Abstract

Audio-tactile Integration in Speech Perception: Effects of Aero-Tactile Information on the Perception of Voicing in American English and Thai

Dolly Goldenberg

2019

This dissertation provides evidence for audio-tactile integration in the perception of speech using aero-tactile stimuli, shows that somatosensory information is integrated with auditory information during speech perception only when it is task relevant, and establishes that aero-tactile information is interpreted by listeners as aspiration during multimodal integration in speech perception.

Three experiments were conducted. The First experiment, outlined in Chapter 2, evaluated the effect of air puffs on two VOT continua, bilabial and velar, and a vowel continuum used as a control. The presence of air puffs was found to significantly increase the likelihood of choosing voiceless responses for the two VOT continua but had no effect on choices for the vowel continuum. At the same time, the responses to the VOT continua were reflective of the distinction function expected according to the acoustic stimuli. This indicates that during the decision-making process, both auditory and aero-tactile inputs were taken into consideration, suggesting that this is indeed an example of multisensory integration.

The second experiment, outlined in Chapter 3, evaluated the effect of aero-tactile

information on the perception of medial stops in American English. This case study was chosen because VOT differences are not typically used for disambiguating stop voicing contrasts in this context. We hypothesized that aero-tactile information is associated with aspiration and concomitant long positive VOT, and thus predicted that it is not expected to shift perception toward voicelessness in the case of medial positions in English. No shift was found in the perception of the continuum for any the participants. However, 40% of the participants in this experiment showed a priming effect where a bias towards voicelessness was found for all responses, regardless of the presence of puffs of air.

The third experiment, outlined in Chapter 4, evaluated the effect of aero-tactile information on perception of an initial VOT continua in Thai. Thai exhibits a three-way voicing contrast, with aspirated voiceless stops, unaspirated voiceless stops, and voiced stops. We hypothesized that the aero-tactile stimuli are perceived as aspiration, and thus predicted that they will shift the perception of voicelessness in Thai only in the case where aspiration is a cue for the voicing distinction. That is, in the comparison between aspirated voiceless stops and unaspirated voiceless stops, but not in the comparison between unaspirated voiceless stops and voiced stops. Indeed, we found that speakers of Thai were affected by the air puffs in the comparison between /p^ha/ and /pa/ but not in the comparison between /pa/ and /ba/.

Audio-tactile Integration in Speech Perception: Effects of Aero-Tactile Information on the Perception of Voicing in American English and Thai

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Dolly Goldenberg

Dissertation Director: Jason A. Shaw

December 2019

Table of Contents

List of Figures 7

List of Tables 8

Acknowledgements 9

Chapter 1: Introduction 12

References 20

Chapter 2: Multimodal Integration in Speech Perception: The Effect of Aero-Tactile Information on Perception of VOT Continua 28

2.1 Introduction 28

2.2 Methods 35

2.2.1 Participants 35

2.2.2 Stimuli 36

2.2.2.1 Acoustic Stimuli 36

2.2.2.2 Tactile (Air Puff) Stimuli 40

2.2.3 Procedure 41

2.2.3.1 Puff Detection Test 42

2.2.3.2 Perturbed Continua Testing 44

2.3 Results 45

2.3.1 Puff Detection Test 45

2.3.2 Perturbed Continua Testing 45

2.3.2.1 Quantifying the effect of puffs on perceived categories 48

2.3.2.2 Comparison of Effect Sizes for the Three Continua 49

2.3.2.3 Analysis of Individual Results 50
2.3.2.4 Analysis of Response Times 51
2.4 Discussion 54
2.5 Conclusion 63
References 64
Chapter 3: The Effect of Aero-Tactile Information on Perception of VOT Continua 75
3.1 Introduction 75
3.2 Methods 78
3.2.1 Participants 78
3.2.2 Stimuli 78
3.2.2.1 Acoustic Stimuli 78
3.2.2.2 Tactile (Air Puff) Stimuli 81
3.2.3 Procedure 81
3.2.3.1 Puff Detection Test 81
3.2.3.2 Perturbed continuum Testing 82
3.3 Results 82
3.3.1 Puff Detection Test 82
3.3.2 Perturbed Continuum Testing 83
3.3.2.1 Quantifying the effect of puffs on the perceived categories of the participants who showed the expected baseline 83
3.3.2.2 Quantifying the effect of puffs on the perceived categories of the participants who did not show the expected baseline 86

3.3.2.3 Assessment of possible priming effect of puffs on the perceived categories of the participants who did not show the expected baseline 88
3.3.2.4 Assessment of a possible learning effect 90
3.4 Bayesian Analysis as a Tool for Quantifying the Variable Behavior 91
3.4.1 Two Patterns of Behavior 91
3.4.2 Quantifying the Differences between the Groups 93
3.5 Discussion 96
3.6 Conclusion 104
References 105
Chapter 4: Audio-Tactile Integration in the Perception of Thai 111
4.1 Introduction 111
4.2 Methods 116
4.2.1 Participants 116
4.2.2 Stimuli 116
4.2.2.1 Acoustic Stimuli 116
4.2.2.2 Tactile (Air Puff) Stimuli 120
4.2.3 Procedure 120
4.2.3.1 Puff Detection Test 120
4.2.3.2 Perturbed continuum Testing 121
4.3 Results 122
4.3.1 Puff Detection Test 122
4.3.2 Perturbed Continuum Testing 122
4.3.2.1 Quantifying the effect of puffs on perceived categories 124

4.3.2.2 Analysis of Individual Results 126
4.4 Discussion 127
4.5 Conclusion 131
References 132
Chapter 5: Audio-Tactile Integration in Speech Perception and the Phonological Representation of Voicing 138
5.1 Aero-Tactile Stimuli as Phonologically Relevant Information 138
5.2 Theoretical Approaches to Phonological Representation 140
5.2.1 Feature-Based Approaches 141
5.2.1.1 The Standard Approach 141
5.2.1.2 Laryngeal Realism 142
5.2.2 Articulatory Phonology: A Gesture-Based Approach 143
5.3 Aero-Tactile Integration and the Phonological Representation of Voicing . 145
References 150
Chapter 6: Conclusion 155

List of Figures

Chapter 2

2.1 Viability test results 38

2.2 The aero-tactile stimulus presentation system 40

2.3a Puff delivery setup 43

2.3b Puff detection test setup 43

2.4 Perceived category boundaries for the initial continua 47

2.5 Comparison of mean $\log_{10}$ response times averaged across participants 54

Chapter 3

3.1 Viability test results 80

3.2 Perceived category boundaries for the Expected Baseline group 84

3.3 Perceived category boundaries for the Primed group 87

3.4 Perceived category boundaries for Primed, experiment vs. post testing 89

Chapter 4

4.1 Viability test results, ... 119

4.2 Perceived category boundaries for the Thai continua 124

Chapter 5

5.1 Gestural Score of the English word palm 145

List of Tables

Chapter 2

2.1 Average VOT durations for American English stops 34

2.2 VOT continua steps showing length of retained aspiration 37

2.3 Output of the GLMM response model 49

2.4 Output of GLMM combining continua to show relative effect sizes 50

2.5 Summary of the individual models 51

2.6 Output of LMM predicting $\log_{10}$ response times 53

Chapter 3

3.1 Voicing continuum steps showing length of retained silent closure 79

3.2 Output of the GLMM response model 85

3.3 Summary of the individual models computed for the participants 86

3.4 Output of the GLMM response model 88

3.5 Output of the GLMM response model 90

Chapter 4

4.1 Voicing continuum steps showing length of aspiration for the continua steps ... 118

4.2 Output of the GLMM response models 126

4.3 Summary of the individual models 127

Acknowledgments:

This dissertation is the culmination of a long journey during which many have guided and accompanied me. I would like to express my sincerest gratitude to each and every one of them, but first and foremost to my advisors and my committee.

To my advisors over the years, Jelena Krivokapić, Ryan Bennett and Jason Shaw: Jelena, you were my first Ph.D. advisor, and the first phonetician in my life. I have learned so much from you and will forever be grateful to you for introducing me to the field. Ryan, you have been my advisor the longest! Thank you for the endless support, comments, discussions, guidance and everything else. Thank you both for continuing your support beyond your tenure at Yale. Jason, you joined this journey quite late, but your contribution to it is undeniable. Thank you for being persistent, for Bayesian reasoning, for the weekly meetings, and for all the useful comments and suggestions.

To the members of my committee, Mark Tiede, Doug Whalen, and Natalie Weber. Mark, you have been my mentor the longest, and it has been an honor and a privilege to know you. Thank you for demonstrating a real passion for research, knowledge and problem solving, for your endless patience, for being the best editor I have ever had, for the best discussions. Thank you for super long EMA experiments, for teaching me the most important thing about data analysis ("garbage in garbage out"). Thank you for the continua, prior to discussing the topic with you I only had end points in mind! Thank you for statistics and programming languages. And for so much more. Doug, thank you for taking me in and making me a part of A93, this has been one of the

most valuable experiences of my life. The contribution of the members of A93 in general and you in particular to phonetics, speech sciences and the linguistic community has been an inspiration. Thank you for your guidance and feedback. Natalie, thank you for agreeing to join the committee so late, for insightful comments and great suggestions.

To everyone else who supported me at The Department of Linguistics at Yale University and Haskins Laboratories: Stephen Anderson, Claire Bowern, Maria Piñango, Chris McDaniel, Joe Cardone, Betty DeLise, Lisa Fresa, Ken Pugh, Tammy Ursini, Wei-Rong Chen, Jocelyn Springfield, Kevin Roon, Hosung Nam and Chris Shadle. Thank you all. Special Thanks to my fellow graduate students for their support and friendship over the years, especially Gregg Castellucci, Shira Calamaro and Chris Geissler.

Thank you everyone who assisted in running the experiments in Bangkok: Sujinat Jitwiriyanont, Rattanasuwan (Max) Rawan, Pittayawat Pittayaporn and the Linguistics Department at Chulalongkorn University.

