mainstream issue of the link between cognitive and sensorimotor functions. This topic, as a number of others, has already been investigated in adaptation or compensation to F0 perturbation (Guo et al., 2017; Hu et al., 2015; Scheerer, Tumber, & Jones, 2015). Last but not least, results by MacDonald et al. (2012) showing that toddlers do not adapt and the associated discussion of this result by Messum and Howard (2012) suggest that the communicative context may also influence adaptation. The question was investigated in birds by Sakata and Brainard (2009) suggesting larger adaptation when the song is produced in presence of another bird but also in humans with other type of perturbations such as speech in noise (Garnier et al., 2010). Social context might thus be relevant to question the real nature of speech targets.

An important topic not developed in this chapter is a systematic analysis of the results of formant perturbation studies in relation to current models of speech production. A joint analysis with the results of other auditory and somatosensory perturbation studies could improve our understanding of feedback and feedforward controls.

Finally, as it is relevant to the link between learning and memory, we would like to emphasize that transfer of adaptation was under-studied so far, despite its potential to bring insight into the nature of speech representations. As already introduced in Houde and Jordan (1998), transfer of learning is an empirical tool to question the nature of speech production units. This approach should be better connected to the equivalent approach developed for perceptual learning (e.g. Chambers et al., 2010). As noted by Cai et al. (2010) patterns of transfer question the way models of speech production represent sensorimotor mapping: both significant

generalization effect, as well as gradient effects should be explained. These models should also be adapted to integrate results from transfer or multiple adaptation studies suggesting that the mapping between auditory and articulatory domains could occur at different linguistic levels and be related in some way to the training word. But more generally, adaptation might be related to the episode of learning, as also discussed in Houde and Jordan (2002) when explaining the long-term effects of adaptation by implicit memory. We strongly believe that understanding the link between sensorimotor learning and memory would be a fruitful path towards understanding of embodied cognition and the links between language and speech. In any case, identifying the condition of specificity vs. generalization of adaptation will clearly contribute to the debate on the nature of speech production representation and to the debate on the nature of internal models and their relation to sensorimotor memories.

7. Conclusion

Twenty years ago, Houde and Jordan introduced formant perturbations in auditory feedback as a new paradigm to explore speech production. This seminal study is cited by papers in various domains: speech production and perception in general, studies using other kind of perturbations related to speech (e.g. pitch alteration, vocal tract perturbation), motor control as well as vocalizations in animals. Moreover, it has inspired a whole research field which is still in expansion. In this review, we scanned all studies citing Houde and Jordan (1998, 2002) and selected 77 articles focused on formant perturbations. The perturbation systems designed for this purpose are reported and described in the review. The main research topics addressed in these studies are also explained, along with their main findings.

The formant perturbation paradigm proved to be insightful in exploring the relationship between speech production and perception. First, the observation of responses to auditory perturbations has shed light on the role of auditory feedback in speech production, and the mechanisms that control it. Experimental findings have been incorporated in speech production models, although some results still need to be modeled. Altering both auditory and somatosensory feedback showed that both modalities

are integrated in the control of speech in a manner that may be specific to the speaker and/or to the task (e.g. the vowel to produce). Associating perceptual categorization tasks and training with formant perturbations revealed a close relationship and mutual influence between speech motor control and phonological categories, mediated by categorical perception.

This relationship between motor control and linguistic units (e.g. phonemes or syllables) has also been explored by observing the generalization of auditory-motor adaptation. Generalization, or transfer, has been observed from a vowel to the same vowel in different words, suggesting the existence of an underlying phoneme representation. While transfer may occur from one vowel to another, supporting the idea of broad generalization in speech learning, the magnitude of transfer seems to depend on some similarity relationship between the training and the transfer utterances. Moreover, simultaneous adaptation to opposite perturbations has been observed in two different vowels and even in the same vowel in different words. This apparent contradiction may represent a challenge for speech production models, as it requires much flexibility in the translation of auditory goals into articulatory gestures, and questions the nature of mental representations interfacing with speech articulation.

Studies in cognitive neurosciences have pinpointed neural correlates of sensory integration and motor control in speech production, in terms of brain regions as well as communication networks and frequency bands. While studying patients with cerebellar degeneration also contributes to this purpose, research in other pathologies, including stuttering, Parkinson's disease, dyslexia, developmental speech disorders, and some autism subtypes, have benefited from formant perturbation experiments in understanding of the main causes and mechanisms underlying these specific disorders. Finally, studying compensation and adaptation in children gives insights in the development of sensorimotor processes at stake in speech production. Effects of communicative situation or social context may also be explored, as it has proven influential in some speech motor control characteristics in adults. Further investigations in children in various communicative contexts could eventually shed light on one of the most intriguing questions in our research field: how does a child learn to speak?

Beyond these core topics associated with Houde and Jordan's paradigm, other questions have emerged in relation to speakers' cognitive functions and social characteristics, as well as learning context. The prompt has been suggested to influence adaptation and transfer pattern. Moreover, Houde and Jordan had already noticed that adaptation was still there when speakers were tested one month later with normal feedback. This observation may suggest that learning is to some extent specific to the context in which it occurs, the testing room for instance. This is consistent with multiple-trace memory models or exemplar-based views (Goldinger, 1998; Hintzman, 1986), according to which each event is recorded in the brain in the form of a trace combining multiples elements from sensory inputs. Being confronted with one of these elements may activate all the traces containing it, and therefore the other elements associated with it. Thus, the specific context of the testing room may reactivate the adaptation that had washed out in other contexts. Investigating retention of adaptation in various time ranges and contexts may pave the way to fruitful research exploring the relationship between speech, learning and memory.

Acknowledgements

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Program (FP7/2007–2013 Grant Agreement no. 339152).

References

Aichert, I., & Ziegler, W. (2013). Segments and syllables in the treatment of apraxia of speech: An investigation of learning and transfer effects. *Aphasiology*, *27*(10), 1180–1199.

Alsius, A., Mitsuya, T., Latif, N., & Munhall, K. G. (2017). Linguistic initiation signals increase auditory feedback error correction. *The Journal of the Acoustical Society of America*, *142*(2), 838–845.

Ans, B., Carbonnel, S., & Valdois, S. (1998). A connectionist multiple-trace memory model for polysyllabic word reading. *Psychological Review*, *105*(4), 678–723.

Berry, J.J., Jaeger, I.V., Wiedenhoeft, M., Bernal, B.A., & Johnson, M.T. (2014). Consonant context effects on vowel sensorimotor adaptation.

In: ISCA (eds.): *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, (pp. 2006–2010).

Berry, J.J., North, C., & Johnson, M.T. (2014). Sensorimotor adaptation of speech using real-time articulatory resynthesis. In *IEEE Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (pp. 3196–3200).

Berry, J.J., North, C., Meyers, B., & Johnson, M.T. (2013). Speech sensorimotor learning through a virtual vocal tract. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, p. 060099). ASA.

Bourguignon, N.J., Baum, S.R., & Shiller, D.M. (2014). Lexical-perceptual integration influences sensorimotor adaptation in speech. *Frontiers in Human Neuroscience*, *8*, 208.

Bourguignon, N.J., Baum, S.R., & Shiller, D.M. (2015). Extrinsic talker normalization alters self-perception during speech. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith (Eds.), *Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS 2015)*. London: International Phonetic Association.

Bourguignon, N.J., Baum, S.R., & Shiller, D.M. (2016). Please say what this word is—Vowel-extrinsic normalization in the sensorimotor control of speech. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(7), 1039–1047.

Brainard, M.S., & Doupe, A.J. (2000). Auditory feedback in learning and maintenance of vocal behaviour. *Nature Reviews Neuroscience*, *1*(1), 31–40.

Brumberg, J.S., Krusienski, D.J., Chakrabarti, S., Gunduz, A., Brunner, P., Ritaccio, A.L., & Schalk, G. (2016). Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task. *Plos One*, *11*(11), e0166872.

Burnett, T.A., & Larson, C.R. (2002). Early pitch-shift response is active in both steady and dynamic voice pitch control. *The Journal of the Acoustical Society of America*, *112*(3), 1058–1063.

Cai, S., Beal, D.S., Ghosh, S.S., Tiede, M.K., Guenther, F.H., & Perkell, J.S. (2012). Weak responses to auditory feedback perturbation during articulation in persons who stutter: evidence for abnormal auditory-motor transformation. *Plos One*, *7*(7), e41830.

Cai, S., Boucek, M., Ghosh, S.S., Guenther, F.H., & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong/iau. *Proceedings of the 8th International Seminar on Speech Production (ISSP)*, (pp 65–68).

Cai, S., Ghosh, S.S., Guenther, F.H., & Perkell, J.S. (2010). Adaptive auditory feedback control of the production of formant trajectories in the Mandarin triphthong/iau/and its pattern of generalization. *The Journal of the Acoustical Society of America*, *128*(4), 2033–2048.

Cai, S., Ghosh, S.S., Guenther, F.H., & Perkell, J.S. (2011). Focal manipulations of formant trajectories reveal a role of auditory feedback in the online control of both within-syllable and between-syllable speech timing. *Journal of Neuroscience*, *31*(45), 16483–16490.

Carbonnel, S., Charnallet, A., & Moreaud, O. (2010). Organisation des connaissances sémantiques: des modèles classiques aux modèles non abstractifs. *Revue de Neuropsychologie*, *2*(1), 22–30.

Casserly, E.D. (2015). Effects of real-time cochlear implant simulation on speech production. *The Journal of the Acoustical Society of America*, *137*(5), 2791–2800.

Caudrelier, T., Perrier, P., Schwartz, J., Rochet-Capellan, A. (2018) Picture naming or word reading: Does the modality affect speech motor adaptation and its transfer? *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, (pp. 956–960).

Caudrelier, T., Perrier, P., Schwartz, J.-L., & Rochet-Capellan, A. (2016). Does auditory-motor learning of speech transfer from the CV syllable to the CVCV word? *Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, (pp. 2095–2099).

Caudrelier, T., Schwartz, J.-L., Perrier, P., Gerber, S., & Rochet-Capellan, A. (2018). Transfer of learning: What does it tell us about speech production units? *Journal of Speech, Language, and Hearing Research*, 61(7), 1613–1625.

Chambers, K.E., Onishi, K.H., & Fisher, C. (2010). A vowel is a vowel: Generalizing newly learned phonotactic constraints to new contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(3), 821–828.

Chon, H., Kraft, S.J., Zhang, J., Loucks, T., & Ambrose, N G. (2013). Individual variability in delayed auditory feedback effects on speech fluency and rate in normally fluent adults. *Journal of Speech, Language, and Hearing Research*, *56*(2), 489–504.

Curio, G., Neuloh, G., Numminen, J., Jousmäki, V., & Hari, R. (2000). Speaking modifies voice-evoked activity in the human auditory cortex. *Human brain mapping*, *9*(4), 183–191.

Daliri, A., Wieland, E.A., Cai, S., Guenther, F.H., & Chang, S.-E. (2018). Auditory-motor adaptation is reduced in adults who stutter but not in children who stutter. *Developmental science*, *21*(2), e12521.

de Bruijn, M.J., ten Bosch, L., Kuik, D.J., Witte, B.I., Langendijk, J.A., Leemans, C.R., & Verdonck-de Leeuw, I.M. (2012). Acoustic-phonetic and artificial neural network feature analysis to assess speech quality of stop consonants produced by patients treated for oral or oropharyngeal cancer. *Speech Communication*, *54*(5), 632–640.

Demopoulos, C., Kothare, H., Mizuiri, D., Henderson-Sabes, J., Fregeau, B., Tjernagel, J., Houde, J.F., Sherr, E.H., & Nagarajan, S. S. (2018). Abnormal speech motor control in individuals with 16p11.2 deletions. *Scientific Reports*, *8*(1), 1274.

Deroche, M.L., Nguyen, D., & Gracco, V.L. (2017). Modulation of speech motor learning with transcranial direct current stimulation of the inferior parietal lobe. *Frontiers in Integrative Neuroscience*, *11*, 35.

Dimov, S., Katseff, S., & Johnson, K. (2012). Social and personality variables in compensation for altered auditory feedback. *UC Berkeley PhonLab Annual Report*, *(6)*6, pp. 259–282.

Doupe, A.J., & Kuhl, P.K. (1999). Birdsong and human speech: common themes and mechanisms. *Annual Review of Neuroscience*, *22*(1), 567–631.

Eckey, A., & MacDonald, E. (2015). Compensations of F0 and formant frequencies in a real-time pitch-perturbation paradigm. *Forschritte der Akustik DAGA'15*, (pp. 1444–1447).

Eliades, S.J., & Miller, C.T. (2017). Marmoset vocal communication: Behavior and neurobiology. *Developmental Neurobiology*, *77*(3), 286–299.

Feng, Y., Gracco, V.L., & Max, L. (2011). Integration of auditory and somatosensory error signals in the neural control of speech movements. *Journal of Neurophysiology*, *106*(2), 667–679.

Frank, A.F. (2011). *Integrating linguistic, motor, and perceptual information in language production*. Doctoral dissertation, University of Rochester.

Garnier, M., Henrich, N., & Dubois, D. (2010). Influence of sound immersion and communicative interaction on the Lombard effect. *Journal of Speech, Language, and Hearing Research*, *53*(3), 588–608.

Goldinger, S.D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Golfinopoulos, E., Tourville, J.A., & Guenther, F.H. (2010). The integration of large-scale neural network modeling and functional brain imaging in speech motor control. *Neuroimage*, *52*(3), 862–874.

Grimme, B., Fuchs, S., Perrier, P., & Schöner, G. (2011). Limb versus speech motor control: A conceptual review. *Motor Control*, *15*(1), 5–33.

Guo, Z., Wu, X., Li, W., Jones, J.A., Yan, N., Sheft, S., Liu, P., & Liu, H. (2017). Top-down modulation of auditory-motor integration during speech production: The role of working memory. *Journal of Neuroscience*, *37*(43), 10323–10333.

Held, R. (1965). Plasticity in sensory-motor systems. *Scientific American*, *213*(5), 84–97.

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67–99.

Hintzman, D. L. (1986). «Schema abstraction» in a multiple-trace memory model. *Psychological Review*, *93*(4), 411–428.

Houde, J.F. (1997). *Sensorimotor adaptation in speech production*. Doctoral dissertation, MIT.

Houde, J.F., & Chang, E.F. (2015). The cortical computations underlying feedback control in vocal production. *Current Opinion in Neurobiology*, *33*, 174–181.

Houde, J.F., & Jordan, M.I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216.

Houde, J.F., & Jordan, M.I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research*, *45*(2), 295–310.

Hu, H., Liu, Y., Guo, Z., Li, W., Liu, P., Chen, S., & Liu, H. (2015). Attention modulates cortical processing of pitch feedback errors in voice control. *Scientific Reports*, *5*, 7812.

Hubl, D., Schneider, R. C., Kottlow, M., Kindler, J., Strik, W., Dierks, T., & Koenig, T. (2014). Agency and ownership are independent components of 'sensing the self'in the auditory-verbal domain. *Brain topography*, *27*(5), 672–682.

Jones, J.A., & Munhall, K.G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Jones, J.A., & Munhall, K.G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *The Journal of the Acoustical Society of America*, *113*(1), 532–543.

Katseff, S., & Houde, J. (2008). Partial compensation in speech adaptation. *UC Berkeley Phonology Lab Annual Reports*, *4*(4), (pp. 445–461).

Katseff, S., Houde, J., & Johnson, K. (2012). Partial compensation for altered auditory feedback: A trade-off with somatosensory feedback? *Language and Speech*, *55*(2), 295–308.

Kelso, J.S., Tuller, B., Vatikiotis-Bateson, E., & Fowler, C.A. (1984). Functionally specific articulatory cooperation following jaw perturbations during speech: evidence for coordinative structures. *Journal of Experimental Psychology: Human Perception and Performance*, *10*(6), 812–832.

Klein, E., Brunner, J., & Hoole, P. (2019). Spatial and temporal variability of corrective speech movements as revealed by vowel formants during sensorimotor learning. In S. Fuchs, J. Cleland & A. Rochet-Capellan (eds.) Speech production and perception: Learning and memory. Peter Lang Publisher (*current book)*.

Krakauer, J.W., Mazzoni, P., Ghazizadeh, A., Ravindran, R., & Shadmehr, R. (2006). Generalization of motor learning depends on the history of prior action. *PLoS biology*, *4*(10), e316.

Krakauer, J.W., Pine, Z.M., Ghilardi, M.-F., & Ghez, C. (2000). Learning of visuomotor transformations for vectorial planning of reaching trajectories. *Journal of Neuroscience*, *20*(23), 8916–8924.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society*, *363*, 979–1000.

Lametti, D.R., Krol, S.A., Shiller, D.M., & Ostry, D.J. (2014). Brief periods of auditory perceptual training can determine the sensory targets of speech motor learning: *Psychological Science*, *25*(7), 1325–1336.

Lametti, D.R., Nasir, S.M., & Ostry, D.J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience*, *32*(27), 9351–9358.

Lametti, D.R., Rochet-Capellan, A., Neufeld, E., Shiller, D.M., & Ostry, D.J. (2014). Plasticity in the human speech motor system drives changes in speech perception. *Journal of Neuroscience*, *34*(31), 10339–10346.

Lametti, D.R., Smith, H.J., Freidin, P.F., & Watkins, K.E. (2018). Cortico-cerebellar networks drive sensorimotor learning in speech. *Journal of Cognitive Neuroscience*, *30*(4), 540–551.

Lane, H., Matthies, M.L., Guenther, F.H., Denny, M., Perkell, J.S., Stockmann, E., Tiede, M., & Zandipour, M. (2007). Effects of short-and long-term changes in auditory feedback on vowel and sibilant contrasts. *Journal of Speech, Language, and Hearing Research*, *50*(4), 913–927.

Li, W., Chen, Z., Yan, N., Jones, J. A., Guo, Z., Huang, X., Cheng, S., Liu, P., & Liu, H. (2016). Temporal lobe epilepsy alters auditory-motor integration for voice control. *Scientific reports*, *6*, 28909.

Maas, E., Mailend, M.-L., & Guenther, F.H. (2015). Feedforward and feedback control in apraxia of speech: Effects of noise masking on vowel production. *Journal of Speech, Language, and Hearing Research*, *58*(2), 185–200.

Maas, E., Robin, D.A., Hula, S.N.A., Freedman, S.E., Wulf, G., Ballard, K.J., & Schmidt, R.A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277–298.

MacDonald, E.N., Goldberg, R., & Munhall, K.G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *The Journal of the Acoustical Society of America*, *127*(2), 1059–1068.

MacDonald, E.N., Johnson, E.K., Forsythe, J., Plante, P., & Munhall, K.G. (2012). Children's development of self-regulation in speech production. *Current Biology*, *22*(2), 113–117.

MacDonald, E.N., & Munhall, K.G. (2012). A preliminary study of individual responses to real-time pitch and formant perturbations. *The Listening Talker: An interdisciplinary workshop on natural and synthetic modification of speech in response to listening conditions*. *2012*, (pp. 32–35).

MacDonald, E.N., Pile, E., Dajani, H., & Munhall, K.G. (2008). The specificity of adaptation to real-time formant shifting. *Proceedings of the International Seminar on Speech Production*, *2008*, (pp. 397–400).

MacDonald, E.N., Purcell, D.W., & Munhall, K.G. (2011). Probing the independence of formant control using altered auditory feedback. *The Journal of the Acoustical Society of America*, *129*(2), 955–965.

Martin, C.D., Niziolek, C.A., Duñabeitia, J.A., Perez, A., Hernandez, D., Carreiras, M., & Houde, J.F. (2018). Online adaptation to altered auditory feedback is predicted by auditory acuity and not by domain-general executive control resources. *Frontiers in Human Neuroscience*, *12*, 91.

Mattar, A.A., & Ostry, D.J. (2007). Modifiability of generalization in dynamics learning. *Journal of Neurophysiology*, *98*(6), 3321–3329.

Max, L., & Maffett, D.G. (2015). Feedback delays eliminate auditory-motor learning in speech production. *Neuroscience letters*, *591*, 25–29.

Max, L., Wallace, M.E., & Vincent, I. (2003). Sensorimotor adaptation to auditory perturbations during speech: Acoustic and kinematic experiments. *Proceedings of the 15th International Congress of Phonetic Sciences*, Futurgraphic Barcelona, Spain, (pp. 1053–1056).

Ménard, L., Perrier, P., & Aubin, J. (2016). Compensation for a lip-tube perturbation in 4-year-olds: Articulatory, acoustic, and perceptual data analyzed in comparison with adults. *The Journal of the Acoustical Society of America*, *139*(5), 2514–2531.

Messum, P., & Howard, I.S. (2012). Speech development: Toddlers don't mind getting it wrong. *Current Biology*, *22*(5), R160–R161.

Mitsuya, T., MacDonald, E.N., & Munhall, K.G. (2014). Temporal control and compensation for perturbed voicing feedback. *The Journal of the Acoustical Society of America*, *135*(5), 2986–2994.

Mitsuya, T., MacDonald, E.N., Munhall, K.G., & Purcell, D.W. (2015). Formant compensation for auditory feedback with English vowels. *The Journal of the Acoustical Society of America*, *138*(1), 413–424.

Mitsuya, T., MacDonald, E.N., Purcell, D.W., & Munhall, K.G. (2011). A cross-language study of compensation in response to real-time formant perturbation. *The Journal of the Acoustical Society of America*, *130*(5), 2978–2986.

Mitsuya, T., Munhall, K.G., & Purcell, D.W. (2017). Modulation of auditory-motor learning in response to formant perturbation as a function of delayed auditory feedback. *The Journal of the Acoustical Society of America*, *141*(4), 2758–2767.

Mitsuya, T., & Purcell, D.W. (2016). Occlusion effect on compensatory formant production and voice amplitude in response to real-time perturbation. *The Journal of the Acoustical Society of America*, *140*(6), 4017–4026.

Mitsuya, T., Samson, F., Ménard, L., & Munhall, K.G. (2013). Language dependent vowel representation in speech production. *The Journal of the Acoustical Society of America*, *133*(5), 2993–3003.

Mollaei, F., Shiller, D.M., Baum, S.R., & Gracco, V.L. (2016). Sensorimotor control of vocal pitch and formant frequencies in Parkinson's disease. *Brain Research*, *1646*, 269–277.

Mollaei, F., Shiller, D.M., & Gracco, V.L. (2013). Sensorimotor adaptation of speech in Parkinson's disease. *Movement Disorders*, *28*(12), 1668–1674.

Munhall, K.G., MacDonald, E.N., Byrne, S.K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, *125*(1), 384–390.

Neufeld, C., Purcell, D., & Van Lieshout, P. (2013). Articulatory compensation to second formant perturbations. *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, p. 060097). ASA.

Niziolek, C.A., & Guenther, F.H. (2013). Vowel category boundaries enhance cortical and behavioral responses to speech feedback alterations. *Journal of Neuroscience*, *33*(29), 12090–12098.

Osu, R., Hirai, S., Yoshioka, T., & Kawato, M. (2004). Random presentation enables subjects to adapt to two opposing forces on the hand. *Nature Neuroscience*, *7*(2), 111–112.

Palethorpe, S., Watson, C.I., & Barker, R. (2003). Acoustic analysis of monophthong and diphthong production in acquired severe to profound hearing loss. *The Journal of the Acoustical Society of America*, *114*(2), 1055–1068.

Pardo, J.S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, *119*(4), 2382–2393.

Parrell, B., Agnew, Z., Nagarajan, S., Houde, J., & Ivry, R.B. (2017). Impaired feedforward control and enhanced feedback control of speech in patients with cerebellar degeneration. *Journal of Neuroscience*, *37*(38), 9249–9258.

Patri, J.-F., Perrier, P., Schwartz, J.-L., & Diard, J. (2018). What drives the perceptual change resulting from speech motor adaptation? Evaluation of hypotheses in a Bayesian modeling framework. *PLoS Computational Biology*, *14*(1), e1005942.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, *116*(4), 2338–2344.

Perrier, P. (2012). Gesture planning integrating knowledge of the motor plant's dynamics: A literature review from motor control and speech motor control. In S. Fuchs, M. Weirich, D. Pape & P. Perrier (eds.). *Speech Planning and Dynamics*, Peter Lang Publishers, pp.191–238.

Pfordresher, P.Q., & Palmer, C. (2006). Effects of hearing the past, present, or future during music performance. *Attention, Perception, & Psychophysics*, *68*(3), 362–376.

Pile, E.J.S., Dajani, H.R., Purcell, D.W., & Munhall, K.G. (2007). Talking under conditions of altered auditory feedback: does adaptation of one vowel generalize to other vowels. *Proceedings of the International Congress of Phonetic Sciences* (pp. 645–648).

Purcell, D.W., & Munhall, K.G. (2008). Weighting of auditory feedback across the English vowel space. *Proceedings of the International Seminar on Speech Production* (Vol. 8, p. 389–392).

Purcell, D.W., & Munhall, K.G. (2006a). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, *120*(2), 966–977.

Purcell, D.W., & Munhall, K.G. (2006b). Compensation following real-time manipulation of formants in isolated vowels. *The Journal of the Acoustical Society of America*, *119*(4), 2288–2297.

Reilly, K.J., & Dougherty, K.E. (2013). The role of vowel perceptual cues in compensatory responses to perturbations of speech auditory feedback. *The Journal of the Acoustical Society of America*, *134*(2), 1314–1323.

Reilly, K.J., & Pettibone, C. (2017). Vowel generalization and its relation to adaptation during perturbations of auditory feedback. *Journal of Neurophysiology*, *118*(5), 2925–2934.

Rochet-Capellan, A., & Ostry, D.J. (2011). Simultaneous acquisition of multiple auditory–motor transformations in speech. *Journal of Neuroscience*, *31*(7), 2657–2662.

Rochet-Capellan, A., Richer, L., & Ostry, D.J. (2012). Nonhomogeneous transfer reveals specificity in speech motor learning. *Journal of Neurophysiology*, *107*(6), 1711–1717.

Sakata, J.T., & Brainard, M.S. (2009). Social context rapidly modulates the influence of auditory feedback on avian vocal motor control. *Journal of Neurophysiology*, *102*(4), 2485–2497.

Sato, M., & Shiller, D.M. (2018). Auditory prediction during speaking and listening. *Brain and Language*, *187*, 92–103.

Sato, M., Troille, E., Ménard, L., Cathiard, M.-A., & Gracco, V.L. (2013). Silent articulation modulates auditory and audiovisual speech perception. *Experimental Brain Research*, *227*(2), 275–288.

Scheerer, N.E., Tumber, A.K., & Jones, J.A. (2015). Attentional demands modulate sensorimotor learning induced by persistent exposure to changes in auditory feedback. *Journal of Neurophysiology*, *115*(2), 826–832.

Schuerman, W.L., Meyer, A.S., & McQueen, J.M. (2017). Mapping the speech code: Cortical responses linking the perception and production of vowels. *Frontiers in Human Neuroscience*, *11*, 161.

Schuerman, W.L., Nagarajan, S., & Houde, J. (2015). Changes in consonant perception driven by adaptation of vowel production to

altered auditory feedback. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith (eds.), *Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS 2015)*. London: International Phonetic Association.

Schuerman, W.L., Nagarajan, S., McQueen, J.M., & Houde, J. (2017). Sensorimotor adaptation affects perceptual compensation for coarticulation. *The Journal of the Acoustical Society of America*, *141*(4), 2693–2704.

Sengupta, R., & Nasir, S.M. (2015). Redistribution of neural phase coherence reflects establishment of feedforward map in speech motor adaptation. *Journal of Neurophysiology*, *113*(7), 2471–2479.

Sengupta, R., & Nasir, S.M. (2016). The predictive roles of neural oscillations in speech motor adaptability. *Journal of Neurophysiology*, *115*(5), 2519–2528.

Sengupta, R., Shah, S., Gore, K., Loucks, T., & Nasir, S.M. (2016). Anomaly in neural phase coherence accompanies reduced sensorimotor integration in adults who stutter. *Neuropsychologia*, *93*, 242–250.

Shadmehr, R., & Mussa-Ivaldi, F.A. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, *14*(5), 3208–3224.

Shih, T., Suemitsu, A., & Akagi, M. (2011). Influences of transformed auditory feedback with first three formant frequencies. *International Workshop on Nonlinear Circuits, Communication and Signal Processing (NCSP'11)*.

Shiller, D.M., Lametti, D., & Ostry, D.J. (2013). Auditory plasticity and sensorimotor learning in speech production. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, p. 060150). ASA.

Shiller, D.M., & Rochon, M.-L. (2014). Auditory-perceptual learning improves speech motor adaptation in children. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1308–1315.

Shiller, D.M., Rvachew, S., & Brosseau-Lapré, F. (2010). Importance of the auditory perceptual target to the achievement of speech production accuracy. *Canadian Journal of Speech-Language Pathology & Audiology*, *34*(3), 181–192.

Shiller, D.M., Sato, M., Gracco, V.L., & Baum, S.R. (2009). Perceptual recalibration of speech sounds following speech motor learning. *The Journal of the Acoustical Society of America*, *125*(2), 1103–1113.

Shum, M., Shiller, D.M., Baum, S.R., & Gracco, V.L. (2011). Sensorimotor integration for speech motor learning involves the inferior parietal cortex. *European Journal of Neuroscience*, *34*(11), 1817–1822.

Smotherman, M., Zhang, S., & Metzner, W. (2003). A neural basis for auditory feedback control of vocal pitch. *Journal of Neuroscience*, *23*(4), 1464–1477.

Sober, S.J., & Brainard, M.S. (2009). Adult birdsong is actively maintained by error correction. *Nature Neuroscience*, *12*(7), 927–931.

Stratton, G.M. (1897). Vision without inversion of the retinal image. *Psychological Review*, *4*(4), 341–360.

Terband, H., & Van Brenk, F. (2015). Compensatory and adaptive responses to real-time formant shifts in adults and children. In M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith (eds.), *Proceedings of the 18th International Congresses of Phonetic Sciences (ICPhS 2015)*. London: International Phonetic Association.

Terband, H., Van Brenk, F., & van Doornik-van der Zee, A. (2014). Auditory feedback perturbation in children with developmental speech sound disorders. *Journal of Communication Disorders*, *51*, 64–77.

Thibeault, M., Ménard, L., Baum, S.R., Richard, G., & McFarland, D.H. (2011). Articulatory and acoustic adaptation to palatal perturbation. *The Journal of the Acoustical Society of America*, *129*(4), 2112–2120.

Tourville, J.A., Cai, S., & Guenther, F.H. (2013). Exploring auditory-motor interactions in normal and disordered speech. *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, p. 060180). ASA.

Tourville, J.A., Reilly, K.J., & Guenther, F.H. (2008). Neural mechanisms underlying auditory feedback control of speech. *Neuroimage*, *39*(3), 1429–1443.

Tremblay, S., Houle, G., & Ostry, D.J. (2008). Specificity of speech motor learning. *Journal of Neuroscience*, *28*(10), 2426–2434.

Tremblay, S., Shiller, D.M., & Ostry, D.J. (2003). Somatosensory basis of speech production. *Nature*, *423*(6942), 866–869.

Trudeau-Fisette, P., Tiede, M., & Ménard, L. (2017). Compensations to auditory feedback perturbations in congenitally blind and sighted speakers: Acoustic and articulatory data. *Plos One*, *12*(7), e0180300.

van den Bunt, M.R., Groen, M.A., Ito, T., Francisco, A.A., Gracco, V.L., Pugh, K.R., & Verhoeven, L. (2017). Increased response to altered auditory feedback in dyslexia: A weaker sensorimotor magnet implied in the phonological deficit. *Journal of Speech, Language, and Hearing Research*, *60*(3), 654–667.

Van Vugt, F. T., & Ostry, D. J. (2018). The structure and acquisition of sensorimotor maps. *Journal of Cognitive Neuroscience*, 30(3), 290–306.

Vaughn, C., & Nasir, S.M. (2015). Precise feedback control underlies sensorimotor learning in speech. *Journal of Neurophysiology*, *113*(3), 950–955.

Villacorta, V.M., Perkell, J.S., & Guenther, F.H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319.

Wei, K., Yan, X., Kong, G., Yin, C., Zhang, F., Wang, Q., & Kording, K.P. (2014). Computer use changes generalization of movement learning. *Current Biology*, *24*(1), 82–85.

Wong, S.M., Domangue, R.J., Fels, S., & Ludlow, C.L. (2017). Evidence that an internal schema adapts swallowing to upper airway requirements. *The Journal of Physiology*, *595*(5), 1793–1814.

Yates, A.J. (1963). Delayed auditory feedback. *Psychological Bulletin*, *60*(3), 213–232.

Zheng, Z.Z., Vicente-Grabovetsky, A., MacDonald, E.N., Munhall, K.G., Cusack, R., & Johnsrude, I.S. (2013). Multivoxel patterns reveal functionally differentiated networks underlying auditory feedback processing of speech. *Journal of Neuroscience*, *33*(10), 4339–4348.

Eugen Klein, Jana Brunner, Phil Hoole

Spatial and temporal variability of corrective speech movements as revealed by vowel formants during sensorimotor learning

Abstract: Previous perturbation studies demonstrate that speakers can reorganize their motor strategies to adapt for articulatory or auditory perturbations (Savariaux, Perrier & Orliaguet, 1995; Rochet-Capellan & Ostry, 2011). However, across most studies we observe a fluctuating amount of inter-individual differences with respect to the adaptation outcome. To evaluate the predictions of the hypotheses put forward to explain these differences, we conducted a multidirectional auditory perturbation study investigating F2 perturbation with native Russian speakers. During participants' production of CV syllables containing the close central unrounded vowel /ɨ/, F2 was perturbed in opposing directions depending on the preceding consonant (/d/ or /g/). The bidirectional shift was intended to encourage participants to produce the vowel /ɨ/ with two different motor strategies and allowed us to investigate intra-individual variation of adaptation patterns as a function of the perturbation direction and the consonantal context. To examine the evolution of the adaptation process, we performed generalized additive mixed modelling (GAMM) on the averaged and individual formant data using the experimental trials as discrete time points. In doing so, we were able to examine sudden changes in participants' adaptation strategies, which appeared as nonlinearities in the F2 curve. Our results suggest that previously formulated hypotheses regarding individual adaptation processes make empirical predictions which are not confirmed by the bidirectional perturbation data. Therefore, we propose a more general hypothesis that the successful adaptation is dependent on speakers' ability to coordinate the perceived auditory errors with appropriate compensatory movements, which is influenced in turn by the complexity of the adaptation task. We discuss this hypothesis in the context of individual adaptation patterns and show that it not only can explain the inter-individual, but also the inter-study variability observed in previous perturbation studies.

Keywords: auditory feedback, real-time perturbations, formants, variability, individual behavior, generalized additive mixed modelling, Russian

1. Introduction

1.1. Perturbation and sensorimotor learning

Picture the situation of taking a photo of beautiful lakeside scenery and accidently dropping your camera into the water. Despite your misfortune, you are lucky and can spot the camera within what appears to be a reachable distance at the lake bottom. Hastily, you try to retrieve the camera but grab a few times beside it before you can actually take hold of it. Or even worse, you realize that the bottom that appeared reachable lies in fact much deeper below the water surface. In the described example, the coordination between your visual input and your hand movements is disrupted by the visual distortions caused by the different reflective angles between the water and the air. The fact that you can eventually grab the camera after a few attempts, assuming the lake bottom is indeed reachable by hand, provides evidence for the flexibility of the human sensorimotor system which is able to adapt for the visual perturbations and to find alternative motor strategies to reach the intended goal.

The same is mostly true for mechanical and auditory perturbations of speech. That is, when you later recall you tale of bad luck to a friend during the conference dinner, and you get upset about the unreasonable repair costs of your camera, you might speak with a mouth full of food. In this case, your articulators' movements might be impeded by pieces of food which will force you to find alternative strategies to intelligibly articulate the words you intend to utter. Or, in another scenario, you may have to increase the loudness of your voice to compensate for the loud conversation happening at the table next to yours.

During experiments applying controlled perturbation, speakers have to produce speech under aggravated conditions, e.g., under blockage of their jaw movements or under altered auditory feedback. As in the initial example with the hand-eye coordination, speakers need to coordinate errors transmitted by their sensory input with appropriate corrective articulator movements to be able to retain intelligibility of their speech. In the case of speech, it is particularly intriguing which sensory channels (e.g., somatosensory, proprioceptive, or auditory) are involved in the process of adaptation. The answer to this question may provide a better understanding of the different types of sensory information relevant for speech

production and ultimately the goals of articulator movements. Thus, the study of perturbed speech provides an empirical means to study the nature of speech sound representations as well as learning processes that occur in speech production.

1.2. Outcome variability in perturbation studies of speech

Despite the general ability of speakers to reorganize their motor strategies to retain the acoustic make-up of the intended speech sounds under aggravated conditions, the outcome of adaptation processes in speech exhibits high inter-individual and inter-study variability. For instance, Gay, Lindblom and Lubker (1981) examined participants' productions of vowels when a bite block was inserted between their teeth. The authors found that speakers were able to adapt to these static perturbations with very little or no practice and produce acoustic outputs equivalent to their unperturbed speech. However, in a study by Savariaux, Perrier and Orliaguet (1995) when speakers' lips were blocked with a tube during the production of the French [u] only six out of 11 speakers were able to partially compensate for the labial perturbation and only one speaker compensated completely by changing the constriction location from a velo-palatal to a velo-pharyngeal region. The remaining four speakers did not compensate at all. Similar variability of the experimental outcomes is also observed across other articulatory perturbation studies, e.g., by Baum & McFarland (1997), Jones & Munhall (2003), and Brunner, Hoole & Perrier (2011). To explain this variability, Savariaux et al. (1995) suggest that the varying degree of adaptation among participants is due to "speaker-specific internal representation of articulatory-to-acoustic relationships".

More recently, it has become possible to study speakers' articulatory-to-acoustic relations by means of real-time perturbation of speakers' auditory feedback. This methodology allows alteration of such acoustic parameters as fundamental frequency (f0; Jones & Munhall, 2000) and vowel formants (F1 and/or F2; Houde & Jordan, 1998; Purcell & Munhall 2006; Villacorta, Perkell & Guenther, 2007), and has the advantage that multiple perturbation conditions can be tested within the same study without participants' awareness of any systematic manipulations. For instance, Rochet-Capellan and Ostry (2011) perturbed the first formant (F1) in the

vowel /ɛ/ in opposing directions depending on the experimental stimulus in which it was embedded (*head* or *bed*), while in a control stimulus (*ted*) the F1 remained unchanged throughout the experiment. The authors found that speakers were overall able to adapt for the three distinct F1 levels which means that during the study participants employed three different motor strategies to produce the vowel /ɛ/. However, as with articulatory perturbation studies mentioned above, there is a noteworthy proportion of speakers, ranging from 10 to 20 % per study, who fail to adapt to auditory perturbations. Roughly speaking, these speakers exhibit two qualitatively different types of adaptation behaviors: either adjusting their response in the same direction as the applied perturbation, or hardly reacting to it.

One of the more recent hypotheses put forward to explain the outcome variability observed in perturbation studies is the idea by Lametti, Nasir and Ostry (2012) that speakers have individual preferences for articulatory or auditory feedback to control their speech production. To empirically evaluate their claim, Lametti et al. (2012) investigated participants in different experimental conditions where the authors either perturbed participants' jaw trajectories without altering their speech acoustics, or perturbed their auditory feedback, or applied both types of perturbation simultaneously. The authors found a negative correlation between the amount of articulatory and auditory adaptation which means that speakers who adapted to articulatory perturbations, adapted to auditory alternations to a lesser degree.

However, Lametti et al.'s (2012) hypothesis conflicts with observations previously made by Ghosh et al. (2010) who investigated the relation between somatosensory and auditory acuity, where acuity stands for the degree to which speakers were sensitive to changes in articulatory and auditory feedback signals. Running contrary to the idea that speakers exhibit individual preferences towards auditory or somatosensory feedback, Ghosh et al. (2010) found that both types of acuity positively correlated with each other as well as with the magnitude of produced sibilant contrasts. In the context of vowels, the latter finding was previously made by Perkell et al. (2004). Furthermore, auditory acuity has been shown to have an influence on the adaptation magnitude during auditory perturbation of vowel formants (Villacorta et al., 2007) as well as during articulatory perturbation of sibilants (Brunner, Ghosh, Hoole, Matthies, Tiede

& Perkell, 2011). In contrast to Lametti et al.'s (2012) hypothesis which predicts that speakers who fail to adapt to auditory perturbations should virtually ignore them, individual differences in auditory acuity provide a way to explain partial compensations which are frequently observed in auditory perturbation studies.

Another explanation for partial compensations was provided by Katseff, Houde & Johnson (2012) who suggest that these are the result of speakers' attempts to integrate the altered auditory signal with the normal somatosensory signal that speakers receive during a perturbation experiment. Similar to other authors (e.g., Sato, Schwartz & Perrier, 2014), Katseff et al. (2012) assume that vowel targets are defined as regions in a multidimensional acoustic-somatosensory space. That is, when during auditory perturbation the acoustic parameters of speakers' speech are diverted from the target, speakers will compensate for the acoustic error. However, their compensation will stop when the discrepancy between the auditory and somatosensory signals becomes too large. Katseff et al. (2012) support their view by the observation that in their study of F1 perturbation the relative compensation magnitude decreased from 100 % for 50 Hz perturbations to 40 % for 250 Hz perturbations. An analogous finding was previously made by MacDonald, Goldberg & Munhall (2010) for F1 and F2 perturbation.

At this point, we would like to add that it is alternatively possible that it is not the discrepancy between the altered acoustic and somatosensory signals that is causing the incomplete compensation, but rather physical restrictions which do not allow participants to compensate beyond a certain physical limit. For instance, it seems plausible that large F1 perturbations could require speakers to push their tongue beyond physical limits imposed by the palate, the upper incisors, or other parts of the vocal tract.

1.3. The role of the adaptation task complexity

Although the hypotheses reviewed above are based on different premises, they mostly ascribe the source of the inter-individual outcome variability to the mechanisms of speakers' internal models of speech motor control. This approach leads to a situation in which each of the proposed hypotheses offers a potential explanation for the inter-individual adaptation

variability in the context of a specific perturbation task; however, none of them does actually provide a general account of the variability that is observed across different experimental tasks or conditions. In this specific situation, it appears necessary to investigate the question of whether the complexity of the adaptation task might have an impact on its outcome. Let us illustrate this point with an example.

It is plausible to assume that the adaptation to bite-block perturbation during production of vowels (e.g., Gay et al., 1981) requires an articulatory adjustment that is more similar to the unperturbed condition compared to the case of lip-tube perturbation during the production of /u/ (Savariaux et al., 1995). During the first task, participants are merely required to lift their tongue more strongly than usual since their jaw, which normally assists at this task, is blocked. Furthermore, the direction of the compensatory tongue movement does not change due to the perturbation. During the lip-tube perturbation, on the other hand, participants have to compensate for blocked lip rounding by retracting their tongue. This articulatory adjustment is less obvious as the articulator used to compensate for the perturbation and its movement direction are less associated with the usual articulatory configuration used to produce the intended sound. As a consequence, the adaptation process may take longer and fewer speakers are able to identify the appropriate articulatory adjustments to compensate for the perturbation. Therefore, in our current study we will also investigate the question of whether the outcome variability can be explained by speakers' inability to coordinate the perceived auditory error with appropriate corrective articulatory movements.

1.4. Current study

To investigate in more detail how speakers translate the altered auditory signal into corrective articulatory movements, we conducted a bidirectional auditory perturbation study with native Russian speakers. Unlike Rochet-Capellan & Ostry (2011), who investigated speakers' adaptation to multiple F1 degrees, we focused our investigation on F2, which is, roughly speaking, an indicator of horizontal tongue displacement. In our experiment, participants had to produce the close central unrounded vowel /ɨ/ embedded in CV syllables /dɨ/ and /gɨ/. Depending on the preceding consonant, F2 in /ɨ/ was perturbed in opposing directions.

The bidirectional perturbation imposed higher adaptation demands on our participants since they had to coordinate their corrective movements in two different ways depending on the perturbation direction. Based on the hypothesis that higher task complexity influences the adaptation process, we expected to observe a high amount of exploratory corrective movements and possibly also spontaneous behavior changes in the course of the experiment.

The combination between the place of articulation (alveolar vs. velar) and the perturbation direction (down vs. up) was counterbalanced between all participants which allowed us to control for the potential influence of articulatory restrictions associated with each syllable on the compensation. To investigate how quickly participants can adapt to abrupt and substantial magnitude changes in perturbation, we increased the perturbation amount in 150 Hz steps across three perturbation phases and excluded ramp trials (gradual changes of perturbation magnitude) from the experiment.

Finally, to understand the spatial and temporal evolution of the adaptation process, we analyzed the formant data with generalized additive mixed models (GAMMs) which allowed us to observe non-linear changes in participants' responses to perturbation. By doing this, we seek to overcome the shortcomings of previous perturbation studies which concentrate on the comparison of speakers' performance between the beginning and the end of the adaptation task, i.e., in most extreme cases during the first and the last trial of the experimental session or more often during the first 15–20 and the last 15–20 trials of the experiment. Unfortunately, this aggregation approach allows only for pairwise time-uncorrelated comparisons (e.g., Feng, Gracco & Max, 2011; Trudeau-Fisette, Tiede, & Ménard, 2017) while the evolution of the adaptation process is often presented only in exploratory scatterplots in earlier studies (e.g., Rochet-Capellan & Ostry, 2011; Lametti et al., 2012; Mitsuya, Munhall & Purcell, 2017).

2. Methods

2.1. Participants

18 native speakers of Russian (14 female and 4 male) without reported speech, language, or hearing disorders participated in the experiment.

All participants were recruited in Berlin. The mean age of the group was 25.8 years (range 20–37). Participants had spent on average three years in Germany prior to the recordings. The study was approved by the local ethics committee and all speakers gave their written consent to participate in the study.

2.2. Equipment

For each experimental session, participants were seated in front of a 19-inch monitor inside a sound attenuated booth. The monitor served to display the stimuli and experimental instructions which were presented in Russian. Participants' speech was recorded with a Beyerdynamic Opus-54 neck-worn microphone and fed back via foam tipped E-A-RTONE 3A earphones (Figure 1). The distance between participants' mouths and the microphone was about 3–5 cm. The earphones attenuated the air-conducted sound by 25–30 dB while the feedback level was amplified relative to the microphone gain to weaken potential effects of air and bone conduction. The feedback volume was fixed across all participants. However, it was not possible to quantify the feedback level in a precise and meaningful manner since actual feedback volume is expected to vary slightly due to such parameters as the length and the size of participants' ear channels. Real-time tracking and formant perturbation were performed with AUDAPTER, which is a C++ audio signal processing application executable within a MATLAB environment (cf. for technical details Cai et al., 2008). The delay of the feedback loop was approximately 14ms. The

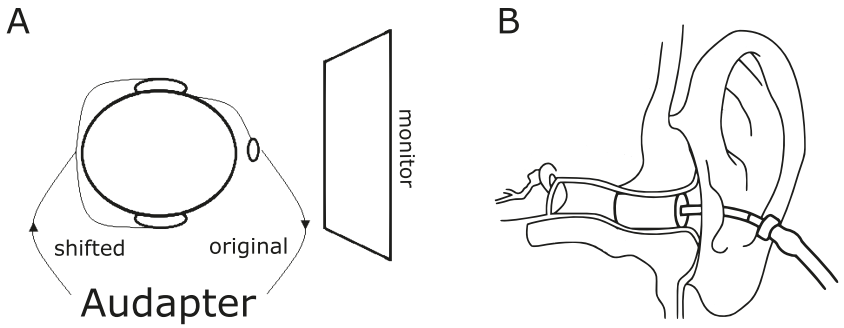

Figure 1: (A) Scheme of the experimental set-up. (B) Foam tipped insert earphones.

original and perturbed audio signals were digitized and saved with a sampling rate of 16 kHz. AUDAPTER also stored data files which contained the formant values (F1, F2, and F3) tracked on each trial.

2.3. Speech stimuli and experimental protocol

For our study we chose Russian since its vowel inventory includes the close central vowel /ɨ/ which is flanked within the F2 space on each side by the two phonemes /i/ and /u/. This constellation allowed us to investigate multiple adaptation in /ɨ/ with bidirectional perturbation of the F2 frequency. The vowel /ɨ/ has a special status in the Russian vowel system since it never appears in word initial position or after palatalized consonants (cf. Bolla, 1981, p. 66).

Each recording session lasted approximately 20–25 minutes and consisted of four experimental phases. Before the start of the first experimental phase, participants completed a few practice trials with unrelated speech material to assure they understood the task and were able to perform it accurately. During a baseline phase, which lasted for 60 trials, no auditory perturbation was applied and participants were able to familiarize themselves with the experimental situation of receiving auditory feedback over earphones. On each trial, which had an approximate duration of 2 seconds, participants were visually prompted to produce one of the four CV syllables /di/, /dɨ/, /gi/, and /gu/. This was done to assess participants' initial F1–F2 formant space. The inter-stimulus interval between the trials was approximately 1.5 seconds. The visual presentation of the stimuli was controlled by a customized MATLAB software package developed at the Institute of Phonetics and Speech Processing, LMU Munich.

During the three following perturbation phases, each of which lasted for 50 trials, participants produced CV syllables containing the close central unrounded vowel /ɨ/ embedded in the context of alveolar and velar consonants /d/ or /g/. Depending on the consonantal context, the F2 was perturbed either downwards or upwards on each trial of each perturbation phase. Within each perturbation phase all stimuli were presented in pseudorandom order. This means that a participant could experience one perturbation direction on one trial and the other direction on the immediately following one; also, the same perturbation direction was never applied on more than two consecutive trials. The interaction between the place of articulation (alveolar vs.

velar) and the perturbation direction (downward vs. upward) was evenly counterbalanced between the 18 participants resulting in two experimental groups (A and B). The perturbation magnitude amounted to 220 Hz during the first perturbation phase and increased in each perturbation phase by 150 Hz. Consequently, the perturbation magnitude was 370 Hz for the second perturbation phase and reached 520 Hz in the last phase of the experiment. The amount of perturbation did not change within each shift phase. There were no ramp trials between the perturbation phases.