Thank you, family and friends who were my family, during this journey: Gal Gur-Arye, Reuma Pollack-Gadassi, Kate Dawson, and Yarden Avital, for your support and love. Evyatar Shaulsky, Zoe Goldenberg-Shaulsky and Amy Goldenberg-Shaulsky, you three are the loves of my life, thank you for your support, love and incredible patience. Evyatar, thank you for your major part in designing and building the air-puffing system and for the beautiful sketch of the system. Thank you for functioning as both an engineer and a main caretaker of our daughters while we were running the experiments in Bangkok.

This dissertation was partially funded by National Institutes of Health (NIH) Grant DC-002717 to Haskins Laboratories and the MacMillan International Dissertation Research Fellowship.

Chapter 1:

Introduction

In multisensory (or multimodal) integration, information from different sensory modalities, such as sight, sound or touch, is integrated by the human perceptual and nervous system into a coherent percept (see Stein & Stanford, 2008; Stein et al., 2009; Bremner et al. 2012; Stein 2012; Spence & Bayne, 2014 for reviews). Multisensory integration is dependent on many factors, including spatial and temporal disparity of the signals, that is, how close they are in space and time (e.g., Slutsky & Recanzone, 2001; Calvert et al., 2004; Zampini et al., 2005, Plöchl et al., 2016, though see Jones & Munhall,1997; Jones & Jarick, 2006; Vroomen & Keetels, 2006 for exceptions); correspondence of the temporal patterns of the signals, that is, how similar is the way the signals are presented and changed over time (e.g., Warren, 1981; Radeau & Bertelson, 1987; Recanzone, 2003); perceptual grouping, that is, organization of the perceptual field into an object and its background, and the relative strength of unimodal versus multimodal perceptual grouping (e.g., King & Calvert, 2001; Sanabria et al., 2004; Harrar & Harris, 2007; Spence, 2015); semantic congruency of the signals (e.g., Laurienti et al., 2004; Molholm et al., 2004, though see Koppen et al., 2008 for an exception); the unity assumption, that is, beliefs about a common distal source of the signals (e.g., Bertelson et al., 1994; Arnold et al., 2005 ; Vatakis & Spence, 2007); perceived causal relation between the signals (e.g., Stetson et al., 2006; Körding et al.,

2007); and cross modal dynamic capture, that is, perception of the various inputs in motion (see Soto-Faraco & Kingstone, 2004; Soto-Faraco et al., 2004 for reviews). Understanding how these various factors combine to modulate multisensory integration under realistic conditions is an important challenge for researchers in the field. This point was raised by Spence (2007) and is still relevant today.

Multisensory integration occurs even though the input from different sensory modalities is processed at different speeds (Eagleman, 2008): for instance, auditory input reaches the cortex in less than half the time of visual input (Molholm et al., 2002). Direct comparisons of processing speeds for haptic input are more difficult, since possible contact points on the skin are distributed over the entire body, not just the area of the eyes and ears. To complicate matters further, the speed of processing is affected by factors such as stimulus intensity (e.g., Colonius, H., & Diederich, 2004). Additional factors such as previous experience (Miyazaki et al., 2006) or the way stimuli are presented (Harrar & Harris, 2008) can affect the correspondence between different sensory signals during the process of integration. How are signals associated with different sensory timing integrated into being perceived as a single coherent event? The answer might be a dynamic recalibration of expectations. Eagleman & Holcombe (2002) and Haggard et al. (2002) demonstrated that participants perceive two events from different modalities (haptic and visual, in this case) as closer temporally then they are in fact because they perceive them as part of the same event: a flash of light that appeared after the participants have pressed a button was perceived as occurring earlier than it did and closer to the button press event. Stetson et al. (2006) suggested that the participants' expectations of the relative timing of motor acts and sensory consequences can shift,

even to the extent that they can switch places: the later event can be perceived as earlier. Similarly, it is possible that sensory inputs that are processed at different speeds but associated with the same event will be part of one coherent percept.

In the perception of speech, multisensory integration is understood as a coherent speech percept constructed from combined inputs from different sensory modalities (see Rosenblum, 2005; Altieri et al., 2011; Kilian-Hütten et al., 2017 for reviews). Most multimodal research in the field of speech perception has concentrated on audio-visual integration (e.g., Sumby & Pollack, 1954; McGurk & MacDonald, 1976; Macleod & Summerfield, 1990; Ross et al., 2006). However, evidence for visuo-tactile and audio-tactile integration in the perception of speech has also been accumulating (Sparks et al., 1978; Reed et al., 1989; Bernstein et al., 1991; Fowler & Dekle, 1991; Gick et al., 2008; Gick & Derrick, 2009; Ito et al., 2009; Derrick & Gick, 2013; Bicevskis, 2015; Goldenberg et al., 2015). In the earlier studies of audio-tactile integration in speech perception the participants either had explicit knowledge of the task (Fowler & Dekle, 1991; Gick et al., 2008), or were trained to make a connection between the tactile and the auditory cues (Sparks et al., 1978; Reed et al., 1989; Bernstein et al., 1991). Later studies were conducted with uninformed and untrained listeners. However, at least in the case of the studies employing aero-tactile integration (Gick & Derrick, 2009; Derrick & Gick, 2013), it is not clear that an effect of tactile information on auditory perception has been established.

The research conducted by Gick & Derrick (2009) and Derrick & Gick (2013) studies the role of aero-tactile information in the perception of speech. It does so by testing the effect of puffs of air on the perception of Voice Onset Time (VOT). VOT is

the temporal lag between the release of a stop consonant and the onset of voicing following or preceding the release (Lisker & Abramson, 1964). It is positive when the release precedes the onset of voicing. Production of a stop consonant with long-lag positive VOT in initial positions (that is, word initially and at the onset of a stressed syllable) is typically accompanied by a stream of air produced by the speaker, also known as aspiration. We argue that the aero-tactile stimuli used in the experiments detailed in this dissertation are being interpreted by the listeners as aspiration.

Gick & Derrick (2009) studied the effect of puffs of air blown on the hand and the neck of listeners on perception of CV syllables in background noise (e.g., /pa/, /ba/). They found that the aero-tactile information affected both the identification of aspirated stops, by enhancing them, and the identification of unaspirated stops, by interfering with them. They concluded that aero-tactile information can be integrated with auditory information in the perception of speech, similar to the way visual information is integrated. In Derrick & Gick (2013), the duo used the same paradigm to test the effect of aero-tactile information from a distal point of contact on the skin. The puffs of air in this study were blown on the ankles of the participants. Comparison of the ankle results to the hand and neck results from Gick & Derrick (2009) did not reveal significant differences, leading Derrick and Gick to conclude that integration is a full-body process.

However, Massaro (2009) claims that in both studies, Gick & Derrick did not establish the existence of audio-tactile integration. The argument is that the participants could have perceived the aero-tactile stimuli unimodally, having interpreted it as aspiration, thus making their decision based on the puffs of air alone when they were provided, without integrating the auditory stimuli during their decision-making. The first

aim of this dissertation is to address Massaro's critique, providing unequivocal evidence for audio-tactile integration in the perception of speech, using aero-tactile stimuli. The auditory stimuli in Gick and Derrick (2009) and Derrick and Gick (2013) was masked by background noise, rendering the acoustic stimuli less informative than it could have been in perfect acoustic conditions. Therefore, it might have been the case that the tactile stimulus was the most prominent signal, and as a result that a unimodal response was made to it. To address this possibility, we constructed voice onset time (VOT) continua ranging from voiceless to voiced sounds (e.g., /pa/ to /ba/) rather than endpoint stimuli only (as in the work by Gick & Derrick).

Using VOT continua in our stimuli design enables us to show the existence of multimodal integration of acoustic and tactile inputs in the perception of speech, rather than a possible unimodal response. Accordingly, in the first experiment detailed below we used a bilabial VOT continuum, a velar VOT continuum, and a control vowel continuum where the varied factor was formant structure. The participants were asked to make a binary choice, deciding between the voiceless and voiced sounds (or between the vowel sounds in the case of the vowel continuum). 50% of the time a small puff of air was blown on the hand of the participants, on the dorsal skin between the thumb and the forefinger. Responses reflecting a shift in the category boundary towards voicelessness in the trials accompanied by puffs of air would support multimodal decision-making. If the choices made by the participants would be unimodal and directed solely by the presence of puffs of air when they are provided, we would expect an overwhelming tendency to choose a voiceless response in the presence of air puffs, not responses shifted towards voicelessness but overall still reflective of the expected category boundary.

Thus, we predict a shift in the category boundary for the VOT continua but not for the vowel continuum, where both sounds being disambiguated are typically produced with a similar amount of air. Such a result would show that aero-tactile information is being integrated with auditory information in the perception of speech and moreover, that this only happens when the tactile information is relevant for the disambiguation being made.

We argue that the tactile information is relevant for some cases of disambiguation but not for others. How does a speaker know when a tactile information is relevant for a perceive sound? This question is related to a bigger question: how does a human know that some properties are relevant for a perceived object while others are not? The scope of these questions goes beyond speech perception and has roots in the Gestalt literature of the early 20th century (see Wagemans et al., 2012, for review). Concepts such as perceptual grouping and front versus background (e.g., King & Calvert, 2001; Sanabria et al., 2004; Harrar & Harris, 2007; Spence, 2015) have been suggested to account for the way humans organize percepts into objects and associate inputs from different sensory modalities with these objects. The predicted shift in the category boundary for the VOT continua but not for the vowel continuum will provide additional support for the involvement of auditory and somatosensory modality in perceptual grouping and will show that this framework is appropriate for the domain of speech perception. It would show that an aspirated sound is perceptually grouped with the tactile percept of a puff of air while an unaspirated sound is not. Specifically, in the case of the vowel continuum, both vowels are typically produced with a similar amount of air. Thus, no difference in perceptual grouping of the two objects/sounds being disambiguated is expected with respect to the tactile input.

The second aim of this dissertation is to investigate the effect of the air puffs on listeners during the process of integration, by utilizing the notion that somatosensory information is integrated with auditory information only when it is task relevant. In the second experiment outlined below we tested the effect of aero-tactile information on the perception of medial stops in American English (e.g., continuum ranging from /ˈa.pa/ to /ˈa.ba/), using the same experimental setting from the first experiment. In non-initial positions VOT differences are not used as a basis for disambiguating stop voicing contrasts. We argue that the aero-tactile information provided during our experiments is associated by the participants with aspiration. Since aspiration is not relevant for the task of disambiguating medial stops in American English, the aero-tactile stimuli associated with it is predicted to have no influence on voicing judgments.

The third aim of this dissertation is to show that aero-tactile information is indeed being interpreted by listeners as aspiration during the process of integration. The third experiment described below satisfies this aim by testing the effect of aero-tactile information on perception of initial VOT continua in Thai. Thai has a three-way voicing contrast for labial and alveolar stops. At both places of articulation there are aspirated voiceless stops, unaspirated voiceless stops, and voiced stops (Lisker & Abramson, 1964). Thus, Thai speakers make use of aspiration in distinguishing aspirated voiceless stops from unaspirated voiceless stops and voiced stops, but not in distinguishing unaspirated stops from voiced stops. If air puffs are associated with aspiration, they are predicted to shift perception of voicelessness in Thai where aspiration is a cue for the voicing distinction, i.e., in the contrast between aspirated voiceless stops and unaspirated voiceless stops, but not in the contrast between unaspirated voiceless stops and voiced stops. The

data from Thai also serve to satisfy the fourth and last aim of this dissertation, expanding the set of languages in which audio-tactile integration in speech perception is shown to operate.

Chapter 1: References

Altieri, N., Pisoni, D. B., & Townsend, J. T. (2011). Some behavioral and neurobiological constraints on theories of audiovisual speech integration: a review and suggestions for new directions. *Seeing and perceiving*, *24*(6), 513.

Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research*, *45*(10), 1275-1284.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & Oc´onnell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal–hearing and hearing–impaired subjects. *Journal of the Acoustical Society of America*, *90*(6), 2971-2984.

Bertelson, P., Vroomen, J., Wiegeraad, G., & Gelder, B. D. (1994). Exploring the relation between McGurk interference and ventriloquism. In *Third International Conference on Spoken Language Processing*.

Bicevskis, K. (2015). *Visual-tactile integration and individual differences in speech perception*. (Master's thesis, The University of British Columbia). Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0166756

Bremner, A., Lewkowicz, D., and Spence, C. (Eds.). (2012). Multisensory development. Oxford: Oxford University Press.

Calvert, G., Spence, C., & Stein, B. E. (Eds.). (2004). *The handbook of multisensory processes*. Cambridge, MA: MIT press.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of cognitive neuroscience*, *16*(6), 1000-1009.

Derrick, D., & Gick, B. (2013). Aerotactile integration from distal skin stimuli. *Multisensory Research*, *26*, 405–416.

Eagleman, D. M., & Holcombe A. O. (2002). Causality and the perception of time. *Trends in cognitive sciences* 6(8). 323–325.

Eagleman, D. M. (2008). Human time perception and its illusions. *Current opinion in neurobiology*, *18*(2), 131-136.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross–modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816–828.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*, 502– 504.

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of the Acoustical Society of America*, *123*(4), EL72-EL76.

Goldenberg, D., Tiede, M. K., & Whalen, D. H. (2015). Aero–tactile influence on speech perception of voicing continua. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*. Glasgow, UK: The University of Glasgow.

Haggard P., Clark S., & Kalogeras J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience 5(*4), 382–385.

Hall, E. T. (1966). *The hidden dimension*. Garden City, NY: Doubleday.

Harrar, V., & Harris, L. R. (2007). Multimodal Ternus: Visual, tactile, and visuo-tactile grouping in apparent motion. *Perception*, *36*(10), 1455-1464.

Harrar, V., & Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. *Experimental brain research*, *186*(4), 517-524.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1245-1248.

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: the relationship between space and time. *Experimental Brain Research*, *174*(3), 588-594.

Jones, J. A., & Munhall, K. G. (1997). Effects of separating auditory and visual sources on audiovisual integration of speech. *Canadian Acoustics*, *25*(4), 13-19.

Kilian-Hütten, N., Formisano, E., & Vroomen, J. (2017). Multisensory integration in speech processing: Neural mechanisms of cross-modal aftereffects. In M. Mody (Ed.), *Neural mechanisms of language* (pp. 105-127). Boston, MA: Springer Science.

King, A. J., & Calvert, G. A. (2001). Multisensory integration: perceptual grouping by eye and ear. *Current Biology*, *11*(8), R322-R325.

Koppen, C., Alsius, A., & Spence, C. (2008). Semantic congruency and the

Colavita visual dominance effect. *Experimental brain research*, *184*(4), 533-546.

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal inference in multisensory perception. *PLoS one*, *2*(9), e943.

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisensory behavioral performance. *Experimental brain research*, *158*(4), 405-414.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Macleod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.

Massaro, D. W. (2009). *Caveat emptor: The meaning of perception and integration in speech perception*. Available from Nature Precedings http://hdl. handle. net/10101/npre, 1.

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

Miyazaki, M., Yamamoto, S., Uchida, S., & Kitazawa, S. (2006). Bayesian calibration of simultaneity in tactile temporal order judgment. *Nature neuroscience*, 9(7), 875.

Molholm, S., Ritter, W., Javitt, D. C., & Foxe, J. J. (2004). Multisensory visual–auditory object recognition in humans: a high-density electrical mapping

study. *Cerebral Cortex*, *14*(4), 452-465.

Plöchl, M., Gaston, J., Mermagen, T., König, P., & Hairston, W. D. (2016). Oscillatory activity in auditory cortex reflects the perceptual level of audio-tactile integration. *Scientific reports*, *6*, 33693.

Radeau, M., & Bertelson, P. (1987). Auditory-visual interaction and the timing of inputs. *Psychological research*, *49*(1), 17-22.

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of neurophysiology*, *89*(2), 1078-1093.

Reed, C. M., Durlach, N. I., Braida, L. D., & Schultz, M. C. (1989). Analytic study of the Tadoma Method: Effects of hand position on segmental speech perception. *Journal of Speech, Language, and Hearing Research*, *32*, 921–929.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 51–78). Malden, MA: Blackwell.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex*, *17*(5), 1147-1153.

Sanabria, D., Soto-Faraco, S., Chan, J. S., & Spence, C. (2004). When does visual perceptual grouping affect multisensory integration?. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(2), 218-229.

Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of

the ventriloquism effect. *Neuroreport*, *12*(1), 7-10.

Soto-Faraco, S., & Kingstone, A. (2004). Multisensory integration of dynamic information. In G. A. Calvert, C. Spence & B. E. Stein (Eds.), *The handbook of multisensory processes*, (pp. 49-67). Cambridge, MA: MIT Press.

Soto-Faraco, S., Spence, C., Lloyd, D., & Kingstone, A. (2004). Moving multisensory research along: Motion perception across sensory modalities. *Current Directions in Psychological Science*, *13*(1), 29-32.

Sparks, D. W., Kuhl, P. K., Edmonds, A. E., & Gray, G. P. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech. *Journal of the Acoustical Society of America*, *63*(1), 246-257.

Spence, C. (2007). Audiovisual multisensory integration. *Acoustical science and technology, 28(*2), 61-70.

Spence, C. (2015). Cross-modal perceptual organization. In Wagemans, J. (Ed.), *The Oxford handbook of perceptual organization*, (pp. 649-664). Oxford: Oxford University Press.

Spence, C., & Bayne, T. (2014). Is consciousness multisensory. In D. Stokes, M. Matthen & S. Biggs (Eds.), *Perception and its modalities*, (pp. 95-132). Oxford: Oxford University Press.

Stein, B.E. (Ed.). (2012). The new handbook of multisensory processing. Cambridge, MA: MIT Press

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hearing research*, *258*(1-2), 4-15.

Stetson, C., Cui, X., Montague, P. R., & Eagleman, D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron*, *51*(5), 651-659.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212-215.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & psychophysics*, *69*(5), 744-756.

Vroomen, J., & Keetels, M. (2006). The spatial constraint in intersensory pairing: No role in temporal ventriloquism. *Journal of experimental psychology: Human perception and performance*, *32*(4), 1063.

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological bulletin, 138*(6), 1172.

Warren, D. H., Welch, R. B., & McCarthy, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics*, *30*(6), 557-564.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & psychophysics*, *67*(3), 531-544.

Chapter 2:

Multimodal Integration in Speech Perception: The Effect of Aero-Tactile Information on Perception of VOT Continua

2.1 Introduction

Multisensory integration in speech perception is the combined use of different sensory modalities in the construction of a speech percept. Most current research on multimodal integration focuses on vision and audition: vision has been demonstrated to enhance the perception of speech when integrated with auditory stimuli in both suboptimal acoustic conditions such as background noise or heavy foreign accent (Sumby & Pollack, 1954; Middelweerd & Plomp, 1987; Reisberg et al., 1987; Macleod & Summerfield, 1990; Ross et al., 2006) and cases of increased cognitive load such as complicated structure or content (Reisberg et al., 1987; Arnold & Hill, 2001). Visual cues have also been demonstrated to facilitate language acquisition both in children (Mills, 1987) and adults acquiring a second language (Hardison, 2007), and to improve the speech perception of individuals with hearing impairments, especially individuals with cochlear implants (e.g., Geers & Brenner, 1994; Grant & Seitz, 2000; Lachs et al., 2001; Kaiser et al.,

2003). Conversely, it has been shown that incongruent visual and auditory cues can interfere with normal perception in adults (McGurk & MacDonald, 1976; Massaro et al., 1993) and infants (Burnham & Dodd, 1996; Rosenblum et al., 1997). This body of evidence suggests that visual and auditory cues are integrated, along with other cues, in the process of speech perception.