Participants were naïve to the purpose of the experiment and were instructed to produce all syllables with prolonged vowels. The prolongation of the vowels maximized the amount of time during which participants were exposed to perturbed vowels. To keep the prolongation duration consistent across participants, they were assisted by a visual go-and-stop signal during their production. The go-and-stop signal had the form of a frame. Between the trials, while the frame stayed red, the response syllable of the upcoming trial appeared on the display and stayed within the frame. When a trial started, the frame color turned green which gave participants the signal to begin with their response.

Following the experimental session, all participants were asked if they noticed anything unusual in their auditory feedback during the experiment. A few of the participants reported that their pronunciation was different from what they are used to or that they perceived an acoustic difference between the syllables /di/ and /gi/. Most participants attributed these pronunciation differences to the effect of listening to their own speech on audio recordings, so when asked if and how these differences affected their production, participants reported to have ignored these. From previous research, however, it is known that participants are not able to voluntarily control their reaction to auditory perturbation even if they are told to ignore it (cf. Munhall, MacDonald, Byrne & Johnsrude, 2009).

All recordings of 18 participants amounted to 3780 trials. The onset and offset of the vowel segment produced on each trial were labeled manually in MATLAB using its graphical input facilities. Subsequently, the formant trajectories were extracted from AUDAPTER's data files based on the labeled onset and offset boundaries. A window with a length of 50 % of each formant trajectory centered at its midpoint was used to compute the formant means produced on each trial.

2.4. Data analysis

All analyses were performed in R (version 3.4.1; R Core Team, 2017). During the data analysis, we first examined the general adaptation pattern that occurred over the course of the experiment in the syllables containing the central vowel /ɨ/. Next, we looked at individual spatial and temporal changes of vowel formants due to the applied perturbation. Finally, by investigating participants' initial F1–F2 vowel space, we evaluated the potential influence of the surrounding sound categories /i/ and /u/ on the individual compensation strategies.

To examine average formant changes in participants' production of the two syllables /dɨ/ and /gɨ/ across the four experimental phases, we fitted a generalized additive model (GAM; Hastie & Tibshirani, 1987). A GAM is a significant extension of a generalized linear regression model which allows the modelling of non-linear relationships between the dependent and independent variables (Wood, 2017a). Therefore, GAMs are much more flexible compared to linear regression models. The non-linear relationships are modelled via complex functions (smooths) which are constructed from ten basis functions (e.g., linear, quadratic, and cubic functions) with an adjustable number of basis dimensions. The number of basis dimensions is a number which indicates the upper limit of how complex the constructed function can be and is estimated directly from the data during the modelling process. That means that the usage of GAMs does not require from the researcher a predefined specification of a certain (non-linear) function as it is derived directly from the data. To prevent overfitting of the data, i.e., modelling of functions which are too complex and therefore might obscure any generalizable patterns in the data, GAMs are estimated using penalized likelihood estimation and cross-validation (cf. for details Wood, 2006). In the case of cross-validation, several subsets of the complete data sample are created always excluding a single data point and the model is refitted to all of these subsets examining how well it predicts the excluded data. One further advantage of GAMs is the possibility to include random effects into the model structure to account for individual response variability across but also within speakers (cf. Baayen, Vasishth, Kliegl & Bates, 2017). To denote the inclusion of random effects in the fitted model, it is dubbed generalized

additive *mixed* model (GAMM). For a hands-on introduction to GAMMs with a focus on dynamic speech analysis see Sóskuthy (2017).

The GAMM offers three main advantages for analyzing the data from the current experiment. First, it is possible to analyze the data as a function of time which allows us to investigate the whole adaptation process rather than just its outcome. Secondly, the non-linearity of parameter smooths does not make any assumptions regarding the temporal or spatial characteristics of the adaptation process. Finally, the parameter smooths can be estimated including random effects which allows us to capture individual variability of the adaptation process.

Prior to building the GAMM model, participants' raw formant frequencies were normalized by subtracting each participants' mean formant frequency produced during the baseline phase for the respective syllable (/di̯/ or /gi̯/). This was done to exclude participant-specific differences regarding their absolute formant magnitudes (e.g., due to gender differences). By means of this normalization, the average F1 and F2 values for /di̯/ and /gi̯/ were set at zero for the baseline phase.

Subsequently, using the *mgcv* package (Wood, 2017b) we fitted one GAMM model for each formant (F1 and F2) with normalized frequencies averaged across all participants and all experimental trials as dependent variable. The data of the unperturbed syllables /di/ and /gu/, which were uttered by participants only during the baseline phase, were not included in the resulting GAMMs. All GAMM models were evaluated, interpreted, and visualized by means of the *itsadug* package by van Rij, Wieling, Baayen & van Rijn (2017).

In the model structure, we included random factor smooths with an intercept split for the perturbation direction (upward vs. downward) in order to assess (potentially non-linear) individual compensation magnitude differences over the course of the experiment. The model also included a fixed effect which assessed the 'constant' effect of the perturbation direction independently from the temporal variation. The resulting models explained 46.6 % and 66.9 % of the variance in the F1 and F2 data, respectively. In comparison, the model which did not include the random smooths (participant-specific temporal variation) but only random intercepts and random slopes explained only 31.2 % of the variance in the F2 data. Maybe somewhat surprisingly, the inclusion of the phase

number (shift 1, shift 2, and shift 3) as an interaction with the perturbation direction did not significantly improve the model fit. We also refitted the F2 model including an interaction between the perturbation direction (upward vs. downward) and the experimental group (A vs. B) which also did not improve the fit. In both cases, the goodness of fit was assessed by the Akaike Information Criterion (AIC; Akaike 1974).

Following the suggestion in Baayen, van Rij, de Cat & Wood (2016), the fitted models were investigated for the presence of autocorrelation in their residuals. Autocorrelation in the present study represents the correlation between the formant frequencies produced by one participant on two consecutive experimental trials. The higher the autocorrelation value is the less amount of information is contributed for the statistical model by each additional experimental trial. Ignoring this issue might result in overconfident estimates of the standard errors, confidence intervals, and p-values. The amount of autocorrelation at lag 1 was relatively moderate in the present data with 0.2 for F1 and 0.17 for F2. The effect of autocorrelation was practically reduced to zero by incorporating AR(1) error models in the specification of the fitted GAMM models. The corrected models explained 23.1 % and 63.4 % of the variance in the F1 and F2 data, respectively. The dropped percentages of the explained variance are due to the refitted models taking into account the autocorrelation which makes their prediction about actual frequency values worse. This is especially true for the F1 model which is an indication that much of the variance in the initial model can be explained by autocorrelated errors rather than by the specified model parameters such as the direction of the applied perturbation. Visual model inspection revealed that the residuals of the adjusted GAMMs followed a normal distribution for F1 and F2 data.

To examine individual spatial and temporal differences of the adaptation process, we extracted F2 curves estimated for each participant by the GAMM model described in the above paragraphs.

In order to evaluate whether the occurrence of certain individual compensation patterns was induced by sound categories surrounding the perturbed vowel, we investigated participants' F1–F2 space using their baseline phase production. For this purpose, we fitted two linear-mixed models using the *lme4* package (Bates, Mächler, Bolker, and Walker, 2015). One model was fitted for each of the two average formant frequencies (F1

and F2) that were produced by participants in the syllables /di/, /di/, /gi/, and /gu/ during the baseline phase. The model structure included the produced syllable and the interaction between the syllable and gender as fixed effects and the formant frequency as dependent variable. Furthermore, both models included an interaction between the syllable and the compensatory pattern observed for each participant (cf. section 3.2 for a detailed discussion of individual compensation patterns). Random intercepts were modeled for each participant as well as random slopes for each produced syllable.

Visual model inspection revealed that the residuals of the chosen models followed a normal distribution for F1 and F2 data. P-values were obtained with the *lmerTest* package by Kuznetsova, Brockhoff, and Bojesen-Christensen (2016).

3. Results

3.1. Overall compensatory behavior

The GAMM estimated for F1 suggested that the applied perturbation did not have a 'constant' effect on the produced F1 values since its average did not significantly differ from the baseline on trials with upward (2.97 Hz, t=1.09, $p > .05$) as well as on trials with downward perturbation (-1.14 Hz, t=-0.32, $p > .05$). These values represent 'constant' F1 differences for the whole experiment since they do not take into account any changes that appeared over time. Taking the temporal variation over the course of the experiment into account, the model did not reveal a F1 difference from the baseline for either of the two perturbation directions (Figure 2A). Furthermore, a direct comparison between trials with applied upward and downward perturbation revealed no significant difference in their F1 curves (Figure 2B). The average F1 difference amounted to 0.96 Hz (95 % CI [-6.03 7.94]) by the end of the first shift phase, 1.90 Hz (95 % CI [-6.14 9.95]) by the end of the second shift phase, and 2.93 Hz (95 % CI [-7.86 13.72]) by the end of the experiment. Random non-linear smooths of the F1 model suggest that there were unsystematic participant-specific F1 changes which are most likely not related to the applied perturbation (Figure 2C).

The absence of systematic compensatory effects in F1 is expected as no F1 perturbation was applied during the experiment. This outcome provides

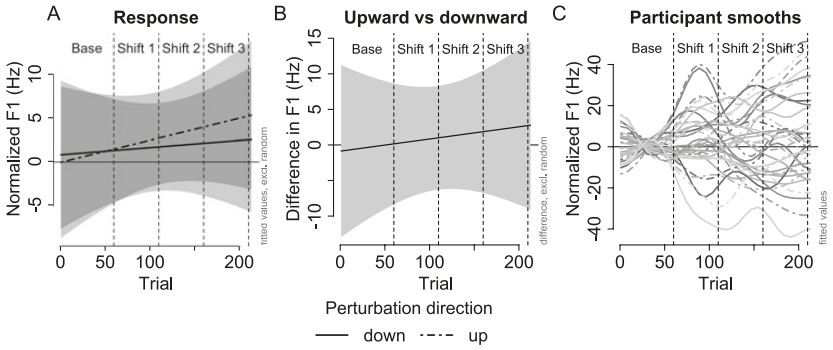

Figure 2: Visual summary of the fitted GAMM model for F1: (A) Average compensatory effects (excluding random participant effects) in F1 for downward and upward perturbation over the course of the experiment. Grey bands represent 95 % confidence intervals. (B) The average difference in F1 between trials produced under opposing perturbation directions over the course of the experiment. Grey bands represent 95 % confidence intervals. (C) Random smooths estimated for each participant for her/his average F1 curve split by the perturbation direction.

additional support for the validity of the applied experimental manipulation and the assumption that any systematic effects found for F2 are due to the application of the bidirectional perturbation. Due to the absence of compensatory effects in F1, we will not discuss this variable any further.

The GAMM estimated for F2 suggested that the applied perturbation had a 'constant' effect on the produced F2 values on trials with upward (-127.58 Hz, t=-5.76, $p < .05$) as well as on trials with downward perturbation (143.14 Hz, t=5.36, $p < .05$). The direction of the 'constant' effect was opposed to the direction of the applied perturbation during upward and downward perturbation. Examining the effect of the perturbation over time, the model revealed that this effect increased for both directions (upward and downward) over the course of the experiment (Figure 3A). On average, however, the effect appears to be stronger for the upward perturbation compared to the downward perturbation. The F2 difference between trials produced under opposite perturbation directions became significant after the baseline phase and increased, as expected, over the three perturbation phases (Figure 3B). The average F2 difference amounted to -131.51 Hz (95 % CI [-183.5

-79.52]) by the end of the first shift phase and to -193.61 Hz (95 % CI [-254.79 -132.44]) by the end of the second shift phase. By the end of the experiment, the average F2 difference reached -261.13 Hz (95 % CI [-345.41 -176.84]). The model suggested that the average compensatory effect in F2 can be modeled by linear functions for both perturbation directions as the estimated degrees of freedom (EDF) for both smooth terms amounted to 1. On the other hand, the random smooths fitted for individual participants exhibited a high degree of non-linearity for upward (EDF= 116.66) and downward (EDF = 97.76) perturbation directions.

The random F2 smooths fitted individually for each participant demonstrate that above and beyond the general tendency to counteract the applied perturbation, participants' adaptation patterns exhibited high variability in both investigated dimensions (formant frequency and time). For instance, the individual smooths refined the general observation that the downward perturbation caused on average weaker compensatory effect

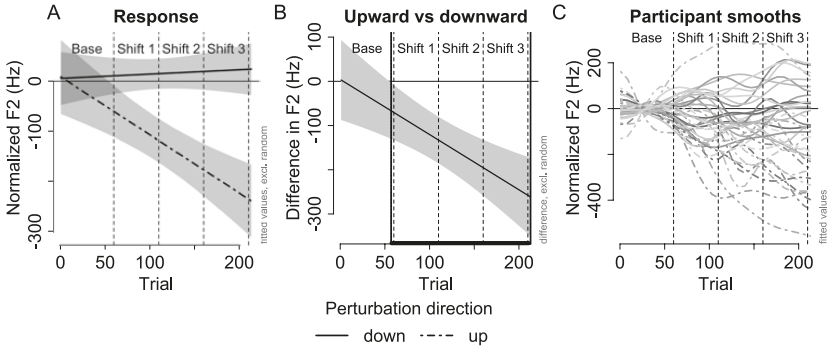

Figure 3: Visual summary of the fitted GAMM model for F2: (A) Average compensatory effects (excluding random participant effects) in F2 for downward and upward perturbation over the course of the experiment. Grey bands represent 95 % confidence intervals. (B) The average difference in F2 between trials produced under opposing perturbation over the course of the experiment. The solid thick line denotes the region where the F2 difference was significant. Grey bands represent 95 % confidence intervals. (C) Random smooths estimated for each participant for her/his average F2 curve split by the perturbation direction.

over time. In Figure 3C, it is apparent that for most participants the solid lines (F2 curves produced under downward perturbation) remained closer to the baseline compared to the dashed lines (F2 curves produced under upward perturbation).

To understand these participant-specific differences, we will examine and discuss individual adaptation patterns in more detail in the next section.

3.2. Individual compensatory patterns

As revealed by the individual F2 curves estimated by the GAMM model, the most distinct characteristic among participants was the magnitude of their compensation for the downward perturbation. Based on this metric, we identified five participants who throughout the experiment were compensating for the downward perturbation to the same extent as for the upward perturbation and 10 participants who compensated less (if at all) for the downward perturbation compared to the upward perturbation. In Figure 4, the first group ('symmetrical' compensation pattern) is represented by participants 3, 9, and 13, while participants 4, 16, and 17 can be considered to represent the second group ('asymmetrical' compensation pattern).

Examining subfigures for participants 3, 9, and 13, we see that the F2 curves for the two syllables /di/ and /gi/ diverged by equal amounts from the baseline as the experiment progressed. For participants 4 and 16, on the other hand, the F2 curve produced under the upward perturbation diverged more strongly from the baseline. In contrast, the F2 curve produced under the downward perturbation appears to have fluctuated around the baseline. For participant 17, the effect of the perturbation direction appears to be flipped with stronger compensation for the downward perturbation.

In addition to the symmetrical and asymmetrical compensation patterns, we identified in the sample three participants who were not able to consistently compensate for the opposite perturbation directions throughout the experiment (see participants 5, 6, and 7 in Figure 4). In summary, all 18 participants who participated in the study exhibited one of the three described adaptation behaviors. Representative data of only nine participants is depicted in Figure 4 due to space limitations.

As revealed by Figure 4, individual adaptation patterns exhibited a lot of spatial and temporal non-linearities. This fact makes it prohibitive to apply plain pairwise comparisons between participants' production during the baseline and the last perturbation phase to assess whether speakers have successfully adapted to the perturbation. In the worst case, this approach risks obfuscating the specific characteristics of the adaptation patterns. Just on grounds of such comparison, participants 5, 6, and 7 would qualify as speakers who failed to compensate in opposite directions. However, examining the evolution of their F2 responses over

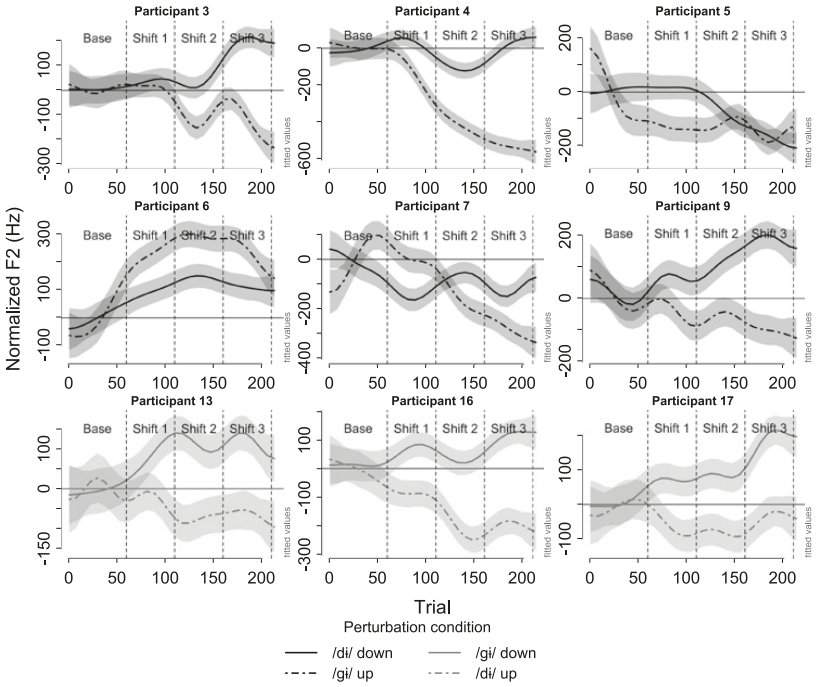

Figure 4: Individual compensatory effects in F2 for downward and upward perturbation across all experimental trials. The F2 curves were estimated by the same GAMM model which is depicted in Figure 3. Please note: individual y-axis scales were applied due to big inter-individual differences of the compensatory magnitude. Vertical dashed lines denote the beginnings and the ends of the experimental phases. After the baseline phase (Base), the perturbation magnitude amounted to 220 Hz (Shift 1), 370 Hz (Shift 2), and 520 Hz (Shift 3).

the course of the experiment it is apparent that all three participants tried to compensate for the applied shifts with participant 7 eventually being able to achieve this goal for the upward but not the downward perturbation direction.

Participant 6, for instance, initially increased F2 in both experimental syllables independently of the perturbation direction. After the second perturbation phase, the produced F2 frequency started to drift again into the negative direction but remained, nonetheless, distinct for both syllables. This pattern provides evidence for the fact that, although she was not counteracting the applied perturbation, participant 6 was able to perceive the auditory errors caused by the downward and upward perturbations and to differentiate between them. Participant 5, on the other hand, differentiated between the two perturbation directions during the first perturbation phase, but changed her compensatory movements for the downward shifts during the second perturbation phase such that she produced the same F2 frequency for both syllables at the end of the experiment.

Analogously to participants 5 and 6, participant 7 was not able to develop an appropriate compensation strategy when the perturbation was first applied. However, she changed her initial incorrect strategy in the course of the experiment. She started to counteract the perturbation during the second perturbation phase and eventually developed two consistently different production strategies by the end of the experiment.

The relative compensation magnitude varied substantially across all participants independently of whether they could successfully compensate for both perturbation directions or not. During the last perturbation phase, for instance, the compensation magnitude fluctuated between 6.6 and 103 % across all participants. Also, the amount of change of the compensation magnitude over the course of the experiment was not identical among the participants. Compare the adaptation patterns of participants 9 and 13 in Figure 4. While for participant 9 the compensation magnitude increased with the increasing perturbation magnitude, the compensation magnitude in participant 13 appears to have reached an absolute compensation limit for both perturbation directions around 100 Hz. Overall, there was a weak negative correlation between the average compensation magnitude and the perturbation magnitude (r = -0.19, t=-10.31, p<.05, 95 % CI [-0.23 -0.16]). This suggests that the

average compensation magnitude slightly decreased as the perturbation magnitude increased.

3.3. Role of the initial F1–F2 space

To explain the occurrence of the symmetrical and asymmetrical compensatory patterns, we investigated the influence of participants' initial F1–F2 space on their compensatory performance.

The mean F1 and F2 frequencies produced by all participants during the baseline phase are summarized in Figure 5 split by participants' gender.

There were no statistically significant within-speaker differences in F1 between the vowels of the four syllables. The average F1 difference between /dɨ/ and /di/ in female participants with asymmetrical compensatory pattern (7 participants) was 19.27 Hz (t = 1.96, p > .05, 95 % CI [93.78 32.63]), 4.17 Hz (t = 1.87, p > .05, 95 % CI [-0.13 7.3]) between /gɨ/ and /dɨ/, and 3.73 Hz (t = 0.44, p>.05, 95 % CI [-11.72 19.18]) between /gu/ and /gɨ/.

In female participants with symmetrical compensatory pattern (4 participants), the average F1 frequencies were lower for every syllable. However, none of these differences was significant (/di/: -2.53 Hz, t= -0.16, p > .05, 95 % CI [-31.47 26.42]; /dɨ/: -3.56 Hz, t = -0.17, p > .05, 95 % CI [-41.2 34.08]; /gɨ/: -2.89 Hz, t= -0.14, p > .05, 95 % CI [-41.31 35.54]; /gu/: -16.49 Hz, t = -1.17, p > .05, 95 % CI [-42.21 9.22]). In female participants who reacted inconsistently to the opposite perturbations (3 participants), the average F1 frequencies were also lower for every syllable compared to the female participants with the asymmetrical compensation pattern. Again, none of these differences was significant (/di/: -9.8 Hz, t = -0.5, p > .05, 95 % CI [-45.61 26.0]; /dɨ/: -23.22 Hz, t = -0.91, p > .05, 95 % CI [-69.78 23.34]; /gɨ/: -23.43 Hz, t= -0.9, p > .05, 95 % CI [-70.97 24.11]; /gu/: -30.53 Hz, t= -1.75, p > .05, 95 % CI [-62.34 1.28]).

The F2 model indicated significant within-speaker differences between F2 values of the investigated vowels. In female participants with asymmetrical compensatory pattern (7 participants) the average F2 difference between /dɨ/ and /di/ was -302.96 Hz (t =-6.3, p < .05, 95 % CI [-390.76 -215.13]), -169.52 Hz (t =-5.26, p < .05, 95 % CI [-228.29 -110.76]) between /gɨ/ and /dɨ/, and -1298.46 Hz (t =-19.08, p < .05, 95 % CI [-1422.68 -1174.34]) between /gu/ and /gɨ/.

In female participants with symmetrical compensatory pattern (4 participants), the average F2 frequencies were lower for every syllable except for /di/. However, none of these differences was significant (/di/: 33.61 Hz, t = 0.47, p > .05, 95 % CI [-95.68 162.89]; /dɨ/: -6.53 Hz, t= -0.07, p > .05, 95 % CI [-188.9 175.8]; /gɨ/: -15.74 Hz, t= -0.16, p > .05, 95 % CI [-194.0 162.54]; /gu/: -5.24 Hz, t = -0.18, p > .05, 95 % CI [-57.98 47.51]). In female participants who reacted inconsistently to the opposite perturbations (3 participants), the average F2 frequencies were lower for every syllable except for /dɨ/ and /gu/ compared to the female participants with the asymmetrical compensation pattern; only the difference for the syllable /gu/ was significant (83.23 Hz, t = 2.33, p < .05, 95 % CI [18.0 148.48]). The remaining three differences were not significant (/di/: -11.31 Hz, t = -0.13, p > .05, 95 % CI [-171.23 148.63]; /dɨ/: 132.12 Hz, t= 1.07, p > .05, 95 % CI [-93.57 357.61]; /gɨ/: -107.42 Hz, t= -0.89, p > .05, 95 % CI [-327.84 113.42]).

The F1 and F2 frequencies produced by male participants were on average lower for every syllable compared to the formants produced by female participants, however, these differences were only significant for F2 values produced for the syllable /di/ (-270.41 Hz, t = -3.58, p < .05, 95 % CI [-408.26 -132.55]).

Overall, the observed F1–F2 space of the vowels /i/, /ɨ/, and /u/ was consistent with previous descriptive studies of Russian vowels (Lobanov, 1971; Bolla, 1981). As expected, there was no statistically significant and no perceivable difference in F1 between the investigated vowels (previous research on formant perception indicates that on average participants do no perceive F1 differences below 50 Hz; Oglesbee & Kewley-Port, 2009). The vowels were differentiated most prominently by F2 with /i/ having the highest and /u/ the lowest values; F2 of /ɨ/ lay between the other two vowel categories. F2 was higher in /dɨ/ compared to /gɨ/ likely due to coarticulation. Furthermore, the initial F1–F2 vowel space did not significantly differ between the three participant groups which exhibited different compensatory patterns during the perturbation phases of the experiment.

4. Discussion

In the current investigation we presented results from a bidirectional auditory perturbation experiment conducted with native speakers of Russian.

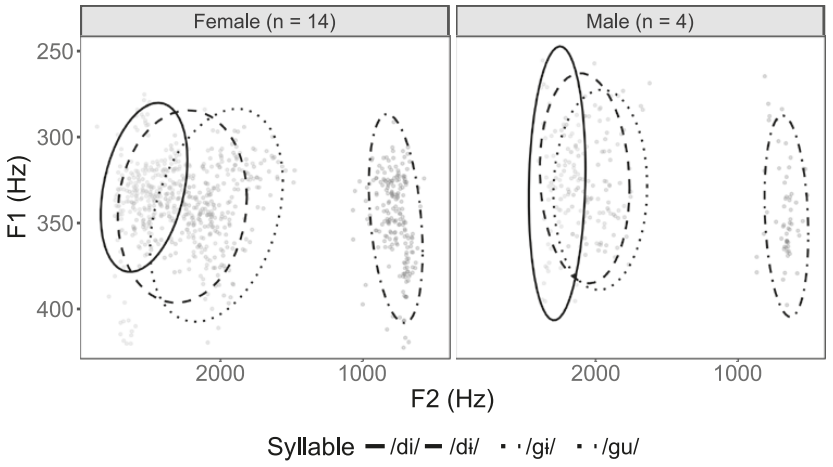

Figure 5: The average F1–F2 vowel space produced by all participants during the baseline phase (no perturbation) for the four syllables /di/, /dɨ/, /gi/, and /gu/. The data is split by participants' gender.

During three perturbation phases of the experiment, participants had to produce the close central unrounded vowel /ɨ/ while its F2 frequency was perturbed in opposing directions depending on the preceding consonant (/d/ or /g/). The bidirectional perturbation was intended to increase the demands associated with the experimental task since participants had to coordinate their corrective movements in two different ways depending on the perturbation direction to produce the target vowel /ɨ/. Based on the recurrent observation that participants counteract the applied auditory perturbation, we expected that the baseline F2 values for the two syllables /dɨ/ and /gɨ/ would diverge over the course of the three perturbation phases since the magnitude of the perturbation increased in opposing directions from one perturbation phase to another. The two consonantal contexts (alveolar vs. velar) were chosen to evaluate the potential influence of physical restrictions on the success of the adaptation outcome.

The average adaptation behavior observed during the study confirmed our main hypothesis. The GAMM model estimated for the normalized F2 frequency suggested that participants were able to adapt simultaneously to two opposing F2 perturbations and employ different strategies to produce

the vowel /i/ depending on the direction of the applied perturbation. These results are qualitatively in line with previous articulatory and auditory perturbation studies which show that most speakers are able to remap their initial articulatory-to-acoustics mapping under aggravated speech conditions (e.g., Gay et al., 1981; Savariaux et al., 1995; Feng et al., 2011).

Furthermore, our results are consistent with findings by Rochet-Capellan & Ostry (2011) who demonstrated that participants are able to simultaneously develop multiple strategies to produce the same target vowel. Adding to these results, our data shows that the results obtained by Rochet-Capellan & Ostry (2011) in the context of F1 are generalizable to F2.

The compensatory effects observed in F2 frequencies for both perturbation directions were absent in our F1 data. This result serves as evidence for the validity of the applied experimental manipulations.

The application of generalized additive mixed modelling (GAMM) allowed us to investigate the evolution of the adaptation process over time. Particularly, we were able to observe participant-specific differences in the spatial and temporal dimensions of the compensatory changes and to understand individual differences in how participants were generally able to cope with the demands of the experimental task.

While one group of participants was almost immediately able to compensate for bidirectional formant perturbations during the first perturbation phase, other participants needed longer periods of time to do so and started to compensate only in the second or the third perturbation phase. Also, a few participants failed to identify the appropriate compensatory adjustments altogether. Since these speakers also tended to change the initial direction of their compensatory movements throughout the experiment, their behavior can be best described as exploratory. In several instances, the directional changes of the compensatory movements were quite abrupt as revealed by the non-linearities of the modelled F2 curves.

Although the required corrective tongue movements were set along the same movement trajectory (forward vs. backward), participants had to figure out two quite different strategies to produce the vowel /i/. In particular, they first had to identify the direction of the applied frequency shift for both experimental syllables and then to adjust their tongue movements appropriately.

It appears plausible that the correct identification of the F2 perturbation direction was not a trivial task since it took different amounts of time for different participants before they eventually started to consistently compensate for the applied perturbations if at all. Most convincing in this regard is the observation that some participants initially followed the perturbation but started to counteract it after a while. These observations suggest that all participants without exception were indeed perceiving the auditory errors caused by the perturbations, however, not all were able to figure out the appropriate articulatory adjustments in order to minimize them. One potential reason for this might be their inability to identify the correct perturbation direction.

This hypothesis can also explain the observation that in typical (unidirectional) auditory perturbation studies, beside a group of participants who counteract the perturbation, there are most likely a few participants who appear to follow the perturbation. Our data suggest that both reactions to auditory perturbations (counteracting and following) can be understood in more general terms as exploratory compensatory behavior where participants wander through the formant space in order to find the appropriate corrective movements to produce the intended acoustic output. In line with this idea is the observation from the current study that no participant actually followed the applied perturbations in both directions (upward and downward).

The observed temporal non-linearities and abrupt directional changes of the compensatory responses challenge the idea that speakers exhibit either auditory or somatosensory feedback preference during speech production (Lametti et al., 2012). Strictly following the idea of feedback preference, speakers with auditory feedback preference should have always reacted consistently to the applied auditory perturbations, i.e., independently of the perturbation direction. At the same time, we should expect that a subset of speakers with a preference for somatosensory feedback should virtually ignore the auditory perturbations. However, the examination of individual adaptation patterns revealed that both assumptions do not appear to be true.

First, among the 18 participants there were no speakers who ignored the applied perturbations, which might have suggested that they disprefer the auditory feedback channel during speech production. Secondly,

and more importantly, for several participants the direction and the magnitude of the compensation were not identical for both perturbation directions, which it should be under the assumption that a speaker exhibits a permanent preference for the auditory feedback channel. Furthermore, several participants were able to acquire the two appropriate compensatory strategies after some practice. That is, the ability to develop a consistent compensatory strategy does not seem to depend on speakers' preference for auditory or somatosensory feedback. The observation that the compensation magnitude was not identical for the two opposite perturbation directions across all participants deserves further attention. The estimated GAMM model suggested that although 17 participants compensated for the upward perturbation, only five of them did it simultaneously for the downward perturbation. (Additionally, a single participant significantly shifted her F2 frequency upwards independently of the applied perturbation direction.) We dubbed this difference in compensatory profiles as asymmetrical and symmetrical compensation patterns.

One potential explanation for the observation that far less participants were able to compensate for the downward perturbation is the articulatory effort associated with the required forward compensatory movement of the tongue. Whereas the backward movement of the tongue from the central position of the /i/ is physically less restricted, the extent of the forward movement is limited by the alveolar ridge and the upper incisors. Since we do not have articulatory data of participants' palatal shapes, this possibility cannot be ruled out completely. However, there is some evidence which undermines this hypothesis.

Taking the introduced idea of physical restrictions further, we have to assume that the forward movement of the tongue in /di/ should be even more restricted compared to /gi/ as the tongue has already a more advanced position in the first syllable. This positional difference between /di/ and /gi/ is supported by the results on participants' initial F1–F2 formant space presented in the section 3.3. Based on this fact, we should expect that the participants who were able to compensate for the downward perturbation were able to do so preferably for the syllable /gi/. However, from the six participants who significantly upshifted their F2 during the perturbation phases, three did it for /di/ and the other three did it for /gi/. That means

that there was no advantage for the syllable /gɨ/ as there ought to be assuming that the distance from the produced vowel to the physical limit (i.e., alveolar ridge/upper incisors) played a crucial role for the success of the forward compensatory tongue movement. This interpretation is further supported by the GAMM modelling as the inclusion of the interaction between the perturbation direction and the syllable did not improve the overall fit.

The idea of the influential role of physical constraints on the compensatory movement can be restated in more abstract terms of somatosensory categories. In these notions, the upper limit of the compensatory movement is no longer assumed to be a physical boundary (i.e., alveolar ridge) but rather a somatosensory category boundary of the neighboring speech sound. In the particular case discussed in the above paragraph, both explanations can be used interchangeably without inducing different interpretations of the results. However, the reformulation of this hypothesis in somatosensory terms allows us to evaluate the potential influence of the somatosensory sound categories on the compensation magnitude in the case of the backward tongue movements since these were generally physically less restricted in both cases of /dɨ/ and /gɨ/.

At first glance, it is conceivable that the higher distance from /ɨ/ towards the somatosensory boundary of /u/ compared to the distance between /ɨ/ and /i/ facilitated the compensation for the upward perturbation. However, as pointed out by Katseff et al. (2012), if we assume that speech sounds are also defined in sensorimotor space, compensatory magnitude should be restricted not only by a neighboring speech sound but foremost by the size of the sensorimotor region of the perturbed category. That means that when speakers deviate too much from the sensorimotor region of the perturbed category, the magnitude of the compensation should decrease. This hypothesis is, however, substantially challenged by the experimental data.

Comparing the compensation magnitudes for the upward perturbation between /dɨ/ and /gɨ/ reveals that participants compensated in both syllables in comparable amounts despite the fact that the somatosensory distance between /dɨ/ and /gu/ was higher compared to /gɨ/ and /gu/. The absence of a difference in the compensation magnitude between /dɨ/ and /gɨ/ was supported by the GAMM modelling as the inclusion of

the interaction between the perturbation direction and the syllable did not improve the overall fit. Furthermore, the examination of individual adaptation patterns revealed that a subset of speakers was able to compensate 100 % of the applied perturbation even when it reached 520 Hz and thereby induced a high degree of somatosensory error. Speaking to the same issue, for some participants who changed their F2 frequency in the same direction as the applied perturbation, the mismatch between the auditory and somatosensory error grew ever higher over the course of the experiment.

Taking all this evidence into account, the emergence of the asymmetric compensatory pattern is difficult to explain in terms of the violation of somatosensory boundaries. Consistent with this idea is the fact that there were no systematic differences in participants' initial F1–F2 formant spaces which could predict their different compensatory profiles (symmetrical, asymmetrical, and non-consistent).

Without resorting to the somatosensory boundaries, we can think of one alternative explanation for the emergence of the asymmetric compensatory pattern. Central to this hypothesis is the idea that the asymmetric pattern emerged due to an asymmetry in the phonemic space of the Russian high vowels. In particular, while /i/ appears only after palatalized consonants in Russian, both /ɨ/ and /u/ follow only non-palatalized ones (cf. Bolla, 1981). The palatalization contrast is an important part of the Russian phonology and is very present for Russian speakers. The most common acoustic feature associated with palatalized consonants is the high F2 frequency at the beginning of the following vowel. This acoustic feature is so important for the perception of palatalization by Russian speakers that even cross-spliced syllables containing non-palatalized consonants and vowels with high initial F2 frequency are perceived as palatalized (cf. Bondarko, 2005). We think that this perceptual effect might have occurred during our experiment.

Since the baseline F2 values of /i/ and /ɨ/ are on average substantially closer to each other compared to /ɨ/ and /u/, it seems reasonable that most participants classified instances of /ɨ/ shifted towards /i/ as phonemic errors of palatalization and corrected for them by lowering their F2. On the other hand, only a few participants reacted to the F2 perturbation of /i/ towards /u/ as it did not induce a change of palatalization status of the perceived syllable. Presumably, those participants who reacted by the same amount

to the downward perturbation as to the upward perturbation were more sensitive to general F2 changes independent of the phonemic status of the perceived syllable. Unfortunately, we do not have participants' perceptual profiles which could settle this question completely.

5. Conclusion

Despite a growing body of research, the factors which induce the inter-individual outcome variability during sensorimotor learning are still much debated. Our investigation has shown that there is merit in varying task parameters within the same experimental session and in analyzing the data of perturbation experiments taking the temporal dimension of the adaptation process into account. By doing this, we could show that the inter- and intra-participant variability present during somatosensory learning in speech is beyond the predictions of the hypotheses which ascribe this variability exclusively to the characteristics of speakers' internal models of speech motor control.

Acknowledgments

We gratefully acknowledge support from DFG grant 220199 to JB. We are also grateful to two anonymous reviewers for their useful suggestions. We thank Felix Golcher for his advice on the statistical modelling of the data. We thank Miriam Oschkinat for her support during data acquisition and Yulia Guseva for her help with the preparation of the manuscript. We also thank all participants who took part in the study.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Baayen, R.H., van Rij, J., de Cat, C., & Wood, S.N. (2016). Autocorrelated errors in experimental data in the language sciences: Some solutions offered by generalized additive mixed models. *arXiv preprint arXiv:1601.02043*.

Baayen, R.H., Vasishth, S., Kliegl, R., & Bates, D. (2017). The cave of shadows: Addressing the human factor with generalized additive mixed models. *Journal of Memory and Language*, *94*, 206–234.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Baum, S.R., & McFarland, D.H. (1997). The development of speech adaptation to an artificial palate. *The Journal of the Acoustical Society of America*, *102*(4), 2353–2359.

Bolla, K. (1981). *A Conspectus of Russian Speech Sounds*. Budapest: Hungarian Academy of Science.

Bondarko, L. V. (2005). Phonetic and phonological aspects of the opposition of 'soft' and 'hard' consonants in the modern Russian language. *Speech Communication*, *47*(1), 7–14.

Brunner, J., Ghosh, S., Hoole, P., Matthies, M., Tiede, M., & Perkell, J.S. (2011). The influence of auditory acuity on acoustic variability and the use of motor equivalence during adaptation to a perturbation. *Journal of Speech, Language, and Hearing Research*, *54*(3), 727–739.

Brunner, J., Hoole, P., & Perrier, P. (2011). Adaptation strategies in perturbed /s/. *Clinical Linguistics & Phonetics*, *25*(8), 705–724.

Cai, S., Boucek, M., Ghosh, S.S., Guenther, F.H., & Perkell, J.S. (2008). A system for online dynamic perturbation of formant trajectories and results from perturbations of the Mandarin triphthong /iau/. In Sock, R., Fuchs, S., Laprie, Y., (Eds), *Proceedings of the 8th International Seminar on Speech Production 2008*, Strasbourg, France, 65–68.

Feng, Y., Gracco, V.L., & Max, L. (2011). Integration of auditory and somatosensory error signals in the neural control of speech movements. *Journal of Neurophysiology*, *106*(2), 667–679.

Gay, T., Lindblom, B., & Lubker, J. (1981). Production of bite-block vowels: Acoustic equivalence by selective compensation. *The Journal of the Acoustical Society of America*, *69*(3), 802–810.

Ghosh, S.S., Matthies, M.L., Maas, E., Hanson, A., Tiede, M., Ménard, L., Guenther, F.H., Lane, H., & Perkell, J.S. (2010). An investigation of the relation between sibilant production and somatosensory and auditory acuity. *The Journal of the Acoustical Society of America*, *128*(5), 3079–3087.

Hastie, T., & Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, *82*(398), 371–386.

Houde, J.F., & Jordan, M.I. (1998). Sensorimotor adaptation in speech production. *Science*, *279*(5354), 1213–1216.

Jones, J.A., & Munhall, K.G. (2000). Perceptual calibration of F0 production: Evidence from feedback perturbation. *The Journal of the Acoustical Society of America*, *108*(3), 1246–1251.

Jones, J.A., & Munhall, K.G. (2003). Learning to produce speech with an altered vocal tract: The role of auditory feedback. *The Journal of the Acoustical Society of America*, *113*(1), 532–543.

Katseff, S., Houde, J., & Johnson, K. (2012). Partial compensation for altered auditory feedback: A tradeoff with somatosensory feedback? *Language and Speech*, *55*(2), 295–308.

Kuznetsova, A., Brockhoff, P.B. & Bojesen-Christensen, R.H. (2016). lmerTest: Tests in Linear Mixed Effects Models. R package version 2.0-30.

Lametti, D.R., Nasir, S.M., & Ostry, D.J. (2012). Sensory preference in speech production revealed by simultaneous alteration of auditory and somatosensory feedback. *Journal of Neuroscience*, *32*(27), 9351–9358.

Lobanov, B.M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608.

MacDonald, E.N., Goldberg, R., & Munhall, K.G. (2010). Compensations in response to real-time formant perturbations of different magnitudes. *The Journal of the Acoustical Society of America*, *127*(2), 1059–1068.

Mitsuya, T., Munhall, K.G., & Purcell, D.W. (2017). Modulation of auditory-motor learning in response to formant perturbation as a function of delayed auditory feedback. *The Journal of the Acoustical Society of America*, *141*(4), 2758–2767.

Munhall, K.G., MacDonald, E.N., Byrne, S.K., & Johnsrude, I. (2009). Talkers alter vowel production in response to real-time formant perturbation even when instructed not to compensate. *The Journal of the Acoustical Society of America*, *125*(1), 384–390.

Oglesbee, E., & Kewley-Port, D. (2009). Estimating vowel formant discrimination thresholds using a single-interval classification task. *The Journal of the Acoustical Society of America*, *125*(4), 2323–2335.

Purcell, D.W., & Munhall, K.G. (2006). Adaptive control of vowel formant frequency: Evidence from real-time formant manipulation. *The Journal of the Acoustical Society of America*, *120*(2), 966–977.

Perkell, J.S., Guenther, F.H., Lane, H., Matthies, M.L., Stockmann, E., Tiede, M., & Zandipour, M. (2004). The distinctness of speakers' productions of vowel contrasts is related to their discrimination of the contrasts. *The Journal of the Acoustical Society of America*, *116*(4), 2338–2344.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved January 5, 2017 from https://www.R-project.org/

van Rij, J., Wieling, M., Baayen, R.H., & van Rijn, H. (2017). itsadug: Interpreting time series and autocorrelated data using GAMMs. *R package version*, 2.3.

Rochet-Capellan, A., & Ostry, D.J. (2011). Simultaneous acquisition of multiple auditory–motor transformations in speech. *Journal of Neuroscience*, *31*(7), 2657–2662.

Sato, M., Schwartz, J.L., & Perrier, P. (2014). Phonemic auditory and somatosensory goals in speech production. *Language, Cognition and Neuroscience*, *29*(1), 41–43.

Savariaux, C., Perrier, P., & Orliaguet, J.P. (1995). Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production. *The Journal of the Acoustical Society of America*, *98*(5), 2428–2442.

Sóskuthy, M. (2017). Generalised additive mixed models for dynamic analysis in linguistics: a practical introduction. *arXiv preprint arXiv:1703.05339*.

Trudeau-Fisette, P., Tiede, M., & Ménard, L. (2017). Compensations to auditory feedback perturbations in congenitally blind and sighted speakers: Acoustic and articulatory data. *Plos One*, *12*(7), e0180300.

Villacorta, V.M., Perkell, J.S., & Guenther, F.H. (2007). Sensorimotor adaptation to feedback perturbations of vowel acoustics and its relation to perception. *The Journal of the Acoustical Society of America*, *122*(4), 2306–2319.

Wood, S.N. (2006). Low-rank scale-invariant tensor product smooths for Generalized Additive Mixed Models. *Biometrics*, *62*(4), 1025–1036.

Wood, S.N. (2017a). *Generalized additive models: An introduction with R*. Chapman & Hall/CRC Texts in Statistical Science.

Wood, S N. (2017b). MGCV: Mixed GAM Computation Vehicle with GCV/AIC/REML smoothness estimation. *R package version*, 1.8–19.

Louise McKeever, Joanne Cleland, Jonathan Delafield-Butt

Aetiology of speech sound errors in autism

Abstract: In looking at speech perception and production, it is vital we understand variation in different populations in order to understand variation in what is perceived as typical speech development; develop bio-markers; and provide effective methods for diagnosis and intervention where required. Research suggests that people with autism experience higher rates of speech sound errors (SSEs) than their peers (Cleland, Gibbon, Peppé, O'Hare, & Rutherford, 2010; Shriberg, Paul, Black, & Santen, 2011), yet the reasons why are unknown. This chapter takes an in-depth look at the current literature on SSEs produced by people with autism, from young children to young adults. It explores why these higher rates occur, moving beyond the previous debate of whether they exist at all in this population. Recent studies using detailed analyses show that children with autism exhibited significantly higher rates of SSEs than typically developing (TD) children, these are discussed in detail alongside a critique of the methods historically used to assess SSEs in this population. This chapter proposes two perspectives that may account for these higher rates of SSEs in autism: a) the speech attunement framework and b) deficits in speech motor control. It explores how both of these perspectives may intersect to produce SSEs in people with autism. Both are discussed in relation to the comorbidities of speech perception issues and motor deficits often found in people with autism. Suggestions are made for future research using sensitive articulatory analysis of speech such as ultrasound tongue imaging or electropalatography. This chapter highlights the need to look equally at both linguistic and motor skills in children with autism to describe accurately the range of cognitive and neurophysiological processes that may affect speech production.

Keywords: autism, speech errors, ultrasound, electropalatography, speech attunement framework, speech motor impairment

1. Introduction

People with autism present with higher rates of speech sound errors (SSEs) than their peers (Cleland, Gibbon, Peppé, O'Hare, & Rutherford, 2010; Shriberg, Paul, Black, & Santen, 2011) yet the reasons why are unknown. While SSEs and related disturbances to speech prosody might be a salient feature of autism on the first encounter, most research has focussed on the

(arguably) more serious nature of social problems in autism. The small amount of research that does exist on SSEs is heterogenic, similar to the presentation of the condition itself. Early studies suggest people with autism have normal speech development or that it is simply delayed (Kjelgaard & Tager-Flusberg, 2001; McCleery, Tully, Slevc, & Schreibman, 2006). However, recent work using sensitive statistical measurements showed that children with autism exhibited significantly higher rates of SSEs than typically developing (TD) children. Moreover, researchers have identified disordered speech development when phonetic and phonological analyses go beyond percentage consonants correct measures (Cleland et al., 2010; Shriberg et al., 2011; Wolk & Brennan, 2013).

The cause of SSEs in autism has not been fully explored in the literature. Up to this stage, researchers have focused on testing whether or not SSEs are a feature of autism, without exploring *why* these may occur. Now that recent evidence demonstrates SSEs are prevalent in autism, this chapter will examine their possible causes. There are currently two perspectives on why SSEs occur in children with autism: (1) the speech attunement framework first described by Shriberg et al. (2011) and (2) the speech motor impairment theory set out by Belmonte et al. (2013). These two perspectives will be discussed in the context of autism and results from the literature will be explored regarding how each framework might intersect.

2. Speech sound errors

Before exploring why SSEs are present in autism, it is important to understand what is expected of normal speech production, what speech sounds errors (SSEs) are and when they occur. Speech production and perception can breakdown at multiple levels, reducing the effectiveness of the final goal of fluent speech (Ferrand, 2014). The neural processing required for speech production and perception is still only partially understood (Baghai-Ravary & Beet, 2013). Speech production has been characterized as one of the most complex motor skills, functioning as multiple subsystems that must effectively coordinate together (Duffy, 2000). For example, the phonatory system, which consists of the laryngeal muscles, vocal folds etc., must work in a coordinated manner to achieve effective voice production. Likewise, the phonatory system must also be coordinated with

other sub-systems (e.g. respiratory system). Speech perception relies on the auditory system in which acoustic signals are transformed into meaningful representation of spoken language (Gandour & Krishnan, 2016). It requires various complex perceptual and cognitive tasks along the auditory pathway. Motor speech representations are important for both perception and production (Ravizza, 2005).

We use the term "Speech Sound Error (SSE)" here to describe difficulties with the production of speech sounds or speech segments (American Speech-Language-Hearing Association, 2017). Common clinical distortions or residual SSEs, such as rhotic or sibilant distortions are relatively minor and generally not associated with language or intelligibility deficits (Shriberg et al., 2011). In contrast, other children may have speech which is unintelligible even to close family members. SSEs are common in early childhood and include articulation errors (motor-based production deficits) and phonological errors (knowledge and use of speech sounds) (Eadie et al., 2015). Articulation errors often come in the form of distortions whereas phonological errors come in the form of substitutions and deletions such as consonant cluster reduction, final consonant deletion, velar fronting and stopping of fricatives. Articulation and phonological errors are not mutually exclusive and both of them can occur in a child's speech profile.

Problems start to arise when speech errors are not resolved during childhood and can then be described as either residual speech sound errors (RSSEs) or persistent speech errors. Residual speech errors arise as "leftovers" from an earlier speech delay (omission or substitution errors) that migrated closer to the norm to become distortions, whereas persistent speech errors are distortions that have been habituated from an early age. Residual speech sound errors often affect late acquired and motorically complex speech sounds such as /s/ and /r/ and are manifested as common clinical distortions of these sounds, for example lateralised /s/ or labiodentalised /r/. These types of errors might be particularly common in people with autism (33 % of verbal adolescents and adults with autism compared to just 1–2 % of the typical population, Shirberg et al., 2001). Why this is the case is not known, though Shriberg and colleagues ascribe it to a difficulty in fine-tuning to the ambient speech model.

3. Aetiology of speech sound errors and autism

Autism is a neurodevelopmental disorder in which there is a frequent co-occurrence of verbal and non-verbal deficits (American Psychiatric Association, 2013). People with autism are known to have persistent deficits in social behaviour, communication, and language, which may be entwined with their difficulties in producing intelligible speech. Evidence on speech impairment in autism is heterogenic. Some researchers have found it is disrupted in children with autism while others have found speech to be either delayed or developmentally appropriate when using perceptual and behavioural checklist assessments (Bartolucci & Pierce, 1977; Kjelgaard & Tager-Flusberg, 2001; Wilkinson, 1998). The literature currently lacks organisation of theoretical concepts, with different studies relying on different methods to measure SSEs.