In recent years, evidence has accumulated suggesting that tactile information may also be integrated with other modalities in the perception of speech. In early studies, the effects of tactile information on perception was demonstrated for participants that either had explicit knowledge of the task (Fowler & Dekle, 1991; Gick et al., 2008), or were trained to make a connection between the tactile and the auditory cues (Sparks et al., 1978; Reed et al., 1989; Bernstein et al., 1991). However, later studies have established that tactile information influences auditory perception of uninformed and untrained listeners as well (Gick & Derrick, 2009; Ito et al., 2009; Derrick & Gick, 2013).

Ito et al. (2009) used a robotic device to pull facial skin, creating patterns of facial skin deformation in listeners, that normally accompany the production of the vowels /ɛ/ and /æ/. They showed that by timing these deformations to auditory stimuli, the perceptual judgments of a synthetic vowel continuum ranging from /ɛ/ to /æ/ were shifted in the expected direction. For example, when the skin was pulled upward (a deformation consistent with /ɛ/) the word *head* was preferred, whereas when the skin was pulled downward (consistent with /æ/) the word *had* was preferred. Crucially, deformations applied rearward (orthogonal to directions consistent with vowel production) had no effect on the perceptual judgments. Ito et al. concluded that somatosensory cues can modulate speech perception, but only when these are congruent

with those expected in production.

Gick & Derrick (2009) studied the effect of applying air puffs to the back of the hand and the center of the neck at the suprasternal notch on auditory perception of a voicing contrast. In their experiment, native speakers of North-American English were asked to determine whether they heard a syllable with an initial voiceless stop or a syllable with an initial voiced stop. The stimuli, the syllables /ba/, /pa/, /da/ and /ta/ produced by a male native speaker of North-American English, were partially masked by white noise in order to increase ambiguity. During some trials, while the participants heard the stimuli, puffs of air were applied to the back of the participant's hand, on their suprasternal notch, or as a control beside and tangent to headphones they wore with no direct contact with hair or skin. The participants were blindfolded; thus, they had no visual information about the application of the air puffs. The duration of the air puffs reflected the duration of the turbulent part of a naturally produced English aspirated consonant. The presence of airflow facilitated the identification of voiceless stops and reduced the identification of voiced stops. Since no such effect was found for the participants in the control group where no direct tactile information was provided, Gick and Derrick concluded that tactile information can modulate speech perception similar to the way vision does.

In a later study, the effect of tactile stimulation of the ankle on auditory perception was tested (Derrick & Gick, 2013). The motivation for using the ankle was two-fold. First, it is a distal location relative to the source of aspiration, farther than the neck and the hand. Thus, while speakers may have experience with feeling air puffs on the back of their hand while they were speaking, or, at least to some extent, with feeling air puffs

on the neck while others were speaking, it is unlikely they have similar experience with feeling air puffs on their ankles. Moreover, even if such experience does exist, it is not frequent or robust, thus it is not likely that participants associate the feeling or a puff of air on their ankle with the production of certain speech sounds. Second, the ankle is distant from the ear, and its representation in the somatosensory cortex is distant from the ear's representation in the somatosensory cortex (Penfield & Rasmussen, 1950). Since comparison of the ankle results to the hand and neck results from Gick & Derrick (2009) did not reveal significant differences, Derrick and Gick concluded that integration is a full-body process and that the association between the felt puff of air and the produced aspirated sound does not depend on direct experience.

The current study aims at providing a solid evidence for audio-tactile integration. Such evidence for multimodal speech perception has been used as a part of the debate about the nature of speech perception, which revolves around the question how speech objects are primarily perceived. Three main answers have been suggested to this question: from an ecological or direct perception point of view, represented in the field of speech by Direct Realism (e.g., Fowler 1981, 1984, 1996), speakers primarily perceive physical events in the actual world - vocal tract gestures. From the point of view of Motor Theory (Liberman et al., 1967; Liberman & Mattingly,1985) speakers primarily perceive abstract representations of vocal tract gestures rather than physical events as such. From a general auditory point of view (e.g., Klatt 1979; Stevens 1981, 1989; Massaro 1987; Diehl & Kluender 1989) the speakers primarily perceive sounds in an acoustic space.

Crucially, the general auditory approaches assume that perception of speech sounds is the same as perception of non-speech sounds. According to this view, the same

mechanisms of audition and perceptual learning are used for perception of all types of sounds. Thus, from this perspective, the primary objects of speech perception may be acoustic or auditory objects (e.g., Klatt 1979; Stevens1981; Stevens 1989; Massaro 1987; Diehl & Kluender 1989, Kingston & Diehl 1994) or acoustic landmarks which convey information about the gestures that produced them (Stevens 2002). These approaches posit an intermediate representation constructed from sensory input. That is, listeners identify acoustic patterns or features by matching them to stored acoustic representations. In contrast with the non-auditory approaches, which assume listeners recover gestures, the auditory approaches assume that listeners perceive "the acoustic consequences of gestures" (Diehl et al., 2004, p. 168) (though see Stevens, 2002). It is assumed that all the relevant information for perception of speech is included in the acoustic signal and is recoverable by general mechanisms of perceptual learning.

An argument in favor of the non-auditory approaches thus comes from research on multisensory integration. This line of research has been used to argue for the independence of speech perception from non-speech auditory perception (see Goldstein & Fowler, 2003; Rosenblum, 2005 for examples and discussion). The argument is that if vision/tactile stimulation is an integral part of the process of speech perception, speech perception cannot be auditory, or at least not exclusively auditory. This argument relies crucially on the interpretation of the experimental findings as supporting multimodal integration in speech perception.

Nonetheless, at least for the air puff studies of Gick & Derrick (2009) and Derrick & Gick (2013), it can be argued that this interpretation is not sufficiently supported by the data. Massaro (2009) claims that it is possible that the participants interpreted the

airflow, when it was provided, as aspiration and relied on this interpretation in making their decision. That is, the criticism is that the participants may have based their responses only on tactile information without any integration with the auditory modality. The possibility that Gick and Derrick's findings were simply the result of a general response to tactile stimuli was tested in Gick & Derrick (2009). A tap condition, in which contact with the same test locations was made using a metal solenoid plunger, established that while aero-tactile stimuli were able to shift speech perception, taps on the skin of the participants did not (see supplemental material, Gick & Derrick, 2009). Derrick & Gick (2013) argue that the results of this test are not just a control for a general attention effect caused by the addition of another type of stimuli, but also suggest that the integration of the tactile signal with the auditory signal is dependent upon it being perceived as "event-relevant, as opposed to merely synchronous" (Gick & Derrick, 2009, p. 406).

However, this test does not rule out Massaro's suggestion that there was no integration, since it is still possible that speech perception during the experiment was unimodal, that is, based solely on aero-tactile information when it was provided, and on auditory information when aero-tactile information was not provided. The stimuli in Gick and Derrick (2009) and Derrick and Gick (2013) was masked by background noise. This made the acoustic stimuli less informative than it could have been in perfect acoustic conditions. Therefore, it might have been the case that the tactile stimulus was the most prominent signal, and as a result a unimodal response was made to it. The current study aims at investigating this question further. Specifically, we use voice onset time (VOT) continua ranging from voiceless to voiced sounds rather than endpoint stimuli only (as in the work by Gick & Derrick). This design enables us to show the existence of

multimodal integration of acoustic and tactile inputs in the perception of speech, rather than a possible unimodal response.

Voice onset time is the interval between the release of a stop consonant and the onset of voicing following or preceding the release (Lisker & Abramson, 1964). In American English stops are habitually produced with a positive VOT. The duration of the positive VOT is longer for voiceless stops than for voiced stops and varies with place of articulation: the more distant the place of articulation from the lips, the longer the VOT. Average VOT durations for American English stops are summarized in Table 2.1. Note that VOT varies with context: it is shorter for stops when following an obstruent than when following a nasal, a glide, or a vowel. For stops in onset positions it is shortest for those in clusters that begin with /s/ (Randolph, 1989).

Place of articulation	VOT length (ms)	
	Voiceless	Voiced
Bilabial	44	18
Alveolar	49	24
Velar	52	27

Table 2.1. Average VOT durations for American English stops (Byrd, 1993).

Our prediction is that if this is not an integrative process, that is, the participants interpret the puffs of air as aspiration and make their decision based solely on this information when it is provided, then the responses for the trials accompanied by air-puffs will reflect this and will be mostly voiceless. However, if instead the results will show a shift in the category boundary in the presence of air puffs this would suggest that aero-tactile information is taken into account along with the auditory information

provided. Such a result would show that participants are using a context-weighted blend of auditory and tactile cues in perceiving and categorizing speech sounds, thus providing an example of multi-sensory integration in the perception of speech.

In addition, a continuum consisting of vowel sounds ranging from /ɛ/ to /ɪ/ in an /hVd/ context was included for use as a control. In contrast with the VOT continua, both edges of the vowel continuum are associated with similar amount of air. The only difference between the endpoints of this continuum is the vowel height, and while higher vowels are produced with a more constricted oral passage, that is, with a more impeded airflow (Jaeger, 1978), both words are approximately associated with a similar amount of airflow. In other words, the participants are asked to categorize based on steps in an acoustic continuum, as they are asked in the case of the VOT continua. However, in this case the additional tactile information is irrelevant and thus not predicted to affect the listeners' decision. The contrast between an effect of air-puffs in the VOT continua and a lack of it in the vowel continuum would provide evidence that the participants' decision is done with respect to relevance, that is, that the aero-tactile information is taken into account only in cases where aspiration (or amount of air produced by the speaker) is relevant for the distinction being made.

2.2 Methods

2.2.1 Participants

42 monolingual native speakers of American English participated in the experiment (24 females; age range 18-56, mean age 28.7, SD = 11.5). The participants were all residents

of Southern Connecticut at the time of the experiment but were born and raised in multiple regions of the US, including the northeast, the northwest, the west coast, the midwest and the south. Their level of education ranged from high school graduates to graduate students. The participants were recruited with flyers and by word of mouth. All were naive to the purpose of the study and had no self-reported speech or hearing defects. They were compensated for their time. All of the participants signed an informed consent form approved by the Yale Human Research Protection Program.

2.2.2 Stimuli

2.2.2.1 Acoustic Stimuli

The stimuli were created by recording a male monolingual native speaker of American English. The speaker, a resident of Southern Connecticut, was born and raised in Arizona and attended college and earned his MA in Ontario, Canada before moving to CT. The speaker produced six tokens of each of the syllables /pa/, /ba/, /ka/, and /ga/. Two eight-step VOT continua were created, one for the bilabial and one for the velar place of articulation. The continua were created by removing the initial burst from the voiceless token and then shortening the aspiration in log-scaled steps, with the final step matching the duration of the voiced token. Aspiration durations for each step of the VOT continua appear in Table 2.2. A nonlinear (logarithmic) step size was chosen because psycho-acoustic perception tends to follow Weber's law (subjective sensation is proportional to the logarithm of the stimulus intensity); e.g., Zwicker & Fastl (2006). See Rosen and Howell (1981) for results on VOT, and Stevens, (2000, p. 228) for a similar effect on the

perception of duration of burst.

Step no.	VOT length (ms)	
	Bilabial continuum	Velar Continuum
1	98	81
2	58	56
3	37	42
4	24	35
5	18	31
6	14	28
7	12	27
8	11	26

Table 2.2. VOT continua steps showing length of retained aspiration (ms).

An additional continuum consisting of vowel sounds ranging from /ɛ/ to /ɪ/ in an /hVd/ context was included for use as a control. It was synthesized from endpoint recordings of a male monolingual native speaker of North-American English producing "head" and "hid", by linearly interpolating F1 and F2 values within the vowel over the eight continuum steps, using an iterative Burg algorithm to shift the location of filter poles and zeros in resynthesis (Purcell & Munhall, 2006).

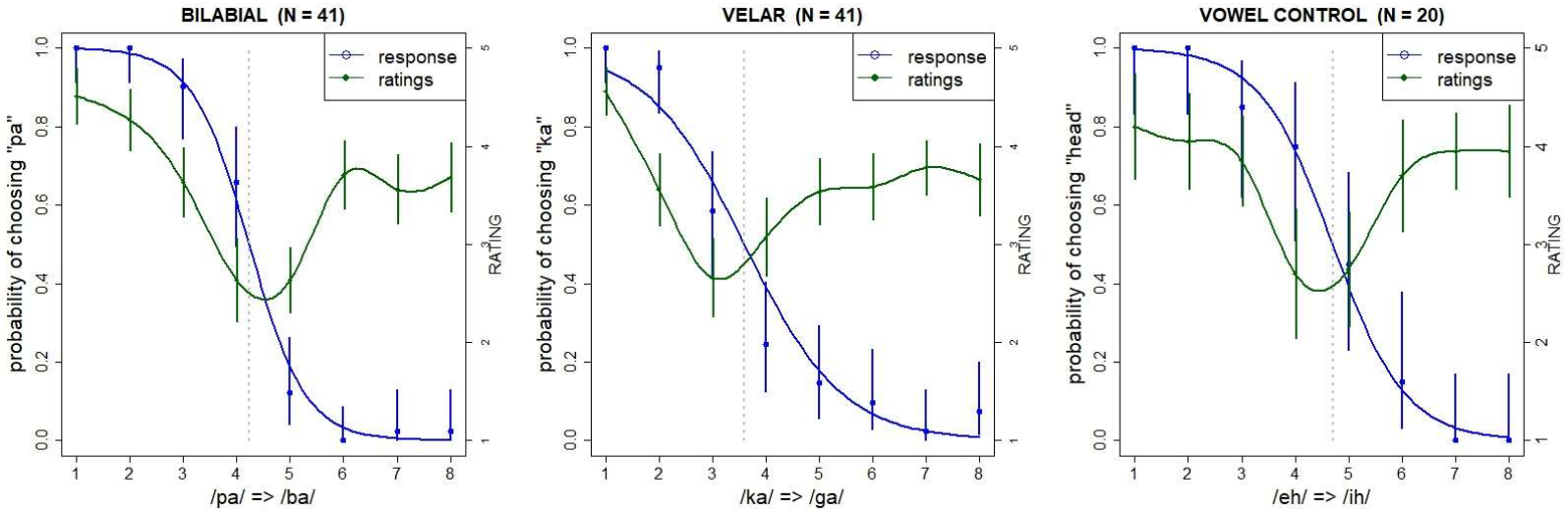

Figure 2.1. Viability test results for the continua: left scale (blue line) shows probability of choosing voiceless or "head" relative to step (dotted vertical line marks 50% crossover point); right scale (green line) shows Likert scale ratings by step. Error bars show 95% confidence intervals.

A pre-test of each continuum was conducted online as a Mechanical Turk task and was used to assess the quality of the stimuli. The test was run with an independent group of participants that did not take part in the main study (N = 41). They were asked to choose whether they heard a /pa/ or a /ba/ (in the bilabial condition), or /ka/ or /ga/ (in the velar condition) and rate the goodness of the token on a five step Likert scale. The sounds from the two continua (/pa/-/ba/ and /ka/-/ga/) were presented in the same test. A similar pre-test was conducted for the vowel continuum in which additional 20 participants were asked to choose whether they heard /hɛd/ or /hɪd/, and to rate the goodness of the token. The order of presentation was randomized in both pre-tests. The results of the pretests are plotted in Figure 2.1.

The bilabial category boundary is approximately centered between its endpoints, that is, its bias (4.2) is close to its midpoint (4.5). The bias was calculated as the 50% crossover point of the psychometric categorization function for the continuum, computed across all listeners. Acuity (a measure of boundary slope) was computed as the difference between the 25% and 75% probabilities for the categorization function. The velar category boundary is not as centralized and is skewed towards voicelessness (bias = 3.6), and its acuity (2.0) is shallower than that of the bilabial (1.1). Finally, the category boundary for the vowel control continuum is also approximately centered (bias = 4.7, acuity = 1.5). The goodness ratings for all three continua are higher at the margins than at the intermediate steps of the continuum, which reflects the fact that the ambiguous sounds were harder to categorize, as expected.

2.2.2.2 Tactile (Air Puff) Stimuli

To deliver air puff stimuli the following equipment was employed. A three-gallon air compressor (Campbell Hausfeld) was connected to a solenoid valve (Parker) used to gate airflow by 1/4-inch polyethylene tubing. The solenoid was toggled by a programmable relay controller device (KMtronic). A pressure transducer (PSC, model 312) and a flow meter (Porter-Parker MPC series) were connected to the tubing in order to monitor pressure and flow data. Solenoid control of airflow and data recording were performed using a custom Matlab (The Mathworks) procedure that was written for this experiment. The tubing was inserted into a soundproof room through a cable port and stabilized using a table microphone stand (see Figure 2.2 for a diagram of the system).

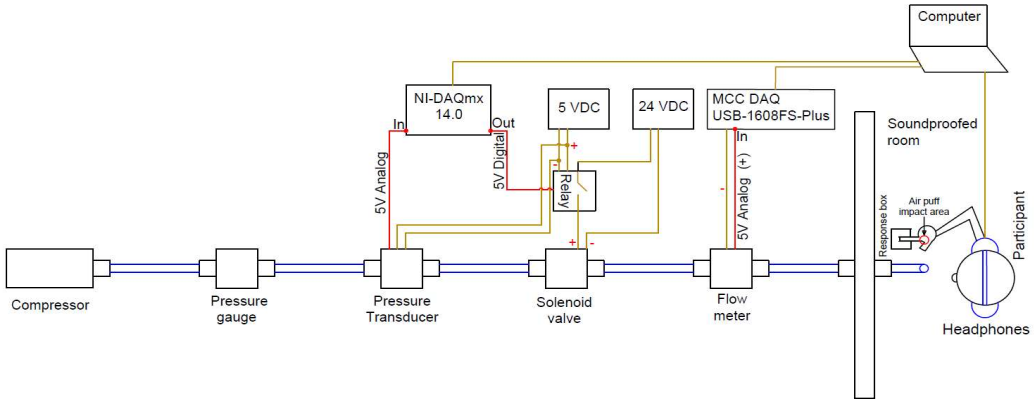

Figure 2.2. The aero-tactile stimulus presentation system

In a given trial the signal to open the air valve solenoid was given by the Matlab procedure, which also controlled acoustic stimulus presentation through the computer's sound card such that the acoustic onset of each of the stimulus was coincident with the onset of the air puff from the tube. Detectable air turbulence exiting the tube was 87 ms in duration for the bilabial condition and 92 ms in duration for the velar condition. These

timings reflect the mean aspiration time (that is, VOT) of the six voiceless tokens that the model speaker produced, thus simulating the temporal properties of the stimuli. The speaker's mean VOTs fall within the VOT range of initial aspirated stops in American English (54-100 ms, Lisker & Abramson, 1967; Cooper, 1991; Byrd, 1993). The airflow at the exit point of the tube was 5 Standard Liters Per Minute (SLPM). Note that this rate is lower than the average airflow of typical speech (8 SLPM, Isshiki & von Leden, 1964), and significantly lower than the average airflow of voiceless stop consonants in CV syllables (about 56 SLPM, Isshiki & Ringel, 1964). The exit point of the tube was placed 5 cm away from the participant's skin, creating an area of initial impact with a diameter of 2-3 cm (similar to Derrick et al., 2009). The air puffs were applied on the dorsal surface of the right hand between the thumb and forefinger (see Figure 2.3a). A microphone placed near the exit of the tube was used to record airflow turbulence during each trial, to verify that air puff stimuli (when scheduled) were delivered with the expected timing.

2.2.3 Procedure

Each experimental session included two parts, an initial test to verify that the air puffs were felt but not heard, seen or otherwise perceived, and the main part, which tested participant responses to the auditory stimuli in the presence and absence of air puffs. Stimuli were presented to the participants through ear-enclosing headphones (Sennheiser HD 202 II).