Interactions between different areas of impairment in autism may cause SSEs. A triad of symptoms associated with autism could impair speech development: social motivation, cognitive (and motor) control, and perceptual control. Social motivation is a set of psychological and biological mechanisms that biases a person to orient to the social world, seek social interactions and maintain social binds. In autism there appears to be a decrease in attention given to social information, causing a cascading effect on the development of social cognitive skills (Chevallier, Kohls, Troiani, Brodkin, & Schultz, 2012). Impaired social motivation means the child may miss vital communicative opportunities in which to develop typical speech. This encompasses cognitive rigidity, a trait associated with speech disorders: one study found children with consistent speech disorder performed worse in cognitive flexibility tasks (Crosbie, Holm, & Dodd, 2009). A piecemeal cognition style alongside these social deficits may also result in autistic traits (Valla, Maendel, Ganzel, Barsky, & Belmonte, 2013). Social motivation impairment may negatively impact the development of neural networks critical to social cognition, e.g. face processing (Sterling et al., 2008). This cognitive style is described as "piecemeal" as the person's attention is either on the individual components of the face or on the physical configuration, losing the important social information in this interaction. This cognitive style may also have a significant effect on processing of speech, where only certain aspects of speech are attuned to,

e.g. phonological elements that are necessary for differentiating meaning may be given preference over phonetic aspects which signal speaker identity. Moreover, suprasegmental aspects of speech, like prosody or pitch, go unnoticed, resulting in the production of unusual-sounding speech.

Perceptual processing may also be a cornerstone in our understanding of autism (Baum, Stevenson, & Wallace, 2015). Sensory representations form the basis of higher-order cognitive representation. However, anomalies in sensory processing in autism are not well understood. Perceptual anomalies in autism may account for commonly found traits in language and speech impairment. People with autism have been found to have difficulty in social orientation to relevant auditory stimuli such as speech (Kuhl et al., 2005; Paul, Chawarska, Fowler, Cicchetti, & Volkmar, 2007). However, significantly more research is required to determine how this affects social and cognitive development.

A link has also been found between deficits in language, literacy and SSEs (Carson et al., 2003; Goffman, 1999; Hayiou-Thomas, Carroll, Leavett, Hulme, & Snowling, 2017; Whitehurst, Smith, Fischel, Arnold, & Lonigan, 1991; Williams & Elbert, 2003). Literacy and SSEs have a complex relationship, presence of speech sound disorders have been found to have a small but significant risk of poor phonemic skills, spelling and word reading. While SSEs alone only have a modest effect on literacy development, when it is part of additional risk factors such as language delay, these can have serious negative consequences. This is consistent with the findings that multiple risks such as SSEs and language delay/disorder can accumulate to predict reading disorder (Hayiou-Thomas et al., 2017).

Previously the primary focus of research in communication in children with autism has been language, prosody and behavioural difficulties (Kjelgaard & Tager-Flusberg, 2001; Owens, 2004; Paul & Norbury, 2012). The presence of speech sound difficulties is now being acknowledged (Cleland et al., 2010; Shriberg et al., 2011; Wolk, Edwards, & Brennan, 2016) but questions remain on the nature of the speech sound errors in autism. It is suggested that children with autism may exhibit speech production that is characteristically different in its organisation from typical speech. This may be due to developmental delay, but may also be due to differences in the underlying psychological or neuromotor structures required to produce speech and to practice it regularly in

everyday experience with social others. Below we discuss the presentation of SSEs in the current literature available.

4. Historical research of SSEs in children with autism

Around 40 years ago, behavioural studies concluded that children with autism had a delayed pattern of acquisition of speech sounds similar to children with intellectual disability (Bartolucci & Pierce, 1977). "Oddities" in speech production were often described following broad phonetic transcriptions (Pronovost, Wakstein, & Wakstein, 1966). Due to the lack of in-depth instrumental analyses or narrow phonetic transcription, these errors revealed little about specific speech patterns and conspicuously absent was the level of detail required to identify the minor articulatory distortions. Speech sound production was often only assessed in addition to other aspects of communication impairment (e.g. receptive and expressive language, social communication deficit, prosody). The main purpose of the analyses in these studies were often other aspects of communication impairment. The analysis of speech was therefore often only assessed in brief using parent questionnaires or short perceptual tests of single words (Pronovost et al., 1966). This severely limited the detail of findings beyond broad diagnostic categories.

During the last century and more recently, smaller, more in-depth case studies used phonological analysis to identify both delayed and disordered phonological processes (Wetherby, Yonclas, & Bryan, 1989; Wolk & Brennan, 2013; Wolk & Edwards, 1993; Wolk & Giesen, 2000). Wolk and Giesen (2000) carried out a phonological analysis of speech elicited with both object naming and spontaneous speech in four children with autism. They found typical but delayed phonological processes. However, they also identified atypical processes such as residual errors, unusual sound changes and chronological mismatch (where phonemes are not acquired in the developmentally typical order). The speech profiles contained evidence of both articulation errors and phonological errors. All children had a diagnosis of a phonological disorder ranging from mild to severe, with one child classed as non-verbal. Even within this very small sample, there is a huge variation, indicating that there are likely different subtypes of speech sound disorders in autism. One of the few studies that focused

exclusively on speech sound behaviour was carried out by Bartolucci and Pierce (1977). Using a picture-naming task (single word analysis) to assess both perception and production, they compared children with autism to TD children and children with intellectual disability. The analysis was limited. Twenty-four consonant sounds were broadly transcribed using the International Phonetic Alphabet and errors were compared in terms of percentage consonants correct. They concluded that verbal children with autism failed to show any atypical traits in the production or perception of speech sounds. Additionally, when looking at the findings in further detail using more sensitive phonetic analysis they found a delayed pattern of acquisition, similar to children with intellectual disability. Conversely, there was a significant difference in the percentage of errors on liquids made by children with autism (11.4 %) compared to children with intellectual disability (4.7 %) and TD children (0 %). Similarly, in their perception task of liquids, there was a significant difference between errors made by children with autism (16.6 %) compared to TD (8.6 %) and intellectual disability (5.5 %). These results indicate an atypical profile rather than delayed speech acquisition. Further analysis of the consistency and frequency of errors may have revealed more about the speech sound patterns of these different groups. These studies tell us that SSEs were being identified in groups of children with autism, but we require measures that are more sensitive in order to determine the causes. Further in-depth analyses are required to determine the *pattern* of errors produced by this particular group at all stages of communication development.

5. Atypical speech in young children with autism

Findings from multiple researchers suggest that children with autism at the prelinguistic stage of communication have different phonatory qualities than their peers and children with other developmental disorders (Schoen, Paul, & Chawarska, 2011). This information could contribute to early identification that would allow intervention to be put in place during the optimal period, i.e. in the first years of life, when the brain is developing rapidly and there is significant neural plasticity.

Atypical speech development appears to be identifiable at an early stage of communication in children with autism. Smaller case studies using

perceptual phonological analysis identified SSEs in young children with autism (under 5 years). Wetherby et al. (1989) analysed the syllables of vocalizations made by three children with autism (under 5 years) in a 30-minute sample of communicative behaviour. They found that the children had a deficient proportion of vocal acts that contained a consonant. This absence of some consonants in communicative acts might be an early warning sign that speech is not developing normally. It is difficult to distinguish whether this is a result of phonological or articulatory issues, thus further analysis of the speech errors made would be required. Samples of communicative interactions are often used in younger cohorts to analyse their phonological development perceptually. One interesting finding from this in-depth analysis is the presence of "atypical vocalizations" in children with autism. Atypical vocalizations were the primary aspect of prelinguistic communication that differentiated children at high risk for autism (9–12 months) from children at low risk (Schoen et al., 2011). To investigate this further Schoen et al. (2011) studied phonological and vocal behaviour using broad phonemic transcription of speech-like utterances and coded non-speech vocalizations without recognisable consonants. They found 30 toddlers (18–36 months) with autism exhibited "atypical vocalizations" and overall a limited number of consonants compared to two groups of TD children (age-matched and language- matched). Whilst the percentage of consonants correct was not different from their peers, the number of speech-like utterances produced was significantly less. The main area of difference between the children with autism and their peers was the presence of "atypical vocalizations". These atypical vocalizations came mainly in the form of high-pitched squeals (Schoen et al., 2011). What this research might suggest is that toddlers with autism do not align their speech to the duration, pitch and phonotactic properties of their ambient language environment.

Toddlers with autism may not tune into the language model of their environment (Sheinkopf, Mundy, Kimbrough Oller, & Steffens, 2000). Their failure to attend to their ambient language environment may negatively affect their ability to acquire spoken language, which in severe cases can mean people with autism remain nonverbal throughout life. In a study of early vocal behaviours in young children with autism (n=15) and children with developmental delays (n=11), the children with autism did not

differ in production of well-formed complex canonical vocalizations, but had significantly more utterances with atypical vocal quality (Sheinkopf et al., 2000). Canonical vocalizations are well-formed consonant-vowel sequences with rapid CV transitions. An impairment in these sequences serves as a sign of speech motor control impairment, which was not the case in this sample. However, the significant presence of atypical vocal quality may be an indicator of speech perception issues. Wallace et al. (2008) reanalysed this data using acoustic analysis and more refined categorization techniques and found that children with autism produced more atypical phonatory qualities than children with developmental delay. On the contrary, Schoen et al. (2011) found that toddlers with autism followed a normal trajectory of phonological development, suggesting no issues with speech development. However, there was a significant presence of atypical vocalizations in their speech, which may be due to a presence of speech attunement issues. Descriptions of vocal profiles differentiating developmental profiles could provide valuable evidence for early biomarkers of autism and could help us explain the origin of the issues in speech production demonstrated at a later age. Further research as to why atypical vocalizations occurs in young children with autism is required.

Current research of SSEs in autism has started to use technology as a means of increasing the sensitivity of analysis of speech, both qualitatively and quantitatively. Shriberg et al. (2001) found a predominance of articulation errors in children with autism using the "PEPPER" software. This software allowed analysis of the type and frequency of consonant and vowel errors in conversational speech. Using this method, they found 33 % of the cohort with autism had at least one type of speech distortion error (residual speech sound errors, such as lateral lisps). These may be an indicator of a disordered speech profile, rather than delayed speech acquisition as previously assumed in earlier research.

An interesting finding from this study was the significant presence of "residual speech sound errors". These occur when speakers older than nine years have two or more of the same type of residual distortion errors (e.g. dentalized sibilants, derhotacization). Thirty-three percent of the children with autism presented with residual speech errors, a significant proportion compared to the expected 1–2 % found in the TD population (Flipsen, 2015). Residual speech sound errors are clinically significant as

they involve sub-phonemic changes in articulatory place and manner and can persist over the individual's lifespan.

Further evidence of SSEs in autism was found by Cleland et al., (2010) who report atypical/non-developmental SSEs in children with autism. They carried out a phonetic and phonological analysis of speech sound production in 69 children with autism. Using standardized clinical perceptual assessments, only 12 % of the sample received a diagnosis of speech delay/disorder. However, when using further in-depth phonological and phonetic analysis, they found 41 % of the group produced speech errors indicative of both speech delay and speech disorder.

The clinical assessment of speech used was a perceptual assessment called the Goldman Fristoe Test of Articulation (GFTA-2; Goldman & Fristoe, 2000). This is one of the few standardized assessments of speech sound behaviours in children. It examines speech sounds in the context of single words. Further research beyond this assessment such as single words of increasing complexity (polysyllables), maximum performance tasks or spontaneous speech may reveal motor constraints which have a substantial negative impact on intelligibility or increase the likelihood of an SSE occurring. Cleland et al. (2010) found non-developmental speech errors occurred despite whether a child's standard score fell within normal range or not on the GFTA (Goldman & Fristoe, 2000). This implies SSEs produced by people with autism may not meet the criteria for a speech disorder in clinical assessments. However, there is more to understand in relation to speech profiles of people with autism, which may reveal information about the different subtypes within autism and whether this aligns with a particular speech profile.

The study by Cleland et al. (2010) is in agreement with previous findings by Kjelgaard & Tager-Flusberg (2001) and Rapin, Dunn, Allen, Stevens, & Fein (2009) that children with autism make a number of SSEs. Cleland et al. (2010) found in their sample that while speech was characterised by developmental phonological errors (gliding, cluster reduction and final consonant deletion), non-developmental errors, indicative of a speech disorder, were also present (e.g. phoneme specific nasal emission and initial consonant deletion). To understand *why* there may be SSEs in the group, Cleland et al. (2010) carried out a battery of standardized assessments in speech, language and non-verbal cognition to determine if there are any

causal links. Interestingly no relationship between speech and language or speech and cognition was identified in this group. This indicates SSEs may be a result of another impeding factor. Cleland et al. (2010) hypothesised that the increase of SSEs may be due to an underlying neuromotor difficulty. Additionally, it could also be due to speech attunement difficulties. Further analysis of auditory perceptual abilities and speech motor abilities is needed to understand the origin of the SSEs in this group.

Wolk and Giesen (2000) carried out a phonetic inventory and process analysis and found in four siblings with autism speech processes indicative of delayed speech development. In addition, they identified atypical processes such as residual articulation errors, unusual sound changes and chronological mismatch in their speech profiles. All four children were significantly delayed in gross motor and fine motor abilities. The combination of residual articulation errors (indicative of motor issues) and unusual sound changes (indicative of perceptual issues) suggests these children appear to have a combination of both speech motor control and speech perception issues. However, they did not find differences in suprasegmental production; children with autism did not produce vocalizations different in fundamental frequency or duration from TD peers, suggesting they are able to tune in to their ambient environment effectively in some ways. These children may be a different subtype of autism. These are limited measures of suprasegmental ability and would require further analysis.

6. Methodological issues of measurement of SSEs

Multiple methodological issues need to be taken into account when assessing SSEs. Firstly, analysis of speech in single-word contexts may be ineffective. It does not examine the effect of complex articulatory gestures during spontaneous speech, which is significantly more motorically complex than single word production (Adams, 1998). Kjelgaard and Tager-Flusberg (2001) investigated language and speech production in eighty-nine children (4;0–14;0 years) with autism. They argued whilst there was significant heterogeneity in the children's language skills, their articulation skills were relatively spared. However, this conclusion is brought into question when noting they also used the Goldman Fristoe Test of Articulation. It required further phonetic and phonological analysis for

Cleland et al. (2010) to identify speech distortions using this assessment alone which was not carried out in this study. Therefore, Kjelgaard and Tager-Flusberg's (2001) assessment of speech sounds may have been inadequate to determine whether their sample of children had SSEs. Clinically the children may not have met the diagnosis for speech sound disorder, but there is value in understanding if SSEs occur in order to gain understanding on speech perception and production in autism.

Perceptual single-word assessments helped identify irregularities in speech sound production in some studies. Rapin et al. (2009) used a single-word assessment, the Photo Articulation Test (Lippke, Dickey, Selmar, & Soder, 1997) to analyse the speech of 62 children with autism. The test yields a score for correct speech sounds produced in naming single word objects. Similar to Cleland et al. (2010) and Shriberg et al. (2011) they found that 28 % of the participants' speech was characterized by persistently and severely impaired speech sound production; this was despite better language comprehension. Additionally, they analysed the spontaneous speech samples and concluded that "several minutes of conversation provides more opportunities for mispronunciations than the single words of the Photo Articulation Test (Rapin et al., 2009). Their assessment was in agreement with the Photo Articulation Test results, finding 28 % of the speech sample was characterized by severely impaired expressive phonologic skills. However, this was not an in-depth analysis, the authors rated each child's speech on a 3-point scale (0= normal to 2= severe impairment). Whilst both these results indicate an abnormality in speech production of some children with autism, again it does not go beyond a quick perceptual analysis. One reason that (Wolk & Giesen, 2000) may have identified SSEs whereas McCleery et al. (2006) and Kjelgaard and Tager-Flusberg (2001) did not, is that they elicited speech using two methods: object naming and spontaneous speech utterances. Additionally, they did not rely on perceptual standardized assessments. It is vital that researchers consider speech in multiple contexts to ensure subtle articulation errors are identified.

The use of ineffective standardised assessments, and issues of the nature of autism can cause difficulty in speech sound assessment (Macrae, 2017). For instance, in children with autism with severe language impairment, there is often difficulty obtaining a speech sample due to expressive language difficulties associated with autism. McCleery et al.

(2006) investigated the consonant production of 14 severely language delayed children with autism and 10 TD children. To assess speech in the context of severe language delay, their assessment involved a communicative inventory providing opportunities for the child to produce voiced and voiceless consonant sounds. All vocalisations, including babbling, were scored in an effort to determine the child's consonant production repertoire. McCleery et al. (2006) concluded that the children with autism showed the same general speech sound production pattern as TD and language-learning impaired children. Interestingly, they acknowledged that the children with autism produced more sounds that were not classified as developmentally normal but did not carry out further analysis on these errors. Transcription and counting of these errors may have revealed an alternative speech pattern in children with autism or clinically significant errors. Furthermore, analysis of the "abnormalities" may have provided indicators of whether the nature of these errors were motoric or phonological, similar to studies of early communication behaviour (Paul, Fuerst, Ramsay, Chawarska, & Klin, 2011; Schoen et al., 2011).

Previous studies investigating speech of children with autism may have misidentified subtle articulation errors because of the imprecise nature of perceptual speech assessment. The assessments described in studies using single word analysis are reliant on these perceptual measures of speech. Precise information about articulatory movements cannot be identified from perceptual analysis alone. Yet this information may reveal more about speech motor control in autism and whether alternative movement strategies exist due to motor impairment and/or attunement issues. Auditory perceptual judgements are susceptible to errors and bias of the listener (Kent, 1996). An example of this is listener normalization where the listener mistakenly recognises phonemes that were not produced by the speaker. Even if errors are identified, suitable transcription techniques are lacking in the ability to distinguish these errors (Kent, 1996). Broad phonetic transcription is reliant on the categories of the IPA chart, even though variation within each category can vary significantly across individuals (Mowrey & MacKay, 1990). Speech assessment needs to look at a speech in multiple contexts, where articulatory gestures are more complex and using more in-depth forms of phonetic and phonological analysis. This

will help to determine whether SSEs in autism are due to speech perception difficulties and/or tuning into speech or to speech motor control issues.

7. Potential causes of SSEs in autism

Several research groups have reported that around a third of children with autism present with oral motor or speech sound abnormalities at various levels of severity (Belmonte et al., 2013; Cleland et al., 2010; Shriberg et al., 2011). We will now discuss two potentially complementary perspectives of *why* this may be the case.

The "speech attunement framework" was originally developed to explain common speech errors in the otherwise typically developing populations (e.g. dentalized sibilants). Shriberg et al. (2011) developed the "speech attunement framework" due to ongoing suggestions that impairments in gross motor, fine motor and oral motor control in people with autism were associated with the speech deficits frequently exhibited. The speech attunement framework posits that a child learning speech needs to attend to their ambient environment or 'tune in' to models in that environment. For example, young children adopt dialect features of their peers by tuning in. In addition, they need to make small and careful adjustments to their speech production to 'tune up' for accurate and socially acceptable speech production (Shriberg et al., 2011). It is also important at this stage to have a maturing speech motor system that ensures adjustments made to speech can be done so with adequate control by the child. A difficulty with speech motor control could intersect with speech attunement and cause the heterogenic speech profiles identified in speakers with autism.

As discussed earlier, people with autism may have a reduced ability and/or motivation to focus on the subtle details of articulation, due to social motivation impairment. This prevents them from making minute adjustments in order to produce speech similar to their social partners and others in their ambient environment (Shriberg et al., 2011). In essence, children with autism are thought not to have the psychological conditions necessary to engage socially with others through language to give the necessary experiences for learning speech. Speech attunement may be affected in people with autism by various combinations of the following conditions:

a) Enhanced auditory capacity, often observed in people with autism (Baum et al., 2015) may lead to earlier "tuning in" when motor maturity has not been achieved. Therefore, SSEs develop due to motor constraints.
b) Constraints in affective social reciprocity, a common trait of people with autism (Chevallier et al., 2012) may delay "tuning in" and any motor speech disorder present may impair the ability to tune up.

Shriberg et al. (2011) investigated whether children with autism had 'speech attunement' issues and a comorbid speech motor disorder, specifically childhood apraxia of speech (CAS). CAS impairs the precision and consistency of speech movements, despite the lack of any neuromuscular deficits. To determine whether the increased presence of SSEs in children with autism was a result of speech attunement issues or CAS, Shriberg et al. (2011) examined the continuous speech of 40 children with autism; 40 TD children; 13 children with speech delay; and 15 individuals with CAS. They used software PEPPER (Programs to Examine Phonetic and Phonological Evaluation Records; Shriberg et al., 2001). They used this software to perceptually and acoustically analyse continuous speech samples, transcribing and prosody-coding subsets of the speech.

This detailed analysis was designed to identify specific signs of CAS or motor speech disorder, e.g. slow speaking and articulation rate, spatiotemporal vowel errors and distorted consonants (American Speech-Language-Hearing Association (ASHA), 2007; Aziz, Shohdi, Osman, & Habib, 2010). Although, it can be argued that the findings did not support a diagnosis of motor speech disorder or CAS, children with autism had voice differences not reported in the CAS group, e.g. inappropriate loudness, abnormally high pitch. Additionally, they had appropriate rate and stress, in direct contrast to symptoms of CAS. Shriberg et al. (2001) use these results as evidence of speech attunement issues, rather than a motor-speech impairment. However, 75 % of children with autism had increased repetitions and revisions; a symptom demonstrated by CAS speakers, with both groups producing these significantly more than TD children. As a result, some indicators of speech attunement were noted:

a) Increased repetitions and revisions, consistent with the description of autistic speech as "disfluent".

b) Misplaced stress, often described as "off" or "singsong" (Peppé, McCann, Gibbon, O'Hare, & Rutherford, 2007). This stress is dissimilar to the well-documented "excessive-equal" stress pattern in apraxia of speech.
c) Inappropriate loudness and pitch.
d) Higher rates of speech delay and speech errors relative to population estimates.

A study by Baron-Cohen and Staunton (1994) found that children with autism whose mothers were non-native English speakers were more likely to develop their mother's non-native accent (83.3 % of the sample) than that of their peers. TD children without social communication difficulties have a strong drive to identify with peers they engage with regularly. A lack of drive to identify with peers, present in autism, could lead to weaknesses in opportunities for speech attunement. We would then expect children with autism to show higher rates of phonetic distortions. Indeed, research by Shriberg et al. (2001) has shown that adults with autism frequently produce articulation distortions, such as sibilant and rhotic distortions. It should be noted that this study included both adolescents and adults, speech distortions did not resolve with age and were classified as "residual speech errors". Cleland et al. (2010) also found a number of older children with similar phonetic distortions. These studies tell us that speech distortions, which may be a result of poor speech attunement in childhood, can continue through to adolescence and adulthood. This is an important finding as children with residual speech sound errors face an increased risk of social, emotional and/or academic challenges relative to their peers with typical speech (Hitchcock, Harel, & Byun, 2015). This likely compounds the social and emotional disadvantages children with autism already have.

One aspect of speech attunement that requires examination is the effect on suprasegmental attributes such as pitch. Tonal languages such as Chinese and Thai rely on the ability to perceive pitch as they involve categorical distinctions of lexical tone. Lexical tones serve a phonemic role, they are vital for speech comprehension and production (Wang, Wang, Fan, Huang, & Zhang, 2017). Wang at el. (2017) found in an event-related potential (ERP) study that 16 children with autism had lexical tone processing that

was impaired and likely had its root cause as a phonological deficit in categorical perception, similar to the findings of Yu et al. (2015). Bonneh, Levanon, Dean-Pardo, lossos & Adini (2011) also found abnormal speech spectrum and fundamental frequency processing in young autistic children who spoke Hebrew. They assessed long-term average spectrum and fundamental frequency variability in 60-second speech samples of 41 children with autism using a picture-naming task. Compared to the control group, the spectra were shallower and there was less harmonic structure in the group with autism. These results imply abnormal processing of auditory feedback or elevated noise and instability in the mechanisms that control phonation. All of which could have a significant impact on the child's ability to tune into speech. Finally, Lyakso, Frolova and Grigev (2016) assessed acoustic features of speech such as fundamental frequency (f0), f0 range, formants, frequency and duration in emotional speech, spontaneous speech and repetitions of words in 60 Russian-speaking children with autism. Similar to previous studies in tonal languages and Hebrew, abnormal prosody was a consistent feature. All children with autism had high values of fundamental frequency, abnormal spectrum and well-marked high frequency. Stressed vowels also had higher values of fundamental frequency. Results indicated speech abnormalities in autism is reflected in their spectral content and fundamental frequency variability. Understanding and producing appropriate stress patterns appears to be difficult for people with autism.

The second perspective for increased prevalence of SSEs in autism is that a subtle, but significant, motor control impairment in autism causes differences in speech production (Adams, 1998; Barbeau, Meilleur, Zeffiro, & Mottron, 2015; Belmonte et al., 2013). This perspective is becoming increasingly attractive as evidence accumulates that motor disruptions in other domains, such as in the purposeful movement of the arms (Crippa et al., 2015; Torres et al., 2013), legs and posture movements in gait (Nayate et al., 2012; Rinehart et al., 2006). Additionally fine motor control during writing and object manipulation (Fuentes, Mostofsky, & Bastian, 2009) are disrupted in children with autism. A recent meta-analysis of motor data in autism suggest motor disruption may be a core feature of autism and not merely a co-morbid or associated condition (Fournier, Hass, Naik, Lodha, & Cauraugh, 2010).

Motor impairment is evident at both gross and fine level in autism. There may be a fundamental underlying problem with motor timing and integration required to produce the correct, efficient kinematic patterns required of skilled movements, including speech (Beversdorf et al., 2001; Gowen & Hamilton, 2013; MacNeil & Mostofsky, 2012; Mostofsky, Powell, Simmonds, & Goldberg, 2009; Whyatt & Craig, 2013). Such disruption to movement early in a child's development is thought to contribute to the broad autism phenotype, disrupting expressive intention and purposeful engagement with others, causing frustration, distress and isolation (Trevarthen and Delafield-Butt, 2013). In verbal expression, articulating fluently requires intricate control and coordination of speech motor mechanisms (Gracco, 1994). Therefore, this perspective proposes the increased rate of SSEs present in children with autism may be a result of common, underlying motor difficulties. Indeed, the residual articulation errors reported by Shriberg et al. (2001) affect the late acquired and articulatory complex speech sounds such as sibilants and rhotics; sounds that require intricate speech motor skills.

Evidence of motor impairment in autism is growing. Neuroanatomical correlates have been proposed for the observed difficulties in motor functioning including abnormalities in the cerebellum (Fatemi et al., 2012), disruption in brain synchronization (Welsh, Ahn, & Placantonakis, 2005), impaired sensory input and multisensory integration (Gowen & Hamilton, 2013). As a result, it is suggested that if general motor abilities are impaired, this could result in a speech motor control impairment (Barbeau et al., 2015). Adams (1998) examined oral-motor and motor-speech production of four young children with autism compared to TD children in both simple and complex phonemic production. Data indicated that children with autism had significantly more difficulty performing oral movements and complex syllable production tasks compared to TD children. These results could indicate a speech motor impairment. However, due to their small sample size, these results are not generalizable. More research of speech motor control compared to general motor abilities is crucial to understanding the case of SSEs.

The connection between speech motor control and general motor abilities has been examined in the TD population. Nip, Green and Marx

(2011) found that TD infants showed a correlation between changes in articulatory movements and development of early communication. Using a motion capture system every three months, the movements of the upper lip, lower lip and jaw were recorded from 24 children (between the ages of 9–21 months). Children who had reduced speech motor control had a delayed trajectory of communication development. Significant associations were identified between orofacial kinematics and the standardized measures of language and cognitive skills, even when age served as a covariate. This initial evidence suggests interactions between cognition, language and speech motor skills during early communication development. Further research is required to identify and quantify causal relations among these co-emerging skills and whether this extends to general motor ability. Alcock (2006) also found that motor control was associated with an existing language impairment, particularly oral motor control. Moreover, Lewis et al. (2011) found children with SSD were slower to complete diadochokinesis tasks and had differences in their oral motor control compared to TD children. These studies provide evidence of an inherent link between speech and general motor capabilities, though whether there is a causative mechanism (in either direction) is unknown.

Little research explores the relationship between speech and general motor impairment in children with autism. Nevertheless, there have been some interesting findings in children with idiopathic speech disorder who have been found to have reduced performance on tasks that involve visual motor control and fine motor control, e.g. grasping and object manipulation (Newmeyer et al., 2007). Peter and Stoel-Gammon (2008) found children with SSEs had deficits in repetitive finger tapping and clapping exercises associated with fine motor control. Lewis et al. (2011) found children with SSE were slower to complete diadochokinesis tasks and a maximum phonation task associated with the competency of speech motor control, compared to TD children. Bradford and Dodd (1994) compared ten phonologically delayed children, ten children with consistent phonological disorder and ten children with inconsistent error patterns. Groups did not differ on simple motor tasks; however, the group with inconsistent error patterns performed significantly worse in timed motor planning tasks and expressive novel-work learning tasks than in the other

two groups. These results provide support for the perspective that inconsistent error patterns are associated with a deficit in some aspects of fine motor planning, a similar pattern that has also been identified in autism (Fournier et al., 2010).

Timing is a fundamental aspect of speech production. Fluent speech requires information to be selected, sequenced and articulated in an accurate and time sensitive manner. A set of quasi-autonomous articulatory systems need to work in coordination (Kotz & Schwartze, 2016; Maassen & Van Lieshout, 2010). Whilst little research has been carried out on speech timing, studies suggest there may be abnormalities in sensorimotor timing in children with autism. Anzulewics, Sobota and Delafield-Butt (2016) found an increase in the speed of fast taps and swipes in children with autism playing an iPad game. Torres, et al. (2013) found an increase in the acceleration-deceleration phases of a reach-to-touch task in children with autism. These tasks demonstrate a subtle, but significant disruption to moment-by-moment control of movement occurring in the region of 30–70 ms, a temporal domain important for speech. Over- and under-compensations of such rapid shifts in force are thought to underpin the overt motor disruptions typically observed (Trevarthen & Delafield-Butt, 2013; Whyatt & Craig, 2013). These compensations may affect basic perception and effect experience resulting in disrupted speech development due to lack of coordination of articulatory systems (Colwyn Trevarthen & Delafield-Butt, 2017). Cook, Blakemore and Press (2013) found sub-second control of velocity and acceleration was affected in individuals with autism in simple arm-swing tasks. This study indicated that fast timing at less than a second (sub-second) required of speech motor control might be disrupted in limb and hand movements in individuals with autism.

Future research needs to look at both linguistic and motor planning skills in children with autism to describe accurately the range of cognitive processes that may be affecting their speech production. These studies above indicate there may be a deficit in motor planning and programming associated with speech sound disorders, but the origin is still unknown (Shriberg & Kwiatkowski, 1994). Therefore, it is important we look at the studies on SSEs in autism in detail to determine what knowledge exists on their nature and causes.

8. Conclusion and future directions

Researchers have identified atypical speech development in children with autism (Shriberg et al. 2001; Cleland et al. 2010; Wolk & Brennan 2013). However, it has been argued these errors are within a sequence of normal development (delayed) rather than atypical (Kjelgaard & Tager-Flusberg 2001; McCleery et al. 2006). Inconsistent outcomes in the literature may be a result of inconsistent and reduced specificity of the perceptual measurements used across studies. This could also reflect the heterogeneity in the population of people with autism in their production of speech. If deficits in speech motor control mirror the deficits in fine motor control, then finer-grained techniques may be needed to identify them. This is important, because even if speech motor control problems are subtle, their existence might indicate that an underlying motor impairment is at the heart of autism. It is unlikely that subtle speech motor control problems will be identified with judgments on the correctness of productions of single words. Instead, one needs speech tasks such as maximum-performance tasks that tax the motor system. Alternatively, it is possible that articulatory analysis will identify qualitative differences in the articulations of children with autism compared to typical speakers. Indeed, articulatory analysis, namely ultrasound tongue imaging, has been used in one study to assess and treat abnormal articulations in children with autism. Cleland, Scobbie, Heyde, Roxburgh and Wrench (2019) (found that ultrasound visual feedback might facilitate speech sound learning. While the study was not focused exclusively on children with autism, three of the children presented with SSEs and autism within the sample and responded to intervention. Although ultrasound tongue imaging is at early stages of development for assessment and intervention, it is a promising method of analysing SSEs in the depth required to identify subtle articulation errors in children with autism.

In conclusion, it is vital to determine *why* SSEs may be occurring and whether such occurrences are a result of disruption to speech attunement, a disruption to speech motor issues, or, more likely, both. Each aspect of speech development and production could affect the other, the two are entwined within the life of the child. Children with autism appear to have less drive to attune to the speech of peers due to particular social

impairments. This may result in a reduction of motivation to produce speech that is intelligible and functional for others to comprehend. This may explain why we see prosodic abnormalities and unusual distortions errors, which do not affect intelligibility, such as phoneme-specific nasal emission (Cleland et al., 2010) and difficulties with articulatory complex speech sounds (Shriberg et al., 2011). Conversely, the disruption in speech motor performance that thwarts its intended meaning for others can itself drive a reduction in motivation to attune, leading to the same set of autistic consequences. Either way, improved aetiological understanding will help to determine principal underlying capacities and therefore routes to more effective intervention. Current research does not provide a clear picture of what theory best applies – if indeed either theory is appropriate without consideration of the other. In addition, it may be that there are subgroups of children with SSEs within the broad autism spectrum. Such ideas require testing. Future research also needs to look equally at both linguistic and motor skills in children with autism to describe accurately the range of mental and neurophysiological process that may be affecting the production of speech. Understanding these questions will help to improve effective speech therapy interventions to target the underlying disruptions that give rise to SSEs at an early age and develop bio-markers for earlier diagnosis of autism.

References

Adams, L. (1998). Oral-motor and motor-speech characteristics of children with autism. *Focus on Autism and Other Developmental Disabilities*, *13*(2), 108–112.

Alcock, K. (2006). The development of oral motor control and language. *Down's syndrome, research and practice. The Journal of the Sarah Duffen Centre*, *11*(1), 1–8.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders : DSM-5*. Arlington, VA : American Psychiatric Association.

American Speech-Language-Hearing Association. (2017). Speech sound disorders: articulation and phonology: Overview. Retrieved June 28, 2017 from http://www.asha.org/PRPSpecificTopic.aslippx?folderid=8589935321&section=Overview

American Speech-Language-Hearing Association (ASHA). (2007). Childhood Apraxia of Speech. Retrieved August 25, 2017, from www.asha.org/policy

Anzulewicz, A., Sobota, K., & Delafield-Butt, J.T. (2016). Toward the autism motor signature: Gesture patterns during smart tablet gameplay identify children with autism. *Scientific Reports*, *6*(1), 31107. doi:10.1038/srep31107

Aziz, A.A., Shohdi, S., Osman, D.M., & Habib, E.I. (2010). Childhood apraxia of speech and multiple phonological disorders in Cairo-Egyptian Arabic speaking children: Language, speech, and oro-motor differences. *International Journal of Pediatric Otorhinolaryngology*, *74*(6), 578–585.

Baghai-Ravary, L., & Beet, S.W. (2013). *Automatic Speech Signal Analysis for Clinical Diagnosis and Assessment of Speech Disorders*. New York, NY: Springer New York.

Barbeau, E.B., Meilleur, A.S., Zeffiro, T., & Mottron, L. (2015). Comparing motor skills in autism spectrum individuals with and without speech delay. *Autism Research*, *8*(6), 682–693.

Baron-Cohen, S., & Staunton, R. (1994). Do children with autism acquire the phonology of their peers? An examination of group identification through the window of bilingualism. *First Language*, *14*(42–43), 241–248.

Bartolucci, G., & Pierce, S.J. (1977). A preliminary comparison of phonological development in autistic, normal, and mentally retarded subjects. *International Journal of Language & Communication Disorders*, *12*(2), 137–147.

Baum, S.H., Stevenson, R.A., & Wallace, M.T. (2015). Behavioral, perceptual, and neural alterations in sensory and multisensory function in autism spectrum disorder. *Progress in Neurobiology*, *134*, 140–160.

Belmonte, M.K., Saxena-Chandhok, T., Cherian, R., Muneer, R., George, L., & Karanth, P. (2013). Oral motor deficits in speech-impaired children with autism. *Frontiers in Integrative Neuroscience*, *7*, 47. doi:10.3389/fnint.2013.00047

Beversdorf, D., Anderson, J., Manning, S., Anderson, S., Nordgren, R., Felopulos, G., & Bauman, M. (2001). Brief report: Macrographia in high-functioning adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *31*(1), 97–101.

Bonneh, Y.S., Levanon, Y., Dean-Pardo, O., Lossos, L., & Adini, Y. (2011). Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in Human Neuroscience*, *4*, 237. doi: 10.3389/fnhum.2010.00237.

Bradford, A., & Dodd, B. (1994). The motor planning abilities of phonologically disordered children. *International Journal of Language & Communication Disorders*, *29*(4), 349–369.

Carson, C.P., Klee, T., Carson, D.K., & Hime, L.K. (2003). Phonological profiles of 2-year-olds with delayed language development predicting clinical outcomes at age 3. *American Journal of Speech-Language Pathology*, 12(1), 28–39.

Chevallier, C., Kohls, G., Troiani, V., Brodkin, E.S., & Schultz, R.T. (2012). The social motivation theory of autism. *Trends in Cognitive Sciences*, *16*(4), 231–239.

Cleland, J., Scobbie, J. M., Roxburgh, Z., Heyde, C., & Wrench, A. (2019). Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback. *Journal of Speech, Language, and Hearing Research*, 62(2), 229–246.

Cleland, J., Gibbon, F., Peppé, S., O'Hare, A., & Rutherford, M. (2010). Phonetic and phonological errors in children with high functioning autism and Asperger syndrome. *International Journal of Speech-Language Pathology*, *12*(1), 69–76.

Cook, J.L., Blakemore, S.J., &Press, C. (2013). Atypical basic movement kinematics in autism spectrum conditions. *Brain*, *136*(9), 2816–2824.

Crippa, A., Salvatore, C., Perego, P., Forti, S., Nobile, M., Molteni, M., & Castiglioni, I. (2015). Use of machine learning to identify children with autism and their motor abnormalities. *Journal of Autism and Developmental Disorders*, *45*(7), 2146–2156.

Crosbie, S., Holm, A., & Dodd, B. (2009). Cognitive flexibility in children with and without speech disorder. *Child Language Teaching and Therapy*, *25*(2), 250–270.

Duffy, J.R. (2000). Motor speech disorders: Clues to neurologic diagnosis. In C.H.Adler & J.E. Ahlskog (Eds.) *Parkinson's Disease and Movement Disorder*s. Totowa, NJ: Humana Press, 35–553.

Eadie, P., Morgan, A., Ukoumunne, O.C., Ttofari Eecen, K., Wake, M., & Reilly, S. (2015). Speech sound disorder at 4 years: Prevalence,

comorbidities, and predictors in a community cohort of children. *Developmental Medicine & Child Neurology*, *57*(6), 578–584.

Fatemi, S.H., Aldinger, K.A., Ashwood, P., Bauman, M.L., Blaha, C.D., Blatt, G.J., Welsh, J.P. (2012). Consensus paper: Pathological role of the cerebellum in autism. *The Cerebellum*, *11*(3), 777–807.

Ferrand, C.T. (2014). *Speech Science: An Integrated Approach to Theory and Clinical Practice*. Michigan: Pearson/Allyn and Bacon.

Flipsen, P. (2015). Emergence and prevalence of persistent and residual speech errors. *Seminars in Speech and Language*, *36*(4), 217–223.

Fournier, K., Hass, C., Naik, S., Lodha, N., & Cauraugh, J. (2010). Motor coordination in autism spectrum disorders: A synthesis and meta-analysis. *Journal of Autism and Developmental Disorders*, *40*(10), 1227–1240.

Fuentes, C.T., Mostofsky, S.H., & Bastian, A.J. (2009). Children with autism show specific handwriting impairments. *Neurology*, *73*(19), 1532–1537.

Gandour, J. T., & Krishnan, A. (2016). Processing tone languages. In G. Hickock, & S. Small (Eds.) *Neurobiology of Language*, Amsterdam: Elsevier, pp. 1095–1107.

Goffman, L. (1999). Prosodic influences on speech production in children with specific language impairment and speech deficits: Kinematic, acoustic, and transcription evidence. *Journal of Speech, Language, and Hearing Research*, *42*(6), 1499–1517.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation 3(GFTA-3)*. San Antonio: Pearson: PsychCorp.

Gowen, E., & Hamilton, A. (2013). Motor abilities in autism: A review using a computational context. *Journal of Autism and Developmental Disorders*, *43*(2), 323–344.

Gracco, V. (1994). Some organizational characteristics of speech movement control. *Journal of Speech and Hearing Research*, *37*(1), 4–27.

Hayiou-Thomas, M.E., Carroll, J.M., Leavett, R., Hulme, C., & Snowling, M.J. (2017). When does speech sound disorder matter for literacy? The role of disordered speech errors, co-occurring language impairment and family risk of dyslexia. *Journal of Child Psychology and Psychiatry*, *58*(2), 197–205.

Hitchcock, E., Harel, D., & Byun, T. (2015). Social, emotional, and academic impact of residual speech errors in school-aged children: A survey study. *Seminars in Speech and Language*, *36*(4), 283–294.

Kent, R.D. (1996). Hearing and believing some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology*, *5*(3), 7–23.

Kjelgaard, M.M., & Tager-Flusberg, H. (2001). An investigation of language impairment in autism: Implications for genetic subgroups. *Language and Cognitive Processes*, *16*(2–3), 287–308.

Kotz, S.A., & Schwartze, M. (2016). Motor-timing and sequencing in speech production. In G. Hickock, & S. Small (Eds.) *Neurobiology of Language*, Amsterdam: Elsevier, pp. 717–724.

Kuhl, P.K., Coffey-Corina, S., Padden, D., & Dawson, G. (2005). Links between social and linguistic processing of speech in children with autism: Behavioural and electrophysiological measures. *Developmental Science*, *8*(1), 1–12.

Lewis, B., Avrich, A., Freebairn, L., Hansen, A., Sucheston, L., Kuo, I., & Stein, C. (2011). Literacy outcomes of children with early childhood speech sound disorders: Impact of endophenotypes. *Journal of Speech, Language, and Hearing Research*, *54*(6), 1628–1643.

Lippke, B.A., Dickey, S.E., Selmar, J.W., & Soder, A.L. (1997). *Photo Articulation Test (PAT-3)*. Austin, Texas. Pro.ed.

Lyakso, E., Frolova, O., & Grigorev, A. (2016). A comparison of acoustic features of speech of typically developing children and children with autism spectrum disorders. *Speech and Computer. 18th International Conference SPECOM*, Budapest, Springer, 43–50.

Maassen, B., & Van Lieshout, P. H. (2010). *Speech Motor Control : New Developments In Basic And Applied Research*. Oxford University Press.

MacNeil, L., & Mostofsky, S. (2012). Specificity of dyspraxia in children with autism. *Neuropsychology*, *26*(2), 165–171.

Macrae, T. (2017). Stimulus characteristics of single-word tests of children's speech sound production. *Language Speech and Hearing Services in Schools*, *48*(4), 219–233.

McCleery, J., Tully, L., Slevc, L.R., & Schreibman, L. (2006). Consonant production patterns of young severely language-delayed children with autism. *Journal of Communication Disorders*, *39*(3), 217–231.

Mostofsky, S., Powell, S., Simmonds, D., & Goldberg, M. (2009). Decreased connectivity and cerebellar activity in autism during motor task performance. *Brain*, *132*(9), 2413–2425.

Mowrey, R., & MacKay, I. (1990). Phonological primitives: Electromyographic speech error evidence. *The Journal of the Acoustical Society of America*, *88*(3), 1299–1312.

Nayate, A., Tonge, B.J., Bradshaw, J.L., McGinley, J.L., Lansek, R., & Rinehart, N.J. (2012). Differentiation of high-functioning autism and Asperger's disorder based on neuromotor behaviour. *Journal of Autism and Developmental Disorders*, *42*(5), 707–717.

Newmeyer, A., Grether, S., Grasha, C., White, J., Akers, R., Aylward, C., & DeGrauw, T. (2007). Fine motor function and oral-motor imitation skills in preschool-age children with speech-sound disorders. *Clinical Pediatrics*, *46*(7), 604–611.

Nip, I., Green, J., & Marx, D. (2011). The co-emergence of cognition, language, and speech motor control in early development: A longitudinal correlation study. *Journal of Communication Disorders*, *44*(2), 149–160.

Owens, R. E. (2004). *Language Disorders: A Functional Approach To Assessment And Intervention* (4th ed.). New York: Pearson Education.

Paul, R., Chawarska, K., Fowler, C., Cicchetti, D., & Volkmar, F. (2007). "Listen my children and you shall hear": Auditory preferences in toddlers with autism spectrum disorders. *Journal of Speech Language and Hearing Research*, *50*(5), 1350–1364.

Paul, R., Fuerst, Y., Ramsay, G., Chawarska, K., & Klin, A. (2011). Out of the mouths of babes: Vocal production in infant siblings of children with ASD. *Journal of Child Psychology and Psychiatry*, *52*(5), 588–598.

Paul, R., & Norbury, C.F. (2012). *Language Disorders from Infancy Through Adolesence; Listening, Speaking, Reading, Writing and Communicating* (4th ed.). St Louis: Elsevier.

Peppe, S., McCann, J., Gibbon, F., O'Hare, A., & Rutherford, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language and Hearing Research*, *50*(4), 1015–1028.

Peter, B., & Stoel-Gammon, C. (2008). Central timing deficits in subtypes of primary speech disorders. *Clinical Linguistics & Phonetics*, *22*(3), 171–198.

Pronovost, W., Wakstein, M.P., & Wakstein, D.J. (1966). A longitudinal study of the speech behavior and language comprehension of fourteen children diagnosed atypical or autistic. *Exceptional Children*, *33*(1), 19–26.

Rapin, I., Dunn, M.A., Allen, D.A., Stevens, M.C., & Fein, D. (2009). Subtypes of language disorders in school-age children with autism. *Developmental Neuropsychology*, *34*(1), 66–84.

Ravizza, S. (2005). Neural regions associated with categorical speech perception and production. In H. Cohen., & C. Lefebvre., (Eds.) *Handbook of Categorization in Cognitive Science*, Elsevier, pp. 601–615.

Rinehart, N.J., Tonge, B.J., Bradshaw, J.L., Iansek, R., Enticott, P.G., & Johnson, K.A. (2006). Movement-related potentials in high-functioning autism and Asperger's disorder. *Developmental Medicine & Child Neurology*, *48*(4), 272–277.

Schoen, E., Paul, R., & Chawarska, K. (2011). Phonology and vocal behavior in toddlers with autism spectrum disorders. *Autism Research*, *4*(3), 177–188.

Sheinkopf, S., Mundy, P., Kimbrough Oller, D., & Steffens, M. (2000). Vocal atypicalities of preverbal autistic children. *Journal of Autism and Developmental Disorders*, *30*(4), 345–354.

Shriberg, L.D., Paul, R., McSweeny, J.L., Klin, A., Cohen, D.J., & Volkmar, F.R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *Journal of Speech, Language, and Hearing Research*, *44*(5), 1097–1115.

Shriberg, L.D., & Kwiatkowski, J. (1994). Developmental phonological disorders I: A clinical profile. *Journal of Speech and Hearing Research*, *37*(5), 1100–1126.

Shriberg, L.D., Paul, R., Black, L.M., & Santen, J.P. (2011). The hypothesis of apraxia of speech in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *41*(4), 405–426.

Sterling, L., Dawson, G., Webb, S., Murias, M., Munson, J., Panagiotides, H., & Aylward, E. (2008). The role of face familiarity in eye tracking of faces by individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *38*(9), 1666–1675.

Torres, E.E.B., Brincker, M., Isenhower, R.W., Yanovich, P., Stigler, K.A., Nurnberger, J.I., & José, J.V. (2013). Autism: The micro-movement perspective. *Frontiers in Integrative Neuroscience*, *7*(7), 32 doi: 10.3389/fnint.2013.00032

Trevarthen, C., & Delafield-Butt, J. (2017). Development of human consciousness In: *Cambridge Encyclopedia of Child Development*. Cambridge University Press

Trevarthen, C., & Delafield-Butt, J.T. (2013). Biology of shared meaning and language development: Regulating the life of narratives. In M. Legerstee, D. Haley, & M. Bornstein (Eds.), *The Infant Mind: Origins of the Social Brain*.New York: Guildford Press, 167–199

Valla, J.M., Maendel, J.W., Ganzel, B.L., Barsky, A.R., & Belmonte, M.K. (2013). Autistic trait interactions underlie sex-dependent facial recognition abilities in the normal population. *Frontiers in Psychology*, *4*, 286. doi:10.3389/fpsyg.2013.00286

Wang, X., Wang, S., Fan, Y., Huang, D., & Zhang, Y. (2017). Speech-specific categorical perception deficit in autism: An event-related potential study of lexical tone processing in Mandarin-speaking children. *Scientific Reports*, *7*, 43254. doi:10.1038/srep43254.

Welsh, J.P., Ahn, E.S., & Placantonakis, D.G. (2005). Is autism due to brain desynchronization? *International Journal of Developmental Neuroscience*, *23*(2–3), 253–263.

Wetherby, A.M., Yonclas, D.G., & Bryan, A.A. (1989). Communicative profiles of preschool children with handicaps: Implications for early identification. *The Journal of Speech and Hearing Disorders*, *54*(2), 148–158.

Whitehurst, G., Smith, M., Fischel, J., Arnold, D., & Lonigan, C. (1991). The continuity of babble and speech in children with specific expressive language delay. *Journal of Speech and Hearing Research*, *34*(5), 1121–1129.

Whyatt, C., & Craig, C. (2013). Sensory-motor problems in autism. *Frontiers in Integrative Neuroscience*, *7*, 51. doi:10.3389/fnint.2013.00051

Wilkinson, K. M. (1998). Profiles of language and communication skills in autism. *Mental Retardation and Developmental Disabilities Research Reviews*, *4*(2), 73–79.

Williams, A.L., & Elbert, M. (2003). A prospective longitudinal study of phonological development in late talkers. *Language Speech and Hearing Services in Schools*, *34*(2), 138–153.

Wolk, L., & Brennan, C. (2013). Phonological investigation of speech sound errors in children with autism spectrum disorders. *Speech, Language and Hearing*, *16*(4), 239–246.

Wolk, L., & Edwards, M.L. (1993). The emerging phonological system of an autistic child. *Journal of Communication Disorders*, *26*(3), 161–177.