2.2.3.1 Puff Detection Test

In the first part of the experiment the participants heard a short tone (500 Hz, 1,000 ms long) in each trial, which was either followed by a 50 ms long air puff, or not followed by a puff. They were presented with two blocks of 50 trials each, in which 25 of the trials were accompanied by air puffs and 25 were not, presented in randomized order. In the first block the participant's right hand was located next to the exit of the tube such that they could feel the puff on the back of their hand (see Figure 2.3a). They were asked to press the "yes" key on a response box with their left hand if they felt or otherwise detected a puff, or the "no" key if they did not. In the second block, the task was the same, but their right hand was positioned on their lap, completely removed from the exit point of the tube (Figure 2.3b). The goal of this part of the experiment was to verify that the participants felt the puff on their hand but did not hear or see or otherwise detect it. In order to reduce the chances of hearing the puff of air, a small desk fan was used to provide a low level of background noise throughout the experiment. The fan was pointed to the wall and away from the participant.

Figure 2.3a. Puff delivery setup: participant right hand placed near outflow of airtube, left hand on response button box. Microphone records air puff delivery for verification of timing.

Figure 2.3b. Puff detection test setup: participant right hand positioned away from outflow of airtube. This test determines whether participant can detect airflow from cues other than tactile hand sensation.

2.2.3.2 Perturbed Continua Testing

In the second part of the experiment, the participant's right hand was located such that they could feel the puff of air on the back of their hand (Figure 2.3a). In this part five blocks were presented during which sounds drawn from one of the three continua were tested: from /pa/ to /ba/, /ka/ to /ga/, or /hɛd/ to /hɪd/. Only one continuum type was used within a given block. Each block included six repetitions of each step of the continuum, for which three instances were accompanied by air puffs and three were not, randomly ordered. Within a session each participant received five blocks each of two different continuum types, resulting in 5 blocks *x* 3 repetitions *x* 2 puff conditions (+/-) *x* 8 continuum steps for a total of 240 separate judgments per continuum type, with 15 per condition at each continuum step. Each participant heard ten blocks: either five velar blocks and five bilabial blocks, five bilabial blocks and five vowel blocks, or five velar blocks and five vowel blocks. Overall, 33 of the participants were tested for the bilabial condition, 32 of the participants were tested for the velar condition, and 19 participants were tested for the vowel condition. In each trial, participants were asked to identify the stimulus they heard and to press the corresponding button on a response box: either "P" or "B" to indicate whether they heard /pa/ or /ba/ during the bilabial blocks, "K" or "G" to indicate whether they heard /ka/ or /ga/ during the velar blocks, and "head" or "hid" to indicate the word they heard during the vowel blocks. The presentation order of the auditory stimuli and the accompanying tactile information (puff present vs. absent) were pseudo-randomized throughout each block. The blocks alternated such that there were no consecutive blocks of the same kind. For half the participants, the right button on the response box indicated a syllable with a voiceless consonant (e.g., "pa"). For the other

half, the right button indicated a syllable with a voiced consonant. A similar counterbalancing was performed for the vowel blocks. In each trial the Matlab control procedure presented the audio stimulus, gated the air puff (or not), and recorded the participant choices from the response box. New trials began 1,000 ms after each button-press response.

2.3 Results

2.3.1 Puff Detection Test

In the first block of the detection test, when their hand was close to the exit point of the tube, participants correctly discriminated puff/no puff conditions at an average rate of 98.1% (s.d. 2.6), with the worst performer at 90%. An exact binomial test confirms that these recognition percentages were well above chance ($p < 0.01$). In the second block, with their hand positioned away from the tube and everything else the same, participants were at chance: 50.4% (s.d. 2.6); best performer 57% (binomial test n.s.). These results confirm that the participants felt the puff of air on their hand, but could not hear, see, or otherwise detect it.

2.3.2 Perturbed Continua Testing

In 387 of the trials (1.9% of the trials) an air puff was requested but not delivered, or not requested but delivered. These trials were excluded from analysis, along with additional 85 trials for which the button-press response time exceeded 8 seconds (~5 s.d.). The data

were then modeled with logistic regression in R (R Core Team, 2016) to estimate the effects of puffs on the perceptual boundary. Figure 2.4 shows the estimated psychometric categorization functions, pooled across speakers, in the presence and absence of air puffs. The vertical axis represents the probability of choosing a voiceless token or /ɛ/ (that is, /pa/ in the case of the bilabial continuum, /ka/ in the case of the velar continuum, or /hɛd/ in the case of the vowel continuum). The horizontal axis shows the 8 steps along the continuum. The baseline condition, without puff, is shown in blue lines with circles, and the condition with air puffs is shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25% and 75% probability points along each curve; the distance between these points gives the acuity (a measure of the slope of the boundary). The shift of the bias to the right in the presence of air puffs in the two VOT continua reflects the fact that there were more voiceless responses in this condition; this contrasts with the control vowel continuum which shows no shift in bias under puffs.

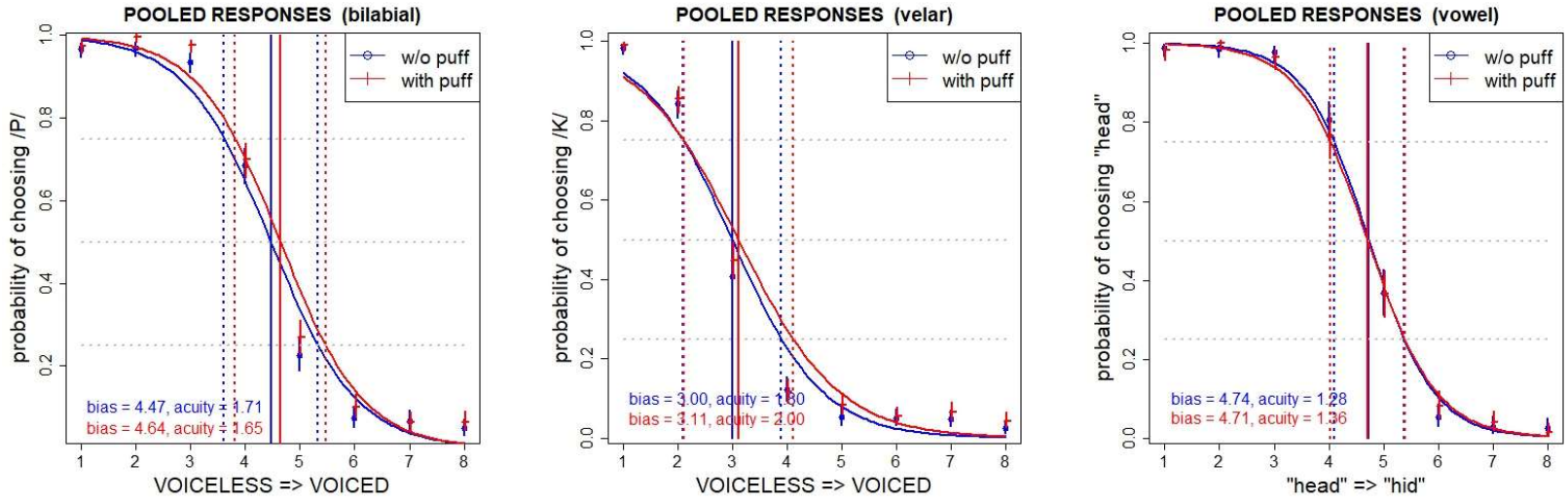

Figure 2.4. Perceived category boundaries, pooled across speakers, with (red) and without (blue) an air puff. Vertical lines show the bias (50%) crossover, which is systematically shifted in the direction of voicelessness for +puff trials in the bilabial (left) and velar (center) continua, but not in the control vowel continuum (right). 95% confidence intervals are indicated for each pooled response.

2.3.2.1 Quantifying the effect of puffs on perceived categories

A generalized linear mixed-effects model (GLMM) computed with the lme4 package (Bates et al. 2015) was used to assess the significance of the puffs contrast for each of the continua separately as they differ in step size, skewness and type (the VOT continua were created by manipulating VOT duration, whereas the vowel continuum was created by manipulating formant structure). In this model[1] the dependent variable (the probability of choosing a voiceless or "head" response) was predicted by the fixed effect of PUFF (-/+) and a continuous covariate of STEP, with random intercepts by participant ID (random slopes by participant were not supported by model comparison, $\chi^2(2) = 0.5094$, $p = 0.775$). The results, summarized in Table 2.3, show a significant shift under +PUFF for the two VOT continua in the direction of voicelessness (bilabial $z = 3.16$**,[2] velar $z = 2.53$*), and no effect of PUFF on the vowel continuum ($z = -0.31$). Marginal R^2 for these models (a measure of effect size), representing the proportion of variance explained by fixed factors alone, was computed using the method of Nakagawa & Schielzeth (2013), as implemented by Lefcheck & Casallas (2014). The effect of STEP was significant for all continuum types. The addition of interaction terms for PUFF and STEP did not improve the fit of the model, in all three cases.

[1] `glmer(RESP ~ PUFF + CSTEP + (1|ID), family=binomial)`

2 `Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1`

Continuum	-Air PUFF (baseline) vs. +Air PUFF			
	Coefficients	z-value	p-value	Marginal R^2
Bilabial	0.244	3.160	0.0016 **	0.733[3]
Velar	0.216	2.533	0.0113 *	0.699
Vowel	-0.037	-0.313	0.7540 n.s.	0.817

Table 2.3. Output of the GLMM response model for each continuum.

For the two VOT continua the effect of +PUFF was to increase the likelihood of a voiceless response; the vowel control continuum was unaffected. Marginal R^2 shows the proportion of variance explained by the fixed factors alone. Note that the fixed factors include, in addition to PUFF, the continuous covariate of STEP. This factor is responsible for much of the explained variance. See footnote 3 for the values of marginal R^2 computed for a model that contains only CSTEP as a fixed factor, without PUFF.

2.3.2.2 Comparison of Effect Sizes for the Three Continua

In order to compare the relative magnitudes of the puff effect we computed a second GLMM on the data combined from all three continua. In this model[4] the probability of choosing a voiceless or "head" response was predicted by the fixed effects of PUFF and CONTinuum type and their interaction, and the continuous STEP covariate, with random slopes for CONT by participant ID (random slopes for PUFF were not supported by

3 Values of marginal R^2 for the following model
glmer(RESP ~ CSTEP + (1|ID), data=d, family=binomial):
Bilabial: 0.731, velar: 0.698, vowel: 0.817

4 glmer(RESP ~ PUFF * CONT + CSTEP + (1+CONT|ID), family=binomial)

model comparison, $\chi^2(3) = 0.4445$, $p = 0.931$). The results are shown in Table 2.4. The relative effect sizes are computed using odds ratios. The interaction terms show the ratio by which the odds ratio of each VOT continuum relative to the Vowel baseline changes for +PUFF, with a larger magnitude observed for the bilabial continuum than the velar.

	coefficients	z-value	p-value	odds ratios	95% confidence intervals
(Intercept)	7.53485	30.22	0.000	1872.162	(1148.363,3052.164)
+PUFF	-0.03315	-0.31	0.758	n.s.	
CONTvel	-2.72139	-6.85	0.000	0.066	(0.030, 0.143)
CONTbil	-0.45373	-1.57	0.117	n.s.	
STEP	-1.59468	-66.77	0.000	0.203	(0.194, 0.213)
+PUFF:CONTvel	0.23953	1.76	0.078	1.271	(0.973, 1.659)
+PUFF:CONTbil	0.23953	2.21	0.023	1.360	(1.043, 1.772)

Table 2.4. Output of GLMM combining continua to show relative effect sizes (using odds ratios). Marginal R^2 for this model is 0.756.

2.3.2.3 Analysis of Individual Results

To assess the degree to which individual participants were sensitive to the air puff effect we computed separate logistic regression models for each, with response predicted by the fixed effect of PUFF and STEP as a continuous covariate.[5] About two thirds of the participants who heard the bilabial continuum showed a shift towards voiceless responses (23/33; binomial test $p < 0.02$), as did about three quarters of the participants who heard the velar continuum (24/32; binomial test $p < 0.01$). About half

[5] `glm(RESP ~ PUFF + CSTEP, family=binomial)`

of the participants who heard the vowel continuum showed small and non-significant shifts towards "head" responses (9/19; n.s.). See Table 2.5 for summary statistics.

Continuum	mean coefficient	s.d. of coefficient	range of coefficient
Bilabial	0.26766	0.479	-0.87388 : 1.66863
Velar	0.21979	0.546	-0.83977 : 0.98542
Vowel	-0.00845	-0.548	-0.99308 :1.02929

Table 2.5. Summary of the individual models computed for the participants

2.3.2.4 Analysis of Response Times

Response times were measured as the duration in milliseconds from the onset of the audio stimulus (which was coincident with the start of the air puff, if present), to the button-press event. For analysis they were log-scaled in order to normalize a right-skewed distribution. Figure 2.5 illustrates the mean response times pooled across participants, by PUFF, CONTinuum type, and STEP along the continuum. An overall effect of CONTinuum type was observed, with bilabial responses slower than velar responses in general, and both significantly slower than vowel control responses.

A linear mixed-effects model[6] computed using lme4 with significance assessed using the lmerTest package (Kuznetsova et al. 2017) in R was used to predict the $\log_{10}$ response time from the fixed effects of PUFF, CONTinuum, and (discrete) continuum STEPs, with random intercepts by participant. The analysis modeled discrete rather than continuous steps along the continuum to investigate how response time interacted with

[6] `lmer(LRT ~ PUFF * STEP * CONT + (1+PUFF+STEP+CONT||ID))`

stimulus, with the expectation that responses to stimuli in the ambiguous range of each continuum would be slower. This expectation is driven by the fact that greater cognitive effort that is required for categorization of ambiguous tokens in comparison to categorization of unambiguous tokes. Model comparison supported the complete interaction between fixed factors and the inclusion of random intercepts for each by participant. Significant results are shown in Table 2.6.

The pattern of main effects confirms that response times are slower for the ambiguous intermediate steps (4, 5, 6), and that responses for the two VOT continua are slower overall than for the vowel control baseline, with the bilabial responses slower than the velar. The negative coefficient for the interaction of +PUFF and the bilabial continuum suggests an overall facilitation effect (responses are faster than baseline), which Figure 2.5 suggests is active on the voiceless end of the continuum (steps 1, 2). This effect was due to the complementary nature of the added information. A similar pattern can be seen in the interaction of STEP with the velar continuum, in that following step 3 responses are also faster than baseline. Step 3 itself is significantly slower, but in this left-skewed continuum this step is closest to the crossover for the velar continuum (see Fig. 4) and can thus be expected to represent its most ambiguous stimulus. Finally, the interaction of steps 5, 6, and 7 with the bilabial continuum shows that these responses were significantly faster than baseline *without* puffs, and significantly slower than baseline *with* puffs.

	coefficients	T-value	P-value	significance
STEP4	0.04363	6.248	0.000	***
STEP5	0.06053	8.575	0.000	***
STEP6	0.03054	4.246	0.000	***
CONTvel	0.1026	14.307	0.000	***
CONTbil	0.1275	18.295	0.000	***
+PUFF:CONTbil	-0.01717	-2.163	0.031	*
STEP3:CONTvel	0.02652	3.248	0.001	**
STEP4:CONTvel	-0.03783	-4.552	0.000	***
STEP5:CONTvel	-0.06216	-7.539	0.000	***
STEP6:CONTvel	-0.03499	-4.247	0.000	***
STEP8:CONTvel	-0.01404	-1.729	0.084	.
STEP5:CONTbil	-0.04550	-5.550	0.000	***
STEP6:CONTbil	-0.02480	-3.022	0.003	**
STEP7:CONTbil	-0.01666	-2.049	0.040	*
+PUFF:STEP5:CONTbil	0.03088	2.759	0.006	**
+PUFF:STEP6:CONTbil	0.0214	1.911	0.056	.
+PUFF:STEP7:CONTbil	0.02473	2.210	0.027	*

Table 2.6. Output of LMM predicting $\log_{10}$ response times from PUFF, CONTinuum, and stimulus STEP along the continuum. The baseline represents -PUFF at STEP1 on the Vowel continuum. Only significant values are shown. Pseudo-R^2 for this model (comparison of fitted vs. observed values) is .447.

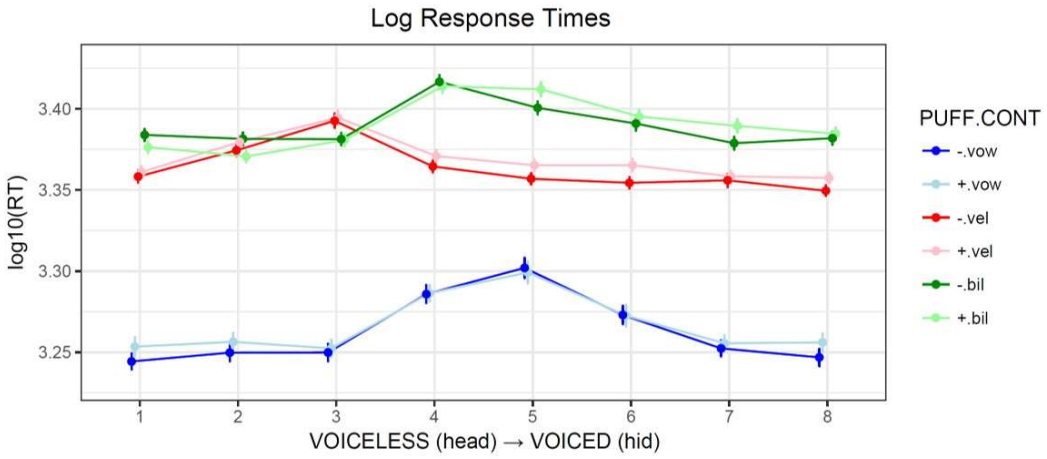

Figure 2.5. Comparison of mean $\log_{10}$ response times averaged across participants, by PUFF (+airflow vs. -airflow), CONTinuum type (VOWel, VELar, BILabial), and continuum STEP. Error bars show the standard error of the mean

2.4 Discussion

The current study found that presence of air puffs significantly increased the likelihood of choosing voiceless responses for the two VOT continua but had no effect on choices for the vowel continuum. The category boundaries for both VOT continua were shifted towards the voiceless end of each continuum in the presence of air puffs. The effect was found to be larger for the bilabial continuum than for the velar continuum, though not significantly so. The observed difference may be due to the unbalanced (left-skewed) velar continuum.

Voicing continua were used rather than endpoints alone to provide evidence for multisensory integration rather than a unimodal response to either the acoustic or the tactile stimuli. Gick and Derrick (2009) and Derrick and Gick (2013) used CV exemplars

in background noise. These masked exemplars provide incomplete acoustic information. In fact, this information might not be sufficient for categorization, which can potentially lead speakers to rely on the tactile information provided to them instead of the insufficient acoustic signal. As Massaro (2009) pointed out, the data from Gick and Derrick is not sufficient to exclude this possibility of unimodal decision-making and conclude that their results reflect multimodal integration. The current study shows that listeners are not relying exclusively on tactile information even in cases where the acoustic information is imperfect such as the intermediate steps of the continuum. The existence of an air-puff alone in each trial was not sufficient for deciding the category: even in the intermediate steps the responses were not overwhelmingly voiceless in the trials accompanied by puffs. Overall, the existence of a puff in a given trial did yield more voiceless responses, but nonetheless, the expected category boundary was evident in trials with and without puffs of air, indicating that the acoustic information represented by the steps of the continuum was taken into consideration by the participants, even when the effects of the air puffs was noticable. We have shown, thus, a case of multimodal integration, as both the acoustic and the tactile stimuli were used by the listeners. Moreover, response time analysis showed that sounds along the continuum were not uniformly affected by the aero-tactile stimuli. This suggests that aero-tactile sensation was processed as a potential additional cue for disambiguation of voiceless from voiced sounds, but weighted by relevance and the degree of ambiguity, in a true multi-sensory integration.