Wolk, L., Edwards, M.L., & Brennan, C. (2016). Phonological difficulties in children with autism: An overview. *Speech, Language and Hearing*, *19*(2), 121–129.

Wolk, L., & Giesen, J. (2000). A phonological investigation of four siblings with childhood autism. *Journal of Communication Disorders*, *33*(5), 371–389.

Yu, L., Fan, Y., Deng, Z., Huang, D., Wang, S., & Zhang, Y. (2015). Pitch processing in tonal-language-speaking children with autism: An event-related potential study. *Journal of Autism and Developmental Disorders*, *45*(11), 3656–3667.

Joanne Cleland and James M. Scobbie

Acquisition of new speech motor plans via articulatory visual biofeedback

Abstract: This chapter describes the concept of categorising persistent Speech Sound Disorder in children as a disorder characterised by erroneous motor plans. Different types of articulatory visual biofeedback are described, each of which is designed to allow children to view their articulators moving in real time and to use this information to establish more accurate motor plans (namely, electropalatography, electromagnetic articulography and ultrasound tongue imaging). An account of how these articulatory biofeedback techniques might lead to acquisition of new motor plans is given, followed by a case study of a child with persistent velar fronting who acquired a new motor plan for velar stops using ultrasound visual biofeedback.

Keywords: visual feedback, articulation, Speech Sound Disorders, electropalatography, ultrasound, electromagnetic articulography

1. Introduction

Children with Speech Sound Disorders (SSD) have difficulty acquiring the speech sounds of their native language in the course of normal development; producing certain sounds incorrectly, substituting them with other sounds or omitting them altogether. SSDs are the most common type of communication impairment; around 11.5 % of eight-year olds (Wren, Miller, Emond, & Roulstone, 2016) have SSDs ranging from common distortions such as lisps and /r/ distortions to speech that is unintelligible even to close family members.

For many children, the cause of their SSD is unknown (though SSDs are also associated with a range of conditions including hearing impairment and cleft palate) and is usually thought to arise from a difficulty acquiring the phonology of their ambient language. Indeed, most children with SSDs have "phonological" impairments (87.5 % in an analysis of caseload referrals by Broomfield & Dodd, 2004). It appears that a lesser number (12.5 % of caseload) have "articulation disorders", in that they more clearly have a problem producing certain (normally late-acquired)

speech sounds. Overall, the problem is thought to be mainly cognitive, so that children have difficulty learning the patterns of their language which often leads them to display the simplification processes representative of an earlier age in typical development, for example by reducing clusters or replacing velars with alveolars, resulting in phonological merger.

In therapy, the resulting homophony motivates remediation in part by confronting children with their inability to signal contrast. There is good evidence that in young children these auditory-based phonological interventions, for example minimal pairs intervention (Law, Garrett & Nye, 2003) are very effective. However, in around half of children with SSDs the problem persists into the school years, and a smaller number still become "intractable", beyond the age of eight. There is growing evidence that these children may not have a purely cognitive phonological disorder, but display (also) subtle motor problems. For example, Wren et al. (2016) found that weak sucking at six weeks of age is a risk factor for SSD at eight years of age. These types of potentially motoric speech impairments need interventions that capitalise on the principles of motor learning (see Maas et al., 2008 for a tutorial). Children with ingrained incorrect motor plans (for example, children who persistently misarticulate certain phonemes) need motor-based techniques for teaching and practicing new articulatory gestures.

In the motor-learning literature, the ontogeny of complex movements is studied by looking at an individual's ability to imitate a novel movement (Paulus, 2014). This is problematic for children who haven't acquired articulatory gestures via the normal auditory route because the main articulator, the tongue, is largely hidden from view. Researchers and clinicians have therefore sought to circumvent this problem by augmenting the acoustic (and tactile) information already available to the speaker through the use of instrumental imaging technologies conveying aspects of vocal tract articulation directly to the speaker, that is, by providing biofeedback.

2. Articulatory feedback approaches

In phonetics the use of instrumental techniques to measure movement of the articulators has a longer history than of sound recordings being used to measure acoustics, beginning with static palatography in the late 18th

century through to cine-Magnetic Resonance Imaging (MRI) in recent years. Techniques like electropalatography (EPG) and electromagnetic articulography (EMA) are well established, with ultrasound and MRI gaining popularity thanks to methodological improvements and falling costs. All of these techniques give researchers data that can be used to create visual images of otherwise invisible articulators, especially the tongue. However, only a small number allow data to be visualised in real time in a way that is immediately meaningful to the viewer, namely EPG, EMA and Ultrasound Tongue Imaging (UTI). Since the 1980s (Dagenais, 1995) the potential for using visualisations of the articulators as a powerful speech therapy tool has been explored. Most of the research to date has focussed on EPG, with a large number of "small n" studies showing its potential as a visual biofeedback (VBF) device (Gibbon, 2013).

EPG is a technique for displaying the timing and location of tongue-palate contact (Hardcastle & Gibbon, 1997). The speaker sees an abstract representation (Figure 1) of linguo-palatal contact, which is very useful for conveying aspects of coronal (and dorsal) consonants (and some vowels) in real time, and is encouraged to use this to modify their own erroneous articulations. It is worth noting that the display in EPG is normalised. All speakers see the same display irrespective of the size and shape of their hard palate. This potentially makes the display easier for the Speech and Language Therapist (SLT) to interpret. Additionally, the anterior third of the EPG palate is displayed in the anterior half of the normalised computer display. This is because the tongue-tip (the part most often in contact with the anterior part of the palate) contains more nerve endings and achieves more fine-grained articulation. While the ⅔ to ½ ratio is arbitrary, the understanding of this visual display is thought to be relatively intuitive (Gibbon & Wood 2010), even for those with cognitive impairment (Cleland et al. 2009).

While EPG shows tongue-palate contact rather than visualising the articulators directly, EMA shows the movements of a small number of specific flesh-points. Sensors are directly attached (glued) to articulators such as the jaw, lips, and (crucially) the tongue, and can be visualised in real time on a computer screen (Figure 1). While EPG shows 62 points of contact on the hard palate, EMA normally tracks a much more limited number of points: usually three sensors attached near to the midsagittal

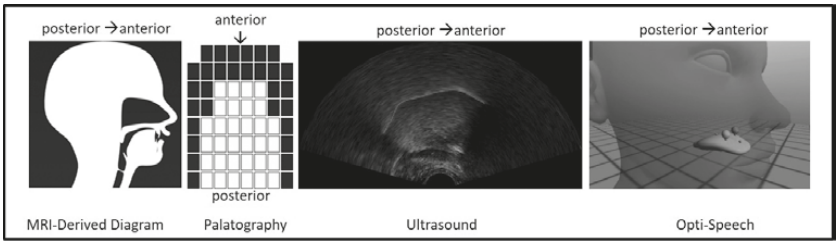

Figure 1: Instrumental articulatory technique displays (not recorded simultaneously). From left to right: MRI-derived animation (produced with permission from Eleanor Lawson), electropalatography, Ultrasound, Opti-Speech (electromagnetic articulography).

tongue tip, then two more on the front of the tongue, about 1.5cm and 3cm posterior (Katz & Mehta, 2015) which is about as far into the anterior oral cavity as can be reached easily. More recent systems, for example the Wave Electromagnetic Speech Research System (NDI, Waterloo, ON) allow three-dimensional tracking of five small sensors affixed to the client's tongue. Software such as "Opti-Speech" (Vick, Mental, Carey, & Lee, 2017) shows the sensors in the context of an avatar (see Figure 1).

EMA has been popular in articulatory phonetics studies because it is one of the few techniques which allows velocity and acceleration of movements to be calculated and interpreted easily, because of the flesh point tracking. However, it is not likely that speakers control speech production in terms of a small number of such points, nor that in experimental studies the most meaningful points are selected, nor studied in a replicable manner. In terms of biofeedback, EMA has not been particularly popular: the equipment is expensive, positioning the sensors on the articulators requires training, and it is potentially invasive, especially for children. However, a small number of studies have shown it to be potentially useful for VBF. Katz and Mehta (2015) evaluated the technique for teaching native speakers of American English to produce the non-English segment [ɖ]. In this study, the Opti-Speech system was used to display the EMA sensors superimposed on an animated avatar showing the tongue in a mid-sagittal head context. Target areas for the sensors were also shown, and on-target articulations were highlighted by changing the sensor colour from red to

green. Results indicated a rapid gain in accuracy associated with visual feedback training. However, extrapolating from these results into the clinical domain should be interpreted with caution for three reasons: firstly, the speakers did not have SSDs; secondly, the speakers were not asked to integrate the new articulation into words; and lastly a similar experiment by Cleland, Scobbie, Nakai, and Wrench (2015) using ultrasound showed that retroflexes were just as easy to teach to English-speaking children using auditory methods as they were with VBF.

To date, just one study has used the Opti-Speech (EMA) system to treat residual speech errors in children and young people. Vick et al., (2017) treated residual /s/ (two children) and /r/ (two children) distortions. Early results showed that it is possible to use the technique to remediate these errors, and that generalisation can occur. However, further research is needed to determine the effectiveness of EMA for treatment of SSDs and also to determine whether clinicians in the field find this technique useable in the practical sense.

In contrast to these studies which use direct EMA displays of the real-time movements of sensors, more recent research has sought to gamify the articulatory information, again in (near) real time. Yunusova et al. (2017) used a single tongue tip sensor to drive a computer game in which the object was for a dragon character to breathe as much fire as possible. The size of the dragon's flames was directly related to the size of the speaker's articulatory working space (AWS). In this case, the augmented VBF was designed with a very specific population in mind: speakers with Parkinson's disease. This particular neurodegenerative condition causes a reduction in articulatory movements (causing dysarthric symptoms such as undershoot) and leads to reduced intelligibility. By providing a metaphor (the fire-breathing dragon) which visually produces more fire in correlation with increasing AWR, speakers with Parkinson's disease were able to use the feedback to increase their intelligibility. Increasing the strength and range of movements which already follow the correct articulatory trajectory is, however, quite different from establishing a correct gesture in replace of an erroneous one (for example, a central fricative produced laterally), or an absent one (for example, in someone who has no velars in their phonetic inventory). Therefore, any gamification of VBF designed for establishing new articulations is likely to need games which relate more

directly to the trajectory of a specific segmental gesture rather than to the global magnitude of change during the production of a word.

In contrast to EPG and EMA, which show a discrete number of points, U-VBF shows an anatomically accurate speaker-specific representation of the tongue. With this technique most of the surface of the tongue is visible in a mid-sagittal view (Figure 1), and interpretation of the images is thought to be relatively intuitive (Bernhardt et al. 2005). In contrast to EPG, the image is an anatomically correct representation of part of the tongue, however, other important anatomical information, such as the relation of the tongue to the hard palate, is not normally visible (Cleland et al., 2019). Moreover, this "raw" ultrasound suffers from artefacts, and the tip of the tongue is often in shadow from the mandible. However, ultrasound has practical advantages over EPG and EMA in that it does not require expensive individual artificial palates or expensive sensors. Moreover, since it involves no intra-oral equipment it is less physically invasive, potentially making it more suitable for children.

Given the practical limitations of EMA most of the clinical studies in the literature have used EPG and, more recently Ultrasound-VBF. Indeed, U-VBF is rapidly gaining popularity, probably because of its lower cost and because more portable high-speed ultrasound systems are now available. To date, 29 small studies have been published in the literature investigating the efficacy of U-VBF (see Sugden, Lloyd, Lam and Cleland, 2019 for a systematic review). Of these studies, 27 were published in the last 10 years and 17 in the last three. While larger clinical trials of both EPG and UTI are needed in the future, it is essential to know theoretically why and how these techniques work because identifying the agents of change (the "active ingredients") in an intervention is essential for refining the intervention and establishing dosage.

None of these instrumental techniques are therapies in their own right (Bacsfalvi et al. 2007); most SLTs use them to supplement traditional techniques, such as articulation therapy (Van Riper & Emerick, 1984) or motor-based intervention (Preston et al., 2013). One key ingredient of articulatory VBF is that it can be used to demonstrate complex articulations that are normally difficult to describe. Describing articulatory movements is an essential part of traditional articulation therapy (Van

Riper and Emerick, 1984). Normally this is done with verbal descriptions, or perhaps diagrams, ranging from impromptu sketches to computer animations.

It is crucial, moreover, to unpick the visual model aspect of EPG/UTI from the biofeedback aspect. That is, we need to know the extent to which a speaker benefits from informative general visual models of articulation, and the extent to which real-time biofeedback of the learner's own tongue during speech production provides crucial additional information.

Considering first the model aspect on its own, studies which investigate the use of an articulatory model to teach new speech sounds are few. Massaro et al. (2008) used a "Talking Head" to teach native English speakers a new vowel [y] and consonant [q]. Talking Heads are artificial animations of speech usually based ultimately on instrumental (e.g. MRI or EMA) data. Some are 3D (e.g. Badin & Serrurier, 2006) and some are 2D (e.g. Kröger et al., 2013), but most attempt to model the movement of the tongue during speech with a cut-away profile or mid-sagittal view of the tongue.

The main application of Talking Heads is usually as a teaching tool for pronunciation training in second language learning (Cleland et al., 2013). However, there is little evidence that this is effective. In the Massaro et al. study (2008) a view of the lips was useful for teaching the high-front rounded vowel [y] but a mid-sagittal Talking Head did not improve learning of the distinction between [k] and the uvular stop [q]. There is a confound here, however, due to one study involving a segment where lip-rounding is the defining feature and one where it is uvular place: lip reading is not only a natural phenomenon but one known to improve perception of speech (see below). Similarly, a study by Fagel and Madany (2008) which used a Talking Head to teach [s] and [z] to children with interdental lisps failed to show an effect. Thus, a visual model alone appears not to be the essential ingredient for success. However, since the above studies did not give the learners any information about closeness to target (e.g. from a human judge or automatic speech recognition), and since articulatory constriction is a key feature of production, further study is required to directly compare an articulatory model against VBF using the same type of display and mediation.

3. Theoretical explanations for the role of biofeedback in learning new articulations

Children who make inappropriate phonetic realisations of certain speech sounds do so because they have an inappropriate motor plan for that sound (Preston et al., 2014; Cleland et al., 2019). Cleland, Scobbie and Wrench (2015) suggest that these erroneous motor plans can be ascribed to one of three categories: 1. It is identical to that of another phoneme, resulting in perceived homophony (as in canonical velar fronting); 2. the motor plan is abnormal or underspecified resulting in something which is perceived as homophonous but is subtly different in some way (as in covert contrast, Gibbon & Scobbie, 1997), for example /t/=[t] and /k/=[t̠]) or; 3. the motor plan is abnormal to the extent that it results in the realisation of an obviously non-native speech sound, for example a lateral lisp in English-speaking children. It is possible that different types of VBF are needed to overcome each of these erroneous motor plans. In the case of category 1, normally a phonological cause would be ascribed, however Cleland et al. (2017) present several cases of children with persistent velar fronting with identical tongue-shapes for /t/ and /k/ but awareness of the error and (initially) an inability to produce a velar articulation of any type. In these, and other cases, the inability to produce the correct articulatory gesture upon imitation is often coupled with a lack of understanding (despite previous intervention) of how the gesture is achieved at all, with one of the children in the Cleland, Scobbie and Wrench (2015) study stating that she thought producing a velar was "impossible" the first time she viewed an ultrasound movie of that segment, highlighting the lack of understanding she had as to the movements required to achieve a velar despite previous therapy targeting this very sound (Cleland et al., 2019).

In addition to a lack of explicit understanding about the movements required to achieve a particular sound, there may be some implicit learning involved in the viewing of tongue movements. In typical audio-visual speech perception, viewing the speaker's lips enhances perception, particularly in noise (Benoît & Le Goff, 1998). Typical speakers integrate lip information into their perceptual system, as shown by the McGurk effect (McGurk & MacDonald, 1976). Clearly whilst lips are easily visible during interactions, the tongue is not. Even so, Badin, Tarabalka,

Elisei, and Bailly (2010) suggest that it is possible to "tongue-read" in the same way as it is possible to lip-read. That is, viewing a Talking Head of tongue movements leads to better discrimination of speech in noise and potentially could be used for learning new articulations. Badin et al. (2010) hypothesise that this is due to a natural, intuitive ability for listeners/viewers to tongue-read, suggesting that this provides support for a perception/production link which could relate to the theory of mirror neurons (Cleland et al., 2019). Mirror neurons are thought to underlie the imitation system, because they are neurons that fire when a person both sees an action being performed (or hears it being performed, in which case they may be called echo neurons) and performs that action themselves. So, in theory, when a person hears a speech sound, the neurons in the motor area required for articulating that speech sound fire. In fact, even passive listening to speech sounds evokes a pattern of motor synergies mirroring those occurring during speech production (D'Ausilio, Bartoli, Maffongelli, Berry & Fadiga 2014). There is emerging evidence that this does not just apply to hearing a speech sound, but also to seeing it. Treille, Vilain, Hueber, Schwartz, Lamalle and Sato (2014) showed activation in the premotor and somatosensory cortices when observing lingual movements from ultrasound, suggesting that demonstration of correct articulatory movements may be a crucial aspect of visual biofeedback. Moreover, using delayed U-VBF might evoke the same process. In this type of feedback, the child (as well as watching the live visual biofeedback) watches their own production replayed after a delay (once they have finished speaking, not to be confused with delayed auditory feedback, which has very short delay times). The SLT then encourages the child to reflect on the correctness of their production. While viewing their own *incorrect* production could potentially have an adverse effect, viewing their own correct production gives a speaker-specific representation of the required articulatory gesture.

Whilst it would be unethical and ethically dubious to compare U-VBF without demonstration to U-VBF with it, it would be feasible to conduct a randomised control trial where one arm of the trial involved the use of an ultrasound-based visual articulatory model, without biofeedback (Cleland et al., 2019). Indeed, a small study of speakers with cleft palate (Roxburgh, 2018) found that the children did just as well with a visual

articulatory model to learn new articulations as they subsequently did with U-VBF. However, this study was limited by a small sample size of just two participants, and that neither had had previous therapy to address the relevant speech problem (i.e. they were not 'intractable', Cleland et al., 2019).

The question remains as to how VBF, or indeed a visual model alone, could lead to acquisition of *new* articulations, especially when, in the case of intractable SSDs, the speakers have been exposed to extensive models of the correct articulation from other speakers, albeit only in auditory form. It seems in this case that the auditory imitation system has failed somehow, perhaps enabling the visual modality to offer useful new information. Indeed, evidence exists that the observation of completely novel behaviour (in this case a previously unseen articulatory movement) generates mirroring activity in the premotor cortex (Cross, Hamilton and Grafton, p. 11, 2006). Moreover, Mattar and Gribble (2005) show that complex motor behaviours, which speech undoubtedly is, are greatly assisted by first observing another engage in the activity. Via this mechanism, models of the new activity are formed in the premotor cortex via the mirror neurons and presumably intensity of neuronal firing increases with practice/exposure. It is not enough to simply watch the new movement repeatedly and expect acquisition of a new motor plan: practice is required by the speaker. (Imagine trying to learn the piano only by watching videos of a pianist's fingers!) Del Giudice, Manera and Keysers (2009, p. 352) explain the mechanisms by which practice of movements leads to acquisition, by looking at grasping: "activity in the premotor cortex leads to a grasping movement. The movement is *seen by the acting individual,* causing activity in neurons in the temporal cortex. This activity is sent to the parietal and premotor cortex, where it finds neurons that are active because the subject is currently performing the action. This leads to Hebbian enhancement of the congruent connections from temporal to parietal and from parietal to premotor neurons representing the same action; incongruent connections do not undergo such enhancement". It is therefore conceivable that seeing a novel speech motor movement leads to development, or otogeny, of the mirror neuron whilst actually *doing* the novel tongue movement yourself leads to Hebbian enhancement, which in turn is enhanced by lingual visual biofeedback. Repeated association of the sound (knowledge of results) with the movement (knowledge of performance) leads to enhancement in

acquisition of the new skill. Of course, this ought to be entirely possible with only the articulatory model, provided the speaker is able to practice accurately, and biofeedback may not be required. However, it is likely that some individuals are unable to make the leap between seeing the new articulation and beginning to practice it themselves, that is, no matter how many times they see it they cannot perform it, or even approximate a performance of it. In this case the speech and language therapist too benefits from the visual feedback as s/he is able to use shaping techniques (Bleile, 2004) to explicitly demonstrate to the speaker that similar motor programmes are already within their grasp.

Evidence for the *biofeedback* aspect of U-VBF comes from experiments on experiential canalised learning. Canalisation is the means by which a developmental process is buffered against perturbations. It ensures that important features of the organism emerge reliably despite great variation between individuals in environmental conditions and genotypic makeup. The classic example is that of ducklings raised in incubators which still spontaneously exhibit the 'correct' preference for their own species' maternal calls, despite never hearing a mother duck. However, if the ducklings are prevented from hearing their own vocalizations, they fail to exhibit selective responses to maternal calls (Gottlieb, 1991) suggesting a key factor is self-produced vocalizations. That is, the speaker must *make* the articulatory movements themselves and evaluate the acoustic output in order to acquire them. Visual biofeedback offers a new modality for learners who have failed to acquire speech sounds via the normal routes. Moreover, in live *bio*-feedback the speaker is able to bootstrap the new visual modality not only onto the auditory modality but also onto the haptic modality to make small adjustments to their articulatory gestures in real time. In the speech therapy clinic this is evidenced by articulatory groping towards the target in the early stages of intervention.

In sum, U-VBF works by first showing the learner what is to them a novel movement, then performance of the new movement leads to Hebbian learning, which is boosted by the visual knowledge of performance provided by U-VBF, this leads to increasing activation of the mirror neuron, laying down of a new general motor programme and hence eventually mastery of the new sound. If the mastery of the new sound is a gradual process then we might expect to detect various types of phonetic gradience in the

short-term longitudinal change, potentially in addition to rapid categorical change. Some evidence of incomplete generalisation of a new articulation is shown in U-VBF studies where post-intervention scores for target segments are lower than 100 % correct. For example, Cleland, Scobbie, Roxburgh, Heyde and Wrench (2019) show that after intervention children with a wide variety of lingual errors show improvements in accuracy of targeted gestures, but no child achieved perfect percent target consonants correct in all phonotactic contexts. However, the approach of categorising segments within words as correct or incorrect obscures the potential subtlety of the process. More important for understanding the pathway to acquisition is the fine detail necessary for a full evaluation of new articulations produced by children as the result of clinical intervention.

For example, consider the two children reported by Cleland et al. (2019) who made progress towards the target, changing posterior (pharyngeal fricatives for sibilants) to anterior articulation, but with incorrect lateral airflow. For these children, the updated motor plan is more accurate, since in it contains more of the correct features of the target, even though the output is still wrong linguistically. The motor plan has therefore changed in a gradient manner, as both children also show progress towards achieving the correct airflow. However, gradient acquisition of targets may manifest differently in each of the three erroneous motor plans 1. Motor plans identical to another sound; 2. Motor plans which are covertly different but perceived as a different sound and 3. Motor plans which result in a non-native sounding phone. Type one is particularly interesting, because in a traditional model these children would be said to have classic substitution errors, thought to be phonological in nature. If this were the case, we would not expect these children to acquire a new articulation in a phonetically gradient manner (though they may acquire it in some phonotactic conditions before others as is the case in typical acquisition of a segment).

What follows is a case study of a child who presented with a classic substitution error who nevertheless shows gradient change during remediation. Rather than presenting only binary information on the correctness of her new articulations, which would obscure more subtle changes, we explore the process in more articulatory detail *during* the therapeutic process.

4. An illustration of gradient acquisition of a new articulation

While typically developing children are usually able to produce velars correctly by the age of three and a half years (Dodd, 2013), those with SSDs may not be able to produce velars till much later. A lack of velars in a child's phonetic inventory has been recognised as a prognostic indicator for a phonological disorder (Grunwell, 1987). Children who persistently fail to differentiate coronal and dorsal articulations may therefore have an underlying motoric deficit. Gibbon (1999) suggests that this may manifest as an "Undifferentiated Lingual Gesture" (ULG), where the tongue moves as a whole, rather than, as expected, by executing gestures using independent parts. Children with UGs show abnormally extensive tongue-palate contact patterns in EPG studies (Gibbon 1999) and (in just one study to date) abnormal dorsal raising in ultrasound (Cleland et al., 2017). This error pattern is motoric, rather than phonological.

While there are studies showing these abnormal articulations, there are no studies showing how articulations change as children initiate a coronal/dorsal differentiation or achieve mastery of it. In several of our previous studies (Cleland et al., 2015b, 2017, 2019) we reported on children who persistently front velars to alveolars, despite being over six years of age. Velar fronting is readily remediated using U-VBF, with some children showing a categorical shift from 0 % velars correct pre-therapy to 100 % post-therapy. Speaker "07F_Ultrax" is reported in Cleland et al. 2015 and 2017. At the time of the U-VBF intervention she was aged 7;6 and presented with velar fronting in the absence of a history of any other errors. Pre-intervention, she produced no correct velars, half-way through intervention she was not perceived to produce any correct velars, but 6 weeks later, at the end of the intervention period, she produced 100 % correct velars in a word list designed to probe this segment in multiple phonotactic positions. She maintained that gain three months later. Prior to intervention she produced both /t/ and /k/ with identical tongue shapes, in other words, a classic merger (see Cleland et al., 2017) appears to have been almost instantly fixed. We turn our attention now to an ultrasound analysis of 07F's productions of alveolars and velars at various time-points in the intervention process.

07F_Ultrax was recorded with simultaneous high-speed ultrasound and audio. The ultrasound was probe-stabilised with a headset (Scobbie, Wrench & Van der Linden, 2008) to allow us to compare tongue shape for /t/ and /k/ directly. Materials were a wordlist containing velars in a wide range of vowel environments and word positions.

Using AAA v2.16 software (Articulate Instruments, 2012) /t/ and /k/ segments were annotated at the beginning of the burst, the nearest ultrasound frame was then selected and a spline indicating the tongue surface fitted to the image using the semi-automatic edge-detection function in AAA software. Splines were then averaged by target segment and compared.

In this case, we are interested in the degree of separation between /t/ and /k/. If 07F presents with merged productions of /t/ and /k/, then we would expect to see no degree of separation between /t/ and /k/ and if she presents with ULGs for both, then we might expect a reduced degree of separation between /t/ and /k/ compared to typically developing children. The difference between /t/ and /k/ can be characterised as maximum radial dorsal difference between these two segments (Figure 2).

Scobbie and Cleland (2017) report the average maximum width of the radial difference between /t/ and /k/ at mid-closure for 30 typically developing children as 11.9mm, 7.5mm and 12.1mm for symmetrical /a/, /i/ and /o/ contexts respectively.

By applying the same measurements (Figure 3) to all the time-points from 07F's data, we can quantify the gradient increase in the degree of separation between /t/ and /k/ at each time point (Figure 4). What is interesting, is that by looking only at percent target consonants correct, 07F appears to make a categorical shift from 0 % to 100 % correct between mid-therapy and post-therapy sessions, whereas in fact she was already beginning to change the production by the mid-therapy session (panel 2) while in the post-therapy session (panel 3) her coronal/dorsal differentiation (6.12mm) actually remained abnormally small. Presumably with practice, as is consistent with the motor learning literature, over time her articulations become more phonetically accurate, until the point where /t/ and /k/ are perceived by a listener as occupying different perceptual categories.

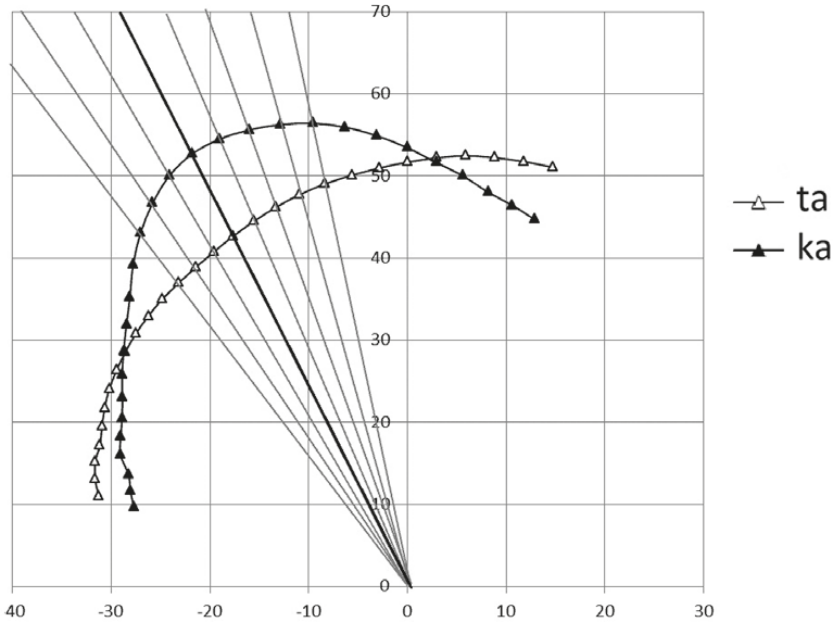

Figure 2: Average /t/ and /k/ from 30 typical children at mid-closure. The diagonal spokes are some of the radial fanlines (emanating from the probe's virtual centre) used for measurement. For each individual child the maximum distance /k/-/t/ along some fanline (in this case, the 4th diagonal line from the left) within the anterior and posterior crossing points of the splines for each individual child is taken as the degree of coronal-dorsal differentiation.

5. Conclusion

Since the 1980s instrumental phonetic techniques have increasingly been applied as biofeedback for learning new articulations in children who have failed to acquire particular phones through the normal route. While EPG has dominated the literature as the technique of choice, and has been shown to be successful for a large number of children, recent studies have focussed on ultrasound visual biofeedback. For the most part VBF is described as a motor-learning approach, though it is often used with children who present with errors described as "phonologically delayed". The case study above shows that even in these cases, evidence of subtle

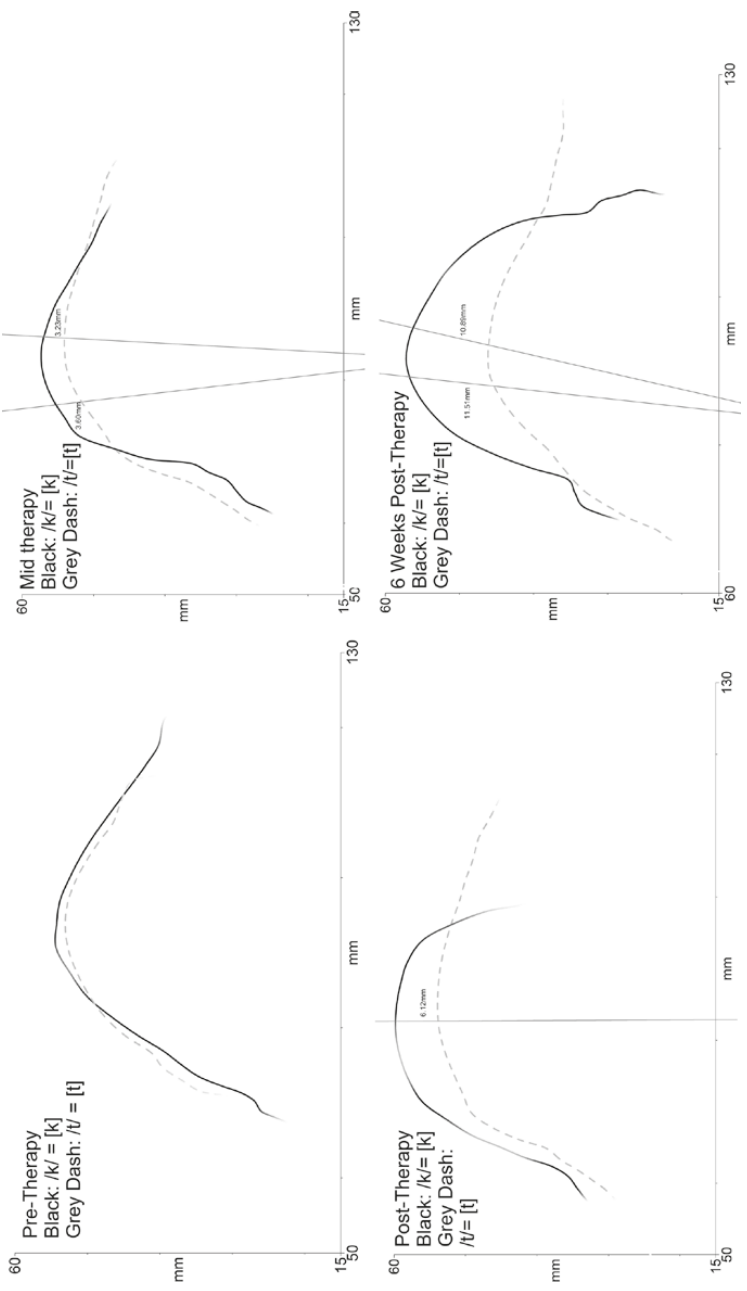

Figure 3: /k/ (black) and /t/ (grey dashes) attempts over time (L-R): pre, mid, post, 6 weeks post intervention. Increased separation between /k/ and /t/ can be seen, but is only at 6 weeks post intervention that /k/ is perceived as distinct from /t/.

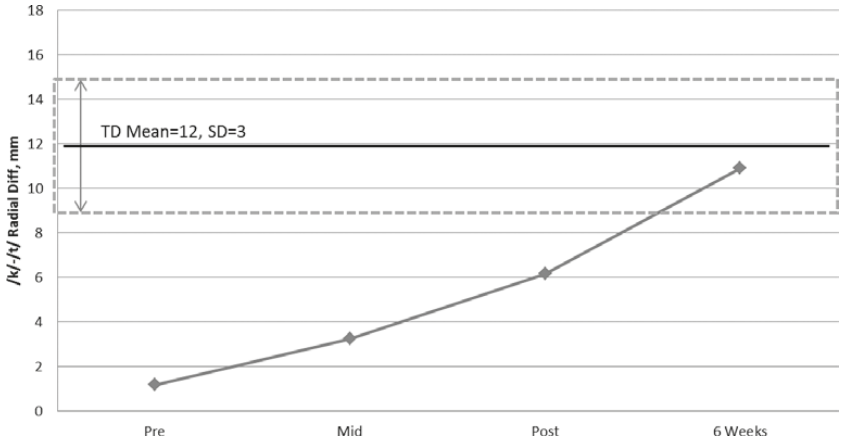

Figure 4: Max radial difference of /k/-/t/ for 07F over time. Y-axis, radial difference between /k/ and /t/, x-axis intervention time point. Grey dashed box: expected radial difference between /k/ and /t/ for typically developing children.

motor-impairments can exist. This calls into question the underlying impairment these children have. However, we wish to caution the reader from drawing the conclusion that all children with "phonological delay" in fact have motor-based problems. Evidence from a large study by Wren et al. (2016) shows that early signs of subtle motor impairment such as weak sucking at six weeks of age, predicts *persistent* SSDs, and not SSDs which remediate in the preschool years. It therefore seems plausible that children with persistent disorders, as exemplified here, are a different subgroup from the outset.

The agents of change in VBF remain underexplored. There are at least four different potential "active ingredients" in VBF therapy that do not exist in traditional approaches: 1. Improved diagnostic information provided by articulatory analysis prior to intervention; 2. An accurate visual articulatory model provided by target patterns/tongue movements; 3. Increased accuracy of positive feedback from the treating SLT made possible by viewing movements; 4. Biofeedback. In reality a combination of all these factors likely impacts on the ability of children to achieve, practice, and ultimately generalise new articulations following biofeedback interventions.

References

Articulate Instruments Ltd 2012. *Articulate Assistant Advanced User Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd.

Bacsfalvi, P., Bernhardt, B.M. & Gick, B. (2007). Electropalatography and ultrasound in vowel remediation for adolescents with hearing impairment. *Advances in Speech-Language Pathology*, *9*(1), 36–45.

Badin, P. & Serrurier, A. (2006). Three-dimensional linear modelling of tongue: Articulatory data and models. In H.C. Yehia, D. Demolin, R. Laboissiére (Eds.), *Seventh International Seminar on Speech Production*, *ISSP7*. Ubatuba, SP, Brazil, UFMG, Belo Horizonte, Brazil, 395–402.

Badin, P., Tarabalka, Y., Elisei, F. & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*(6), 493–503.

Benoît, C. & Le Goff, B. (1998). Audio-visual speech synthesis from French test: Eight years of models, designs and evaluation at the ICP. *Speech Communication*, *26*(1–2), 117–129.

Bernhardt, B., Gick, B., Bacsfalvi, P. & Adler-Bock, M. (2005). Ultrasound in speech therapy with adolescents and adults. *Clinical Linguistics and Phonetics*, *19*(6–7), 605–617.

Broomfield, J. & Dodd, B. (2004). Children with speech and language disability: Caseload characteristics. *International Journal of Language and Communication Disorders*, *39*, 303–324.

Bleile, K.M. (2004). *Manual of articulation and phonological disorders: Infancy through adulthood*. Cengage Learning.

Cleland, J., McCron, C., & Scobbie, J.M. (2013). Tongue reading: Comparing the interpretation of visual information from inside the mouth, from electropalatographic and ultrasound displays of speech sounds. *Clinical Linguistics and Phonetics*, *27*(4), 299–311.

Cleland, J., Scobbie, J.M., Heyde, C., Roxburgh, Z., & Wrench, A.A. (2017). Covert contrast and covert errors in persistent velar fronting. *Clinical Linguistics and Phonetics*, *31*(1), 35–55.

Cleland, J., Scobbie, J.M., Nakai, S., & Wrench, A.A. (2015a). Helping children learn non-native articulations: the implications for ultrasound-based clinical intervention. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow, UK: University of Glasgow. Paper number 0698,

Retrieved February, 2, 2017 from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0698.pdf

Cleland, J., Scobbie, J.M., Roxburgh, Z., Heyde, C., & Wrench, A.A. (2019). Enabling new articulatory gestures in children with persistent speech sound disorders using ultrasound visual biofeedback. *Journal of Speech, Language, and Hearing Research 62*(2*),* 229–246.

Cleland, J., Scobbie, J.M., & Wrench, A.A. (2015b). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics and Phonetics*, *29*(8–10), 575–597.

Cleland, J., Timmins, C., Wood, S.E., Hardcastle, W.J. & Wishart, J.G. (2009). Electropalatographic therapy for children and young people with Down's syndrome. *Clinical Linguistics and Phonetics*, *23*(12), 926–939.

Cross, E.S., Hamilton, A.F.D.C., & Grafton, S.T. (2006). Building a motor simulation de novo: observation of dance by dancers. *Neuroimage*, *31*(3), 1257–1267.

Dagenais, P. (1995). Electropalatography in the treatment of articulation/phonological disorders. *Journal of Communication Disorders*, *28*(4), 303–329.

D'Ausilio, A., Maffongelli, L., Bartoli, E., Campanella, M., Ferrari, E., Berry, J., & Fadiga, L. (2014). Listening to speech recruits specific tongue motor synergies as revealed by transcranial magnetic stimulation and tissue-Doppler ultrasound imaging. *Philosophical Transactions of the Royal Society B*, *369*(1644), 20130418.

Dodd, B. (2013). *Differential diagnosis and treatment of children with speech disorder*. Chichester: John Wiley & Sons.

Fagel, S. & Madany, K. (2008). A 3-D virtual head as a tool for speech therapy for children. In *Ninth Annual Conference of the International Speech Communication Association*. Brisbane, Australia, 2643–2646.

Gibbon, F.E. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research*, *42*(2), 382–397.

Gibbon, F.E (2013). *Bibliography of electropalatographic (EPG) Studies in English (1957–2013)*. Retrieved November, 12, 2018 from http://www.articulateinstruments.com/EPGrefs.pdf

Gibbon, F., & Scobbie, J.M. (1997). Covert contrasts in children with phonological disorder. *Australian Communication Quarterly*. (Autumn), 13–16.

Gibbon, F.E. & Wood, S.E. (2010). Visual feedback therapy with electropalatography. In: Williams, A. L., McLeod, S. and McCauley, R.J. (Eds.) *Interventions for speech sound disorders in children*. Baltimore: Paul H. Brookes Pub, pp. 509–532.

del Giudice, M.D., Manera, V., & Keysers, C. (2009). Programmed to learn? The ontogeny of mirror neurons. *Developmental Science*, *12*(2), 350–363.

Gottlieb, G. (1991). Experiential canalization of behavioral development: Theory. *Developmental Psychology*, *27*(1), 4–13.

Hardcastle, W., & Gibbon, F. (1997). Electropalatography and its clinical applications. In M. Ball, & C. Code (Eds.), *Instrumental Clinical Phonetics* London: Whurr, pp. 149–193.

Katz, W.F., & Mehta, S. (2015). Visual feedback of tongue movement for novel speech sound learning. *Frontiers in Human Neuroscience*, *9*, 612.

Kröger, B.J., Gotto, J., Albert, S., & Neuschaefer-Rube, C. (2013). *A visual articulatory model and its application to therapy of speech disorders: a pilot study*. Universitätsbibliothek Johann Christian Senckenberg.

Law, J., Garrett, Z. & Nye, C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *Cochrane Database of Systematic Reviews*, *3*. Art. No.: CD004110.

Maas, E., Robin, D.A., Hula, S.N.A., Freedman, S.E., Wulf, G., Ballard, K.J., & Schmidt, R.A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology*, *17*(3), 277–298.

Massaro, D., Bigler, S., Chen, T., Perlman, M., & Ouni, S. (2008). Pronunciation training: The role of ear and eye. *Ninth Annual Conference of the International Speech Communication Association*, 22–26 September, Brisbane, Australia, 2623–2626.

Mattar, A.A., & Gribble, P.L. (2005). Motor learning by observing. *Neuron*, *46*(1), 153–160.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264* (5588), 746–748.

Paulus, M. (2014). How and why do infants imitate? An ideomotor approach to social and imitative learning in infancy (and beyond). *Psychonomic Bulletin & Review*. *21*(5), 1139–1156.

Preston, J.L., Brick, N., & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology*, *22*(4), 627–643.

Preston, J.L., & Leaman, M. (2014). Ultrasound visual feedback for acquired apraxia of speech: A case report. *Aphasiology*, *28*(3), 278–295.

Roxburgh, Z. (2018). *Visualising articulation: real-time ultrasound visual biofeedback and visual articulatory models and their use in treating speech sound disorders associated with submucous cleft palate*. Unpublished doctoral dissertation, QMU Edinburgh, UK.

Scobbie, J.M. & Cleland, J. (2017). Dorsal crescents: Area and radius-based mid-sagittal measurements of comparative velarity. Paper presented at Ultrafest VIII, Potsdam, 4th–6th October 2017. Potsdam: University of Potsdam.

Scobbie, J.M., Wrench, A.A., & van der Linden, M. (2008). Head-probe stabilisation in ultrasound tongue imaging using a headset to permit natural head movement. In *Proceedings of the 8th International Seminar on Speech Production*, Strasbourg, 373–376.

Sugden, E., Lloyd, S., Lam, J., & Cleland, J. (2019). Systematic review of ultrasound visual biofeedback in intervention for speech sound disorders. *International Journal of Language and Communication Disorders*. https://doi.org/10.1111/1460-6984.12478

Treille, A., Vilain, C., Hueber, T., Lamalle, L., & Sato, M. (2017). Inside speech: Multisensory and modality-specific processing of tongue and lip speech actions. *Journal of Cognitive Neuroscience*, *29*(3), 448–466.

Van Riper, C., & Emerick, L.L. (1984). *Speech correction: An introduction to speech pathology and audiology*. Englewood Cliffs, NJ: Prentice-Hall.

Vick, J., Mental, R., Carey, H., & Lee, G.S. (2017). Seeing is treating: 3D electromagnetic midsagittal articulography (EMA) visual biofeedback for the remediation of residual speech errors. *The Journal of the Acoustical Society of America*, *141*(5), 3647–3647.

Wren, Y., Miller, L.L., Peters, T.J., Emond, A., & Roulstone, S. (2016). Prevalence and predictors of persistent speech sound disorder at eight years old: Findings from a population cohort study. *Journal of Speech, Language, and Hearing Research*, *59*(4), 647–673.

Yunusova, Y., Kearney, E., Kulkarni, M., Haworth, B., Baljko, M., & Faloutsos, P. (2017). Game-based augmented visual feedback for enlarging speech movements in Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *60*, 1818–1825.

Marion Dohen

Do manual gestures help the learning of new words? A review of experimental studies

Abstract: We all produce manual gestures when we speak and these gestures have been shown to play an important role in the act of communicating. The aim of this chapter is to further investigate the specific role played by manual gestures in combined semantic and lexical learning by reviewing the experimental evidence provided by the literature. Nineteen articles met our selection criteria. They explore the effect of manual gestures in learning new words in both typically developing and speech and language impaired participants. Even though it was not an exclusion criterion, none of the studies dealt with adults: all tested children of various ages. Several research questions are addressed: 1. Is there a general advantage of using manual gestures in learning new words? 2. Is there a specific effect of manual gestures vs. other additional cues? 3. Is there a differential effect on learning to comprehend and to produce the newly learned words? 4. Do different types of gesture have different effects? 5. Does testing at different points in time yield different results? 6. Does producing the gesture during training matter? 7. Do manual gestures help generalize the use of newly learned words to new contexts? Hypotheses on the reasons why gestures would play a positive role for word learning are then suggested.

Keywords: gestures, memory, novel word learning, child language acquisition, speech production, speech and language disabilities

1. Introduction

Manual gestures are part of communication. We all move our hands and arms while we speak and researchers have argued that these gestures "are an integral component of the communicative act of the speaker" (Kendon, 2004; p. 359). According to the growth point theory, gestures and speech stem from a common thought process (McNeill, 1992; McNeill & Duncan, 2000; McNeill et al., 2008). They would even be controlled by the same motor system (Gentilucci & Dalla Volta, 2008). The brain integrates both signals when perceiving a communicative act (e.g., Özyürek et al., 2007) even though the networks involving the processing

of the two modalities do not overlap completely (Bernardis, Salillas, & Caramelli, 2008).

Before being able to speak, babies begin communicating intentionally using gestures, more specifically pointing gestures. This is probably due to the fact that manual gestures are mastered more easily by infants than speech (e.g., Goodwyn & Acredolo, 1993). Gestures are indeed holistic whereas speech is sequential. From a motor point of view, the hands are easier to control than the oral/vocal system required for speaking. One could then argue that manual gestures could be used by infants simply before they can speak and not actively play a role in speech and language development. Iverson and Thelen (1999) proposed a model describing the co-development and entrainment of the arm/hand and oral/vocal motor systems from birth to around 18 months. The model describes how the two systems and their development are closely related. The first gestural communicative acts have also been shown to predict the onset of the first words (e.g., Iverson & Goldin-Meadow, 2005; Goldin-Meadow, 2007). Gesture use is predictive of later vocabulary size (Rowe, Özçalışkan, & Goldin-Meadow, 2008). The type of gesture even predicts the class of words acquired (Kraljević, Cepanec, & Šimleša, 2014). Later on, babies start combining gestures and words to create utterances and this stage has also been shown to be predictive of the first multi-word utterances (e.g. Capirci et al., 1996; Goldin-Meadow & Butcher, 2003; Iverson & Goldin-Meadow, 2005; Rowe & Goldin-Meadow, 2009). Iverson, Capirci and Caselli (1994) showed that, as late as 16 months, the majority of the children they observed still had a clear preference for gestural communication, even though they had equivalent gestural and verbal repertoires (see also, Caselli et al., 2012). From two to 3;6 years of age, it has been shown that the production of iconics and beat gestures was correlated with language development (Nicoladis, Mayberry, & Genesee, 1999; Mayberry & Nicoladis, 2000) and more specifically with verbal vocabulary development (Acredolo & Goodwyn, 1988).

Taken together, this research suggests that children naturally produce manual gestures to communicate and that these gestures play a role in language acquisition, not only before speech onset but also later (Özçalışkan & Goldin-Meadow, 2005). Purposely encouraging infants to communicate using symbolic manual gestures from 11 months of age

has been shown to have positive effects on later language development (Goodwyn, Acredolo, & Brown, 2000 but see Johnston, Durieux-Smith, & Bloom, 2005 for contradictory evidence). Kahn (1981) also tested this in "nonverbal, hearing, retarded" children but found the effect to be highly dependent on individuals.

Some studies also show that children use gestures for speech and language comprehension (Morford & Goldin-Meadow, 1992). Parental gestures help children map meanings on new words (Clark & Estigarribia, 2011). 18-month-olds manage to interpret gestures and words indifferently as labels for object categories, but 26-month-olds seem to have a preference for words (Namy & Waxman, 1998, see also Suanda et al., 2013).

All this put together suggests that manual gestures play a role in acquiring speech and language (Capirci & Volterra, 2008) even though it still remains unclear what this exact role is. The aim of this chapter is to better comprehend various elements of this role in the specific field of word learning. To learn a new word, one has to map both a meaning and a lexical form to a concept, which can be respectively labeled as semantic and lexical learning. "Word learning is a complex task that requires (…) to create new semantic and lexical representations, then link these new representations and integrate them with existing phonological, lexical, and semantic representations" (Kapalková, Polišenská, & Süssová, 2016, p. 59).

Gestures have been shown to support different types of learning (e.g., Kelly, Manning, & Rodak, 2008; Goldin-Meadow, 2011) and long-term memorization (e.g., Church, Ayman-Nolley, & Mahootian, 2004). More specifically, there is evidence that manual gestures could help lexical learning alone. Gestures have indeed been shown to promote the learning of words in a foreign language in children (Tellier, 2008 but see Rowe, Silverman, & Mullan, 2013) and adults (Macedonia & von Kriegstein, 2012; Kelly et al., 2014; Macedonia & Repetto, 2016). Rowe, Silverman and Mullan (2013) put forward the fact that this effect is dependent on the individual. Gogate, Bahrick and Watson (2000) observed that mothers naturally use gestures when they teach new words to their infant. Evidence also suggests that manual gestures facilitate lexical access in adults (e.g., Rauscher, Krauss, & Chen, 1996; Krauss & Hadar, 1999).

This chapter will focus on semantic and lexical learning combined, such as in the situation in which a child learns a new word from her native language. It will provide a review of the experimental evidence on the role of manual gestures in word learning. The first question it will address is whether or not the existing evidence suggests a positive role of adding manual gestures to spoken words for learning them in both typical individuals (section 3.2.) and in people with speech and language impairments of various types (section 3.3.). We will then (second question) analyze whether using manual gestures has a differential effect compared to using other additional cues (section 3.4.). The third question will examine whether there are differences between receptive and expressive learning in terms of types of effects of adding gestures (section 3.5.). We will then question whether the type of gesture has an influence (fourth question, section 3.6.). The fifth question will examine whether the effect is immediate and if it holds over time (section 3.7.). We will also examine whether producing the gesture vs. simply observing it makes a difference (sixth question, section 3.8.). Finally, the seventh question will tackle generalization of learning (section 3.9.). A discussion will then suggest several hypotheses to explain the potential positive effect of manual gestures on word learning.

2. Methodological considerations

2.1. Terminology and acronyms

In the following, expressive and receptive learning will be distinguished. Expressive (or productive) learning refers to being able to produce the learned word upon testing. Receptive (or comprehensive) learning refers to being able to comprehend the learned words upon testing. For the sake of space and clarity, the following acronyms will be used in the text: CI: Cochlear Implant; T21: Trisomy 21; SLI: Specific Language Impairment; TD: Typically Developing.

2.2. Inclusion criteria

This analysis reviews only articles written in English in order for the reader to be able to directly access their content. It explored only journal articles. Only experimental studies directly controlling training and testing

material and procedures were included. We decided to exclude observational studies, even if they describe valuable data: it is indeed difficult to evaluate the size of the effects when the material used cannot be controlled and varies from one participant to the other. No further restriction was made on methodological aspects e.g., number of participants, age, language tested and type of population (typical and clinical). To be included, studies could test receptive and/or expressive learning as well as generalization. We also chose to review studies both directly evaluating the effect of gesture vs. none, those comparing different types of gestures and those comparing the use of gesture vs. other additional cues.

2.3. Exclusion criteria

Studies analyzing the role of gestures in learning words in a foreign language or learning new pseudo-words for already known words were excluded from this review in order to be more homogenous in terms of cognitive processes involved in the task performed by the participants. They are commented on in the introduction. This analysis excluded conference articles.

3. Description and analysis of selected studies

3.1. General description of the sample of studies reviewed

The final sample of studies reviewed here consists of 19 articles describing a total of 20 experimental studies relevant to the topic. Some articles describe two studies whereas some studies are analyzed in two articles from different points of views. Even though this was not a selection criterion, all the studies found dealt with children. Fourteen studies included TD children and eight included children with various disabilities (DIS) involving speech impairments: children with T21 (three studies, N=21), children with SLI/Developmental Language Disorder (three studies, N= 57), deaf and hearing-impaired children (three studies, N=38) and children with cerebral palsy (one study, N=3). Table 1 provides an overview of the characteristics of the populations involved in the studies in terms of number of participants and ages. One can note the strong variability in the number of participants included in the different studies ranging from 4 to 120 as well as in their ages ranging from a mean age

Table 1: Overview of the populations analyzed in the 20 experimental studies in terms of number of participants and age (TD = Typically Developing children; DIS = children with various disabilities involving speech impairments; mos. = months; yrs. = years; age in years: yrs.;mos.).

	Number of participants		Age of participants	
	TD	DIS	TD	DIS
Total	527	119		
Mean	35.1	14.9	44.1 mos. (3;8 yrs.)	82.3 mos. (6;10 yrs.)
Standard deviation	29.7	9.2	31.9 mos. (2;8 yrs.)	40.3 mos. (3;4 yrs.)
Minimum	10	4	8.45 mos.	42.3 mos. (3;6 yrs.)
Maximum	120	33	128 mos. (10;8 yrs.)	158.4 mos. (13;2 yrs.)

of 8.45 months to 13;2 years. Eleven studies dealt with English (N=401), four with Dutch (N=127), three with German (N=100) and one with Slovak (N=18). Table 2 provides details on all these studies such as number of participants, age, language, experimental design, and type of gestures tested.

3.2. Is there a general advantage of adding gestural cues for learning new words in TD children?

The aim here is to provide a first very general overview of the results of the studies concerning the efficiency of adding manual gestures to learn new words. A total of eight studies directly compared word learning with and without manual gestures (three between-subject designs and five within-subject designs) in TD children. As a whole, they tested 261 participants. Five studies (Capone & McGregor, 2005; Booth, McGregor, & Rohlfing, 2008; McGregor et al., 2009; de Nooijer et al., 2014; Lüke & Ritterfeld, 2014) involving a total of 212 children put forward a significant positive advantage of adding manual gestures during training to learn new words either expressively, receptively or both. Two studies (Bird et al., 2000; van Berkel-van Hoof et al., 2016) involving a total of 29 children found no difference between conditions: new words were learned equally well expressively and/or receptively whether they were trained alone or alongside a manual gesture. There is no clear effect of language or age on the effect of gesture on word learning (see

table 2). One study (Ting, Bergeson, & Miyamoto, 2012) put forward a disadvantage of adding manual gestures: words were learned less well when trained with a manual gesture rather than alone. This specific study is however quite different from those cited above. It involved much younger children (8.5 months) and this implied using specific methods very different from those used in the others. During training, some infants were familiarized with the target words using videos in which they could see a person uttering the words while others saw the person speaking and gesturing the words. Upon testing, the infants saw videos with a speaker uttering passages including the familiarized words vs. ones with the same speaker uttering passages with words not used during training. Preference was evaluated through looking durations. The infants trained with manual gestures showed no preference for the videos with the familiarized words whereas those in the word only condition did. Even if it was important to include this study in the present review for it to be exhaustive, because of the reasons presented above, it was decided to put aside this study when tackling the following research questions. One could indeed argue that differences in the effects observed could directly result from the great methodological differences corollary to involving infants.

As a whole, the studies reviewed suggest that adding a manual gesture to the word during training improves word learning performances. A potential explanation why some studies found no effect of adding a gesture to learn new words could be that the participants were children and that the gestures used during training were produced by adults. It may be the case that, as suggested by de Nooijer and colleagues (2014), gestures produced by peer-models would be more efficient. Imitating other's actions may indeed be easier when the actions are modeled by peers of similar ages (see Schunk, 1987, mixed results however). It may also be the case that children have more facilities identifying themselves with the person modeling the action if the latter is a peer (Liuzza, Setti, & Borghi, 2012). On the other hand, this explanation is contrary to the fact that language acquisition is of course guided by interaction with adults (primarily the parents). Note however that as soon as children attend day care or school, language acquisition is also largely influenced by communicative interactions with peers.

Table 2: Summary of the information on the articles reviewed: reference (for the sake of conciseness and when there was no ambiguity, all references with more than two authors are stated as 1st author et al., year), description of the population, mean age of participants (standard deviation, range), experimental design, gesture type tested, modality during training (observation and/or imitation), type and number of words learned, existing or invented, known or not to the participants, modality during training (observation and/or imitation), number of training sessions and frequency, modality of recall (expressive and/or receptive learning), testing time: immediate and/or delayed (delay after end of training), control of gesture production during recall, summary of the results (only significant results are reported), language used in study.

Reference	Population – no. of part. (no. of f.)	Age of part.: m. (sd. - rng.)	Experimental design (no. of gestures presented if relevant)	Gesture modality for training	Words learned: type (no.)	Exist. or inv.
Booth et al., 2008	TD - 80 (39 f.) G: 16 (9 f.) GP: 16 (8 f.) GT: 16 (9 f.) GM: 16 (7 f.) BL: 16 (6 f.)	29.31 mos. (0.89 – 28–31)	**btw.:** [a] GAZE - GAZE+POI - GAZE+POI+T - GAZE+POI+T+M - BSL	obs.	n. (3)	inv.
Bird et al., 2000	T21 - 10 (? f.) TD - 10 (? f.) [b]	T21: 42.3 mos. (25–62) TD: 21.8 mos. (14–30)	**w/in.:** WD - ARB[c](4) - WD+ARB	obs. + free imit.	n. (6)	inv.
Capone & McGregor, 2005	TD - 19 (13 f.)	28.7 mos. (0.99 – 27–30)	**w/in.:** WD - WD+SHP (2) - WD+FNC (2)	obs.	n. (6)	inv.
Capone, 2007	TD[d] - 18 (12 f.)	28.72 mos. (1.02 – 27–30)				
Capone Singleton, 2012	TD - 16 (8 f.)	32.63 mos. (4.02 – 27–42)	**w/in.:** WD - WD+SHP (1) - WD+FNC (1) - WD+POI	obs.	n. (3)	inv.

Words known?	Word modality for training	No. of training ses./Frequency	Recall modality	Testing time	Gesture prod. during recall	Summary of results	Language
no	obs.	1	expr. + rec. + rec. cat. general.	imm. + deld. (3–5 d.)	no	**expr.:** GAZE~GAZE+POI~ GAZE+POI+T~ GAZE+POI+T+M~BSL - imm. > deld. **rec. & rec. general.:** GAZE+POI, GAZE+POI+T GAZE+POI+T+M > BSL	English
no	obs. + free imit.	3 / ?	expr. + rec.	imm.	no	**expr.:** TD > T21 - WD~ARB~WD+ARB **rec.:** T21: WD+ARB>WD ~ARB - TD: none	English
no	obs.	3 / daily	expr. + rec.	imm. + deld. (~ 9.5 d.)	yes (Capone, 2007)	**expr.:** uncued resp.: WD+SHP>WD~WD+FNC - cued resp.: WD+SHP~ WD+FNC>WD **rec.:** WD+SHP>WD+FNC~ WD~chance	English
no	obs.	3 / ~ every 2 d.	expr. + cat. general. (expr. + rec.)	deld. (~ 4.1 d.)	no	**All tests:** WD+SHP>WD+POI~ WD+FNC	English

(*continued on next page*)

Table 2: Continued

Reference	Population – no. of part. (no. of f.)	Age of part.: m. (sd. - rng.)	Experimental design (no. of gestures presented if relevant)	Gesture modality for training	Words learned: type (no.)	Exist. or inv.
de Nooijer et al., 2014	TD - 53 (31 f.)	8.6 yrs. (0.6)	**w/in.:** DEF - DEF+PANT[e](6) - DEF+PANT+IMIT (6) - DEF+ACT	obs. and/ or imit.	v.[f] (24)	exist.
Giezen et al., 2013 (study 2)	CI - 8 (2 f.) prelingually deaf	6;11 yrs. (9 mos. - 5;9–8;1)	**w/in.:** WD - ARB (8) - SSS	obs.	n. (8)	inv.
Kapalková et al., 2016	TD - 18 (12 f.)	2 yrs.(24–34 mos.)	**btw.:** WD+ICO[g] (10) - WD+PIC	obs. + imit.	? (10)	inv.
Kohl et al., 1979	H - 4 (? f.)3 CP, 1 T21	13.2 yrs. (11.1–16.1)	**w/in.:** WD - PartSgn - CompSgn	obs.	n. (18) + v. (6) + prep. (6)	exist.
Lüke & Ritterfeld, 2014 (study 1)	TD - 20 (5 f.)	4;9 yrs. (3;4–5;11)	**w/in.:** WD+ICO[h] (3) - WD+ARB (3) - WD	obs.	n. (9)	inv.
Lüke & Ritterfeld, 2014 (study 2)	SLI[i] - 20 (7 f.) WD+ICO: 10 (5 f., 4 bil.) WD: 10 (2 f., 5 bil.)	4;7 yrs. (3;4–5;7)	**btw.:** WD+ICO (9) - WD	obs.	n. (9)	inv.
McGregor et al., 2009	TD[j] - 40 (21 f.) WD: 13 (7 f.) WD+G: 12 (8 f.)WD+P: 15 (5 f.)	1;8–2;0 yrs. WD: 20.68 mos. (0.95) WD+G: 21 mos. (1.54) WD+P: 21.26 mos. (1.38)	**btw.:** WD - WD+ICO[k] (1) - WD+PHO	obs.	Under	exist.

Words known?	Word modality for training	No. of training ses./Frequency	Recall modality	Testing time	Gesture prod. during recall	Summary of results	Language
no	obs.	2 / daily	def. recall + rec.	imm.	no	**def. recall:** DEF+PANT>DEF, DEF+ACT for loc. v. (vs. abs. & obj.) **rec.:** none	Dutch
no	obs.	1	rec.	imm.	no	WD~ARB~SSS	Dutch
no	obs. + imit.	15 / 4 times a wk.	expr.	deld. (T1: 1 d.; T2: 2 wk.; T3: 6 wk.)	no	WD+ICO>WD+PIC T1>T3 - T1~T2 - T2~T3	Slovak
no	obs.	15 / daily (Mond. - Sat.)	expr. + rec.	imm.	yes	**expr.:** for 1 part. only PartSgn, CompSgn > WD **rec.:** CompSgn~PartSgn>WD	English
no	obs.	1	expr. + rec.	imm.	no	**expr.:** none, no correct labelings **rec.:** WD+ICO~WD+ARB> WD	German
no	obs.	3 / weekly	expr. + rec.	imm. / deld. (1 wk.)	no	**imm.:** expr. & rec.: none **deld.:** T1 & t2: expr.: WD+ICO>WD rec.: none	German
no except 4 part.	obs.	1	rec. + general.	imm. / deld. (2–3 d.)	no	**rec.:** WD+ICO~WD>WD+PHO - imm.~deld.>pre-test **rec. general.:** WD+ICO~WD>WD+PHO - deld.>imm.~pre-test **WD+ICO:** deld.>pre-test, deld.~imm., imm.~pre-test	English

(*continued on next page*)

Table 2: Continued

Reference	Population – no. of part. (no. of f.)	Age of part.: m. (sd. - rng.)	Experimental design (no. of gestures presented if relevant)	Gesture modality for training	Words learned: type (no.)	Exist. or inv.
Mollink et al., 2008	HI - 14 (10 f.) All wore hearing aids	5;11 yrs.(13 mos. - 4;4–8;3)	**w/in.:** CTL - WD - WD+SGN[l] (16) - WD+CLR[m]	obs.	n. (64)	exist.
Mumford & Kita, 2014	TD - 120 (57 f.) WD+MG: 36 WD+ESG: 32 WD: 33	41.48 mos. (3.13 – 36–47)	**btw.:** WD - WD+MG (5) - W+ESG (5)	obs.	v. (5)	inv.
O'Neill, Topolovec, & Stern-Cavalcante, 2002 (expe. 1)	TD - 40 (22 f.) DES: 20 (12 f.) POI: 20 (10 f.)	DES: 33.9 mos.(0.98, 32–35) POI: 34.8 mos. (1.11, 33–36)	**btw.:** WD+ICO[n] (5) - WD+POI	obs.	adj. (5)	exist.
O'Neill, Topolovec, & Stern-Cavalcante, 2002 (expe. 2)	TD - 32 (16 f.) DES: 16 (8 f.) POI: 16 (8 f.)	DES: 39.8 mos.(1.76, 37–43) POI: 40.6 mos. (1.75, 37–43)	**btw.:** WD+ICO[o] (4) - WD+POI		adj.[p] (4)	
Romski & Ruder, 1984	T21 - 10 (? f.)	5:7 yrs. (14.94 mos., 3:11–7:10)	**w/in.:** CTL[q] - WD - WD+SGN[r] (4)	obs. + enactment of actions on obj.	n. (12) + v.[s] (12)	exist.
Ting et al., 2012	TD - 20 (? f.) WD+SGN: 10 WD: 10	WD+SGN: 8.5 mos. (1, 7–9.5) WD: 8.4 mos. (0.93, 7.1–9.5)	**btw.:** WD - WD+SGN[u] (4)	obs.	n. (4)	exist.

Words known?	Word modality for training	No. of training ses./Frequency	Recall modality	Testing time	Gesture prod. during recall	Summary of results	Language
no	obs. + imit.	3 / weekly	expr.	deld. (T1: 1 wk; T2: 5 wk.)	no	**condition:** WD+SGN> WD+CLR~WD>CTL **test time:** T1>T2 **iconicity:** T1: strong~weak - T2: strong>weak	Dutch
no	obs.	1	rec. general.	imm.	no	WD+MG>WD~WD+ESG	English
var.	obs.	1	rec. general.	imm.	yes	Tendency towards WD+ICO>WD+POI	English
no						WD+ICO>WD+POI	
no[4]	obs.	m.=23.1 (rng.=10–48) / daily (weekdays)	expr. + rec. + rec. general.	deld. (?)	yes	**expr.:** few resp. - WD~WD+SGN **rec.:** not sig. but ad. for 5 part. dis. for 2 part. none for 3 part. **expr. general.:** WD>WD+SGN **rec. general.:** WD+SGN>WD - rec.>expr.	English
no	obs.	1	rec.	imm.	no	**looking time:** WD+SGN: trained~untrained WD: trained>untrained	English

(*continued on next page*)

Table 2: Continued

Reference	Population – no. of part. (no. of f.)	Age of part.: m. (sd. - rng.)	Experimental design (no. of gestures presented if relevant)	Gesture modality for training	Words learned: type (no.)	Exist. or inv.
van Berkel-van Hoof et al., 2016	52 (25 f.)HI[v] - 16 (? f.) SLI[w] - 17 (? f.)TD - 19 (? f.)	10;8 yrs. (8.47 mos., 9–11)	**w/in.:** WD - WD+ICO[x] (10)	obs. + imit.	n. (20)	inv.
Vogt & Kauschke, 2017a	SLI - 20 (10 f.) TD AM - 20 (10 f.)TD LM - 20 (11 f.)No exposition to gesture or sign	SLI: 4;6 yrs. (0;7) TD AM: 4;5 yrs. (0;3)TD LM: 3;3 yrs. (0;16)	**w/in.:** WD+ICO (var.) - WD+ATT[y] (1)	obs.	n. (var.)[z] + v. (var.)[aa]	exist.
Vogt & Kauschke, 2017b						

Abbreviations (in alphabetical order): **ad.:** advantage – **adj.:** adjective – **beg.:** beginning – **bil.:** bilingual – **btw.:** between-subject design – **cat.:** category – **d.:** day – **def.:** definition – **deld.:** delayed – **dis.:** disadvantage – **exist.:** existing – **expe.:** experiment – **expr.:** expression – **f.:** female – **general.:** generalization – **imm.:** immediate – **imit.:** imitation – **inv.:** invented – **m.:** mean – **mo.:** month – **mos.:** months – **n.:** noun – **no.:** number – **obj.:** object – **obs.:** observation – **part.:** participant – **prep.:** preposition – **prod.:** production – **rec.:** reception – **rng.:** range – **sd.:** standard deviation – **ses.:** session – **sig.:** significant – **v.:** verb – **var.:** variable – **w/in.:** within-subject design – **wk.:** week – **yr.:** year – **yrs.:** years.
Acronyms (in alphabetical order): **AM:** Age-matched group (individually matched in chronological age (+/- 9 mos.) and gender) – **CI:** Cochlear Implant (prelingually deaf) – **CP:** Cerebral Palsy – **HI:** Hearing Impaired – **T21:** Trisomy 21 – **H:** Handicapped – **LM:** Language-matched group (individually matched on grammar comprehension, receptive and expressive vocabularies (nouns and verbs), word definition and nonword repetition (scores +/- 1/2 sd)) – **SLI:** Specific Language Impairment – **SLN:** Sign language of the Netherlands – **TD:** Typically Developing.

Words known?	Word modality for training	No. of training ses./Frequency	Recall modality	Testing time	Gesture prod. during recall	Summary of results	Language
no	obs. + imit.	3 / w/in. 1 wk.	rec.	Beg. of ses. 2 (T1) and 3 (T2) + 1 ses. within same wk. (T3)	no	**HI:** WD+ICO>WD - T1<T2<T3 - ad. gets larger over time **SLI:** WD+ICO~WD - T1<T2<T3 **TD:** same as SLI	Dutch
no	obs.	3 / every 2–3 d.	expr. + rec.	imm. after 1st training ses. (T1)/ deld. (T2, 2–3 d.)	no	**expr.:** WD+ICO>WD+ATT - T2>T1>pre-test T1: v.: WD+ICO>WD+ATT, n.: WD+ICO~WD+ATT T2: v.: WD+ICO~WD+ATT, n.: WD+ICO>WD+ATT **rec.:** WD+ICO>WD+ATT -T2>T1>pre-test	German
			def. prod.	deld. (2–3 d.)		**General tendency**: WD+ICO>WD+ATT - LM<SLI~AM	

(*continued on next page*)

Table 2: Continued

Experimental conditions: ACT: action enactment (part. asked to create a ges.) – **ARB:** arbitrary sign (no iconic resemblance with referent) – **ATT:** attention-directing gesture – **BSL:** baseline – **CompSgn:** complete signing (all words are signed), signs from Signed English Dictionary, 1 sign for each word – **CTL:** control (no training: only pre-test and post-test) – **DEF:** verbal definition – **ESG:** end-state gesture (depicts shape or lines formed by action) – **FNC:** function gesture (dynamic symbol) – **ICO:** iconic gesture – **IMIT:** gesture imitation – **MG:** manner gesture (depicts action of the hand) – **PANT:** pantomime – **PartSgn:** partial signing (signing of keywords only) – **PHO:** photo – **PIC:** picture – **POI:** pointing at target – **SGN:** sign – **SHP:** shape gesture (static symbol) – **SSS:** Sign Supported Speech – **WD:** word

[a] GAZE: Experimenter gazes at target – T: Experimenter additionally extends arm till touches object – M: Experimenter additionally pushes object across the table – BSL: Experimenter looks at table midway from target and foil

[b] Group matching: mental age - T21 trained in manual signs - TD no

[c] Signs produced with both hands symmetrically

[d] Subgroup of Capone & McGregor (2005)

[e] Gesture observation only

[f] Verbs of 3 types: locomotion (loc.), object-manipulation (obj.), abstract (abs.)

[g] Gestures based on Slovak Sign Language

[h] Iconics constructed or adopted from German Sign Language (visible feature of characters' head or neck)

[i] Group matching: age, sex, bilingualism

[j] Group matching: chronological age, total number of words, number of spatial terms

[k] Experimenter holds right hand over left and moves right hand under the left

[l] Signs from the Sign Language of the Netherlands – 2 sub-conditions: strong and weak iconicity

[m] Experimenter names a color in addition to naming the picture

[n] Demonstrating property described by adjective, performed on toy

[o] Demonstrating property described by adjective, performed on toy

[p] Adjectives describing tactile properties applied to animal names familiar to the participants

[q] Control for natural acquisition of the trained words (only pretest and postest)

[r] Signs from Signing Exact English (Gustason, Pfetzing, & Zawolkow, 1975)

[s] Transitive verb, representing object manipulation

[t] Individualized selection for each participant to select unknown words

[u] Signs from Signing Exact English (Gustason, Pfetzing, & Zawolkow, 1975) – 2 signs out of 4: iconic; 2 other arbitrary

[v] Hearing aids of different types – exposed or users of the SLN

[w] Some exposed to SSS

[x] Gestures invented accordingly to SLN formational principles and depicting "a defining feature" of the character

[y] Attention-directing gesture: raised forefinger in front of upper body

[z] Shape of animal – no. of nouns taught: TD AM: 6; SLI & TD LM: 4

[aa] Manner and/or path of movement – no. of verbs taught: TD AM: 6; SLI & TD LM: 4

3.3. Is there a general advantage of adding gestural cues for learning new words in children with speech and language deficits of various types?

A total of seven studies directly compared word learning with and without manual gestures (one between-subject design and six within-subject designs) in children with speech and language difficulties. They tested a total of 99 children. Giezen, Baker and Escudero (2013; N=8) found no effect of adding a manual gesture for word learning in children with CI. Van Berkel-van Hoof and colleagues (2016; N=17) found the same result for children with SLI. All other studies found an advantage of using manual gestures for learning new words, whether it be expressively or receptively or both, in a total of 74 participants (children with T21: Kohl, Karlan, & Heal, 1979; Bird et al., 2000; Romski & Ruder, 1984 – children with SLI: Lüke & Ritterfeld, 2014 – children with hearing impairments: Mollink, Hermans, & Knoors, 2008; van Berkel-van Hoof et al., 2016 – children with cerebral palsy: Kohl, Karlan, & Heal, 1979). Among the latter studies, Bird and colleagues (2000) and van Berkel-van Hoof and colleagues (2016) also included a group of TD children for which they found no effect of adding manual gestures to the learning of new words. The lack of a positive effect in Giezen, Baker and Escudero (2013; children with CI) could be explained by the fact that they used only one training session and only immediate and no delayed testing. The number of taught words was also very important (64) which could result in a floor effect. The authors interestingly put forward that, even if there was no positive effect, there was no negative effect either. Using manual gestures thus did not interfere with word learning.

The results of the reviewed studies therefore suggest that manual gestures could help children with speech and language difficulties learn new words, maybe even more so than for TD children. Van Berkel-van Hoof and colleagues (2016) actually found a positive effect of gestures for word learning only for hearing impaired children (vs. TD children and SLI children). They hypothesize that "Because these children are bimodal bilinguals, they process augmentative signs through the phonological loop as they do speech" (p. 346). Gestures may be more effective to help learn new words when the participants are used to using and/or seeing gestures.

Bird and colleagues (2000) indeed found a positive effect of signs for word learning in children with T21 familiar with signing but not in TD children. Van Berkel-van Hoof and colleagues (2016) also found a positive effect of iconic gestures for children with hearing impairments familiar with sign language whereas they found no such effect in children with SLI and TD children. Note however that this hypothesis is not backed by the fact that a number of other studies did find positive effects of gestures for word learning in TD children who had never been exposed to signs. There may be interferences with other factors such as age and duration of training, but these are difficult to analyze because of the variability in methodologies used and populations tested. A crucial point is that none of the studies found a negative effect of adding manual gestures for word learning: either there was no effect or a positive one.

3.4. Are manual gestures more (or less) efficient than other additional cues for word learning?

It could be the case that providing any additional cue, whether it be a manual gesture or something else, could improve word learning. This section examines in more detail the studies comparing the effect of using manual gestures to that of using other additional cues. Booth, McGregor and Rohlfing (2008) compared several conditions: 1. using pointing to the object to learn its label; 2. additionally touching it; 3. additionally moving it across the table (in TD children). There was no advantage of the two latter conditions compared to the former which all yielded similar positive effects for receptive word learning compared to a word alone condition. This suggests that the advantage solely emerged from using a pointing gesture since adding other cues did not further improve the effect.

Kapalková, Polišenská and Süssová (2016) analyzed expressive word learning in two groups of TD participants: one of them learned the words alongside manual gestures and the other with pictures. Even though participants managed to learn the new words in both groups, performances were significantly better in the gesture group.

McGregor and colleagues (2009) compared the learning of the preposition ‘under’ in three groups of TD children: one with the word only, one with an additional manual gesture, and one with a photograph. The

results show that performance improved from pre-test to immediate testing after training to a similar extent in all groups. A further analysis, however, showed a manual gesture advantage when comparing performance at pre-test and delayed post-test. This suggests that manual gestures promote learning more than photographs, not immediately after training, but after a two- to three-day delay. Manual gestures would thus be more efficient for maintaining word learning over time.

Mollink, Hermans and Knoors (2008) compared adding a sign to the spoken word during training to adding a color (labeled verbally by the experimenter during learning) to providing the word alone in children with hearing impairments. They found that receptive learning performances were better for the word + sign condition than for the two other conditions and not different for the two other conditions. This suggests that adding a color does not help promote word learning more than learning the word alone whereas adding a gesture does.

The results of the above four studies suggest that adding a manual gesture to the word during training is not equivalent to adding any other cue. It appears that the gesture plays a different role than other additional cues such as pictures.

3.5. Is there a differential effect of gesture on expressive vs. receptive learning? Interaction with number of training sessions

As stated in section 2, two types of learning can be distinguished: expressive and receptive learning. The aim of this section is to differentially examine the effect of adding manual gestures to the learning of new words for expressive and receptive learning.

Out of the seven studies cited to address the question in section 3.2. and directly comparing the use of gesture vs. none for word learning in TD children, four tested receptive learning only (McGregor et al., 2009; Ting, Bergeson, & Miyamoto, 2012; de Nooijer et al., 2014; van Berkel-van Hoof et al., 2016) and four evaluated receptive and expressive learning (Bird et al., 2000; Capone & McGregor, 2005; Booth, McGregor, & Rohlfing, 2008; Lüke & Ritterfeld, 2014).

Concerning receptive learning, four studies (Capone & McGregor, 2005; Booth, McGregor, & Rohlfing, 2008; McGregor et al., 2009; Lüke

& Ritterfeld, 2014) put forward an advantage of using gestures (vs. none) to learn new words, whereas the three others (Bird et al., 2000; de Nooijer et al., 2014; van Berkel-van Hoof et al., 2016) found no effect. The number of training sessions does not seem to explain the fact that some studies found no effect: two of the studies finding no effect used three training sessions and one of them used two, whereas three studies finding a positive effect of using manual gestures used only one training session and another study only three sessions. Another hypothesis to explain differences in the results could be the number of words learned. de Nooijer and colleagues (2014) taught 24 words to the participants and van-Berkel-van Hoof and colleagues (2016) 20 and found no effect, whereas all other studies finding a positive effect of adding manual gestures to learn new words taught between one and nine words.

Recall that all the studies testing expressive learning also analyzed receptive learning. Only one study (Bird et al., 2000) found no effect of gestures on both expressive and receptive learning. Booth, McGregor and Rohlfing (2008) as well as Lüke and Ritterfeld (2014) found that whichever condition (gesture or none), the participants did not manage to learn the new words expressively, even though they did receptively with an advantage for the gesture condition. As stated by Booth, McGregor and Rohlfing (2008), this may be due to insufficient training (only one session in both studies). Expressive learning would thus require more training than receptive learning. This hypothesis is corroborated by the results of Capone and McGregor (2005) who found a positive effect of iconic gestures underlying shape (vs. iconic gestures depicting function and no gesture) for learning new words after three training sessions. Note that when the participants did not manage to provide an expressive response, the experimenter provided a gestural cue. For these cued responses, the authors found a positive effect of both types of iconic gestures over no gesture. In the no gesture condition, the participants did not manage to provide any expressive responses even though they managed to learn some words receptively just as in Booth, McGregor, & Rohlfing (2008) and Lüke & Ritterfeld (2014). Capone Singleton (2012) obtained similar results comparing shape gestures, function gestures and pointing gestures: shape gestures showed an advantage over function and pointing gestures for expressive word learning after three training sessions. Overall, these

observations corroborate the fact that receptive learning is faster than expressive learning and that adding gestures can promote faster expressive learning even though such learning still takes longer than receptive learning. Note however that Vogt and Kauscke (2017a) found an advantage of iconic gestures over attention-directing gestures even at the end of the first training session for expressive learning. Booth, McGregor and Rohlfing (2008) also suggest that the lack of a positive effect of gestures on expressive learning may be due to the fact that the participants were not asked to produce the words during training (also the case in Lüke & Ritterfeld, 2014 and Capone & McGregor, 2005). All these arguments are further corroborated by the study by Kapalková, Polišenská and Süssová (2016) who found positive effects of using iconic manual gestures (vs. picture support) on expressive word learning after 15 sessions in a paradigm in which participants produced the words during training.

Out of the seven studies addressing the effect of adding a manual gesture on learning new words in children with speech and language difficulties (section 3.3.) and directly comparing the use of gesture vs. none, two tested receptive learning only, one expressive learning only and four both receptive and expressive learning. All studies, except Giezen, Baker, & Escudero (2013) and Lüke & Ritterfeld (2014), found a positive effect of adding manual gestures to the learning of new words receptively for at least one group of children with speech and language impairments. Note however that the effect was not significant in Romski and Ruder (1984): it was positive for only five out of 10 children with T21. Van Berkell-van Hoof and colleagues (2016) found a gestural advantage for receptive learning for children with hearing impairments but not for those with SLI.

Bird and colleagues (2000) found no effect of adding manual gestures to the learning of new words expressively in children with T21 even though they did for receptive learning. Lüke and Ritterfeld (2014) and Mollink, Hermans and Knoors (2008), on the other hand, report a positive effect respectively in children with SLI and hearing impairments. The effect is also positive in Kohl, Karlan, and Heal (1979) for children with T21 or cerebral palsy but it does not reach significance. Results from Romski and Ruder (1984) are unclear. It appears that a positive effect for expressive learning was obtained more often in children with disabilities compared to TD children (see above). To build on the discussion above, it is also the

case that the studies dealing with children with disabilities often included more training sessions than those dealing with TD children.

In a nutshell, the effect of manual gesture on receptive learning appears to be influenced by the number of words taught: when too many words are taught no advantage for gestures appears. Concerning expressive learning, some studies find no effect of gesture, but this is mainly due to the fact that the children did not manage to learn the words expressively whichever the condition mainly because of insufficient training (floor effect). With more training, children managed to learn new words expressively with an advantage when manual gestures were present during training (though see Vogt & Kauscke, 2017a, for an advantage of gesture on expressive word learning after only one training session). It seems that the positive effect of manual gestures on expressive word learning is greater for children with speech and language impairments though the studies involving such participants generally included more training sessions than those with TD children.

3.6. Does gesture type matter? Does iconicity matter?

The aim of this section is to analyze whether the experimental evidence puts forward a specific advantage of adding different types of manual gestures to the learning of new words. Specifically, one could hypothesize that if the gesture puts forward an iconic resemblance with the referent it could be more beneficial for learning the new word labeling the referent: "if a sign is more iconic than the spoken word, its form conveys information about a word's meaning and may thus assist a child in mapping new words to meanings" (Bird et al., 2000, p. 260).

Lüke and Ritterfeld (2014) found that both iconic and arbitrary gestures were equally beneficial to word learning in TD children aged 4;9 years. This could appear as contrary to the results of Namy and Waxman (1998) who compared the ability of children from 18 to 26 months to learn either word or arbitrary gestural labels. They found that 18-month-olds learned word or gestural labels indifferently whereas 26-month-olds learned word labels more easily and needed extra training to learn gestural labels. Bird and colleagues (2000) also found that 21.8-month-olds were not able to expressively learn arbitrary gestural signs alone as labels. Marentette and Nicoladis (2011) found that children aged 40 to 60 months could learn

iconic gesture labels for objects but not arbitrary gesture labels. This difference in findings could suggest that, even if children have difficulties learning only an arbitrary gestural label (without a word) for an object, the arbitrary gesture could however help them learn a corresponding and simultaneously presented word (at least receptively). Lüke and Ritterfeld (2014) hypothesize that "arbitrary gestures may have enhanced the interest of the child in the presented words in contrast to words introduced without gestures". This finding is however contradictory to that of Bird and colleagues (2000) who found no difference between a word only and a word + arbitrary sign condition on expressive and receptive word learning (mean age = 21.8 months). This difference in findings could be due to the age of the participants: the ones in the Bird et al. (2000) study may have been too young to manage to learn arbitrary signs even though Namy and Waxman (1998) found that 18-month-olds can learn words and arbitrary gestural labels. This hypothesis is backed by the fact that Bird and colleagues (2000) did find a beneficial effect of arbitrary signs for receptive learning of new words in older children with T21 (mean age = 42.3 months). Note however that the participants with T21 were trained in using signs prior to the study whereas TD children were not. Giezen, Baker and Escudero (2013) found no benefit in adding arbitrary signs to learn new words in 6;11-year-old children with CI. The scores were however close to 100 % and the absence of a gesture benefit could be due to a ceiling effect.

Another surprising observation, taking into account the results of Marentette and Nicoladis (2011), is that Lüke and Ritterfeld (2014) found no advantage of iconic gestures over arbitrary ones. This could be due to a ceiling effect for iconic gestures (as suggested by the authors themselves). Vogt and Kauschke (2017a) ran an interesting follow-up experiment to their main study comparing iconic and attention-directing gestures. It compared the use of iconic gestures versus arbitrary ones. Even though sample size was small (18 TD children) and impeded reaching statistical significance (according to the authors themselves), the results suggest that both expressive and receptive performances were higher for the iconic than the arbitrary gesture condition.

Mollink, Hermans and Knoors (2008) found a positive effect of adding signs to words for receptive word learning in hearing impaired children.

They analyzed this effect as a function of sign iconicity. Even though the effect was the same for signs with strong iconicity than for those with weak iconicity one week after training, the results after five weeks show that learning was better for strongly iconic signs. The interesting thing is that at five weeks, the performance decreased compared to one week only for weakly iconic signs but remained the same for strongly iconic signs suggesting that iconicity helped longer memory retention. This finding is corroborated by that of van Berkel-van Hoof and colleagues (2016) who found a positive effect of adding iconic signs for receptive word learning in hearing impaired children. Note however that they did not find a beneficial effect of iconic gestures over none in 10;8-year-old TD children and children with SLI. Lüke and Ritterfeld (2014) did find a beneficial effect of iconic gestures over none in 4;7-year-old children with SLI.

Capone and McGregor (2005) and Capone Singleton (2012) compared the effect of using iconic gestures underlying shape to ones underlying function. They found that shape gestures were more efficient in promoting expressive and receptive word learning than function gestures. This suggests that type of iconicity could be as important as iconicity itself. In both studies the shape gestures were static symbols and the function gestures were dynamic symbols. Even if performances on words trained with a function gesture were generally not better than those for words trained with no gesture, the authors put forward an interesting finding. In expressive learning testing, when the participants did not manage to produce the word, the experimenter provided the gestural cue. In these cases, function gestures functioned as good as shape gestures to help the children produce the new words upon testing, suggesting that function gestures may have a beneficial effect even though they are not as effective as shape gestures. Capone Singleton (2012) speaks of a shape bias already put forward by other researchers (Kemler Nelson et al., 2000; Smith et al., 2002). Mumford and Kita (2014) also compared different types of iconic gestures for learning new verbs receptively: ones underlying the manner of an action (dynamic) and the other its resulting end state (shape, static). Contrary to the above-mentioned studies, the authors found a beneficial effect over word only training solely for the dynamic manner gestures and not for the static end state gestures. This may be due to a difference in the ages of the participants: the participants were aged 41.48 months on average whereas

those in the Capone studies were aged around 28 months. Actually, it appears that the shape bias mentioned above would wear off with age (Imai, Gentner, & Uchida, 1994). Also note that in the Mumford & Kita (2014) study, the children learned verbs whereas they learned nouns in the Capone studies. Vogt and Kauscke (2017a) compared iconic gestures underlying path and/or manner to ones underlying shape. They found a larger advantage for path-manner than shape gestures for immediate learning of verbs but similar effects for both gesture types for immediate learning of nouns. On the other hand, after a two- to three-day delay from the end of training, they found a larger advantage for shape than path-manner gestures for nouns but similar effects for both gesture types for verbs. This result helps understand the differences between the Mumford & Kita study and the Capone studies.

Vogt and Kauscke (2017a) found an advantage of iconic gestures over attention-directing gestures (raised forefinger in front of upper body) even at the end of the first training session. The authors conclude that: "it is the iconicity of the gestures (that is the resemblance to the referent), rather than the item-specific encoding of both auditory and visual information to a lexical form, that helps learning" (p. 22). O'Neill, Topolovec and Stern-Cavalcante (2002) analyzed the generalization of the use of newly learned adjectives to qualify other objects than those used during training with a similar distinctive quality referred to with the adjective. For example, during training, the children were presented with a 'lumpy cat' and taught the adjective 'lumpy'. Upon testing they were presented with a 'lumpy turtle' and a 'smooth turtle' and asked to designate the 'lumpy' one. During the learning phase, some adjectives were learned with a descriptive gesture and others with a pointing gesture. In a first experiment, they found no difference between the two conditions except for the adjectives describing non-visual properties of objects (descriptive gesture advantage) and concluded that: "gesture may play a more important role in the learning of less visually detectable properties" (p. 255). In a second experiment, using lower frequency adjectives describing only non-visible properties, they found that descriptive gestures were more efficient in promoting learning than pointing gestures. An interesting finding is a higher frequency of mention of nontarget properties (resp. less expression of uncertainty during testing) by the participants in the pointing than in the descriptive gesture

condition, but only in experiment 1. Capone Singleton (2012) compared iconic gestures underlying shape to pointing. She found that shape gestures were more efficient in promoting expressive and receptive word learning than pointing. Booth, McGregor and Rohlfing (2008) however found that the use of pointing by the experimenter during training yielded better receptive learning than not using any gesture. It may therefore be the case that attention-directing gestures are also beneficial for word learning, but less than iconic gestures.

Overall, the studies reviewed here suggest that different types of manual gestures can have differential effects. Pointing gestures appear to be helpful for word learning compared to no gesture even though there is some evidence that gestures having an iconic resemblance with the referent would be more effective. "the use of descriptive gestures during the teaching of novel adjective terms appears (…) to have helped children to isolate the particular property intended by the speaker in a manner not possible when point gestures were used instead" (O'Neill, Topolovec, & Stern-Cavalcante, 2002).

Results of the studies comparing arbitrary and iconic signs do not all agree but do suggest in general that iconics are more beneficial. Several studies also put forward differences between iconic gestures with a bias of shape over function for nouns and path-manner over shape gestures for verbs. Capone and McGregor (2005) suggest that the role of gestures is to draw "attention to an important aspect of the word learning problem (shape, function or both), thereby reinforcing salient semantic content of the spoken language" (p. 1478). As suggested by Vogt and Kauschke (2017a), iconic gestures may facilitate the association with the lexical form. Type of iconic gesture could interact with type of word learned (e.g., nouns vs. verbs) and also the word itself. For example, one could hypothesize that some words are better represented by static shape iconic gestures (e.g., hands shaped as a round to illustrate the word 'ball') and others by dynamic function gestures (e.g., fingers miming cutting to illustrate the word 'scissors').

3.7. Does testing time matter?

Some of the studies reviewed tested the participants immediately after training, others after various delays and others at both times. Immediate

testing tackles fast mapping, the initial stage of word learning "in which a first connection of a word and referent is retained (Carey, 2010; Carey & Bartlett, 1978)" (Lüke & Ritterfeld, 2014, pp. 203–204). Delayed testing examines slow mapping, the retention of meaning and label association in memory when the "child forms a robust and more sophisticated lexical representation of the word (Carey, 2010; Horst, Parsons, & Byron, 2011)" (Lüke & Ritterfeld, 2014, p. 204). It is possible that a gestural effect could appear both at the stages of fast and slow mapping or only after some time as observed in other learning tasks (e.g., Cook, Mitchell, & Goldin-Meadow, 2008; Cherdieu et al., 2017).

A total of five studies (iconic gestures: Capone & McGregor, 2005; Lüke & Ritterfeld, 2014; Mumford & Kita, 2014; Vogt & Kauschke, 2017a; pointing: Booth, McGregor, & Rohlfing, 2008) put forward a beneficial effect of using iconic gestures to fast map new words receptively but not expressively (except Vogt & Kauschke, 2017a) in TD children (no gestural benefit: Bird et al., 2000).

Lüke and Ritterfeld (2014) found similar results for children with SLI (positive effect of gestures on receptive word learning). Bird and colleagues (2000) found the same for children with T21 and Kohl, Karlan and Heal (1979) for one child with T21 and three children with cerebral palsy. Giezen, Baker and Escudero (2013) however found no positive effect of using gestures during training for immediate receptive word learning in children with CI. Note that they used arbitrary signs whereas Lüke and Ritterfeld (2014) used iconic gestures (see section 3.6. for discussion on this topic). Bird and colleagues (2000) also used arbitrary signs but the children were exposed five times more to each label than those in the Giezen, Baker, & Escudero (2013) study. Finally, note that the testing scores in the Giezen, Baker, & Escudero (2013) study were relatively high (above 80 % correct responses) suggesting a possible ceiling effect.

Booth, McGregor and Rohlfing (2008) found a positive effect of pointing gestures on receptive word learning in TD children both immediately and after a three-to five-day delay. The same was obtained for iconic gestures by Capone and McGregor (2005; 11.5-day delay). Kapalková, Polišenská and Süssová (2016) found an effect of testing delay on general expressive word learning, all conditions (picture vs. iconic gesture support) put together with no interaction. Note however that all testing

sessions were delayed (one day after end of training vs. two weeks and six weeks), performances being better after one day than after two or six weeks. McGregor and colleagues (2009) however find a larger effect of gesture (over speech only) on the receptive acquisition of the preposition 'under' only after two to three days and not at immediate testing and only for generalization (not for trained pairs of objects). Note however that in all conditions including the speech only condition, the experimenter modeled the 'under' relationship on objects during training. Even though this is not a manual gesture per se, it may act as a gesture, which would explain the results. The authors also analyzed the correlation between short-term and long-term performances and found, only for the gesture group, that "children who demonstrate modest gains on the immediate post-test build on those gains for a more impressive performance at delayed post-test" (p. 819) and this only for unlearned combinations: "The gesture advantage was revealed by the children's ability to follow under instructions given the untrained generalization items" (p. 820). Lüke and Ritterfeld (2014) found the same result in children with SLI, the positive effect of gesture only emerging for expressive (and not receptive) learning after a one-week delay. Vogt and Kauschke (2017a) found no effect of condition x testing time on performance in expressive and receptive word learning in TD and SLI children but this study did not include a 'no gesture' condition, it only compared the use of iconic and attention-directing gestures.

Van Berkel-van Hoof and colleagues (2016) only used delayed testing at several time points and found no advantage for iconic gestures over speech in all cases in TD and SLI children even though word learning performance improved over time (three testing time points). This improvement is probably due to the fact that there was extra training between testing times. Note that they did, however, find a positive effect of using gestures during training in hearing impaired children and that this advantage increased in magnitude over time. Mollink, Hermans and Knoors (2008) also found positive effects of adding signs to words for receptive word learning in hearing impaired children one week and five weeks after the end of training (no fast mapping testing). In contrast to van Berkel-van Hoof and colleagues (2016), they found that instead of increasing, performance decreased with time. This discrepancy however probably stems from the fact that they tested children after one and five weeks whereas

van Berkel-van Hoof and colleagues (2016) tested children only one or two days after the end of training.

To summarize, in general, upon immediate testing, there is a gestural effect essentially for receptive learning in TD children as well as in children with speech and language impairments. This effect generally holds for delayed testing. Some studies however find no immediate advantage but do find a gestural advantage upon delayed testing, especially for expressive learning. Note that Brown and colleagues (2012) as well as McGregor (2014) found that performances in recall of newly learned words were better after 12 or 24 hours than immediately.

3.8. Is observing the gesture during training enough or does producing the gesture work better?

This section examines whether simply observing a gesture promotes word learning or if also producing it during the training phase could be more helpful. Some studies have indeed shown that imitating the gesture during training improves the beneficial effect of manual gestures for learning words in a foreign language (e.g., Macedonia, Bergmann, & Roithmayr, 2014).

A total of five studies (Capone & McGregor, 2005; Booth, McGregor, & Rohlfing, 2008; McGregor et al., 2009; Lüke & Ritterfeld, 2014; Mumford & Kita, 2014) found a positive effect of adding a gesture during training to learn new words through observation of gesture in TD children. The same observation was made by Kohl, Karlan and Heal (1979) for one child with T21 and three with cerebral palsy, by Lüke and Ritterfeld (2014) for children with SLI and by Mollink, Hermans and Knoors (2008) for children with hearing impairments. One could therefore conclude that producing the gesture during training is not necessary for a positive effect of gesture on word learning to emerge. Note that in all the above-mentioned studies (except Capone & McGregor, 2005 and Mollink, Hermans, & Knoors, 2008), the positive effect was only observed for receptive (and not expressive) learning. The results of studies in which the participants were explicitly asked to or were allowed to imitate the gestures (Bird et al., 2000; de Nooijer et al., 2014; Kapalková, Polišenská, & Süssová, 2016; van Berkel-van Hoof et al., 2016) actually do not obtain better results in terms of expressive learning. McGregor and colleagues (2009) only found

a positive effect of gesture after a two- to three-day delay. Even though the children did not imitate the gesture during training, they did enact the 'under' relationships learned on actual objects. This could have a similar effect as performing a gesture, which would explain the lack of a gesture advantage over word-alone training upon immediate testing: the children actually observed something close to a manual gesture even in the word only condition. The fact that a positive effect does emerge after delay could suggest that actual actions on objects even though positive for word learning are not as effective as manual gestures.

Some interesting observations come from studies in which imitation was not forced and which analyzed the correlation between imitation and word learning performances. In Bird et al. (2000), the participants were not required to, but could, imitate the gestures during training. Correlational analysis showed no correlation between imitation vs. none and word learning performance for children with T21 but did find moderate to high correlation for TD children. De Nooijer and colleagues (2014) tested receptive verb learning and included both a condition in which participants only observed the gesture and one in which they were also asked to imitate the gesture. They did not find any advantage of imitating the gesture rather than just seeing it.

It could also be the case that producing the gesture upon recall could facilitate the latter. Only one study directly controlled for gesture production during testing. O'Neill, Topolovec and Stern-Cavalcante (2002) found a positive correlation between descriptive gesture production during testing and receptive performance in TD children only for two out of the five adjectives learned in experiment 1 but for all adjectives in experiment 2. Kohl, Karlan and Heal (1979) also controlled for sign production during recall in children with disabilities but did not correlate that measure to word learning performance. The same can be said of Romski and Ruder (1984) for children with T21 and Capone (2007) for TD children. Mollink, Hermans and Knoors (2008) asked the hearing-impaired participants to repeat the word during training but no indication is given as to whether they imitated the signs.

An interesting study relevant to the present question is that of de Nooijer and colleagues (2013). It was not included in the review itself because it did not directly compare the use of gestures during learning with that of other

cues. In their study, participants (N=120; mean age = 10 years) learned verbs always associated with a gesture during training. The focus was on analyzing the effect of gesture imitation (or not) by participant during learning and/or recall. The verbs were of three types: abstract, locomotion or object-manipulation. They found a positive effect of imitation (over none) only for the object-manipulation verbs. They suggest that gesture imitation would be crucial only when the action imitated is goal-directed.

Imitation of the spoken word also appears to be a crucial factor. Bird and colleagues (2000) indeed found significant correlations between spontaneous imitations of the word during learning and the expressive and receptive learning performances. There was no analysis of specific gestural imitations, which were infrequent in the TD participants.

The study by de Nooijer and colleagues (2014) included an extra interesting condition in which they asked the children to invent a gesture for the word. They found that the children invented gestures very similar to those the experimenters had invented for the other experimental conditions (sometimes simplified). However, no clear positive or negative effect of this condition over the others (no gesture or gesture from the experimenter) emerged.

In a nutshell, producing the manual gesture during training does not appear to be absolutely necessary since several studies find a gestural advantage even if the children only observed the gesture during training. A few, but not all, studies however showed a positive correlation between production of the manual gesture during training and better word learning performances. Results are still too sparse to draw a strong conclusion. One study also showed that producing the gesture during testing results in better word recall.

3.9. Analyzing generalization

Another interesting aspect of learning is generalization. Once the words are learned in one context, it is indeed important to be able to generalize their use to other contexts. Category generalization corresponds to the capacity to be able to use nouns labeling objects for other objects of the same category (for example, being able to use the noun for the same object in a different color than during training). For action verbs, generalization

corresponds to the ability to use the verb to describe a similar action in another context (for example, being able to use the verb to describe the action but performed on different objects than during training). For adjectives, generalization is the ability to use the newly learned word to qualify a different object having the same property than that designated by the learned adjective.

Booth, McGregor and Rohlfing (2008) found that TD children managed to receptively generalize the use of new words to objects from the same category more when the words were trained with a pointing gesture. Capone Singleton (2012) obtained the same results for iconic gestures underlying shape, both receptively and expressively. McGregor and colleagues (2009) found the same but only receptively and only after a two- to three- day post-training delay. In the two latter studies, the participants learned nouns for objects in three conditions: with an iconic gesture depicting the object's shape, with one illustrating the object's function or with none. They were then tested on the generalization of the use of the noun for another object of the same category (similar shape and same function). They did this more efficiently when they had learned the noun with a shape gesture than with no gesture (not true for function gestures, see above for discussion on this potential shape bias).

O'Neill, Topolovec and Stern-Cavalcante (2002) found that descriptive gestures were more efficient then pointing to receptively generalize the use of adjectives to new objects by TD children. In Experiment 2, the adjectives described non-visible properties of the insides of objects and participants were asked to generalize the adjectives to objects of different shapes and colors but with the same non-visible property.

Romski and Ruder (1984) found that signs helped children with T21 receptively generalize the use of new verbs and nouns but impeded expressive generalization. Participants learned verb/noun combinations describing actions performed on objects and were then tested on how they generalized the use of the nouns and verbs when they were combined differently than during training.

Mumford and Kita (2014) compared the effect of adding manner or end-state gestures to verbs during learning with a control condition with no gesture. The participants were then tested receptively on their generalization performances. Children were taught a verb alongside a gesture

focusing on the manner of the action or one focusing on the end-state or none. They were then asked to select from two videos which one corresponded to the verb label. The materials were different but one of the videos displayed the same manner as during training and the other one the same end-state. They found that participants generalized more often based on manner when they had seen a manner gesture during training (no preference bias for end-state gestures and no gestures). The authors draw several conclusions from their findings: 1. the fact that the results are different for manner and end-state gestures suggests that the "gestural content" plays a role; 2. the fact that children manage to generalize suggests that "gestures do not simply help children to associate a word with a scene in general"; 3. "iconic gestures provide a sketch of abstract semantic representations of verbs, which help children carry out fast mapping (…) of newly encountered verbs and correctly apply the verbs to novel complex scenes".

To summarize, several studies find that manual gestures play a positive role in generalizing the use of newly learned words to new contexts. McGregor and colleagues (2009) suggest that: "gesture input promoted more robust knowledge of the meaning of *under*, knowledge that was less tied to contextual familiarity and more prone to consolidation" (p. 824).

4. Summary and explanatory hypotheses for the role of manual gestures in word learning

This chapter reviewed a total of 19 articles describing 20 experimental studies examining the effect of using manual gestures during training on word learning. Even though this was not a criterion for selection, they all tested children of various ages who were either typically developing (TD) or with speech and language difficulties from various origins: hearing impairment (HI), specific language impairment (SLI), trisomy 21 (T21), cerebral palsy (CP). Based on these studies and their results, several research questions were addressed in order to try and better understand the potential role of gestures in word learning.

The first question (section 3.2.) was very general and asked whether manual gestures, whichever their type, actually improved word learning performances in TD children. Five out of eight studies involving a total of

212 children and directly comparing a word + gesture condition to a word only condition found better word learning performances in the word + gesture condition. Two studies, testing a total of 29 children, found no difference between conditions. The results of all the studies reviewed therefore suggest that supplementing a word with a manual gesture during training is beneficial to word learning (see section 3.2. for more details).

Section 3.3. examined the same research question in children with speech and language impairments. Six out of seven studies involving 74 children and directly comparing a word + gesture condition to a word only condition found better word learning performances in the word + gesture condition for at least one group of children with speech and language impairments. Two studies, testing a total of 25 children, found no difference between conditions. In these two latter studies, the children were not disadvantaged by manual gestures either. All this suggests that using manual gestures to teach new words to children with speech and language impairments could be useful. The studies reviewed indeed show that using gestures at worst does not have a positive effect and at the best promotes word learning, but in no cases impedes it. The evidence reviewed here actually suggests that children with speech and language impairments would benefit even more from manual gestures than TD children (see section 3.3. for more details). This could have implications for speech and language intervention.

Section 3.4. tackled the question of the specificity of gestures over other additional cues for word learning. The results of the studies comparing the use of manual gestures, be it pointing or iconic representational gestures, to other cues, such as pictures, all suggest that there is a specificity of manual gestures which promote word learning more than other additional cues.

Section 3.5. further examined the potentially different effects manual gestures could have on expressive vs. receptive learning. Expressive learning refers to the ability to produce the word when asked and receptive learning refers to the ability to comprehend the word after training. In general, performances in receptive learning are better than those in expressive learning. Receptive learning appears to be positively influenced by manual gestures as long as not too many words are taught. Expressive learning was shown to take longer than receptive learning and requires more training sessions. A manual gesture advantage thus only appears when a sufficient

number of training sessions is used. Finally, the studies reviewed suggest that the positive effect of manual gestures on expressive word learning is greater for children with speech and language impairments even though it is also true that these studies generally used more training sessions than those involving TD children (see section 3.5. for details).

Section 3.6. aimed at analyzing whether different types of gestures had different effects. The studies reviewed suggest there are differences between manual gestures of different types in terms of effect on word learning. Pointing appears to be helpful for word learning even if the positive effect would be weaker than for iconic gestures. Even if the results do not all agree, iconics would be more beneficial than arbitrary signs even if the latter do also have a positive effect. It also appears that type of iconicity (e.g., underlining shape vs. function) would have differential effects even though this appears to interact with the type of words learned (e.g., nouns vs. verbs) (see section 3.6. for more details).

Section 3.7. tackled the question of testing time and the potential differences between gestural effects for immediate vs. delayed testing. In immediate testing, a positive effect of manual gesture most often appears only for receptive learning in TD children as well as in children with speech and language impairments. This beneficial effect generally holds in time and still can be observed upon delayed testing. The positive effect sometimes appears only after a certain delay (at least one night) especially for expressive learning (see section 3.7. for details).

Section 3.8. examined whether actually producing the gesture, vs. just observing it, during training yielded different results. The results of the studies reviewed suggest that gesture production during training is not mandatory for a beneficial effect of manual gestures to appear. Some results however suggest that producing the gesture during training would result in a greater positive effect of manual gestures. Production of the gesture during testing could also yield better recall performances (see section 3.8. for details).

Finally, section 3.9. analyzed potential effects of manual gestures on generalization performances. Some results suggest that manual gestures could enhance generalization of the newly learned words to other contexts.

In the following, we will address different hypotheses to explain why gestures would play a beneficial role in word learning based both on personal thoughts and on those proposed by other authors.

4.1. Do gestures simply function as attention attractors to the word learning context?

A hypothesis to explain why gestures facilitate word learning could be that using a gesture during training would function as an attention getter: the gesture would help focus the learner's attention on the object of the training, i.e. the word pronounced by the experimenter. Joint attention (Tomasello, Carpenter, & Liszkowski, 2007), which is important for learning, would be enhanced by the gesture. If this is the case, one could expect that gestures would not facilitate learning more than any other means of attracting attention. Some of the studies reviewed in this chapter provide information relative to this question.

Booth, McGregor and Rohlfing (2008) compared the use of pointing towards the labeled object and the use of gaze towards the latter. They found that learning performances are better in the pointing than in the gaze condition. They also controlled for child attention by analyzing looking time of the participants when the experimenter labeled the object during learning. They found that the participants' looking time towards the target object did not differ between conditions. This suggests that the level of attention was the same across conditions and that, even so, the use of pointing enhanced receptive word learning. The authors conclude that: "socio-pragmatic factors come to play a larger role than perceptual-attentional factors in word learning by the time the children reach 2 ½ years of age" (p. 198). This suggests that there is something more to manual gestures than just drawing attention to the word learning context. Note however that O'Neill, Topolovec and Stern-Cavalcante (2002) found a percentage of correct responses in the pointing condition very close to chance suggesting that learning did not occur in the pointing gesture condition whereas it did in the descriptive gesture condition in Experiment 1. Performances were however higher than chance in the pointing condition in Experiment 2.

Bird and colleagues (2000) did not find better expressive and receptive word learning performances when words were associated with an arbitrary gesture than when they were associated with no gesture. If gestures solely functioned as attention getters, arbitrary gestures should also attract the participants' attention.

Kapalková, Polišenská and Süssová (2016) found that providing a manual iconic gesture during training yielded better expressive learning performances compared to providing a picture. McGregor and colleagues (2009) also found that an iconic gesture was more efficient than a photograph for learning the meaning of the preposition 'under'. It however seems that gesture and picture/photographs should equally function as attention getters. Capone Singleton (2012) compared the effect of shape gestures, function gestures and pointing gestures and found that shape gestures yielded better expressive learning and category generalization performances than function and pointing gestures. Even if one could argue that a representational gesture could function as a stronger attention getter than a pointing gesture, there is no reason why a representational gesture underlining shape would focus attention more than one underlying function. The authors suggest that "Whereas both pointing and iconic gestures can draw attention to an abject, the iconic gesture may also orient children to attend to or strengthen their inferences about specific features and their connection to the word label" (p. 289–290).

Vogt and Kauschke (2017a) found a positive effect of iconic gestures over an attention directing gesture (raised forefinger in front of upper body) both for fast and slow mapping expressive and receptive learning of new words by children with SLI as well as age-matched and language-matched TD children: "iconic gestures provide an advantage over and above focusing children's attention" (p. 21). Note that receptive and expressive learning also occurred in the attention-directing gesture condition even though to a lesser extent than in the iconic gesture condition. The advantage of iconics vs. the attention-directing gesture further depended on testing time and word type: it was only significant for verbs for fast mapping and for nouns for slow mapping.

Put together, all this evidence suggests that either gestures have an additional role than just attracting the learner's attention, even more so for iconic representational gestures, or that gestures attract attention more than other cues as hypothesized by McGregor and colleagues (2009): "gestures are interesting, and thus draw more attention to moments of training" (p. 822).

4.2. Memory enhancement

Manual gestures could function as additional traces to help the learner memorize the new words more efficiently. Several of the studies reviewed

in this chapter provide insight on this hypothesis. McGregor and colleagues (2009) put forward a gestural advantage only after two- to three-days and not immediately. Lüke and Ritterfeld (2014) also found a gestural advantage only after a one-week delay and not immediately for expressive learning. Finally van Berkel-van Hoof and colleagues (2016) found that the gestural advantage grew over time. All this evidence suggests that gestures help learners memorize new words longer than when the words are trained without a gesture.

Mollink, Hermans and Knoors (2008) compared word learning performances after one and five weeks. They found a gestural advantage but no differential effect between strong and weak iconicity signs after one week. After five weeks, performance decreased but only for weak iconicity gestures and not for strong iconicity ones. This suggests that iconicity favors longer memorization.

Studies on word and sentence list recall also provide information on the question. In these studies, participants are generally provided with a list of words or sentences that they are instructed to try to memorize. Cohen and Otterbein (1992) showed that pantomimic and non-pantomimic gestures both favored the recall of sentence lists when these did not form a narrative. Feyereisen (2006) found a similar result for representationals and non-representationals and Thompson (1995) for iconics. Igualada, Esteve-Gibert and Prieto (2017) showed that three-to five-year-old children recalled words better when these were presented with beat gestures than when they were presented alone. All these studies suggest that when manual gestures are present during encoding of the memory, it is better encoded. The same phenomenon could be at work for word learning.

4.3. Cognitive load minimization

Manual gestures may also minimize the cognitive load involved in word learning as suggested by Goldin-Meadow and colleagues (2001). This study did not directly tackle word learning. Participants first viewed a sequence of letters they were asked to memorize, they then had to solve a math problem after what they were asked to recall the sequence of letters. Those who could gesture during math problem solving recalled more letters than those who were prevented from gesturing. The authors suggest that gesturing lessened cognitive load during problem solving freeing space

for memorizing the sequence of letters. McGregor and colleagues (2009) suggest that gesture "externalized a meaningful aspect of the referent in the visual world. By making that meaning more obvious, gesture may free cognitive-linguistic resources for processing the word itself and, perhaps the other lexical and syntactic elements involved" (p. 823). Gestures could help reinforce the link between the lexical form and the concept it refers to by putting forward a distinctive property of the meaning the word refers to, depicting it and attracting the learner's attention to it. Illustrating this link could free part of the cognitive load involved in finding this property and processing this link. This could free more cognitive load to actually learn the lexical form associated with it.

5. Conclusion

This review of experimental evidence of the role of manual gestures for word learning shows that even if there is not a gestural advantage in all studies, the majority of them show that words are learned more efficiently when they are associated with a manual gesture during training in typically developing children. This effect appears to be even stronger in children with speech and language impairments. Manual gestures are more efficient than other additional cues such as pictures. Even if manual gestures appear to play a positive role for both expressive and receptive word learning, receptive learning is faster. The gestural advantage seems to be present for different types of manual gestures, it appears to be strongest for gestures bearing a physical resemblance with the referent. Some studies find a gestural advantage even immediately after training while others find this advantage only after a delay.

The evidence suggests that manual gestures play a deeper role than just attracting the learner's attention. Gestures could facilitate word learning by enhancing memorization and/or alleviating cognitive load.

Acknowledgments

This work was funded by the French National Research Agency through An@tomy2020 project (ANR-16-CE38-0011). The author would also like to thank the two reviewers who read this chapter and provided insightful comments to improve it.

References

Acredolo, L., & Goodwyn, S. (1988). Symbolic gesturing in normal infants. *Child Development*, *59*, 450–466.

Bernardis, P., Salillas, E., & Caramelli, N. (2008). Behavioural and neurophysiological evidence of semantic interaction between iconic gestures and words. *Cognitive Neuropsychology*, *25* (7–8), 1114–1128.

Bird, E.K.-R., Gaskell, A., Babineau, M.D., & Macdonald, S. (2000). Novel word acquisition in children with Down syndrome: Does modality make a difference? *Journal of Communication Disorders*, *33* (3), 241–266.

Booth, A.E., McGregor, K.K., & Rohlfing, K.J. (2008). Socio-pragmatics and attention: Contributions to gesturally guided word learning in toddlers. *Language Learning and Development*, *4* (3), 179–202.

Bornstein, H., Hamilton, L., Saulnier, K., & Roy, H. (1976). *The signed English dictionary for pre-school and elementary levels*. Washington, D.C.: Gallaudet College Press.

Brown, H., Weighall, A., Henderson, L.M., & Gaskell, M.G. (2012). Enhanced recognition and recall of new words in 7-and 12-year-olds following a period of offline consolidation. *Journal of Experimental Child Psychology*, *112* (1), 56–72.

Capirci, O., Iverson, J.M., Pizzuto, E., & Volterra, V. (1996). Gestures and words during the transition to two-word speech. *Journal of Child Language*, *23* (3), 645–673.

Capirci, O., & Volterra, V. (2008). Gesture and speech: The emergence and development of a strong and changing partnership. *Gesture*, *8* (1), 22–44.

Capone, N.C. (2007). Tapping toddlers' evolving semantic representation via gesture. *Journal of Speech, Language, and Hearing Research*, *50* (3), 732–745.

Capone Singleton, N.C. (2012). Can semantic enrichment lead to naming in a word extension task? *American Journal of Speech-Language Pathology*, *21* (4), 279–292.

Capone, N.C., & McGregor, K.K. (2005). The effect of semantic representation on toddlers' word retrieval. *Journal of Speech, Language, and Hearing Research*, *48* (6), 1468–1480.

Carey, S. (2010). Beyond fast mapping. *Language Learning and Development*, 6, 184–205.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29.

Caselli, M.C., Rinaldi, P., Stefanini, S., & Volterra, V. (2012). Early action and gesture "vocabulary" and its relation with word comprehension and production. *Child Development*, *83* (2), 526–542.

Cherdieu, M., Palombi, O., Gerber, S., Troccaz, J., & Rochet-Capellan, A. (2017). Make gestures to learn: Reproducing gestures improves the learning of anatomical knowledge more than just seeing gestures. *Frontiers in Psychology*, *8*, 1689.

Church, R.B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning? *Bilingual Education and Bilingualism*, *7* (4), 303–319.

Clark, E.V., & Estigarribia, B. (2011). Using speech and gesture to introduce new objects to young children. *Gesture*, *11* (1), 1–23.

Cohen, R.L., & Otterbein, N. (1992). The mnemonic effect of speech gestures: Pantomimic and non-pantomimic gestures compared. *European Journal of Cognitive Psychology*, *4*(2), 113–139.

Cook, S.W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesturing makes learning last. *Cognition*, *106*(2), 1047–1058.

de Nooijer, J.A., van Gog, T., Paas, F., & Zwaan, R.A. (2013). Effects of imitating gestures during encoding or during retrieval of novel verbs on children's test performance. *Acta Psychologica*, *144* (1), 173–179.

de Nooijer, J.A., van Gog, T., Paas, F., & Zwaan, R.A. (2014). Words in action: Using gestures to improve verb learning in primary school children. *Gesture*, *14* (1), 46–69.

Feyereisen, P. (2006). Further investigation on the mnemonic effect of gestures: Their meaning matters. *European Journal of Cognitive Psychology*, *18*(2), 185–205.

Gentilucci, M., & Dalla Volta, R. (2008). Spoken language and arm gestures are controlled by the same motor control system. *The Quarterly Journal of Experimental Psychology*, *61*(6), 944–957.

Giezen, M.R., Baker, A.E., & Escudero, P. (2013). Relationships between spoken word and sign processing in children with cochlear implants. *Journal of Deaf Studies and Deaf Education*, *19* (1), 107–125.

Gogate, L.J., Bahrick, L.E., & Watson, J.D. (2000). A study of multimodal motherese: The role of temporal synchrony between verbal labels and gestures. *Child Development*, *71* (4), 878–894.

Goldin-Meadow, S. (2007). Pointing sets the stage for learning language—and creating language. *Child Development*, *78* (3), 741–745.

Goldin-Meadow, S. (2011). Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2* (6), 595–607.

Goldin-Meadow, S., & Butcher, C. (2003). Pointing toward two-word speech in young children. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet*, Mahwah, NJ: Erlbaum, pp. 85–107.

Goldin-Meadow, S., Nusbaum, H., Kelly, S.D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12* (6), 516–522.

Goodwyn, S.W., & Acredolo, L.P. (1993). Symbolic gesture versus word: Is there a modality advantage for onset of symbol use? *Child Development*, *64* (3), 688–701.

Goodwyn, S.W., Acredolo, L.P., & Brown, C.A. (2000). Impact of symbolic gesturing on early language development. *Journal of Nonverbal Behavior*, *24* (2), 81–103.

Gustason, G., Pfetzing, D., & Zawolkow, E. (1975). *Signing exact English*. Silver Spring, MD: Modern Signs Press.

Horst, J.S., Parsons, K.L., & Bryan, N.M. (2011). Get the story straight: Contextual repetition promotes word learning from storybooks. *Frontiers in Psychology*, *2*, 17.

Igualada, A., Esteve-Gibert, N., & Prieto, P. (2017). Beat gestures improve word recall in 3- to 5-year-old children. *Journal of Experimental Child Psychology*, *156*, 99–112.

Imai, M., Gentner, D., & Uchida, N. (1994). Children's theories of word meaning: The role of shape similarity in early acquisition. *Cognitive Development*, *9* (1), 45–75.

Iverson, J.M., Capirci, O., & Caselli, M.C. (1994). From communication to language in two modalities. *Cognitive Development*, *9* (1), 23–43.

Iverson, J.M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science*, *16* (5), 367–371.

Iverson, J.M., & Thelen, E. (1999). Hand, mouth and brain. The dynamic emergence of speech and gesture. *Journal of Consciousness Studies*, *6* (11–12), 19–40.

Johnston, J.C., Durieux-Smith, A., & Bloom, K. (2005). Teaching gestural signs to infants to advance child development: A review of the evidence. *First Language*, *25* (2), 235–251.

Kahn, J.V. (1981). A comparison of sign and verbal language training with nonverbal retarded children. *Journal of Speech, Language, and Hearing Research*, *24*, 113–119.

Kapalková, S., Polišenská, K., & Süssová, M. (2016). The role of pictures and gestures as a support mechanism for novel word learning: A training study with 2-year-old children. *Child Language Teaching and Therapy*, *32* (1), 53–64.

Kelly, S.D., Hirata, Y., Manansala, M., & Huang, J. (2014). Exploring the role of hand gestures in learning novel phoneme contrasts and vocabulary in a second language. *Frontiers in Psychology*, *5*, 673.

Kelly, S.D., Manning, S.M., & Rodak, S. (2008). Gesture gives a hand to language and learning: Perspectives from cognitive neuroscience, developmental psychology and education. *Language and Linguistics Compass*, *2* (4), 569–588.

Kemler Nelson, D.G., Frankenfield, A., Morris, C., & Blair, E. (2000). Young children's use of functional information to categorize artifacts: Three factors that matter. *Cognition*, *77* (2), 133–168.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kohl, F.L., Karlan, G.R., & Heal, L.W. (1979). Effects of pairing manual signs with verbal cues upon the acquisition of instruction-following behaviors and the generalization to expressive language with severely handicapped students. *AAESPH Review*, *4* (3), 291–300.

Kraljević, J.K., Cepanec, M., & Šimleša, S. (2014). Gestural development and its relation to a child's early vocabulary. *Infant Behavior and Development*, *37* (2), 192–202.

Krauss, R.M., & Hadar, U. (1999). The role of speech-related arm/hand gestures in word retrieval. In R. Campbell & L. Messing (Eds.), *Gesture, speech, and sign*, Oxford: Oxford University Press, pp. 93–116.

Liuzza, M T., Setti, A., & Borghi, A.M. (2012). Kids observing other kids' hands: Visuomotor priming in children. *Consciousness and Cognition*, *21* (1), 383–392.

Lüke, C., & Ritterfeld, U. (2014). The influence of iconic and arbitrary gestures on novel word learning in children with and without SLI. *Gesture*, *14* (2), 204–225.

Macedonia, M., Bergmann, K., & Roithmayr, F. (2014). Imitation of a pedagogical agent's gestures enhances memory for words in second language. *Science Journal of Education*, *2* (5), 162–169.

Macedonia, M., & Repetto, C. (2016). Brief multisensory training enhances second language vocabulary acquisition in both high and low performers. *International Journal of Learning, Teaching and Educational Research*, *15* (3), 42–53.

Macedonia, M., & von Kriegstein, K. (2012). Gestures enhance foreign language learning. *Biolinguistics*, *6* (3–4), 393–416.

Marentette, P., & Nicoladis, E. (2011). Preschoolers' interpretations of gesture: Label or action associate? *Cognition*, *121* (3), 386–399.

Mayberry, R.I., & Nicoladis, E. (2000). Gesture reflects language development: Evidence from bilingual children. *Current Directions in Psychological Science*, *9* (6), 192–196.

McGregor, K.K. (2014). What a difference a day makes: Change in memory for newly learned word forms over 24 hours. *Journal of Speech, Language, and Hearing Research*, *57* (5), 1842–1850.

McGregor, K.K., Rohlfing, K.J., Bean, A., & Marschner, E. (2009). Gesture as a support for word learning: The case of under. *Journal of Child Language*, *36*, 807–828.

McNeill, D. (1992). *Hand and mind: what gestures reveal about thought*. Chicago: University Of Chicago Press.

McNeill, D., & Duncan, S.D. (2000). Growth points in thinking-for-speaking. In D. McNeill (Ed.), *Language and Gesture*, Cambridge: Cambridge University Press, pp. 141–161.

McNeill, D., Duncan, S.D., Cole, J., Gallagher, S., & Bertenthal, B. (2008). Growth points from the very beginning. *Interaction Studies*, *9* (1), 117–132.

Mollink, H., Hermans, D., & Knoors, H. (2008). Vocabulary training of spoken words in hard-of-hearing children. *Deafness & Education International*, *10* (2), 80–92.

Morford, M., & Goldin-Meadow, S. (1992). Comprehension and production of gesture in combination with speech in one-word speakers. *Journal of Child Language*, *19* (3), 559–580.

Mumford, K.H., & Kita, S. (2014). Children use gesture to interpret novel verb meanings. *Child Development*, *85* (3), 1181–1189.

Namy, L.L., & Waxman, S.R. (1998). Words and gestures: Infants' interpretations of different forms of symbolic reference. *Child Development*, *69* (2), 295–308.

Nicoladis, E., Mayberry, R.I., & Genesee, F. (1999). Gesture and early bilingual development. *Developmental Psychology*, *35* (2), 514–526.

O'Neill, D.K., Topolovec, J., & Stern-Cavalcante, W. (2002). Feeling sponginess: The importance of descriptive gestures in 2- and 3-year-old children's acquisition of adjectives. *Journal of Cognition and Development*, *3* (3), 243–277.

Özçalışkan, Ş., & Goldin-Meadow, S. (2005). Gesture is at the cutting edge of early language development. *Cognition*, *96* (3), B101–B113.

Özyürek, A., Willems, R.M., Kita, S., & Hagoort, P. (2007). On-line integration of semantic information from speech and gesture: Insights from event-related brain potentials. *Journal of Cognitive Neuroscience*, *19* (4), 605–616.

Rauscher, F.H., Krauss, R.M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science*, *7* (4), 226–231.

Romski, M.A., & Ruder, K.F. (1984). Effects of speech and speech and sign instruction on oral language learning and generalization of action+ object combinations by Down's syndrome children. *Journal of Speech and Hearing Disorders*, *49*, 293–302.

Rowe, M.L., & Goldin-Meadow, S. (2009). Early gesture selectively predicts later language learning. *Developmental Science*, *12* (1), 182–187.

Rowe, M.L., Özçalışkan, Ş., & Goldin-Meadow, S. (2008). Learning words by hand: Gesture's role in predicting vocabulary development. *First Language*, *2 8*(2), 182–199.

Rowe, M.L., Silverman, R.D., & Mullan, B.E. (2013). The role of pictures and gestures as nonverbal aids in preschoolers' word learning in a novel language. *Contemporary Educational Psychology*, *38* (2), 109–117.

Schunk, D.H. (1987). Peer models and children's behavioral change. *Review of Educational Research*, *57* (2), 149–174.

Smith, L.B., Jones, S.S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on- the-job training for attention. *Psychological Science*, *13* (1), 13–19.

Suanda, S.H., Walton, K.M., Broesch, T., Kolkin, L., & Namy, L.L. (2013). Why two-year-olds fail to learn gestures as object labels: Evidence from looking time and forced-choice measures. *Language Learning and Development*, *9* (1), 50–65.

Tellier, M. (2008). The effect of gestures on second language memorisation by young children. *Gesture*, *28* (2), 219–235.

Thompson, L.A. (1995). Encoding and memory for visible speech and gestures: A comparison between young and older adults. *Psychology and Aging*, *10* (2), 215–228.

Ting, J.Y., Bergeson, T.R., & Miyamoto, R.T. (2012). Effects of simultaneous speech and sign on infants' attention to spoken language. *The Laryngoscope*, *122* (12), 2808–2812.

Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, *78* (3), 705–722.

van Berkel-van Hoof, L., Hermans, D., Knoors, H., & Verhoeven, L. (2016). Benefits of augmentative signs in word learning: Evidence from children who are deaf/hard of hearing and children with specific language impairment. *Research in Developmental Disabilities*, *59*, 338–350.

Vogt, S.S., & Kauschke, C. (2017a). Observing iconic gestures enhances word learning in typically developing children and children with specific language impairment. *Journal of Child Language*, *44* (6), 1458–1484.

Vogt, S.S., & Kauschke, C. (2017b). With some help from others' hands: Iconic gesture helps semantic learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*, *60*, 3213–3225.

Pamela Fuhrmeister

Interference in memory consolidation of non-native speech sounds

Abstract: For several decades, researchers have been investigating the challenges of and constraints on learning the speech sound inventory of a second language in adulthood. Commonalities that have emerged from these findings include the immense individual variability reported in non-native speech sound learning studies and the rare attainment of native-like proficiency in perception or production of second language speech sounds. While numerous studies have shed light on various aspects of this challenging process, many questions about the extent and nature of these difficulties remain. A nascent line of research suggests that some of the difficulty in non-native speech sound learning could be attributed to various sources of interference that disrupt the memory consolidation process, thus interfering with the retention of learned phonetic information. It is well documented in the broader learning literature that interference from competing stimuli or subsequently learned skills can disrupt memory consolidation processes. However, this phenomenon has received little attention in the speech literature, and the potential sources of interference in the speech domain have yet to be identified. In this review, I discuss how integrating theories of memory consolidation with non-native speech sound learning models can more accurately capture patterns of learning observed in the non-native speech sound learning literature, specifically patterns showing failures of memory consolidation due to interference.

Keywords: sleep, second language learning, adults, memory consolidation, non-native speech sounds

1. Introduction

Adult second language learners face many challenges, especially when attempting to master the speech sounds of a non-native language. Although a second language learner must gain proficiency in a number of linguistic domains (e.g., morphology, syntax, semantics), acquiring perceptual sensitivity to the speech sounds of another language is an important step in language acquisition. Indeed, several studies support the notion that speech perception abilities can facilitate higher levels of language learning

(e.g., lexical acquisition). For example, native speech perceptual abilities in infancy have been shown to predict language development in early childhood (Tsao et al., 2004; Kuhl et al., 2008), and in adulthood, superior perceptual discrimination of non-native speech contrasts can facilitate learning of lexical items that contain those contrasts (Silbert et al., 2015). Thus, mastery of the speech sound inventory of a language may be a crucial step in language acquisition more generally. Unfortunately, however, most studies of second language learners report rather poor perceptual abilities for difficult phonetic contrasts (e.g., Bradlow et al., 1999; Flege, 2003), even when learning commenced in childhood (Pallier et al., 1997). As such, several types of training paradigms have been devised in an attempt to optimize non-native speech sound learning, specifically for difficult speech sound contrasts. Paradigms that have been employed in training studies differ in terms of whether learning takes place in an implicit (Lim & Holt, 2011; Vlahou et al., 2012; Wade & Holt, 2005) or explicit manner (e.g., Earle et al., 2017; Earle & Myers 2015a, 2015b), or whether participants were trained on tokens with limited variability or high variability (e.g., tokens produced by multiple talkers or occurring in multiple phonological contexts, Logan et al., 1991; Lively et al., 1993, 1994; Bradlow et al., 1997, 1999). Other paradigms utilized either natural speech or exaggerated versions of speech to make differences between stimuli more salient (McCandliss et al., 2002; Golestani & Zatorre, 2004; Swan & Myers, 2013). Although each of these training paradigms has indeed demonstrated learning, the majority of these studies report only moderate success, and most second language learners ultimately fail to attain native-like perception or production of non-native speech sounds (e.g., Bradlow et al., 1999; Piske et al., 2001; MacKay et al., 2001).

Most studies probing plasticity in the speech system focus on the initial processes of *learning* speech sounds; however, language acquisition involves more than the initial encoding of stimuli. In order for an individual to develop and maintain language proficiency, critical aspects of a language, such as phonemes and lexical items, need to be consolidated into long-term memory for later retrieval. Stable long-term memory representations may then facilitate retention and generalization of learned speech sounds. Until recently, the role of memory in speech sound acquisition had been underexplored, and therefore, many questions about the

memory functions underlying this process remain. Due to the general lack of success in non-native speech sound learning, it is becoming more apparent that consistent failures of memory consolidation may underlie the challenge of learning non-native speech sounds, and this merits further exploration. More specifically, training conditions that facilitate stronger initial learning of speech sounds and limit exposure to interference after training and before consolidation takes place may be key factors in promoting long-term changes in the speech system. The current chapter begins with a review of models of non-native speech sound learning, followed by a discussion of memory consolidation theories and experimental evidence for those. In light of this evidence, I conclude with an alternative interpretation of some work on non-native speech sound learning and discuss how considering strength of learning and interference in memory consolidation may have explanatory power for some of the challenges reported in these studies.

2. Constraints on non-native speech sound learning: Perceptual similarity of first language speech sounds

By the first few months of life, infants demonstrate perceptual sensitivity to many different speech sounds that are found in world languages (Eimas et al., 1971). However, by the end of the first year of life, infants lose the ability to discriminate certain phonetic contrasts that do not occur in the ambient language (e.g., Werker & Tees, 1984; Kuhl et al., 2006). While this warping of perceptual space appears to facilitate language acquisition in early childhood (Tsao et al., 2004; Kuhl et al., 2008), the process of perceptual reorganization can greatly constrain the acquisition of non-native speech categories later in life. In particular, perceptual similarity of native and non-native speech sounds can cause non-native speech sounds to be more difficult to perceive as distinct categories (e.g., Best et al., 2001). For example, native English speakers often struggle to perceptually distinguish dental and retroflex voiced stop consonants found in Hindi. The perceptual similarity of these sounds to the English alveolar /d/ category and their similarity to each other make these exceptionally challenging for learners to disambiguate. Indeed, several models of non-native speech sound learning account for such difficulties by considering the relationship

between native and non-native speech sounds. For example, the perceptual assimilation model (e.g., Best, 1994; Best & Tyler, 2007) and the native language magnet model (Kuhl, 1994; Kuhl et al., 2008) both posit that non-native speech sounds that are perceptually similar to native phonemes will be harder to perceive than perceptually dissimilar sounds. However, the two models differ on which dimensions of the speech signal are considered important for perception. The perceptual assimilation model focuses on naive listeners' perception of non-native speech sounds as a result of articulatory similarity between native and non-native speech sounds (Best, 1994; Best & Tyler, 2007). Specifically, this model predicts that naive listeners will assimilate unfamiliar non-native speech sounds to the perceptual category in the native language that is produced by the most similar articulatory gesture (Best & Tyler, 2007). For example, native English speakers often map dental and retroflex voiced stop consonants found in Hindi onto the alveolar /d/ sound found in English. As a result, certain speech sounds are more difficult to perceive, while speech sounds without a similar native language category are perceived more easily (Best et al., 2001). For instance, English contains no speech categories similar to click sounds found in Zulu, and native English speakers typically discriminate these sounds accurately (Best et al., 1988). In contrast to articulatory gestures, the critical dimension of the native language magnet model is acoustic space. This model concentrates on the developmental processes underlying the acquisition of native speech sounds (Kuhl, 1994; Kuhl et al., 2008) and postulates that infants take advantage of statistical learning in order to acquire the speech sound categories of their native language. After sufficient exposure, infants' perceptual space becomes "warped," and prototypes for native language speech categories emerge. These prototypes act as magnets that attract perceptually similar speech sounds. Analogous to the perceptual assimilation model, this magnet effect would result in some non-native speech sounds (i.e., perceptually similar sounds) being more difficult to learn than others (perceptually dissimilar sounds). While similar in some ways to the perceptual assimilation model and the native language magnet model, Flege's (1995) speech learning model takes a slightly different approach. First, this model focuses on experienced adult second language learners and proposes that difficulties in second language speech production stem from perceptual obstacles. In

other words, one cannot produce what one cannot perceive. An additional component of this model predicts that non-native speech sound learning becomes more difficult over the lifespan; however, it deviates from traditional critical period hypotheses, as it predicts a gradual decrease in non-native speech production abilities over the lifespan, rather than an abrupt decline after puberty. While this model emphasizes adult second language learners' speech production abilities, it is similar to the models described above in that it attributes difficulties with the second language sound system to similarities with the native language.

Attention to dimension models add to these models by clarifying the processes required for acquiring non-native speech categories. Attention to dimension models propose that a speaker of a language has learned to direct attention to relevant parts of the acoustic signal for his or her native language and that learning new speech categories requires a learner to attend to previously unattended dimensions of the signal (Francis & Nusbaum, 2002). Specifically, native speakers of a language have learned to direct attention to relevant acoustic cues in the speech signal, and they have simultaneously learned to ignore other cues. Learning to reweight acoustic cues or to direct attention to different parts of the speech signal presents a challenge for learners. This model differs from the others described above in that it explains how a listener's perceptual space changes as a result of experience with the native language (i.e., via selective attention to meaningful features).

While each of these models focuses on distinct phases of the acquisition process (e.g., infants, naive listeners, or experienced second language learners) and accounts for challenges in different ways (e.g., development, articulatory representations, production, or dimensions of the acoustic signal), some similarities emerge from a close comparison of them. For example, these models do not necessarily attribute poor second language perception and production abilities to a critical period or a loss of neural plasticity; rather, difficulties are attributed to prior experience with the first language and how that experience shapes perception and production of new speech categories. Clearly, the native language exerts constraints on perception and production of second language speech sounds, and these constraints are especially strong when speech sounds in the native and non-native languages are close in perceptual space. These models certainly

capture a great deal of the non-native speech sound learning process, especially the relative difficulty of different speech sounds as a result of native language background. However, many non-native speech training studies observe considerable variability even among individuals of the same language background (e.g., Golestani & Zatorre, 2004, 2009; Myers & Swan, 2012; Yi et al., 2014), which suggests that factors beyond the native language constrain this process.

3. Memory consolidation in speech and language learning

To account for individual variability in speech sound learning, it may be helpful to consider how models of memory consolidation predict retention, forgetting, or elaboration of learned perceptual information. Most models of memory consolidation posit that two separate memory systems work in tandem to consolidate different types of learning. Specifically, the fast, hippocampus-mediated and often sleep-dependent system serves to consolidate declarative or explicit memories in order to integrate them into existing networks of knowledge, while the slow, hippocampus-independent system associated with implicit or non-declarative learning induces local changes to neuronal circuitry as a result of continued experience and does not typically rely on sleep (e.g., complementary learning systems, McClelland et al., 1995; McClelland, 1998; see also Marshall & Born, 2007; Dudai, 2004). With extensive exposure over longer periods of time, the slow, procedural learning system is able to discover regularities in the input from the environment. As a result, this memory system is able to sort information into categories, and this knowledge about categories can aid generalization to new contexts or exemplars (see McClelland et al., 1995 for discussion). In connectionist models, rapid sequential learning (i.e., learning one task immediately after another), leads to what is called catastrophic interference (McCloskey & Cohen, 1989). In other words, networks must completely forget or overwrite the information they have learned in order to accommodate new information. According to McClelland (1998), catastrophic interference underscores the need for the rapid, hippocampus-mediated consolidation system: the hippocampus acts as a temporary memory store and allows for memory traces to be selectively consolidated into long-term memory and integrated into existing

networks of information. Crucially, this process of selective consolidation via the hippocampus does not result in information being overwritten.

A rich literature on memory consolidation suggests that sleep can improve performance on a task even in the absence of further practice through a period of off-line consolidation (Stickgold et al., 2000; see Marshall & Born, 2007 for review). Sleep has been shown to facilitate abstraction of information from episodic traces, and this integration of information into existing knowledge networks allows learning to generalize to novel contexts (Davis et al., 2009). Sleep-mediated consolidation has also been shown to help protect newly learned information from interference (e.g., Ellenbogen et al., 2006, 2009; Drosopoulos et al., 2007) or recover learning that decayed throughout the course of a day (Fenn et al., 2003). Several studies have additionally investigated the contributions of sleep-dependent memory consolidation to word learning (e.g., Tamminen & Gaskell, 2013; Dumay & Gaskell, 2007; Davis et al., 2009). Consistent with McClelland's complementary learning systems account, a study by Davis et al. (2009) found that newly learned words become integrated into the existing lexicon after a period of sleep. In this study, participants learned novel words that overlapped with existing lexical items by several phonemes (e.g., *cathedruke-cathedral*), and these novel words only showed evidence of lexical competition with existing words (e.g., *cathedral*) on a lexical decision task after a period of sleep had occurred. Additionally, brain activation measured by functional magnetic resonance imaging (fMRI) in this study revealed activation of the hippocampus in response to the novel lexical items before sleep and cortical activation after sleep. This suggests that the newly learned lexical items were temporarily stored in the hippocampus and became re-represented in cortical areas after sleep, which is in line with predictions from McClelland's complementary learning systems account. In addition to offline gains in the absence of further practice, sleep between two periods of practice in word learning can benefit long-term retention of novel words (Mazza et al., 2016), and even daytime naps or short periods of sleep can improve or stabilize learning of tasks, including word learning (Lahl et al., 2008; Heim et al., 2017). The extant literature shows clear benefits of sleep for word learning; however, of interest to the current chapter is whether similar findings are also observed in speech learning.

Indeed, some studies have found benefits of sleep for speech or auditory learning tasks (see Earle & Myers, 2014 for review). For example, Fenn et al. (2003) carried out an experiment in which participants learned to understand computer-synthesized versions of native language words. In this study, participants who slept between training and test improved in the absence of further practice. Moreover, a degradation of performance was observed for participants who were trained in the morning, but this loss was recovered following sleep. Thus, sleep may help recover performance that has degraded throughout the course of a day, as well as facilitate generalization to novel contexts. In fact, a later study by the same group suggests that sleep promoted generalization of information learned with a large amount of variability, but not when learning took place with a closed set of tokens (Fenn et al., 2013). In similar fashion, Xie et al. (2017) tested listeners on generalization of learning of one Mandarin-accented talker to a novel Mandarin-accented talker. Like the Fenn et al. (2003) study, Xie and colleagues (2017) found that performance on the untrained talker degraded throughout the day for a group trained in the morning. Thus, sleep was necessary for generalization to a novel talker and was not simply observed with the passage of time. Although these studies have found performance gains following sleep, other work on sleep in auditory learning and perceptual learning of speech have not observed any added benefits of sleep. For example, Roth and colleagues (2005) found that sleep was not necessary for improvement on a speech in noise task, but rather, the passage of time was sufficient to induce performance gains. Similarly, Eisner and McQueen (2006) found no additional benefit of sleep for perceptual learning of speech. In a lexically-guided perceptual learning task, participants were exposed to lexical items containing non-canonical productions (an ambiguous fricative between /f/ and /s/) of certain speech sounds embedded in a lexical context that served to disambiguate the speech sound. After exposure, participants were asked to categorize tokens on a non-word continuum from /f/-/s/. In this paradigm, a shift in the category boundary, (i.e., categorization of more ambiguous tokens in a non-word context consistent with the lexical bias in the exposure condition) indicates perceptual learning. Indeed, participants in this study showed a category boundary shift consistent with their exposure condition both immediately after training and 12 hours later. Interestingly,

a group that slept during the 12-hour interval showed no greater learning effect than a group that remained awake during the day throughout the post-training interval.

Several differences could explain the discrepancies in the findings of these studies. First, different learning systems may underlie the various tasks employed in these studies. For example, the Eisner and McQueen (2006) study measured adaptation to episodic representations of talker-specific idiosyncrasies in speech production (that is, how that talker produced a particular speech sound in a non-standard way). Recent evidence suggests that adaptation to talker-specific, episodic information (i.e., details of a specific talker's voice) occurs rapidly and is stable over time, while tasks involving more abstract representations may emerge only after consolidation (Brown & Gaskell, 2014). This could explain why both groups in the study by Eisner and McQueen (2006) remained stable over time, regardless of sleep. The fact that participants in the study by Fenn and colleagues (2003) were trained and tested on completely different words may explain why these participants benefitted from sleep. Abstract representations facilitate generalization to novel contexts (in this case, generalizing knowledge of synthetic speech sounds to new lexical contexts), and because abstract representations manifest only after offline consolidation, improvement after sleep in this context is not surprising. However, studies by Earle and Myers (2015a) and Earle et al. (2017) have found overnight improvement on tasks even when no generalization was needed (i.e., the same tokens were used in training and testing).

Sleep may be especially advantageous when learning involves the formation of new representations, rather than adapting existing representations to accommodate atypical exemplars. For instance, several recent studies have found benefits of sleep in non-native speech sound learning. Earle and Myers (2015b) trained participants to learn the Hindi dental/retroflex contrast on a closed set of tokens but found that sleep enabled generalization to stimuli produced by novel talkers. Earle et al. (2017) even found that duration of sleep predicted overnight improvement on the tasks used to assess non-native speech sound learning. Specifically, they found that the amount of slow wave sleep predicted overnight gains on identification of the speech sounds, while total sleep duration predicted participants' ability to discriminate the speech sounds. Importantly, this study suggests

that individual differences in sleep duration may account for some of the individual variability typically observed in training studies. On the other hand, an additional study by this group found some surprising limits to the benefits of sleep consolidation (Earle & Myers, 2015a). This study likewise trained participants on the Hindi dental/retroflex contrast and observed improvement after sleep, but this advantage only held if participants had been trained in the evening. This study consisted of two groups of participants who were trained on the contrast. One group was trained in the morning hours and one in the evening hours. Each group returned approximately 12 and 24 hours later for reassessment. Surprisingly, only the evening-trained participants showed improvement following an overnight interval. The authors reasoned that this discrepancy could be a result of interference from native language exposure. Specifically, participants trained in the morning had a day's worth of input from their native language prior to sleep, while the evening-trained group presumably had much less. Subsequent experiments in this study indicated that the lack of overnight improvement seen for the morning trained group stemmed from exposure to perceptually similar native language speech sounds.

Other recent studies have similarly found that exposure to certain stimuli or engagement in certain tasks can interfere with learning or consolidation of newly acquired skills or representations. For example, alternating perceptual training with speech production practice has been shown to attenuate or interfere with learning. In a study by Baese-Berk and Samuel (2016), native Spanish-speaking participants were trained to learn a difficult, non-native Basque speech contrast in the laboratory. Training consisted of an ABX discrimination task in which participants heard three sounds and were asked to indicate whether the third was more similar to the first or second sound presented. For this study, some participants were asked to repeat the third sound presented on each trial out loud before indicating their decision. Surprisingly, the group that repeated the trained sound showed very little learning of the contrast on a perceptual post-training assessment. Furthermore, production of any sounds, even sounds that were dissimilar to the trained contrast, seemed to attenuate learning. This suggests that the cause of this interference was not solely a result of exposure to the participants' own poor productions of the non-native contrast. Rather, it raises questions about the mechanisms underlying

speech perception and production in the development of speech sound representations.

Exposure to phonological variability before or after training may additionally attenuate learning and disrupt consolidation. For example, a study by Fuhrmeister and Myers (2017) examined whether native English-speaking participants trained on a non-native, Hindi dental/retroflex contrast would benefit from additional exposure to the contrast in a different vowel context. In this study, one group of participants heard the contrast in only one vowel context in minimal pair non-words (/d̪ug/ and /ɖug/) throughout training and testing. Another group heard the contrast in two different vowel contexts in assessments (/d̪ug/ and /ɖug/ vs. /d̪ig/ and /ɖig/; assessments consisted of a pretest, immediate posttest, and a delayed posttest), but they had identical training to the other group (i.e., /d̪ug/ and /ɖug/ only). Notably, the participants who were exposed to the Hindi sounds in two vowel contexts performed significantly worse than the group exposed to one vowel context only on the tasks involving the contrast in the trained vowel context, despite having more total exposure to the contrast. Additionally, participants exposed to the contrast in two vowel contexts showed no evidence of overnight improvement on the stimuli in the trained vowel context, while those who heard the sounds in only one vowel context did improve after an overnight interval. These findings suggest that exposure to novel speech sounds in different vowel contexts may interfere with learning or consolidation of the contrast in a trained vowel context, even if that extra exposure is limited (i.e., at test only). It is also possible that the learning of the trained vowel context was less stable for participants exposed to two different vowel contexts, which may have prevented further improvement as a result of sleep. As can be seen, memory consolidation influences speech learning in the following ways:

- Sleep helps consolidate newly formed representations of both natural and synthetic speech sounds as indicated by performance improvement (Earle & Myers, 2015a; Earle et al., 2017) and generalization (Earle & Myers, 2015b; Fenn et al., 2003, 2013).
- Not all types of perceptual speech learning tasks show improvement after sleep (Roth et al., 2005; Eisner & McQueen, 2006), and this may depend on task difficulty, whether new representations are being

formed, or whether existing representations are being expanded to accommodate new exemplars.
- Exposure to certain stimuli (e.g., perceptually similar native language speech sounds) following training may interfere with learning of novel speech sounds (Earle & Myers, 2015a).
- Training conditions (e.g., exposure to phonological variability, Fuhrmeister & Myers, 2017; production of speech sounds or words, Baese-Berk & Samuel, 2016) may destabilize or attenuate learning, which may affect the consolidation process.

Although sleep appears to facilitate memory consolidation of speech in a variety of ways, many questions remain. For example, it is unclear what types of stimuli might interfere with non-native speech sound learning or under what conditions interference effects could be avoided. However, it may be possible to carry over insights from other domains in order to inform these questions and make predictions about speech learning.

4. Failures of consolidation: Interference effects in learning

In order to fully understand how new memories are formed, it is important to examine cases in which consolidation fails. Over a century ago, Müller and Pilzecker (1900) proposed that memories exist in an initially labile state, in which they are subject to interference from subsequently learned tasks. In their studies, they tested explicit recall of strings of unrelated digits and observed that their participants were not able to recall one list as well when tested 24 hours later if they had learned a subsequent list immediately following practice on the first list. Walker's (2005) model for procedural memory consolidation similarly assumes that newly acquired memory traces are fragile and must undergo a process of stabilization before becoming resistant to interference. This model is comprised of three main stages: acquisition, consolidation-based stabilization, and consolidation-based enhancement. Walker (2005) argues that the stabilization stage depends on the passage of time only, while the enhancement stage relies on sleep. Two dissociable systems have been proposed in the memory literature, which presents a challenge for extending theories of memory consolidation to other domains. The declarative memory system underlies explicit learning of facts or episodes (sometimes referred to as the

memory for "what"), while the procedural memory system serves memory of implicitly acquired actions or procedures (the memory of "how") (e.g., Squire, 2004). Although Walker's model was originally intended for procedural memory and Müller and Pilzecker's account for declarative memory (though their account predates this term), some findings suggest procedural and declarative memory systems may not be as dissociable as once thought (Poldrack et al., 2001). In addition, studies including both declarative and non-declarative tasks lend support to the notion that the consolidation of a newly acquired skill or memory can be disrupted if an interfering task or stimulus is introduced before the memory has stabilized. Furthermore, it remains unclear whether speech category formation can be neatly classified as either procedural or declarative learning, and this process may be different at different points throughout the lifespan or under different learning conditions. Therefore, for the remainder of this review, I will draw on the literature of learning and memory processes for both declarative and procedural tasks and will reflect on the importance of a stabilization period following initial encoding in order to help new memory traces become resistant to interference. This section reviews a series of studies that have examined interference effects in several domains of learning, including the time course required for stabilization and consolidation of memory traces and the strength of initial learning or encoding. The goal of drawing on this literature is to make predictions about how speech sound learning may be facilitated by mitigating interference or adhering to a training paradigm or schedule that is more conducive to consolidation and long-term retention.

A seminal study in the motor learning domain demonstrated interference with a behavioural task, in which participants learned to move a two-hinged handle to a target while compensating for perturbation (Brashers-Krug et al., 1996). Participants who learned to compensate for perturbation experienced a disruption of consolidation if, immediately following training, they were trained to compensate for perturbation in the opposite direction. Another group of participants completed identical training on the first task, but their second task consisted of moving the handle to a target in the absence of any perturbation. These participants showed no interference effect from the second task. An additional group completed the two training sessions with perturbation in opposing

directions but waited four hours between the two training episodes. This group also showed performance improvement on the first task, indicating that a period of four hours was sufficient to stabilize learning of the first task, making it immune to disruption from a second task. Similarly, a study by Walker et al. (2003) demonstrated interference effects in a finger tapping sequence task unless six hours had passed between the two training sessions. In addition to behavioural tasks, the application of transcranial magnetic stimulation (TMS) as a source of interference has been tested (Muellbacher et al., 2002). In this motor learning study, participants practiced a finger movement sequence and were assessed on their improvement in speed. When TMS was administered immediately following training, participants showed no retention of the behavioural gains observed during the learning phase. However, if a period of six hours had lapsed before TMS was applied, no interference was observed. These studies support the consolidation hypothesis and Walker's consolidation model by demonstrating that the passage of time is necessary to stabilize newly encoded motor memories. Furthermore, the type of task that follows learning may dictate whether learning on the first task is disrupted. These findings may be able to make important predictions in the speech domain. For example, if learners train on non-native speech sounds and are exposed to speech sounds in their native language before the stabilization period has concluded, consolidation of the non-native sounds may be obstructed. Similarly, it may also be the case in the speech domain that not all stimuli or tasks interfere equally. If that is indeed the case, it will be important to identify which types of stimuli or tasks (e.g., native language exposure, Earle & Myers, 2015a; speech production, Baese-Berk & Samuel, 2016) are able to interfere with speech sound learning.

Similar task and timing effects to those found in motor learning have been observed in visual perceptual learning tasks. For example, in a visual hyperacuity task, participants saw two presentations of three dots arranged vertically on a screen and were asked to indicate whether the middle dot was offset in the either the first or second group of dots presented (Seitz et al., 2005). Following training, participants completed training on another task: in one task, the presentation of the dots was the same except the offset was presented in the opposite direction, and other tasks varied the spatial location and the orientation of the dots (i.e., the

dots were presented horizontally). Crucially, only the participants who were trained with the opposite offset direction experienced interference. Visual perceptual learning of stimuli that were presented at different spatial locations or in different orientations did not interfere with initial encoding of the task, which the authors attributed to the retinotopic specificity of spatial location and orientation. Another critical finding in this paper was that participants who waited one hour before training on the opposite direction did not demonstrate any attenuation of learning on the first direction. An additional visual perceptual learning study using a line orientation detection task found a period of 3.5 hours to be sufficient to eliminate retrograde interference from a second visual task (Shibata et al., 2017). These findings from visual learning studies provide further support that fragile memory traces remain susceptible to interference until a period of stabilization has passed. Like the motor learning study by Brashers-Krug et al. (1996), the study by Seitz et al. (2005) shows that not all tasks have the potential to interfere with consolidation of a previously learned task. Evidence from these two domains, namely vision and motor learning, suggests that domain-general processes may underlie consolidation of learning and may therefore be applicable to the speech domain.

In further support of the consolidation hypothesis, one study investigating interference from consecutive tasks in patients with amnesia found surprisingly similar stabilization effects, despite the fact that declarative memory consolidation deficits are a hallmark of amnesia. Dewar et al. (2009) had individuals with amnesia learn word lists, and these participants showed a graded advantage in recall after the presentation of interfering stimuli at different time points. Participants experienced a delay between the initial learning session and the presentation of interfering stimuli, and longer delays facilitated recall of the original word lists more effectively than shorter delays. This suggests that even individuals with amnesia who have declarative memory consolidation deficits can benefit from a stabilization period following learning.

If non-native speech sound learning processes parallel those of visual, motor, and word learning, the stabilization phase prior to consolidation may be crucial to learning and retention of non-native speech sounds. If the stabilization phase is disrupted from exposure to conflicting stimuli or

practice on an interfering task, this may hinder consolidation and retention of novel speech sounds.

Although many studies have found robust support for the consolidation hypothesis, results from other studies challenge its reliability. For example, Goedert and Willingham (2002) trained participants on two implicit motor tasks with the goal of testing whether these memories undergo consolidation and become resistant to interference from learning a similar task. The researchers first utilized a serial reaction time task in this study, in which participants saw a sequence of circles appear in boxes on a screen and pressed a button corresponding to each box after a circle was presented. In this paradigm, participants are unaware that the sequence is not random but consists of an underlying pattern; therefore, learning is implicit. The second task used in this study was a task in which participants learned a new visuomotor mapping. Participants were instructed to point at a target on a screen while wearing prism glasses that displaced their vision. Training for each task followed a traditional interference paradigm (train on task A, train on task B, test on task A), and participants were trained on different sequences and visual displacements for task B at varying intervals following training on task A. Unlike several previous studies, this study did not find evidence that the motor memories had been consolidated and become resistant to interference, as even 24 hours was not sufficient to protect against interference from task B. A study employing similar visuomotor tasks by Caithness et al. (2004) additionally found that memories for one task were susceptible to interference from another task even 24 hours later. Walker et al. (2003) demonstrated similar effects in a finger tapping task. Interestingly, participants who trained on task A, waited 24 hours, and performed task A again before learning task B did not retain their learning of task A. Walker and colleagues (2003) posited that reactivation of consolidated memories can shift them into labile states, causing them to become susceptible to interference once again. Caithness et al. (2004) speculated that performance on task B in their study may have been sufficient to reactivate memory traces of task A, which allowed task B to interfere with task A. Goedert and Willingham (2002) largely attributed this ostensible lack of consolidation to task differences or neural structures underlying the specific task used in their study.

An open question is what these memories require in order to be transferred into a stable state and resist subsequent interference. It is possible that certain tasks or types of learning undergo a different consolidation process than others or are not consolidated at all. It appears that non-native speech sound learning can indeed undergo consolidation as evidenced by improvement in the absence of further practice (Earle & Myers, 2015a, 2015b; Earle et al., 2017), and it may be the case that a stabilization or consolidation period would protect newly formed phonetic category representations from interference.

5. Stability and strength of learning

The studies reviewed in the last section, which encompass both procedural and declarative tasks, provide important evidence for consolidation theories proposed by Müller and Pilzecker (1900) and Walker (2005): most newly acquired memory traces need to undergo a period of stabilization in order to become resistant to interference. In some cases, the presence of interfering stimuli during the stabilization phase may be strong enough to disrupt the consolidation process entirely. An important question to address in speech learning studies will be how to minimize interference during the stabilization phase or to identify training conditions in which information may be consolidated in spite of interference.

In addition to the passage of time during a stabilization phase, strength and stability of initial learning may be an important factor in determining whether information is consolidated. Ebbinghaus (1885) first proposed that increasing the repetition of practice trials in a task may lead to better retention of the information 24 hours later. In addition to the early findings by Ebbinghaus, several recent studies in the visual domain lend support to this idea. For example, a study by Hauptmann et al. (2005) found that participants who practiced a visual task until performance reached asymptote improved after a period of sleep, while those who did not practice to this criterion failed to show overnight improvement. Tucker and Fishbein (2008) trained participants on a series of declarative memory tasks and had some take a nap following training. Interestingly, only the high-performing participants in training benefitted from sleep, suggesting that stronger learning can facilitate overnight improvement. In a study by

Shibata and colleagues (2017), participants who overlearned (continued to practice after the point of mastery) a visual perceptual task did not experience interference from a second task, suggesting that hyper-stable learning can accelerate or even obviate the need for a stabilization phase following learning. Taken together, these studies suggest that benefits of consolidation may depend on how strongly information is initially learned.

Conversely, a few studies have found that sleep preferentially enhances recall of weakly learned information or performance on more difficult tasks. For example, Drosopoulos and colleagues (2007) had participants memorize word pairings and manipulated how strongly the pairings were learned. Of the participants who only weakly learned the pairings, participants who slept forgot significantly fewer word pairings when tested two days later as compared to a wake group. However, sleep and wake groups were comparable if the information was strongly learned during training. Although sleep did seem to benefit the group that did not learn the information as rigorously to begin with, these findings also support the benefits of strong initial encoding. Even though the wake group did not sleep immediately after learning, they performed equivalently to the sleep group. In addition, it seems difficult to rule out ceiling effects in this study, as the participants in the strong encoding group performed at over 95 % accuracy. Similarly, a study using a procedural motor learning task found superior benefits of sleep for the most difficult task during training, as measured by an increase in speed (Kuriyama et al., 2004). Analogous to the Drosopoulos et al. (2007) paper, the participants who learned the easier tasks still outperformed the group that learned the more difficult task, although the benefits from sleep were not as drastic.

As shown above, the mixed evidence presented here implies a complicated relationship between strength of learning and consolidation. Some studies suggest that strongly encoded information is advantageous for consolidation, while others show stronger sleep-related benefits for weakly learned information. Critically, in the studies showing benefits of sleep consolidation for weakly learned information, participants who trained on easier tasks or trained on the same tasks to a higher criterion demonstrated superior overall performance, which should be considered along with the superior benefits of sleep for weakly learned information. Additionally, ceiling effects arguably cannot be completely ruled out in these studies. It

is also possible that the benefits of sleep are observed in a u-shaped trajectory because the qualitative changes associated with sleep-mediated consolidation do not always manifest behaviourally. For example, newly learned information may need to reach a minimum level of stability in order to trigger consolidation, and sleep may show the strongest influence on these memories as far as behavioural changes can be observed. However, sleep has been shown to induce qualitative changes to memories, such as the ability to generalize to new contexts (e.g., Fenn et al., 2013), increased automaticity of a task as measured by electrophysiological components (Atienza et al., 2004), and differential functional activation patterns in response to stimuli after sleep (Davis et al., 2009). For example, the study by Atienza et al. (2004) trained participants to discriminate auditory tone patterns. Some participants slept after training, while another group was sleep deprived. They found improved behavioural performance for both groups, regardless of sleep; however, an electrophysiological component that responds to the involuntary switching of attention was elicited only in the participants who slept after training. Using fMRI, Davis et al. (2009) found changes in functional brain activation following sleep in participants who learned new words. Specifically, they found activation in the hippocampus before sleep consolidation, but after sleep, activation was observed in cortical areas. This indicates that the memory traces of the new words underwent qualitative changes in how they were represented in the brain. Therefore, sleep may indeed benefit learning or qualitatively reorganize information, even if these changes are not always evident in behavioural performance. All things considered, the benefits of strong initial learning seem clear: stronger encoding or overlearning typically results in better overall behavioural performance, and it can protect against interference from subsequent learning and potentially bestow benefits equivalent to those of sleep for long-term retention. This may be an important consideration for learning situations in which sleep following training is not possible.

6. Elucidating findings from non-native speech sound learning studies in the context of interference and stability

Concepts such as stability of learning and interference in memory consolidation may offer a more comprehensive account of some findings

from non-native speech sound learning. That fact that parallels emerge from several domains of learning (e.g., visual, motor, and word learning) may indicate that domain-general encoding, stabilization, and consolidation processes underlie many different types of learning, including speech sound learning. In fact, several studies reviewed above can be viewed through the lens of interference theories. For example, the finding by Earle and Myers (2015a), that native language exposure interfered with consolidation of a non-native, Hindi contrast, could be explained both by non-native speech sound learning theories and interference theories; however, a more comprehensive explanation could be arrived at by considering these theories together. Although theories of non-native speech sound learning (such as those reviewed above) differ on certain details and areas of focus, most attribute difficulties in non-native speech sound learning to perceptual similarity of native language speech sounds. In line with these theories, the stability and robustness of native-language phonetic categories may greatly enhance the difficulty of learning perceptually similar non-native categories. Additionally, both Müller and Pilzecker's (1900) consolidation hypothesis and Walker's (2005) procedural memory consolidation model postulate a necessary stabilization phase after learning takes place. If native language exposure immediately follows training on non-native speech sounds before the new speech category representations have had time to stabilize, native language input would interfere with these memory traces and impede or prevent consolidation from taking place. Results from Earle & Myers (2015a), in which native language exposure disrupted consolidation of a non-native phonetic contrast, diverge from the synthetic speech study by Fenn and colleagues (2003), in which sleep was able to recover information that was degraded (or possibly interfered with) throughout the course of a day. However, Ebbinghaus (1885) and other studies reviewed above support the view that strength and stability of learning is crucial to consolidation, and this notion may account for the discrepancy observed in these studies. Learning novel acoustic mappings to existing speech categories (that is, learning how the unusual synthetic speech signal maps to well-developed English phonology), as was done in the Fenn et al. (2003) paper, is arguably less difficult than establishing entirely new

perceptual categories. It is reasonable to speculate that both the time course of learning and consolidation and the strength of initial learning work in tandem to selectively consolidate memories. Learning may have been more stable for the synthetic speech task in Fenn et al. (2003) than the non-native speech sounds learned in Earle and Myers (2015a), which would explain why sleep-mediated consolidation was able to recover learning of synthetic speech that had decayed throughout the day but not the developing representations of non-native speech sounds.

Strength and stability of learning may further elucidate studies finding disruptions of learning or consolidation as a result of speech production (Baese-Berk & Samuel, 2016) or phonological variability (Fuhrmeister & Myers, 2017). Neither of these studies was designed according to the typical interference paradigm (learn task A, learn task B, test on task A); however, it appears that speech production and exposure to phonological variability resulted in representations that were less stable and less able to benefit from consolidation. Ultimately, the precise cause for attenuated perceptual learning following speech production remains unclear. Motor theories of speech perception posit that articulatory gestures underlie perceptual representations of speech (see Galantucci et al., 2006 for review). According to this view, it is possible that activating motor representations interferes with developing representations of speech categories. An additional possibility is that engaging native language phonological categories in any modality diminishes the strength and stability with which the novel categories are learned. Exposure to phonological variability may similarly reduce stability of learning: according to attention to dimension models of non-native speech sound learning, learners must direct their attention to relevant acoustic cues, which are, in many cases, different from the relevant cues for the first language (Francis & Nusbaum, 2002). Presentation of novel speech sounds in different phonological contexts may not allow the learner to quickly discover the acoustic cues that are necessary to distinguish different speech categories, as formant transitions sometimes change based on the vowel that follows a consonant. If the learner receives conflicting information for different phonological contexts, this would likely result in learning that is less stable, which may not benefit from consolidation, at least in the short term.

7. Promoting consolidation of non-native speech sounds

With models of interference and stability in memory consolidation in mind, it may be possible to improve training programs in order to support consolidation of novel phonetic categories. First, it is necessary to determine what types of stimuli can interfere with or destabilize this process. Specific tasks or stimuli that have the potential to interfere with non-native speech sound learning have not been extensively investigated. Nevertheless, by examining the extant literature, it appears that native language phonology is one source of interference (Earle & Myers, 2015a). Future perceptual training paradigms may induce more robust learning if they attempt to minimize exposure to native language phonology until a sufficient stabilization period has passed, especially if training takes place earlier in the day. If it is not possible to minimize native language exposure, learners may benefit from longer or more intensive training, in order to strengthen or stabilize learning. As seen in the study by Shibata and colleagues (2017), hyper-stable learning was resistant to subsequent interference, and non-native speech sound learning may show a similar pattern. Although some similarities between speech sound learning and visual or motor learning exist, this may be an area where speech diverges from other domains. Specifically, there are few everyday activities that come in conflict with the visual and motor tasks utilized in the studies reviewed above—for instance low-level line orientation detection tasks or compensation for perturbation in motor learning. On the other hand, is difficult to avoid speech in the real world, which presents challenges for experimental design. For example, participants who learn non-native speech categories in the laboratory most likely have immediate access to interfering or conflicting stimuli, such as their own speech production or acoustic speech input from listening to other talkers. Unless this is experimentally controlled (i.e., participants stay in the laboratory for an extended period of quiet time following training), it is difficult to account for the events that happen after a learning session. Due to practical limitations, this has yet to be explored.

As discussed above, work by Baese-Berk and Samuel (2016) suggests articulation or production of speech sounds also seems to interfere with developing perceptual representations of speech. Interestingly however,

work by Bradlow and colleagues (1997) suggests that production accuracy can be enhanced by perceptual training alone, and this effect is stable over time (Bradlow et al., 1999). A similar study by Neufeld (1979) corroborates these results: participants who received perceptual training only were later able to produce words in a second language without a detectable non-native accent. It may be the case that perceptual training alone can induce concomitant improvements in speech production. Because relatively few studies have examined the relationship between speech perception and production and its influence non-native speech sound learning, future research will ultimately be needed to determine at what point in the learning trajectory and in what capacity production of new speech sounds is beneficial to the learner. With the available evidence, however, it seems that minimizing speech production in training, at least in the early stages of learning, may result in optimal outcomes for both perception and production of second language speech sounds.

Additionally, differing acoustic cues (e.g., the different formant trajectories associated with a dental stop in the context of an /i/ compared to an /u/ vowel) may be detrimental to stable learning and consolidation of non-native speech sounds. Thus, it may be advisable to limit phonological variability in training or during the stabilization phase following training. While several studies have found advantages of high-variability training procedures, these studies have typically taken place over the course of several weeks (e.g., Logan et al., 1991; Lively et al., 1993, 1994; Bradlow et al., 1997, 1999). In that case, the slow, procedural learning system may have had ample time to discover the regularities in the input (McClelland et al., 1995). This would also explain the enhanced generalization abilities as a result of this type of training; participants may have developed more robust abstract representations of the phonological categories, allowing for generalization to novel talkers or phonological contexts. While this may be the case, the efficacy of this training paradigm may be limited to certain situations. In particular, intensive training over long periods of time may not always be possible. For example, some second language classes meet only once per week, and this frequency may not be sufficient for learners to take advantage of high-variability training. As seen in Fuhrmeister and Myers (2017), even minimal exposure to phonological variability during a testing phase only (i.e., training was identical) attenuated learning of a

non-native contrast in the trained vowel context. Additionally, no overnight improvement was observed for those participants. This suggests that exposure to variability may not be optimal in certain cases, especially when training sessions are sparse. Especially in such situations, it is important that learners can consolidate newly acquired information to begin developing representations for non-native phonological categories. A more efficient training method may involve evening training sessions that occur close to sleep (to minimize native language exposure afterwards) that include only limited variability in the stimulus presentation. In fact, Earle and Myers (2015b) found generalization of non-native speech sounds to a novel talker following an interval of sleep, even though their training tokens consisted of sounds spoken by a single talker and presented in a single vowel context. Thus, sleep consolidation processes facilitated generalization to the sounds spoken by a new talker. Based on the evidence presented, striking a delicate balance between stimulus variability and proximity of training to sleep may promote consolidation of the trained information to long-term memory, which is essential to building novel phonological categories. Even so, second language learners experience different cues in the real world, and they need to be able to integrate information across them. Ultimately, future research will need to determine which factors promote abstraction over different acoustic cues.

Although we have some evidence as to what types of stimuli have the potential to interfere with non-native speech sound learning, many questions remain open. For example, visual and motor learning studies have often found that not all tasks interfere equally. In the study discussed above by Seitz and colleagues (2005), visual stimuli that differed in spatial location and orientation did not interfere with a perceptual learning task. Because primary visual cortex is retinotopically organized, Seitz et al. (2005) reasoned that different neurons were responding to visual stimuli in different spatial locations and orientations, whereas the same neurons responded to the task in which only the direction of offset differed and location and orientation remained the same. These same neurons had to be overwritten in order to learn the second task. It is unknown whether any potential speech analogs for such a task exist. Because primary auditory cortex is tonotopically organized, however, it is possible that training on similar stimuli at a different frequency (e.g., varying talker gender) would

not interfere with training on the first frequency range. However, speech may differ from other types of learning due to the complexity of the signal. In addition, findings from motor learning suggest that training on an opposite force, as in Brashers-Krug et al. (1996), interferes with learning of the initial direction, and word learning studies using an A-B A-C paradigm indicate that a new pairing (C) with an original stimulus (A) interferes with recall of the original pairing (A-B). Speech correlates of such findings are less obvious, and future research will need to address these questions.

Next, the time course of consolidation and susceptibility to interference in non-native speech sound learning should be considered. If the consolidation hypothesis or Walker's (2005) model can be applied to speech learning, it is essential to determine under what conditions a stabilization phase is necessary, and when so, how long this stabilization period needs to be in order to protect memory traces from subsequent interference. In the visual, motor, and word learning studies reviewed above, several time frames ranging from a few minutes to several hours have proven successful in protecting information against interference; however, some studies found interference even after 24 hours had passed between training sessions on each task. This suggests that domain or task differences may be responsible for some of the varied results obtained in these studies. Based on the results of Earle and Myers (2015a), it seems clear that non-native speech learning would benefit from a stabilization period; however, this has yet to be explored in the speech domain. Ultimately, future research will need to elucidate the time course of stabilization and consolidation in the speech domain.

8. Interference and second language speech learning in naturalistic contexts

While this review has primarily focused on memory consolidation and interference in the context of laboratory learning of non-native speech sounds, these concepts may be relevant to naturalistic learning environments, as well. For example, an individual's amount of first language use (among other factors) has been found to predict speech production accuracy in the second language (Flege et al., 1997; Piske et al., 2001). One possible explanation for these results is that habitual

interference from the first language obstructs developing second language speech category representations, especially if the stabilization phase is consistently disrupted. Further support for this idea comes from studies measuring language proficiency following immersion programs. Immersion programs have largely been successful for second language acquisition, including acquisition of speech sounds (e.g., Anderson, 2004; Cheour et al., 2002; Freed et al., 2004). Immersion settings present few opportunities for interference from the native language, which may allow memory traces of second language speech sounds to stabilize and more efficiently be consolidated into long-term memory. In addition, the procedural memory system can likely develop more robust category representations with the time and amount of exposure afforded by the immersion setting to discover regularities in the second language speech system. In fact, the study by Freed et al. (2004) compared native English-speaking college students learning French in a classroom setting in the home country, an intensive summer immersion program in the home country, and a semester-long study abroad program in France. While students in the study abroad and summer immersion programs outperformed the classroom learners after the period of study, students in the study abroad program reported much more first language use (English) outside the classroom than the summer immersion group. Consistent with first language use accounts, students in the study abroad group made fewer gains in proficiency than the students in the immersion program. Walker et al.'s (2003) findings on reconsolidation may similarly explain why less frequent first language use contributes to better speech perception and production in the second language. In this study, they found that practice on a task that had already been consolidated through sleep could reactivate the memory trace of that task, causing it to return to a labile state. If participants learned a second task after reactivating memory traces of the first task, interference from the second task on the first was observed. If upon reactivation, memories return to a labile state subject to interference, using the second language may reactivate memories and transfer them to this labile state. If frequent and intermittent use of the first language interferes with the reactivated second language memory traces, it may have the power to interfere with reconsolidation of the second language memory traces, especially if first language memory traces are much more robust. Thus, studies of

interference in memory consolidation may have explanatory power even in naturalistic language learning settings.

9. Speech sound learning across the lifespan

A common empirical finding in studies of second language acquisition is that second language speech sound learning typically decreases with age. Many have attributed this to a putative critical period (e.g., Granena & Long, 2012); however, others have found a more linear decline in second language speech production abilities throughout the lifespan that would be less indicative of a critical period (e.g., Flege et al., 1995). Although a complete discussion of this issue is outside the scope of this chapter, some speculations can be made when considering both models of memory consolidation and non-native speech sound learning. Assuming no strictly defined critical period exists but non-native speech sound learning becomes increasingly more difficult throughout the course of the lifespan, ideas put forward in the learning and memory literature may be applicable. For example, the same system may underlie acquisition of speech sounds in infancy, childhood, and adulthood; however, adults cannot approach the speech sound learning task in the same way as an infant or child because their prior experience and interactions with the environment are vastly different (see Best & Tyler, 2007 for discussion). Infants do not have well-established first language speech categories, while adults have developed quite robust speech categories in the native language after years or decades of exposure. In fact, Burnham et al. (1991) found that speakers of a language become more categorical in their perception of native language speech sounds over the life span, and Baker et al. (2008) observed that children are less likely than adults to assimilate second language speech sounds to first language categories. This may imply that increasingly more stable first language categories are less malleable than less-stable categories, such as those in childhood. This could have several implications for non-native speech sound learning. As predicted by non-native speech perception and learning models, the native language exerts a powerful influence on the perception and ability to learn non-native speech categories. Additionally, theories of learning and memory (e.g., Ebbinghaus, 1885) and findings in the visual domain by Shibata et al. (2017) suggest that stable or strongly

learned information can cause proactive interference. In this way, theories from both of these domains could be taken together to imply that native language speech categories interfere with non-native speech sound learning in a graded manner: stronger native language categories may increase the difficulty in developing new categories. In addition, children may be less susceptible to interference in certain types of learning prior to puberty. In a study by Dorfberger and colleagues (2007), children before and after the onset of adolescence learned a procedural motor task. Younger children showed no advantage in learning or retaining the sequence; however, when they were trained on an additional, opposing sequence, 9- and 12-year-olds demonstrated no evidence of interference of the second sequence on the first. Seventeen-year-olds, on the other hand, did show an interference effect. This finding lends support to critical period hypotheses but in a different way than they are traditionally depicted. Proponents of a critical period typically focus on child advantages in learning, but this finding suggests adults may be as good as children at learning certain tasks; however, their consolidation processes may differ. Some studies investigating second language speech learning in children and adults have found superior perceptual learning in adults initially, but after some time children not only caught up to the adults in performance but actually surpassed them (Snow & Hoefnagel-Höhle, 1978). This finding may be consistent with the Dorfberger et al. (2007) study: interference from the first language may not have influenced children's learning of novel speech sounds, while it may have obstructed adults' learning over time.

Simultaneous and sequential bilinguals may additionally offer some insight into memory consolidation of speech sounds over the lifespan. For example, children who begin learning a second language in early childhood seldom have a detectable non-native accent in either language; however, late-onset bilinguals often do (e.g., Flege et al., 1995). Thus, it appears that infants and young children can learn the speech sound inventory of two languages simultaneously without interference from either language. Assuming the slow, procedural memory system subserves this process (McClelland et al., 1995; Dudai, 2004), this system would discover regularities in both sound systems simultaneously, even if speech categories in both languages are close in perceptual space. For example, an infant learning Spanish and English would learn that the distribution of the

bilabial voiced stop consonant clusters around a lower voice onset time than the same category in English. The lack of experience with a specific language's speech sound inventory may additionally facilitate the acquisition of speech categories in two languages simultaneously.

Furthermore, evidence from bilinguals who learned languages sequentially may shed light on this process. For example, a study by Antoniou et al. (2012) found that early sequential bilinguals who were dominant in their second language were influenced by their dominant (second) language in their perception of consonants in both languages. This parallels previously discussed findings showing effects of first language use on second language speech perception and production: less frequent first language use (or more frequent second language use) may be a strong factor in the development of second language speech categories. In other words, while important, age of initial acquisition of a language is not necessarily the determining factor in how well the speech system of that language will be learned. Ultimately, future research will be needed to elucidate relative influences of age of acquisition and amount of language use on the development of second language speech category representations.

10. Conclusion

Although a comprehensive account of the non-native speech sound learning process has yet to be established, many insights can be gained from considering findings from domain-general learning and memory studies. In particular, factors such as interference during a critical stabilization phase following learning and the strength of initial encoding may be important considerations when designing training paradigms for learning novel speech categories or when interpreting findings from this field. Whether models of memory consolidation can be applied to speech sound learning without modification remains unclear; however, they make concrete predictions for future studies and appear to have a great deal of explanatory power within the current literature.

Acknowledgements

I would like to thank Emily Myers for extensive discussions about this topic and feedback on a previous version of this manuscript, as well as

Rachel Theodore and Erika Skoe for their helpful comments on a previous version of the manuscript. This work was supported by NSF IGERT DGE-1144399 to the University of Connecticut.

References

Anderson, R.T. (2004). Phonological acquisition in preschoolers learning a second language via immersion: A longitudinal study. *Clinical Linguistics & Phonetics*, *18*(3), 183–210.

Antoniou, M., Tyler, M.D., & Best, C.T. (2012). Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode? *Journal of Phonetics*, *40*(4), 582–594.

Atienza, M., Cantero, J.L., & Stickgold, R. (2004). Posttraining sleep enhances automaticity in perceptual discrimination. *Journal of Cognitive Neuroscience*, *16*(1), 53–64.

Baese-Berk, M.M., & Samuel, A.G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language*, *89*, 23–36.

Baker, W., Trofimovich, P., Flege, J.E., Mack, M., & Halter, R. (2008). Child—adult differences in second-language phonological learning: The role of cross-language similarity. *Language and Speech*, *51*(4), 317–342.

Best, C.T. (1994). The emergence of native-language phonological influences in infants: A perceptual assimilation model. *Haskins Laboratories Status Report on Speech Research*. *SR-107*/108, 1–30.

Best, C.T., McRoberts, G.W., & Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of Acoustical Society of America*, *109*(2), 775–794.

Best, C.T., McRoberts, G.W., & Sithole, N.M. (1988). Examination of perceptual reorganization for nonnative speech contrasts: Zulu click discrimination by English-speaking adults and infants. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 345–360.

Best, C.T., & Tyler, M.D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M.J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of*

language experience in speech perception and production. Amsterdam: John Benjamins, pp. 13–34.

Bradlow, A.R., Akahane-Yamada, R., Pisoni, D.B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long- term retention of learning in perception and production. *Perception & Psychophysics*, *61*(5), 977–985.

Bradlow, A.R., Pisoni, D.B., Akahane-Yamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of Acoustical Society of America*, *101*(4), 2299–2310.

Brashers-Krug, T., Shadmehr, R., & Bizzi, E. (1996). Consolidation in human motor memory. *Nature*, *382*(6588), 252–255.

Brown, H., & Gaskell, M.G. (2014). The time-course of talker-specificity and lexical competition effects during word learning. *Language, Cognition and Neuroscience*, *29*(9), 1163–1179.

Burnham, D.K., Earnshaw, L.J., & Clark, J.E. (1991). Development of categorical identification of native and non-native bilabial stops: infants, children and adults. *Journal of Child Language*, *18*(2), 231–260.

Caithness, G., Osu, R., Bays, P., Chase, H., Klassen, J., Kawato, M., Wolpert, D.M. & Flanagan, J.R. (2004). Failure to consolidate the consolidation theory of learning for sensorimotor adaptation tasks. *Journal of Neuroscience*, *24*(40), 8662–8671.

Cheour, M., Shestakova, A., Alku, P., Ceponiene, R., & Näätänen, R. (2002). Mismatch negativity shows that 3–6-year-old children can learn to discriminate non-native speech sounds within two months. *Neuroscience Letters*, *325*(3), 187–190.

Davis, M. H., Di Betta, A. M., Macdonald, M. J., & Gaskell, M. G. (2009). Learning and consolidation of novel spoken words. *Journal of Cognitive Neuroscience*, 21(4), 803-820.

Dewar, M., Garcia, Y.F., Cowan, N., & Sala, S.D. (2009). Delaying interference enhances memory consolidation in amnesic patients. *Neuropsychology*, *23*(5), 627–634.

Dorfberger, S., Adi-Japha, E., &Karni, A. (2007). Reduced susceptibility to interference in the consolidation of motor memory before adolescence. *PLoS One*, *2*(2), e240.

Drosopoulos, S., Windau, E., Wagner, U., & Born, J. (2007). Sleep enforces the temporal order in memory. *PLoS One*, *2*(4), e376.

Dudai, Y. (2004). The neurobiology of consolidations, or, how stable is the engram? *Annual Review of Psychology*, *55*, 51–86.

Dumay, N., & Gaskell, M.G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, *18*(1), 35–39.

Earle, F.S., Landi, N., & Myers, E.B. (2017). Sleep duration predicts behavioural and neural differences in adult speech sound learning. *Neuroscience Letters*, *636*, 77–82.

Earle, F.S., & Myers, E.B. (2014). Building phonetic categories: an argument for the role of sleep. *Frontiers in Psychology*, *5*, 1192.

Earle, F.S., & Myers, E.B. (2015a). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(6), 1680–1695.

Earle, F.S., & Myers, E.B. (2015b). Overnight consolidation promotes generalization across talkers in the identification of nonnative speech sounds. *The Journal of the Acoustical Society of America*, *137*(1), EL91–EL97.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Leipzig: Verlag von Duncker und Humblot.

Eimas, P.D., Siqueland, E. R., Juscyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Eisner, F., & McQueen, J.M. (2006). Perceptual learning in speech: Stability over time. *The Journal of the Acoustical Society of America*, *119*(4), 1950–1953.

Ellenbogen, J.M., Hulbert, J.C., Jiang, Y., & Stickgold, R. (2009). The sleeping brain's influence on verbal memory: boosting resistance to interference. *PLoS One*, *4*(1), e4117.

Ellenbogen, J.M., Hulbert, J.C., Stickgold, R., Dinges, D.F., & Thompson-Schill, S.L. (2006). Interfering with theories of sleep and memory: sleep, declarative memory, and associative interference. *Current Biology*, *16*(13), 1290–1294.

Fenn, K.M., Margoliash, D., & Nusbaum, H.C. (2013). Sleep restores loss of generalized but not rote learning of synthetic speech. *Cognition*, *128*, 280–286.

Fenn, K.M., Nusbaum, H., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language of perceptual learning. *Nature*, *425*(6958), 614–616.

Flege, J.E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*. Timonium, MD: York Press, pp. 229–273.

Flege, J.E. (2003). Assessing constraints on second-language segmental production and perception. In A. Meyer & N. Schiller (Eds.), *Phonetics and Phonology in Language Comprehension and Production, Differences and Similarities*. Berlin: Mouton de Gruyter, pp. 319–355.

Flege, J.E., Frieda, E.M., & Nozawa, T. (1997). Amount of native-language (L1) use affects the pronunciation of an L2. *Journal of Phonetics*, *25*(2), 169–186.

Flege, J.E., Munro, M.J., & MacKay, I.R. (1995). Factors affecting strength of perceived foreign accent in a second language. *The Journal of the Acoustical Society of America*, *97*(5), 3125–3134.

Francis, A.L., & Nusbaum, H.C. (2002). Selective attention and the acquisition of new phonetic categories. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(2), 349–366.

Freed, B.F., Segalowitz, N., & Dewey, D.P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26*(2), 275–301.

Fuhrmeister, P. Myers, E.B. (2017). Non-native phonetic learning is destabilized by exposure to phonological variability before and after training. *The Journal of the Acoustical Society of America*. *142*(5), EL448–EL454.

Galantucci, B., Fowler, C.A., & Turvey, M.T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Goedert, K.M., & Willingham, D.B. (2002). Patterns of interference in sequence learning and prism adaptation inconsistent with the consolidation hypothesis. *Learning & Memory*, *9*(5), 279–292.

Golestani, N., & Zatorre, R.J. (2004). Learning new sounds of speech: reallocation of neural substrates. *Neuroimage*, *21*(2), 494–506.

Golestani, N., & Zatorre, R.J. (2009). Individual differences in the acquisition of second language phonology. *Brain and Language*, *109*(2–3), 55–67.

Granena, G., & Long, M.H. (2013). Age of onset, length of residence, language aptitude, and ultimate L2 attainment in three linguistic domains. *Second Language Research*, *29*(3), 311–343.

Hauptmann, B., Reinhart, E., Brandt, S. A., & Karni, A. (2005). The predictive value of the leveling off of within-session performance for procedural memory consolidation. *Cognitive Brain Research*, *24*(2), 181–189.

Heim, S., Klann, J., Schattka, K.I., Bauhoff, S., Borcherding, G., Nosbüsch, N., Struth, L., Binkofski, F.C., &Werner, C.J. (2017). A nap but not rest or activity consolidates language learning. *Frontiers in Psychology*, *8*, 665.

Kuhl, P.K. (1994). Learning and representation in speech and language. *Current Opinion in Neurobiology*, *4*(6), 812–822.

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D.,Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*(1493), 979–1000.

Kuhl, P.K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science*, *9*(2), F13–F21.

Kuriyama, K.,Stickgold, R., & Walker, M.P. (2004). Sleep-dependent learning and motor-skill complexity. *Learning & Memory*, *11*(6), 705–713.

Lahl, O., Wispel, C., Willigens, B., & Pietrowsky, R. (2008). An ultra short episode of sleep is sufficient to promote declarative memory performance. *Journal of Sleep Research*, *17*(1), 3–10.

Lim, S.J., & Holt, L.L. (2011). Learning foreign sounds in an alien world: Videogame training improves non-native speech categorization. *Cognitive Science*, *35*(7), 1390–1405.

Lively, S.E., Logan, J.S., & Pisoni, D.B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic

environment and talker variability in learning new perceptual categories. *The Journal of Acoustical Society of America*, *94*(3), 1242–1255.

Lively, S.E., Pisoni, D.B., Yamada, R.A., Tohkura, Y., & Yamada, T. (1994). Training Japanese listeners to identify English /r/ and /l/. III. Long-term retention of new phonetic categories. *The Journal of Acoustical Society of America*, *96*(4), 2076–2087.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of Acoustical Society of America*, *89*(2), 874–886.

MacKay, I.R., Meador, D., & Flege, J.E. (2001). The identification of English consonants by native speakers of Italian. *Phonetica*,*58*(1–2), 103–125.

Marshall, L., & Born, J. (2007). The contribution of sleep to hippocampus-dependent memory consolidation. *Trends in Cognitive Sciences*, *11*(10), 442–450.

Mazza, S., Gerbier, E., Gustin, M.P., Kasikci, Z., Koenig, O., Toppino, T.C., & Magnin, M. (2016). Relearn faster and retain longer: Along with practice, sleep makes perfect. *Psychological Science*, *27*(10), 1321–1330.

McCandliss, B.D., Fiez, J.A., Protopapas, A., Conway, M., & McClelland, J.L. (2002). Success and failure in teaching the [r]-[l] contrast to Japanese adults: Tests of a Hebbian model of plasticity and stabilization in spoken language perception. *Cognitive, Affective, & Behavioural Neuroscience*, *2*(2), 89–108.

McClelland, J.L. (1998). Complementary learning systems in the brain: A connectionist approach to explicit and implicit cognition and memory. *Annals of the New York Academy of Sciences*, *843*(1), 153–169.

McClelland, J.L., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McCloskey, M., & Cohen, N.J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.) *Psychology of learning and motivation*, Vol. 24, Academic Press. pp. 109–165.

Muellbacher, W., Ziemann, U., Wissel, J., Dang, N., Kofler, M., Facchini, S., Boroojerdi, B.,Poewe, W., & Hallett, M. (2002). Early consolidation in human primary motor cortex. *Nature*, 415(6872), 640–644.

Müller, G.E., & Pilzecker, A. (1900). *Experimentelle Beiträge zur Lehre vom Gedächtnis*. (Vol. 1). JA Barth.

Myers, E.B., & Swan, K. (2012). Effects of category learning on neural sensitivity to non-native phonetic categories. *Journal of Cognitive Neuroscience*, *24*(8), 1695–1708.

Neufeld, G.G. (1979). Towards a theory of language learning ability. *Language Learning*, *29*(2), 227–241.

Pallier, C., Bosch, L., & Sebastián-Gallés, N. (1997). A limit on behavioural plasticity in speech perception. *Cognition*, *64*(3), B9–B17.

Piske, T., MacKay, I.R., & Flege, J.E. (2001). Factors affecting degree of foreign accent in an L2: A review. *Journal of Phonetics*, *29*(2), 191–215.

Poldrack, R.A., Clark, J., Pare-Blagoev, E.J., & Shohamy, D. (2001). Interactive memory systems in the human brain. *Nature*, *414*(6863), 546–550.

Roth, D.A.E., Kishon-Rabin, L., Hildesheimer, M., & Karni, A. (2005). A latent consolidation phase in auditory identification learning: time in the awake state is sufficient. *Learning & Memory*, *12*(2), 159–164.

Seitz, A.R., Yamagishi, N., Werner, B., Goda, N., Kawato, M., & Watanabe, T. (2005). Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(41), 14895–14900.

Shibata, K., Sasaki, Y., Bang, J.W., Walsh, E.G., Machizawa, M.G., Tamaki, M., Chang, L.-H. & Watanabe, T. (2017). Overlearning hyper-stabilizes a skill by rapidly making neurochemical processing inhibitory-dominant. *Nature Neuroscience*, *20*(3), 470–475.

Silbert, N.H., Smith, B.K., Jackson, S.R., Campbell, S.G., Hughes, M.M., & Tare, M. (2015). Non-native phonemic discrimination, phonological short term memory, and word learning. *Journal of Phonetics*, *50*, 99–119.

Snow, C.E., & Hoefnagel-Höhle, M. (1978). The critical period for language acquisition: Evidence from second language learning. *Child Development*, *49*(4), 1114–1128.

Squire, L.R. (2004). Memory systems of the brain: a brief history and current perspective. *Neurobiology of Learning and Memory*, *82*(3), 171–177.

Stickgold, R., LaTanya, J., & Hobson, J.A. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience*, *3*(12), 1237–1238.

Swan, K., & Myers, E. (2013). Category labels induce boundary-dependent perceptual warping in learned speech categories. *Second Language Research*, *29*(4), 391–411.

Tamminen, J., & Gaskell, M.G. (2013). Novel word integration in the mental lexicon: Evidence from unmasked and masked semantic priming. *The Quarterly Journal of Experimental Psychology*, *66*(5), 1001–1025.

Tsao, F.M., Liu, H.M., & Kuhl, P.K. (2004). Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child Development*, *75*(4), 1067–1084.

Tucker, M.A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep*, *31*(2), 197–203.

Vlahou, E.L., Protopapas, A., & Seitz, A.R. (2012). Implicit training of nonnative speech stimuli. *Journal of Experimental Psychology: General*, *141*(2), 363–381.

Wade, T., & Holt, L.L. (2005). Incidental categorization of spectrally complex non-invariant auditory stimuli in a computer game task. *The Journal of the Acoustical Society of America*, *118*(4), 2618–2633.

Walker, M.P. (2005). A refined model of sleep and the time course of memory formation. *Behavioural and Brain Sciences*, *28*(1), 51–64.

Walker, M.P., Brakefield, T., Hobson, J.A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*(6958), 616–620.

Werker, J.F., & Tees, R.C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, *7*(1), 49–63.

Xie, X., Earle, F.S. & Myers, E.B. (2017). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience*, *33*(2), 196–210.

Yi, H.G., Maddox, W.T., Mumford, J.A., & Chandrasekaran, B. (2014). The role of corticostriatal systems in speech category learning. *Cerebral Cortex*, *26*(4), 1409–1420.

Lisa Morano, Louis ten Bosch, Mirjam Ernestus

Looking for exemplar effects: testing the comprehension and memory representations of r'duced words in Dutch learners of French

Abstract: In this study, we tested whether second language (hereafter L2) learners can encode in the form of exemplars phonetic variation that does not occur regularly in their native language (hereafter L1). Three groups of Dutch learners of French performed a long-term repetition priming lexical decision task in which words were repeated. The second occurrence (target) of an experimental word either matched or mismatched the pronunciation of its first occurrence (prime). When a target matched its prime, both tokens had a completely devoiced or a completely voiced high vowel in their first syllable. When a target mismatched its prime, the prime had a devoiced high vowel in its first syllable, while the target had a voiced high vowel in its first syllable, and *vice versa*. In condition AA and in condition BB we reused the same token (albeit different tokens per condition) in case of a repetition match. In condition AB, we used different tokens for prime and target. The results show that L2 learners are able to encode phonetic information that does not occur regularly in their L1 in the form of exemplars, showing that exemplars are formed before the L2 phonological filter applies, but only under very limited conditions: when the prime is difficult to process and when the matching and mismatching tokens are easily distinguishable. Contrary to our expectations, we also found that mismatching devoiced primes significantly accelerated the recognition of the voiced B targets. We hypothesize that this latter result comes from a higher activation of abstract representations after difficult primes. Our results therefore show different processing patterns for identical testing conditions using different tokens (conditions AA and BB). These results question the use of exemplars in everyday speech comprehension, adding to the growing body of evidence that exemplar effects only arise in very restricted unnatural conditions.

Keywords: memory, exemplar models, second language learning, adults, reduced words, French, comprehension

1. Introduction

Many researchers now assume that the mental lexicon is hybrid in nature (Pierrehumbert, 2002; McLennan, Luce, & Charles-Luce, 2003; Goldinger, 2007), containing, for each word, both an abstract representation of the word's pronunciation (*i.e.* a string of abstract symbols such as phonemes), and a cloud of exemplars (i.e. occurrences encountered by the listener, each encoding fine acoustic characteristics such as speech rate, the speaker's voice, but also phonetic details). Indeed, purely abstractionist or purely exemplarist models of speech comprehension both fail to account for all the findings in the literature. For example, listeners' ability to adapt to a speaker's specific way of talking such as a lisp (i.e. perceptual learning; e.g., Norris, McQueen, & Cutler, 2003), or listeners' ability to generalize a phonological rule to new words (e.g., Cristia, Mielke, Daland, & Peperkamp, 2013), cannot be explained if their mental lexicons only contain exemplars of previously encountered tokens without any degree of abstraction. Evidence for exemplars, on the other hand, comes from priming experiments (e.g., Tulving & Schacter, 1990), in which it has repeatedly been shown that native listeners recognize words faster or more accurately when they occur for the second time in the experiment (as "targets") than when they occur for the first time (as "primes") especially if the two tokens share fine, phonologically irrelevant, acoustic characteristics such as information about the speaker's voice (i.e. both the prime and the target are uttered by the same person; e.g., McLennan & Luce, 2005). These specificity (or exemplar) effects suggest that the participants stored the first occurrences of the words with at least some degree of acoustic detail, that is, in the form of exemplars.

Nearly all experiments investigating exemplars have been conducted with native (L1) listeners. Exemplar research has barely studied second language (L2) learners. Nevertheless, there is much to gain from research with L2 listeners. First, if exemplars play a substantial role in speech comprehension, as most researchers currently assume, the findings obtained with L1 listeners should generalize to L2 listeners, as it is unlikely that listeners use two different mechanisms for speech comprehension in a L2 and in their L1. Second, research with non-native listeners may provide information about which acoustic details are exactly stored in exemplars.

Are exemplars faithful representations of the acoustic signal or are they affected by the listener's linguistic knowledge? That is, for L2 listeners, are exemplars formed before or after their L1 phonological filter (Troubetzkoy, 1939) applies?

It has been shown that L2 learners' abstract representations diverge from those of natives. Pallier, Colomé, and Sebastián-Gallés (2001) found that even highly proficient Spanish-Catalan listeners treat all minimal pairs specific to Catalan as homophones in a lexical decision task with medium-term auditory implicit repetition priming: for Spanish-Catalan bilinguals, [netə] 'granddaughter', primed equally well [netə] and [nɛtə] 'clean', and *vice versa*. Proficiency appears to play an important role, as was shown in another study. Darcy, Dekydtspotter, Sprouse, Glover, Kaden, McGuire, and Scott (2012) tested intermediate and advanced American English learners of French on two front *vs*. back rounded vowel contrasts in French (/y/-/u/ and /œ/-/ɔ/), which do not occur in English. In a lexical decision task with implicit repetition priming, both the intermediate and advanced learners patterned like the natives on the /œ/-/ɔ/ contrast, albeit with slower reaction times, while the intermediate learners, but not the advanced learners, treated the /y/-/u/ minimal pairs as homophones. This suggests that the intermediate learners did not distinguish /y/ and /u/ in their lexical representations, while the advanced learners did. These studies suggest that L2 phonological variation that is irrelevant in listeners' L1 is not immediately stored in listeners' L2 abstract representations, and that it may, or may not, eventually be stored abstractly at higher proficiency levels.

Exemplars in L2 listeners need not be different from exemplars in L1 listeners since L2 listeners have been shown to remain sensitive to L1 irrelevant contrasts provided the task employed could be performed without requiring lexical processing such as a phoneme categorization task (Sebastián-Galles & Baus, 2005; Diaz, Mitterer, Broersma, & Sebastian-Galles, 2012). That is, L2 listeners are able to perform simple low-level tasks in phonetic mode but as soon as linguistic processing is required, such as for a lexical decision task, then their L1 phonological filter prevents them from processing the stimuli in a native-like fashion (with the notable exception of Darcy et al.'s, 2012, results). If exemplars are formed before the phonological filter applies, L2 exemplars can thus well encode L1

irrelevant variation. If exemplars are formed after the phonological filter applies, L2 exemplars probably encode less L1 irrelevant variation. Our research question was the following: Are L2 intermediate learners able to encode, in the form of exemplars, fine linguistic details about the properties of the prime that are not relevant in their L1, and to subsequently use them for speech comprehension (i.e. to comprehend the target)?

As previously mentioned, very little exemplar research has been carried out in L2. We could only find two studies reporting exemplar effects for L2 listeners. Trofimovich (2005) tested American English learners of Spanish in an immediate repetition task. The participants first listened to a list of 36 prime words uttered by three male and three female speakers (the study phase). The participants then performed a 3–4 minute distractor task, followed by an immediate repetition task (the test phase) in which all the primes were repeated (as targets) either in the same voice as during the study phase, or in a different voice from the opposite gender, along with new words. These tasks were performed twice: once in English and once in Spanish, the task order being counterbalanced over all the participants. In their L2, the participants were faster at repeating the words previously heard in the same voice than words which had not been presented during the study phase, but they were equally fast at repeating words heard for the first time in the experiment as words previously heard in the experiment in a different voice. The participants thus treated L2 words repeated in a different voice just as new items in the test phase.

In their L1, Trofimovich's participants showed priming but no exemplar effects: the participants were faster at repeating English words already heard in the study phase than words which had not been presented in the study phase, but it did not matter whether those words were uttered in the study phase in the same or in a different voice. Although Trofimovich's study did not replicate previous studies which found exemplar effects for native listeners (e.g. Craik & Kirsner, 1974; Palmeri et al., 1993; Luce & Lyons, 1998), it shows that exemplar effects can be found for L2 learners.

Further evidence that L2 listeners can store exemplars was provided by Winters, Lichtman, and Weber (2013). The authors tested three groups of listeners in German: English monolinguals, English learners of German, and German monolinguals in an old/new auditory categorization task. The stimuli were monosyllabic consonant-vowel-consonant (CVC) German

words, which varied in frequency of occurrence (low, medium, high), and were uttered by five female voices in one block and five male voices in another block (the order being counterbalanced over participants). Within each block, half of the words were repeated either with the same or a different voice. The authors found that target words presented in the same voice as their primes were classified correctly more often than target words presented in a different voice, irrespective of the listener group.

L2 listeners are thus able to store details about the speaker's voice in the form of exemplars. This may not come as a surprise since L2 listeners already have ample experience in processing indexical variation in their L1, and it has been shown that the ability to use consistent information about a speaker's voice across items is easily transferable to L2 speech perception (Bradlow & Pisoni, 1999). The question is whether L2 listeners not only store in exemplars indexical information but also phonetic variation that occurs regularly in their L2 but not in their L1. While exemplar effects encoding indexical variation have already been attested by Winters, Lichtman, and Weber (2013) and Trofimovich (2005), to our knowledge, no previous study has found exemplar effects encoding L2 phonetic variation that does not occur regularly in the listener's L1. In this study, we tested whether exemplar effects in L2 listeners can also be found when manipulating regular phonetic variation instead of indexical (or speaker) variation.

One way to study exemplar effects for regularly occurring L1 specific phonetic variation instead of indexical variation is to focus on pronunciation variants of words resulting from reduction. Reduction is the weakening or deletion of phonemes or even whole syllables, occurring in informal connected speech, compared to the words' canonical pronunciations, that is the pronunciations of words in isolation (Ernestus & Warner, 2011). Most previous experiments investigating exemplar effects by manipulating linguistic variation focused on categorical variation, substituting one allophone with another allophone (e.g. [ɛ] with [e] in Pallier et al., 2001; and [t] and [d] with [ɾ] in McLennan, Luce & Charles-Luce, 2003). It could be argued that in these experiments listeners stored several abstract representations (one for each word pronunciation variant) rather than different exemplars. Using categorical variation therefore makes it difficult to attest for the role of exemplars.

Reduction reflecting continuous variation, on the other hand, cannot be stored abstractly. Such reduction may result in an infinite number of realizations, which all activate the same abstract pronunciation variant of the word. Reduction reflecting continuous variation is thus an interesting characteristic to manipulate in order to test for unambiguous exemplar effects. To our knowledge, no previous study has done so.

In our study, we investigated the reduction phenomenon of phrase-medial high vowel devoicing. In casual French, in a noun phrase like *la cité* ([la.si.te] 'the city'), the /i/ can be more or less devoiced (up to completely) as the voicing (i.e. vibration of the vocal folds) fails to be re-established in time after the devoiced consonant /s/ (Torreira & Ernestus 2010). Furthermore, phrase-medial high vowel devoicing in French is a gradient phenomenon. In their corpus study, Torreira and Ernestus found that the high vowels were more devoiced or completely absent after certain consonants, the higher the speech rate, and the further away the vowel was from the end of the accentual phrase. Given that the same variables predict presence and amount of voicing, absence of voicing is the end of a continuum that is reached in extreme devoicing conditions. This phenomenon has never been reported for Dutch, suggesting that it is part of the sound pattern of French but not of Dutch. Consequently, if Dutch learners of French show exemplar effects in an experiment that manipulates phrase-medial high French vowel devoicing, we can conclude that L2 learners can also store, in the form of exemplars, L2 specific sound patterns.

We wished to use a task that requires deep processing of the stimuli to approach everyday speech processing. In our study, we used a lexical decision task. Although it can be argued that a lexical decision task is a very artificial task to investigate speech comprehension, it ensures a deeper linguistic processing than an old/new categorization task (or continuous recognition memory task) or a shadowing task, which are often used in exemplar studies (e.g. Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993; Goldinger, 1996; Bradlow, Nygaard, & Pisoni, 1999; Trofimovich, 2005; Mattys & Liss, 2008; Winters et al., 2013). The words' forms need to be accessed to elicit responses from the participants: to decide whether a stimulus is a real word or not the participants need to access what the word means, even vaguely.

We tested Dutch intermediate learners of French in a lexical decision task in French in which the experimental words contained a high vowel following a voiceless consonant. The experimental words were all repeated either as a pronunciation match (i.e. both the high vowel of the prime and that of the target were devoiced, or both were voiced) or as a pronunciation mismatch (i.e. when the high vowel of the prime was devoiced, the vowel of the target was voiced and vice versa). If participants react faster to a target when it matches than when it mismatches the pronunciation of its prime, we can conclude that L2 participants show exemplar effects, indicating that they are able to store, in the form of exemplars, phonetic information that does not occur regularly in their L1, and to later on reuse those exemplars to comprehend the next token of the word.

We ran the same experiment three times. In condition AB, we used different recordings for prime (a voiced or devoiced token A) and target (a voiced or devoiced token B). As already pointed out by Hanique et al. (2014), using two different tokens (or recordings) for prime and target represents a more ecologically valid testing condition than using identical tokens, given that in daily life, we never hear the exact same token twice: in a conversation, if a person repeats a word, she will produce a new token that will vary slightly from the first one.

We compared this condition with two conditions in which the prime and target were identical in case of a match (like in nearly all the previous studies on exemplar effects): one using only the tokens used in the first condition as primes (condition AA), and one using only the tokens used in the first condition as targets (condition BB).

2. Method

2.1. Participants

We tested 120 Dutch university students who had studied French for four to seven years in high school (intermediate level, or B1–B2 levels of the Common European Framework of Reference for Languages, CEFR, Council of Europe, 2011) and who were paid for their participation. The participants were between 18 and 29 years old (mean: 21.74), 95 were female and 105 were right-handed. None of the participants reported any

hearing problems. The participants were randomly assigned to one of the three conditions (AB, AA, BB).

2.2. Materials

Our experimental words were selected from the vocabulary of two beginners' textbooks used in French classes at Dutch secondary schools (*Franconville* and *Grandes Lignes*). They were bisyllabic words containing a high vowel (/i/, /y/, or /u/) following a voiceless consonant in their first syllable (cf. Appendix 1). Out of all possible words, we selected the 24 most frequent words, with a preference for those containing /i/ and /y/ as these vowels are more constricted than /u/, which allows them to be more easily devoiced than /u/ (Meunier, Meynadier, & Espesser, 2008)[1]. The frequency of occurrence of our experimental words in the movie subtitles corpus of Lexique 3.81 (New, Pallier, Ferrand, & Matos, 2001) ranged from 0.71 (per million words) for *cycliste* 'cyclist' to 107.92 for *sujet* 'subject'(mean: 31.40, cf. Appendix 1), that is, they were fairly frequent words (most of them ranging between the median at 8 occurrences per million words and the third quartile at 43 occurrences per million), which is normal for beginners' vocabulary words.

We also selected 78 bisyllabic frequent words, without particular restriction, from the aforementioned beginners' textbooks to be used as existing-word fillers. Finally, we created 102 bisyllabic pseudo-word fillers by adding a phonotactically legal syllable to the first syllable of all the experimental and existing-word fillers already selected.

1 One of the reviewers attracted our attention to the fact that the participants may not process the devoiced vowel at all despite the remaining durational and formant cues signaling the presence of the vowel (as it has been shown to happen for German natives listening to Japanese accented German; Zimmerer, Rei, & Reetz, 2013). In that case, three items could be confused with other French words (*purée* 'mashed potatoes' could be confused with *pré* 'meadow'; *pilote* 'pilot' with the reduced form of *pelote* 'woolen ball'; and *poulet* 'chicken' with *plaie* 'wound'). However, the occurrence frequencies of the possibly confounded words (*pré* 'meadow'; *pelote* 'woolen ball'; *plaie* 'wound') are all lower than the occurrence frequencies of our stimuli, making it unlikely that our participants knew these words.

All the stimuli, preceded by their definite determiners, were recorded in a sound attenuated booth with a head mounted microphone at 44.100 Hz by the first author of this paper, a female French native speaker from Caen. The easiest way to obtain fully devoiced high vowels in the first syllables appeared to have the speaker produce all the experimental words without their determiners. In this way, for the "devoiced" (that is: 'containing a devoiced high vowel in the word's first syllable') recordings, the speaker could comfortably whisper the first syllables and then voice the second syllable, while for the "voiced" (that is: 'containing a voiced high vowel in the world's first syllable') recordings, she could just speak out loud the whole words. The first vowel of the devoiced stimuli was always completely devoiced and the first vowel of the voiced stimuli was always fully voiced (cf. Figure 1). The speaker also recorded all the experimental words with their determiners. The best devoiced and voiced recordings without determiners were then each paired with their closest voiced recordings with determiner in terms of intonation and duration. The final stimuli were obtained by cross-splicing the voiced determiners with the devoiced and voiced recordings without determiners.

We created two tokens for each voicing type, meaning that for each experimental word we obtained four tokens: a voiced token A, a voiced token B, a devoiced token A, and a devoiced token B. Tokens A were on average 805 ms long (804 ms for the voiced ones, SD = 106, and 806 ms for the devoiced ones, SD = 124) and tokens B were on average 811 ms long (796 ms for the voiced ones, SD = 134, and 826 ms for the devoiced ones, SD = 136). Note that for the B tokens, it is not the case that the devoiced form was always longer than the voiced one (*cf.* Appendix 1 for the durations of all individual tokens). The existing-word fillers and the pseudo-word fillers were not cross-spliced but two tokens were recorded per word-type. The average duration of the existing-word fillers was 719 ms (SD = 120) and of the pseudo-word fillers 739 ms (SD = 128).

Finally, all the stimuli were scaled to 70 dB of average intensity. All the stimulus recording, editing, and scaling was performed in Praat (Boersma & Weenink, 2017).

The lexical decision task consisted of two blocks of 132 trials each. Twelve of the experimental words were presented in the first block and 12 in the second block. Within each block, the experimental words were

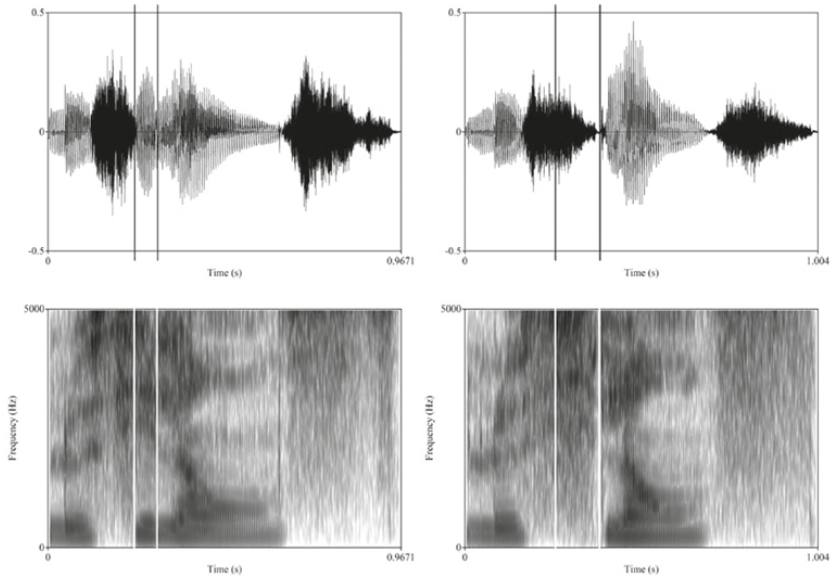

Figure 1: Waveforms (top panels) and spectrograms (bottom panels) of the target word *le silence* 'the silence': voiced token A on the left, and devoiced token A on the right. The high-vowel /i/ boundaries are indicated by the vertical lines.

repeated either as a variant match (i.e. both prime and target had either voiced or devoiced vowels) or as a variant mismatch (i.e. when the prime was voiced, the target was devoiced, and vice versa). The prime and target were separated by seven to 98 trials (average: 65), replicating the lags used in the first and third experiments of Hanique, Aalders, and Ernestus (2014). Although these lags are not as long as the ones used by Goldinger (1996), who found exemplar effects one week after presentation of the prime, they are long enough to ensure that our results could not stem from the participants holding the primes in their working memories until they could process the target.

The remainder of the trials per block included 36 bisyllabic real-word fillers (of which six were repeated), and 48 bisyllabic pseudo-word fillers (of which 18 were repeated). Finally, six real-word fillers and six pseudo-word fillers were used for practice trials, with two real-word fillers and two pseudo-word fillers being repeated. The practice trials

Table 1: Average absolute temporal differences (in milliseconds) between primes (voiced and devoiced) and targets (voiced and devoiced) per condition. Standard deviations are given between parentheses.

Condition	*Match*	*Mismatch*
AB	44 (25)	53 (36)
AA	0	49 (30)
BB	0	50 (35)

were the same for all the participants, and they were very similar in frequency of occurrence and phonological structure to the stimuli in the experiment.

We created five pseudo-randomizations of the trials: a block never started with an experimental word; there were never two experimental words in a row; there were never more than eight pseudo-word fillers in a row; and a prime and a target were never separated by more than 100 trials. For each pseudo-randomization, we then created four different stimulus lists that kept the trial order obtained by pseudo-randomization constant and differed only regarding the voicing type of the experimental words. In each of the four stimulus lists, the primes and targets of half of the experimental words occurred in the same pronunciation variant (six voiced ones, and six devoiced ones), and those of the other half showed a difference in voicing (six voiced primes followed by devoiced targets, and six vice versa). Consequently, across all four stimulus lists created from one pseudo-randomization, each experimental word was tested for each of the four possible matching and mismatching combinations. Each of the 20 lists created in total were randomly assigned to two participants per condition.

In Condition AB, we used different recordings (or tokens) for the primes and the targets, so that even in case of a match, the prime (token A) and the target (token B) were different recordings. As shown in Table 1, in the condition AB, the primes and targets matched in pronunciation variant but diverged in terms of duration. In condition AA and in condition BB, we only used the tokens A and B, respectively, so that in case of a match, prime and target were the same token and thus did not differ in duration (hence the zeros in Table 1).

2.3. Procedure

The participants were tested individually in a sound attenuated booth equipped with headphones, a mouse, and a button box with stickers *JA* 'yes' / *NEE* 'no' on the buttons. The participants first signed a consent form and filled in a language background questionnaire, before doing the lexical decision task. The lexical decision task was presented with PsychoPy (Peirce, 2007). The participants were instructed to indicate as fast as possible with the button-box, using their dominant hand, whether the word they heard over the headphones was a real word in French or not. The instructions insisted that the participant did not need to know the exact meaning of the word in order to press the 'yes' button but that they had to be certain that the word occurred in French. The next trial initiated 1000 ms after the participant's answer or 3500 ms after the onset of the preceding stimulus in case the participant did not react. In order to increase motivation and discourage guessing, the participants received feedback in percentage accuracy at the end of each block. The whole experiment session lasted a little less than half an hour.

3. Results

One participant in condition AB and one participant in condition BB were removed from the dataset since their accuracy on the experimental words in the lexical decision task was below chance level (43.75 % and 33.33 %, respectively).

We analysed all the data from this study using the software R (R Development Core Team, 2007). All the trials to which the participants did not react were discarded (ten out of the 5664 experimental word trials). Accuracies were analysed by means of a linear mixed effects model for logistic regression (Jaeger, 2008), for which the dependent variable was the probability of a correct response. Reaction times (RTs; measured from word offset) to correct trials within 2.5 standard deviations from the targets' grand mean (345 ms; discarding 52 data points out of 1768; 3 % of the data) were analysed by means of mixed effects regression models (Baayen, Davidson, & Bates, 2008). Prior to analysis, all RTs and stimulus durations were log-transformed. Our dependent variable for the linear

mixed effect model was thus the log-transformed RT. We used item and participant as crossed random effect factors.

Our predictors of interest were Voicing (a categorical predictor indicating whether the first high vowel of the stimulus was voiced or devoiced), Condition (AB, AA, and BB) or Token (A or B), and Repetition match (i.e. whether the prime and target of the experimental word were of the same pronunciation variant). Since Condition and Token overlap considerably in terms of the variation they explain, for each model reported, we compared two variants of our best model: one using Condition and one using Token in order to select the best of the two predictors. We retained in our final model the predictor which lowered the Akaike Information Coefficient (AIC) of the model by at least two points.

Our control predictors were: log Stimulus duration, Trial number (i.e. the position of the trial in the experiment, in order to control for learning or fatigue effects), Distance (lag) between prime and target (in number of intervening trials), log RT to the previous trial (so as to control for local speed effects), and log RT on the prime. The continuous and discrete numerical predictors, that is, all the control predictors, have been centred around the mean.

We first fitted a simple main effects model with all the predictors relevant to the dependent variable. Interactions were then tested between the predictors of interest only. To obtain the most parsimonious yet adequate model, only predictors and interactions which showed significant effects (i.e. t or z with an absolute value exceeding 1.96) were retained in the final models. Predictors which were significant in an interaction, but not as main effects were kept in the models as well. Once the fixed effect structure was finalized, random slopes on item and participant were tested for all fixed effects. A random slope was kept in the final model exclusively when supported by likelihood ratio tests (i.e. $p<0.05$). Finally, following Baayen (2008), to ensure no significant effect was driven by outliers, the final RT model was refitted: RTs with residual standard errors more than 2.5 standard deviation units were excluded from the dataset of the final statistical model (49 data points were removed out of 1716; 3 % of the dataset). No predictor lost significance as a result of

this refitting of the model[2]. The p values reported were obtained with the lmerTest package version 2.0–36 (Kuznetsova, Brockhoff, & Christensen, 2017).

3.1. Accuracy data

The participants' accuracy was relatively high although not at ceiling (83.92 % overall, with 85.52 % accuracy for the pseudo-word fillers, 86.13 % for the real-word fillers, and 75.70 % for the experimental words). Participants' lower accuracy on the experimental words was probably due to the fact that the experimental words were less frequent than the real-word fillers and thus less familiar to the participants.

First occurrences

We first verified whether the participants were sensitive to the devoicing manipulation. To do so, we looked at the participants' accuracy on the primes only (N=2824), since the participants' accuracy on the targets might have been influenced by whether the targets matched or mismatched their primes. The results are presented in Table 2. The participants were significantly more accurate on the voiced (75.79 %) than on the devoiced (66.42 %) tokens A, as indicated by a simple effect of Voicing (cf. Table 2), while the difference was not statistically significant for tokens B (75.80 % accuracy on the voiced tokens and 73.49 % on the devoiced ones), as shown by releveling the variable and rerunning the model (β = 0.16, S.E.= 0.23, z= 0.67, p>0.1), and as indicated by the significant interaction between Voicing and Token (cf. Table 2). We also found a significant random slope of Voicing on Item, which indicates that the effect of Voicing was significantly larger for some items than others.

2 One of the reviewers suggested that we use the Median Absolute Deviation (MAD; Leys, Ley, Klein, Bernard, & Licata, 2013) to prune our data instead of first discarding outliers 2.5 Standard Deviations from the targets' mean RT and then discarding again outliers deviating more than 2.5 standard units from the predicted values before re-fitting the model. An analysis of our RT data using the MAD is provided in Appendix 2. Importantly, both analyses find the same predictors significant. Thus, both analyses come to the same conclusions.

Table 2: Statistical model fitting the probability of a correct response to the primes. N = 2824. Standard error is indicated by SE. The intercepts represent devoiced A tokens' first occurrences. Predictors and random slopes that did not reach significance at the 5 % level were not retained in the model and are not listed in the table.

Fixed effects		**B**	**SE**	**z**	**p<**
(intercept)		0.96	0.29	3.36	0.001
Token	B	0.46	0.19	2.41	0.05
Voicing	voiced	0.63	0.19	3.26	0.01
Voicing * token	voiced * B	-0.48	0.21	-2.29	0.05
Random effects			**Variance**	**SD**	
Item	Intercept		1.65	1.28	
	voicing		0.49	0.70	
Participant	Intercept		0.40	0.63	

In sum, the participants were thus clearly sensitive to the devoicing manipulation for the tokens A, but not for the tokens B. That is, to the participants, the devoiced and voiced tokens A were more distinguishable from one another than the devoiced and voiced tokens B, although this was more the case for some experimental words than for others.

Second occurrences

Given that the participants were sensitive to the devoicing manipulation (at least for the tokens A), we can now investigate whether the participants were more accurate on matching than on mismatching targets. When only considering the targets whose primes were answered to correctly (N=2041), there appeared to be no effect of Repetition match on accuracy, neither as a main effect nor in interaction with Condition or Token.

3.2. Reaction Time data

The RT data suggest priming across all conditions (cf. Figure 2): when the participants correctly classified both the prime and the target of the experimental word as real words, they were on average 106 ms faster on the target (345 ms) than on the prime (451 ms). Note that all RTs are from word offset.

We analysed statistically the RTs to the targets answered to correctly, provided their primes had also been answered to correctly. The results

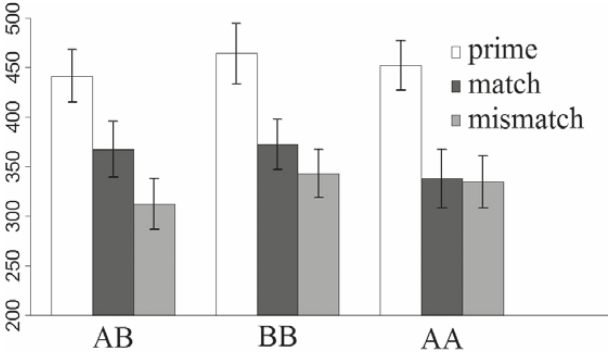

Figure 2: Reaction times (in milliseconds) from word offset for the experimental primes and targets (in match and mismatch cases) when both have been answered to correctly, by condition. Error bars: 95 % confidence intervals. N = 3454.

are presented in Table 3. Almost all our control predictors showed significant effects. The participants were faster at answering targets when they also answered quickly on the previous trial; when they had recognized the prime quickly; when the number of intervening trials between prime and target was low; and when the stimuli were short.

More importantly, all of our factors of interest also showed significant effects. The effect of Repetition match differed between the conditions AB *vs.* AA ($\beta = 0.17$, S.E. = 0.05, $z = 3.27$, $p<0.01$) and BB *vs.* AA ($\beta = 0.15$, S.E. = 0.05, $z = 3.00$, $p<0.01$), as shown by releveling the variable and rerunning the model. Given that the conditions AB and BB thus patterned together against the condition AA (cf. Figure 2), it is not surprising that Token of the target (A or B) was a much better predictor than Condition (the model with Token had an AIC ten points lower than the AIC of the model using Condition).

We also found a main effect of Voicing (see Table 3), without an interaction of Voicing with Token: participants were slower at processing devoiced targets, independently of whether the targets were token A or token B (cf. Figure 3). That is, contrary to the *Accuracy data,* which showed that the participants were only sensitive to devoicing for the tokens A, the RT data show that the participants were sensitive to the devoicing manipulation for both tokens A and tokens B. L2 listeners were thus sensitive to L1

Table 3: Statistical model fitting the log-transformed response times (measured from word offset) to the targets provided their corresponding primes had been answered to correctly. N = 1667 after removal of the outliers. Standard error is indicated by SE. The intercept represents the reaction time to a devoiced target A mismatching its prime. Predictors and random slopes that did not reach significance at the 5 % level were not retained in the model and are not listed in the table.

Fixed effects		β	SE	t	p<
(intercept)		5.76	0.05	117.40	0.001
Repetition match	Match	-0.09	0.04	-2.05	0.05
Token	B	-0.02	0.05	-0.41	n.s.
Voicing	Voiced	-0.16	0.04	-3.96	0.001
Number of trials between prime and target		0.001	0.0006	2.23	0.05
Stimulus duration (ms logged)		-1.00	0.14	-6.93	0.001
RT to the preceding trial (ms logged)		0.14	0.02	7.17	0.001
RT to the prime (ms logged)		0.26	0.02	11.05	0.001
Repetition match * Voicing	match * voiced	0.10	0.04	2.31	0.05
Repetition match * token	match * B	0.12	0.04	2.68	0.01
Random effects			**Variance**	**SD**	
Item	Intercept		0.02	0.14	
	Voicing		0.02	0.14	
Participant	Intercept		0.03	0.16	
	RT to the prime		0.02	0.13	
	RT to the preceding trial		0.01	0.10	
Residual			0.17	0.42	

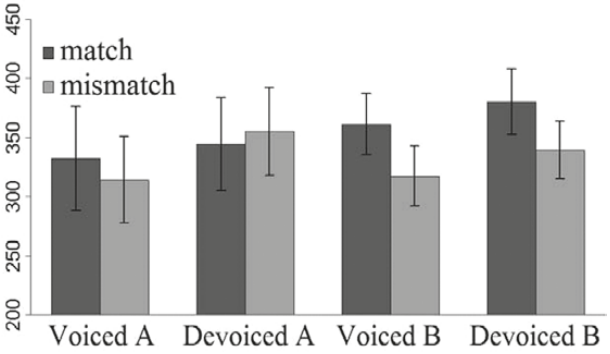

Figure 3: Reaction times (in milliseconds) from word offset for the experimental targets which have been answered to correctly both at prime and target, grouped by voicing and by token. Error bars: 95 % confidence intervals per bar. N=1727.

irrelevant information. Interestingly, the significant main effect of Token without an interaction of Voicing and Token indicates that the tokens B were processed significantly faster (i.e. were easier to comprehend for the participants) than the tokens A, independently of whether the tokens were voiced or not.

Repetition match was significant in interaction with Token on the one hand (as previously mentioned) and with Voicing on the other hand. The three-way interaction was not significant ($\chi^2(2) = 0.79$, $p>0.1$). The significant simple effect of Repetition match indicates that when the target was the devoiced token A, the participants were faster at answering the target when it matched its prime than when it mismatched it prime.

The significance of Repetition match in the other three cases (i.e. when a devoiced B target matched its prime, when a voiced B target matched its prime, and when a voiced A target matched its prime), is difficult to assess from Table 3 given the separate significant simple effects of Voicing and Token on the one hand, and their significant interactions with Repetition match on the other hand. In order to understand the overall effect of Repetition match, we analysed the different contrasts using releveling. By releveling, the model does not change, but the mathematical formulation makes it possible to determine the simple effects in the other three cases. We placed alternatively on the intercept of the model reported

in Table 3, the voiced tokens A, the voiced tokens B, and the devoiced tokens B. Only for the voiced tokens B did we find a significant effect of Repetition match ($\beta = 0.17$, S.E. = 0.03, $z = 5.02$, $p<0.001$), indicating that the participants were significantly slower when the voiced targets B matched their primes (either voiced primes A or voiced primes B). In other words, the participants were significantly faster when the voiced targets B mismatched their primes than when the voiced targets B matched their primes. For both the voiced tokens A and the devoiced tokens B, the main effect of Repetition match was not significant.

In sum, the participants were sensitive to the devoicing manipulation as they were less accurate on the devoiced than on the voiced primes A, and they were slower on both the devoiced A and B targets than on the voiced A and B targets. Repetition match showed no effect in the *Accuracy data,* possibly because of lack of statistical power. In the RT data, the A and B tokens patterned differently regarding the effect of repetition match: the devoiced A tokens were answered to faster when they were preceded by a matching prime, while the voiced B tokens were answered to significantly faster when they were preceded by a mismatching prime. In other words, devoiced primes always shortened the participants' RTs on the targets, while voiced primes never led to any significant differences in RTs between a matching and a mismatching target.

4. General discussion

This study investigated whether L2 learners show exemplar effects for variation in the acoustic signal that they are not familiar with from their L1. If exemplars are formed after the L1 phonological filter applies, L2 exemplars do not differ from L1 exemplars regarding indexical variation, but only regarding L1 irrelevant linguistic variation.

We tested Dutch intermediate learners of French in a lexical decision task in which words were repeated (i.e. using long-term implicit repetition priming) in the same (match) or in a different (mismatch) pronunciation variant. Our experimental words were French words whose first vowel was voiced in one pronunciation variant (voiced word tokens) and devoiced in the other (voiceless word tokens). Vowel devoicing is not a characteristic of Dutch and thus linguistically irrelevant for Dutch native listeners.

In order to investigate whether the match effect is not only present under the conditions normally tested in exemplar experiments, but also under more ecologically valid conditions, we tested three conditions. In two conditions (AA and BB), the prime and target were identical tokens in the pronunciation match case. These two conditions follow the vast majority of the previous literature on exemplar effects (which reuses the same token). In a third condition (AB), the primes and targets were always different instantiations, so that, even when an experimental word was repeated as a pronunciation variant match, it was nevertheless a different token, just like in everyday conversations.

Our data suggest an exemplar effect for the devoiced A targets, since the devoiced A tokens were answered to faster when they were preceded by devoiced A primes than by voiced A primes. This match effect shows that L2 listeners are able to encode and store in the form of exemplars phonetic variation that does not occur regularly in their L1 (vowel devoicing). Exemplars thus seem to be formed before the phonological filter applies and to faithfully represent the acoustic signal. The information they encode is probably the same for both native and non-native listeners.

If exemplars are formed before the phonological filter applies, one may wonder whether exemplars are part of the mental lexicon. This question has also been raised by Goldinger (2007), Cutler, Eisner, McQueen, and Norris (2010), Ramus, Peperkamp, Christophe, Jacquemot, Kouider, and Dupoux (2010), and Nijveld, ten Bosch, and Ernestus (2015), among several authors, who hypothesise that exemplars are stored in episodic memory, which is a general type of memory (Tulving, 1985). Episodic traces are detailed memory representations which are context-dependent in the sense that they encode specific events (e.g. listening to a word, watching a movie, hurting one's toe) with their context (e.g. which voice uttered the word, in which row one was seated, how early it was). If exemplars are faithful representations of the acoustic signal, they are likely to be part of episodic memory.

The significant interaction we found between Repetition match and Token indicates that our participants used different processes to comprehend the B and the A tokens. Although conditions AA and BB both used identical tokens for matching primes and targets, they did not pattern in the same way in the participants' RT behaviour on the targets. Rather, the BB condition patterned with the AB condition. In both conditions,

there were no exemplar effects. It is thus not the fact that the prime and the target were identical that led to exemplar effects. These results are in contrast with all previous studies on exemplar effects, including Hanique et al.'s (2014), which showed that exemplar effects can also arise when the prime and target are different tokens in the match condition.

Various explanations have been put forward to explain why exemplar effects arise in certain conditions and not in others. One hypothesis is that exemplar effects occur when speech processing is slow, such as when listening to dysarthric speech (Mattys & Liss, 2008), or when real words need to be distinguished from very real-word-like pseudowords (McLennan & Luce, 2005). This time-course hypothesis (McLennan & Luce, 2005) can explain the presence *versus* absence of exemplar effects as the participants were slower on the targets A than on the targets B.

The time-course hypothesis, however, cannot account for mismatch effects. Our data showed one mismatch effect. Participants responded more slowly to voiced B tokens when they were preceded by voiced than devoiced tokens. This raises the question of where this effect comes from. This is an important question since it may provide some insight into the conditions leading to exemplar effects, and therefore to the nature of exemplar effects. The difference in results between conditions AA (match effect for devoiced tokens) and BB (mismatch effect for voiced tokens) is the most interesting one, since both conditions used identical tokens for prime and target and it is therefore not obvious what drives the difference in response pattern.

It may be the case that the difference in response pattern is due to subtle acoustic differences between the set of A tokens and the set of B tokens. The voiced and devoiced tokens were probably more different from each other in condition AA than in condition BB. The selection of the tokens for the primes and target for condition AA was made before the selection of the tokens for condition BB and from the same pool of recordings. Consequently, for the cross-splicing of tokens B, the first author had fewer recordings to choose from than for the cross-splicing of tokens A, which probably caused voiced and devoiced tokens A to be better matched than voiced and devoiced tokens B on other acoustic characteristics than devoicing. This was definitely true for stimulus duration (cf. Appendix 1): the voiced and devoiced tokens A only differed by 2ms

on average, while the voiced and devoiced tokens B differed by 28.5 ms on average[3].

To further investigate potential differences between the voiced and devoiced tokens which might have caused our asymmetric results in the AA and BB conditions, we conducted a post-hoc spectral comparison of all voiced and devoiced tokens, using the differences along the Mel Frequency Cepstral Coefficients alignment path, time warped. The results are summarized in Appendix 3. We found that the voiced and devoiced A tokens differed more from each other than the voiced and devoiced B tokens. However, this difference was not significant ($t(45) = -0.27$, $p>0.1$) probably because of lack of statistical power. Consequently, it is possible that the difference between voiced and devoiced vowels stood out less clearly for the B tokens than for the A tokens, especially given the accuracy differences found between the voiced and the devoiced primes: the participants were about 9 %, and significantly more accurate on the voiced than on the devoiced primes A, but only 2 % more accurate on the voiced than on the devoiced primes B, and this latter difference was not statistically significant.

The participants' significantly lower accuracies on the devoiced A primes compared to all other primes, in combination with their significantly lower RTs on both the A and B devoiced targets compared to the voiced targets could explain our pattern of results. On the one hand, the difficulty of processing of both the A and B devoiced tokens could have led the participants' abstract representation to reach a higher level of activation (as activation only increases over time, e.g. Norris & McQueen, 2008) than after the processing of a voiced prime (for which activation stopped to increase as the word was recognized earlier in time). When a voiced target then followed a devoiced prime, the ease of processing of the voiced forms combined with the high activation of the abstract representation, led to a quicker answer on a mismatching than on a matching target. On the other hand, the fact that the devoiced A primes were particularly difficult to comprehend could have led to stronger individual

3 This difference in stimulus duration probably stems from a difference in the duration of the high vowel (cf. Figure 1). Importantly, 28.5 ms are above the threshold of just noticeable differences for vowel duration (Quené, 2007; Nooteboom & Doodeman, 1980).

memory traces (or exemplars) being encoded for the devoiced A than for the devoiced B primes. In turn, these highly activated exemplars would then be easy to retrieve and to match to the particularly distinguishable devoiced A targets. When both the prime and target were devoiced tokens, the participants could thus more easily use the exemplar formed with the prime in condition AA than in condition BB. This would explain why there was only a match (exemplar) effect in condition AA with devoiced targets.

This explanation of our asymmetric results would be in line with other studies which propose that listeners may display exemplar effects only under testing conditions that encourage participants to rely on their recent (or episodic) memory. Luce and Lyons (1998) found exemplar effects in an old/new categorisation task, which explicitly requires the participants to make use of their recent memory, but not in a lexical decision task. Hanique et al. (2014) only found exemplar effects in a lexical decision task when it was crystal clear to the participants that tokens were repeated (when the percentage of repeated tokens was high and the number of intervening trials between the prime and the target remained low). Moreover, they only found exemplar effects when manipulating only linguistic and not both linguistic and indexical variation within one experiment. Thus, if the stimuli included too much variation, like the tokens B in our experiment did, no exemplar stood out from the other episodic traces, and consequently no exemplar could be reused in the matching conditions.

Other types of variation have been shown to influence the presence of exemplar effects. For example, confusability between vowels categories has been shown to hinder the benefits of High-Variability training on vowels' identification (Wade, Jongman, & Sereno, 2007), while High-Variability training benefits are traditionally explained with more exemplars creating a more robust category as a cloud than individual exemplars. It thus seems that to produce effects, exemplars need to be clearly recognized or labelled by the listener as belonging to two separate clouds or categories.

So far, we have explained our results within models assuming hybrid lexicons. Some other recent models of speech perception answer the problem of the lack of invariance of the speech signal by focusing on how listeners integrate incoming information from the input with their own predictions over the same speech signal, depending on the situation. For example, in their 'ideal adapter' framework, Kleinschmidt and Jaeger (2015) propose

that listeners constantly learn from details in the speech signal to immediately adapt their expectations about the incoming input. Whereas this framework accounts well for adaptations to differences among individual speakers stemming from regular and suprasegmental variation within the speech input, it is less clear which predictions it would make with regard to adaptation to irregular phonetic variation. In our study, participants probably noticed that words were repeated, however, they could certainly not predict whether the target would match or mismatch its prime. In the absence of certainty, we may expect listeners not to adapt, and thus to rely on their abstract representations, representing the full forms. Consequently, voiced B targets should benefit from a matching voiced prime (meeting the listeners' long-term expectations of the listeners). However, this is not what we found. In the AB and BB conditions, a mismatching prime speeded the recognition of its voiced target. Our design, however, is not best suited to test the predictions of the 'ideal adapter' model. More studies manipulating irregular phonetic variations with more predictable stimuli are needed to test predictive models of speech perception.

Finally, our results strongly support Hanique et al. (2014)'s claim that exemplars probably play a very limited role in everyday speech comprehension given that in our study, not only exemplar effects arose in very limited conditions, but we also found significant mismatch effects (i.e. the use of abstract representations), even in the very conditions which were expected to trigger exemplar effects. It is currently assumed that exemplars are used for speech comprehension. However, given Hanique et al.'s result, our results, and the many null results reported in the exemplar literature (e.g. Luce & Lyons, 1998; McLennan et al., 2003; Mattys & Liss, 2008; Hanique et al., 2014, Nijveld et al. 2015), it is quite clear that exemplar effects are not so robust. Researching the exact conditions which can consistently trigger exemplar effects is essential in order to find which role exemplars actually play in everyday speech perception.

5. Conclusion

Exemplar effects can also be found for L2 learners, even when the prime and target encode phonetic information that does not occur regularly in the learners' L1. This shows that exemplars can encode information that the phonological filter usually discards, and exemplars must therefore be

formed before the phonological filter applies. Exemplars are thus probably not part of the mental lexicon. Interestingly, we also found that participants displayed different response patterns when presented with different tokens of the same words in exactly the same testing conditions. This finding particularly questions the robustness of exemplar effects. Hanique et al. (2014) already warned that exemplars are probably not used in everyday speech comprehension given the limited conditions under which exemplar effects arise. Our study supports this conclusion and extends it to L2 listeners for whom the conditions under which exemplar effects arise appear even more limited.

Acknowledgements

This work was supported by the European Research Council under Grant ERC-2011-StG and the Netherlands Organization for Scientific Research under VICI grant 277-70-010, both awarded to the third author.

Appendix 1: Experimental word-types (with their translations) classified by their high vowel, with their token durations (in ms) and frequencies of occurrence (per million words) as reported for movie subtitles in the database Lexique3. Standard Deviations from the mean are reported between parentheses.

		A		**B**		**Frequency**
High-vowel	**Word-types**	voiced	devoiced	voiced	devoiced	Freqfilm2
/i/	le chinois *the Chinese language*	660	730	601	694	21.88
	la cité *the city*	753	650	728	682	14.55
	le citron *the lemon*	681	655	613	599	8.10
	le cycliste *the cyclist*	943	923	919	884	57.46
	le kilo *the kilo*	921	871	907	955	24.77
	le pilote *the pilot*	944	897	979	895	70.70
	la piscine *the swimming pool*	933	1028	910	972	85.08
	le silence *the silence*	898	966	925	1019	18.76
	le ticket *the ticket*	903	865	963	959	0.71
/y/	la cuisine *the kitchen*	691	655	632	665	19.91
	la culture *the culture*	853	884	925	921	25.73
	la fumée *the smoke*	660	710	668	693	5.19
	le futur *the future*	883	958	846	903	29.10
	la purée *the mashed potatoes*	933	1028	910	972	22.19
	le succès *the success*	821	763	811	862	14.85
	le sujet *the subject*	700	787	762	741	32.33
	le surnom *the nickname*	740	752	654	794	22.05
	la tulipe *the tulip*	765	763	762	871	5.74

Appendix 1: Continued

		A		B		Frequency
/u/	la couleur *the colour*	967	1003	1012	1071	105.53
	le couloir *the corridor*	766	683	704	740	39.58
	le courage *the courage*	725	713	694	709	107.92
	la poubelle *the garbage (can)*	690	677	641	632	6.20
	le poulet *the chicken*	644	626	616	664	13.62
	la poupée *the doll*	868	804	903	872	1.53
Average (SD)		**806 (110)**	**808 (130)**	**795 (136)**	**824 (137)**	**31.40 (31)**

Appendix 2: Statistical model fitting the log-transformed response times (measured from word offset) to the targets whose corresponding primes have been answered to correctly. N = 1647 after removal of the outliers that are 2.5 absolute deviations lower or higher than the median. Standard error is indicated by SE. The intercept represents the reaction time to a devoiced target A mismatching its prime.

Fixed effects		**β**	**SE**	**t**	**p<**
(intercept)		5.74	0.05	113.00	0.001
Repetition match	Match	-0.12	0.04	-2.82	0.01
Token	B	-0.05	0.05	-1.10	n.s.
Voicing	Voiced	-0.18	0.04	-4.35	0.001
Number of trials between prime and target		0.002	0.0006	2.69	0.01
Stimulus duration (ms logged)		-1.13	0.16	-7.09	0.001
RT to the preceding trial (ms logged)		0.17	0.03	6.23	0.001
RT to the prime (ms logged)		0.30	0.02	15.89	0.001
Repetition match * voicing	match * voiced	0.12	0.04	2.90	0.01
Repetition match * token	match * B	0.16	0.04	3.68	0.001
Random effects			**Variance**	**SD**	
Item	Intercept		0.02	0.15	
	Voicing		0.02	0.14	
	RT to the preceding trial		0.007	0.08	
Participant	Intercept		0.04	0.19	
	Stimulus duration		0.13	0.37	
	RT to the preceding trial		0.01	0.11	
Residual			0.17	0.42	

Appendix 3: Average spectral differences along the Mel Frequency Cepstral Coefficient alignment path between primes (voiced and devoiced) and targets (voiced and devoiced) time warped per condition. Standard Deviations are reported between parentheses.

Condition	Prime	Target	Condition	Spectral differences
AB	Voiced A	Voiced B	Match	755 (109)
			Match	813 (120)
	Devoiced A	Devoiced B		
	Voiced A	Devoiced B	Mismatch	1091 (207)
			Mismatch	1092 (256)
	Devoiced A	Voiced B		
AA	Voiced A	Voiced A	Match	0
	Devoiced A	Devoiced A	Match	0
	Voiced A	Devoiced A	Mismatch	
	Devoiced A	Voiced A	Mismatch	1044 (197)
BB	Voiced B	Voiced B	Match	0
	Devoiced B	Devoiced B	Match	0
	Voiced B	Devoiced B	Mismatch	
	Devoiced B	Voiced B	Mismatch	1028 (218)

References

Baayen, H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, MA: Cambridge University Press.

Baayen, H., Davidson, D., & Bates, D. (2008). Mixed-effects with crossed random effects for subject and items. *Journal of Memory and Language*, *59*(4), 390–412.

Boersma, P., & Weenink, D. (2017). *Praat: doing phonetics by computer [Computer program]. Version 6.0.31*. Retrieved August, 21, 2017 from http://www.praat.org/

Bradlow, A., Nygaard, L., & Pisoni, D. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219.

Bradlow, A., & Pisoni, D. (1999). Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085.

Council of Europe. (2011). *Common European framework of reference for languages: Learning, teaching, assessment*. Retrieved September 3, 2017 from https://www.coe.int/en/web/common-european-framework-reference-languages/

Craik, F., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, *26*(2), 274–284.

Cristia, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of implicitly learnt sound patterns, *Laboratory Phonology*, *4*(2), 259–285.

Cutler, A., Eisner, F., McQueen, J., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In Fougeron C., Kühnert B., d'Imperio M., & Vallée N. (Eds.), *Papers in Laboratory Phonology 10*, Berlin: Mouton de Gruyter, pp.91–111.

Darcy, I., Dekydtspotter, L., Sprouse, R., Glover, J., Kaden, C., McGuire, M., & Scott, J. (2012). Direct mapping of acoustics to phonology: On the lexical encoding of front rounded vowels in L1 English–L2 French acquisition. *Second Language Research*, *28*(1), 5–40.

Díaz, B., Mitterer, H., Broersma, M., & Sebastián-Gallés, N. (2012). Individual differences in late bilinguals' L2 phonological processes: From acoustic-phonetic analysis to lexical access. *Learning and Individual Differences*, *22*(6), 680–689.

Ernestus M., & Warner N. (2011). An introduction to reduced pronunciation variants [Editorial]. *Journal of Phonetics 39*(SI), 253–260.

Goldinger S. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *22*(5), 1166–1183.

Goldinger S. (2007). A complementary-systems approach to abstract and episodic speech perception, In *Proceedings of the 16th International Congress of Phonetic Sciences*, 49–54.

Hanique I., Aalders E., & Ernestus M. (2014). How robust are exemplar effects in word comprehension? *The Mental Lexicon*, *8*(3), 269–294.

Jaeger F. (2007). Categorical data analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, *59*(4), 434–446.

Kleinschmidt D., & Florian J. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review*, *122*(2), 148–203.

Kuznetsova A., Brockhoff P., & Christensen R. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26.

Leys C., Ley C., Klein O., Bernard P., & Licata L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, *49*(4), 764–766.

Luce P., & Lyons E., (1998). Specificity of memory representations for spoken words. *Memory and Cognition*, *26*(4), 708–715.

Mattys S., & Liss J. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. *Perception and Psychophysics*, *70*(7), 1235–1242.

McLennan C., & Luce P. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321.

McLennan C., Luce P., & Charles-Luce J. (2003). Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 539–553.

Meunier C., Meynadier Y., & Espesser R. (2008). Voyelles brèves en parole conversationnelle, In *Proceedings of Journées d'Etude sur la Parole (JEP)*, 97–100.

New B., Pallier C., Ferrand L., & Matos R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE, Retrieved December, 4, 2016 from http://www.lexique.org. *L'Année Psychologique*, *101*(3), 447–462.

Nijveld A., ten Bosch L., & Ernestus M. (2015). Exemplar effects arise in a lexical decision task, but only under adverse listening conditions, The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*. Glasgow: University of Glasgow.

Nooteboom S., & Doodeman G. (1980). Production and perception of vowel length in spoken sentences. *The Journal of the Acoustical Society of America*, *67*(1), 276–287.

Norris D., & McQueen J. (2008). ShortlistB: A Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Norris D., McQueen J., & Cutler A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

Pallier C., Colomé A., & Sebastián-Gallés N. (2001). The influence of native-language phonology on lexical access: exemplar-based vs. abstract lexical entries, *Psychological Science*, *12*(6), 445–449.

Palmeri T., Goldinger S., & Pisoni D. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

Peirce J. (2007). PsychoPy – Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13.

Pierrehumbert J. (2002). Word-specific phonetics. In Gussenhoven C., & Warner N. (Eds.), *Laboratory Phonology VII*, Berlin, Germany: Mouton de Gruyter, pp. 101–139.

Quené H. (2007). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*(3), 353–362.

R Development Core Team. (2007). *R: A Language and Environment for Statistical Computing*, Retrieved September, 14, 2016 from http://www.R-project.org. Vienna, Austria: R Foundation for Statistical Computing.

Ramus F., Peperkamp S., Christophe A., Jacquemot C., Kouider S., & Dupoux E. (2010). A psycholinguistic perspective on the acquisition of phonology. In Fougeron C., Kühnert B., d'Imperio M., & Vallée N. (Eds.), *Laboratory Phonology 10: Variation, Phonetic Detail and Phonological Representation*. Berlin, Germany: Mouton de Gruyter, pp. 311–340.

Sebastián-Gallés N., & Baus C., (2005). On the relationship between perception and production in L2 categories. In Cutler A. (Ed.), *Twenty-First Century Psycholinguistics: Four Cornerstones*, New York, NY: Erlbaum, pp.279–292.

Torreira, F. and Ernestus, M. (2010). Phrase-medial vowel devoicing in spontaneous French. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (Interspeech 2010)*, Makuhari, Japan, 2006–2009.

Trofimovich P. (1995). Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics*, *26*(4), 479–504.

Troubetzkoy N. (1939/1969). *Principles of Phonology*. Translation Baltaxe Christiane. Berkeley and Los Angeles, CA: University of California Press.

Tulving E., & Schacter D. (1990). Priming and human memory systems, *Science*, *247*(4940), 301–306.

Tulving E. (1985). How many memory systems are there? *American Psychologist*, *40*(4), 385–398.

Wade T., Jongman A., & Sereno J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2–3), 122–144.

Winters S., Lichtman K., & Weber S. (2013). The role of linguistic knowledge in the encoding of words and voices in memory. In Voss E., Tai S.D., & Li Z. (Eds.), *Selected Proceedings of the 2011 Second Language Research Forum: Converging Theory and Practice*, Sommerville, MA: Cascadilla Proceedings Project, pp. 129–138.

Zimmerer F., Yasuda R., & Henning R., (2013). Architekt or Archtekt? Perception of devoiced vowels produced by Japanese speakers of German, *In Proceedings of the 14th Annual Conference of the International Speech Communication Association (Interspeech)*, 417–420.

Speech Production and Perception

Edited by Susanne Fuchs and Pascal Perrier

Vol.	1	Susanne Fuchs / Melanie Weirich / Daniel Pape / Pascal Perrier (eds.): Speech Planning and Dynamics. 2012.
Vol.	2	Anne Hermes: Articulatory Coordination and Syllable Structure in Italian. 2013.
Vol.	3	Susanne Fuchs / Daniel Pape / Caterina Petrone / Pascal Perrier (eds.): Individual Differences in Speech Production and Perception. 2015.
Vol.	4	Louis-Jean Boë / Joël Fagot / Pascal Perrier / Jean-Luc Schwartz (eds.): Origins of Human Language: Continuities and Discontinuities with Nonhuman Primates. 2017.
Vol.	5	Jessica Di Napoli: The Phonetics and Phonology of Glottalization in Italian. 2018.
Vol.	6	Susanne Fuchs / Joanne Cleland / Amélie Rochet-Capellan (eds.): Speech production and perception: Learning and memory. 2019.

www.peterlang.com