Although participants were not instructed to answer as quickly as possible, analysis of response times did reveal significant differences between continua and within

continua. The intermediate steps of the continua, that is, the ambiguous stimuli between the two endpoints, were the hardest for participants to categorize, as expected. This was suggested by the longer response times associated with these steps, for all three continua. as longer response times generally indicate a greater cognitive load (e.g., DeLeeuw & Mayer, 2008). This is consistent with the fact that linguistic ambiguity was found to affect measures of cognitive load directly (Engonopulos et al., 2013) and that processing ambiguous tokes on a VOT continuum has been shown to be particularly sensitive to effects of cognitive load (Mattys and Wiget, 2011).

For the two VOT continua in general response times were slower than the vowel control baseline. Crucially, the response times for the VOT continua did not show a uniform response to air puffs, shown most clearly by the bilabial continuum. As illustrated in Figure 2.5 and shown by the results in Table 2.6, air puffs had a *facilitatory* effect at the voiceless end of the continuum (encoded by the negative +PUFF:CONTbil interaction; t = -2.2*); i.e., responses were faster with puffs. This effect was caused by the complementary nature of the added information. Conversely, air puffs at the voiced end of the continuum had an *inhibitory* effect (encoded by the positive +PUFF:STEP:CONTbil interaction for steps 5 (t = 2.8**), 6 (t = 1.9·), and 7 (t = 2.2)). In this case, the added information was contradictory. The pattern of results indicates that an air puff cue is evaluated together with the concurrent audio stimulus and weighted by the ambiguity of the latter.

We have mentioned, in the introduction, that evidence for multisensory integration has been used to argue in favor of certain approaches for the primary objects of speech perception. The argument is that if non-acoustic information, tactile in the current case, is

an integral part of the process of speech perception, speech perception cannot be auditory, or at least not exclusively auditory. A counterargument that has been discussed in the literature is that association with visual and other sensory information might be learned from experience but is not inherently part of the auditory primitives of speech perception (e.g., Massaro 1987; Diehl & Kluender 1989; Kluender 1994). Rosenblum (2005) offers a few arguments against auditory primitives that are associated with other modalities at later stages: first, multisensory integration has been shown in pre-linguistic infants (Rosenblum et al., 1997). Second, multisensory integration has been shown to operate at an early stage of online perception, before phonetic categorization and possibly before phonetic feature extraction (Summerfield 1987; Green 1998; Rosenblum & Gordon 2001). Rosenblum argues further that evidence for multisensory integration at an early stage of speech processing is consistent with evidence for multisensory integration in other domains (for discussion see Shimojo & Shams 2001; Stoffregen & Bardy 2001. But see Remez et al., 1998). Rosenblum also argues that multisensory integration has been shown in contexts where participants had no speech experience associated with the task (Fowler & Dekle 1991). However, in the experiment conducted by Fowler and Dekle the participants were aware of the task thus it is not clear that this is indeed a counter argument for learned association. The ankle data from Derrick & Gick (2013) may be an example for such a context, since humans have no speech experience associated with sensing a puff or air on their ankle, or at least not a frequent or robust experience associated with such a sensation.

Based on the evidence cited above, Rosenblum argues that speech perception is modality neutral. Specifically, he argues for gestural objects that have spatial and temporal

dimensions but are not specified along any sensory dimension. According to this view the sensory dimensions are the medium through which perceivers recover the gestures, and the objects of speech perception themselves are of a higher order than just auditory, visual or tactile. The idea is that perception is sensitive to underlying gestural primitives instantiated in any modality. This view, which is consistent with Direct Realism (e.g., Fowler 1981, 1984, 1996) and Motor Theory (Liberman et al., 1967; Liberman & Mattingly,1985), is supported by the cited evidence for the automaticity and ubiquity of multisensory integration. However, it is not the only view that is consistent with such evidence. It may be the case that the primary objects of speech perception do have a sensory content, but they are specified for more than one modality. That is, it may be the case that they are not just auditory, but multimodal in nature. The evidence presented here suggests that tactile information is considered during the perception of speech. However, it does not rule out the option that the integration of the additional tactile modality operates in later stages of online perception.

The lack of an obvious connection between aero-tactile stimulation on the hand and speech perception in the current experiment contrasts with the direct somatosensory link posited by Ito et al. (2009). In their experiment they determined that perception of vowels is affected by deforming the skin on the face of the participant in the same way the skin moves when these vowels are produced. Crucially, deformations applied orthogonal to the up and down directions used in the production of these vowels had no effect. This kind of direct link between somatosensory stimulus and speech perception is not reflected in the current study, as air puffs were applied on the back of hand of the participants, a location that does not directly relate to the creation of aspiration during

the production of stop consonants. Nonetheless, the results presented here confirm that aero-tactile stimulation can also shift perception, though only when the cue is relevant (vowel perception was unaffected). In both types of studies then, tactile information affected speech perception only when the cues applied were congruent with the ones expected in production of the perceived sounds.

In addition to addressing the critique against Gick & Derrick (2009) and Derrick & Gick (2013) and providing evidence for integration of auditory and tactile input in the perception of speech, the current work extends the work of Gick and Derrick in two ways. First, rather than a between-subject design, here a within-subject design was used in which each participant served as their own control. Thus, the comparison between the perception of the VOT continua with and without tactile stimuli was done within participant, and not across groups of participants. This allowed a direct comparison between the responses of the same individual to the same auditory stimuli with and without aero-tactile stimulus. Second, a vowel continuum was used as a control. Since aero-tactile sensation is hypothesized not to be relevant for distinguishing /ɛ/ from /ɪ/, effects observed on the VOT continua but not on the vowel continuum shows that the obtained results were not just an artifact of puffs alone, but rather a context-sensitive effect, indicating a true multi-sensory phenomenon. Moreover, since this was a within-subject design, the comparison between the VOT continuum and the vowel continuum was done within participant. That is, the participants that heard vowel blocks were sensitive to the effect of aero-tactile stimulation when the acoustic stimuli were taken from a VOT continuum, and at the same time showed no such sensitivity when the acoustic stimuli were taken from a vowel continuum. As discussed above, these results

are consistent with Ito et al. (2009), showing that while tactile cues can indeed modulate perception, they do so only when congruent with the production contrast being disambiguated.

While statistically significant, the effect of puffs found in this study was not observed for all the participants. Population estimates of audio-visual integration susceptibility vary widely and range between 26% and 98% of the tested population (Nath and Beauchamp, 2012). In the current study, between two thirds (in the bilabial continuum) and three quarters (in the velar continuum) of the participants showed susceptibility to puffs in their responses. These clear majorities contrast with participants who showed some effect of puff on their response to the vowel continuum (about half), though of these shifts, none were significant. The absence of effect on the VOT continua for some of the participants may stem from lack of statistical power, given the small size of the effect and further division of the data into participant-sized bins, though for most of the participants a significant effect was found even after the division of the data. Finally, it is possible that some of the participants were not affected by the aero-tactile stimuli because of the relatively low airflow (5 SLPM), in comparison to the average airflow of voiceless stop consonants in CV syllables (about 56 SLPM, Isshiki & Ringel, 1964). Although the puff detection test has confirmed that these participants have felt the puff, it is possible that they did not interpret it as related to aspiration since the airflow was incongruent with the typical airflow of speech. It might be the case that the threshold for judging stimuli as speech related varies across different dimensions and that the airflow threshold was not met for these participants.

The current study did not test the length of the integration window, as it did not

vary the relative timing of the auditory stimuli and the tactile stimuli. However, it has been shown previously that this window operates asymmetrically. Derrick et al. (2009) and Gick et al. (2010) found for audio-tactile stimuli that integration extends to 200 ms when air puff follows audio but only 50 ms when air puff precedes audio. Bicevskis (2015) studied visuo-tactile integration by presenting participants with video of faces producing the syllables /pa/ and /ba/, without an air puff, or accompanied by an air puff occurring synchronously with the visual stimuli or at different timings, up to 300 ms before and after the stop release. Bicevskis found that the integration window for visuo-tactile stimuli is also asymmetric: when an air puff followed visual stimuli the integration window extended to 300 ms, but when it preceded visual stimuli the integration window only extended to 100 ms. These windows extend farther than the audio-visual integration window reported by Munhall et al. (1996) (0 ms to 180 ms) and van Wassenhove et al. (2007) (-30 ms to 170 ms) for McGurk phenomena but exhibit the same properties of asymmetry. The asymmetry appears to be ordered by the relative speed in which each modality is processes in the case of tactile sensation and audition; i.e., tactile sensation is slower than audition. However, auditory input is processed faster than visual input (Molholm et al., 2002). Munhall et al. (1996) suggest that knowledge of the natural world may play a role in validating the range over which integration is permitted to occur; e.g., thunder is expected to follow lightning, and air turbulence is typically heard before it is felt. Thus, relative timings of potential speech cues that violate these expectations are less likely to be integrated.

Finally, although we have observed an effect of distal aero-tactile stimulation on speech perception, we have not provided an explanation for why the phenomenon

occurs. It is possible that humans have sufficient exposure as children to speech produced by others who are in close proximity to them. In 1966, Hall defined four spaces encircling every person. The most inner space, the intimate space, is characterized as the spaces closest to the body, up to 45 cm away from it. This is a space reserved for sexual partners and children. This distance is sufficiently short for aspirated stops to be felt on the skin of a child or a partner. Children are also found in close proximity to others during social interaction with their peers: Aiello and Jones (1971) studied the proxemic behavior of children ages 6-8 and found that the mean distance between children during social interaction differed by sex and sub-culture, but overall ranged between 5.3 and 13.5 inches, a distance sufficiently short for aspirated stops to be felt on the skin. Aiello & Aiello (1974) found that personal space grows bigger as children grow older, suggesting that the chance of being exposed to felt aspiration at younger age is larger than it is in conversations at later stages of life. Because such stimulation would not be particularly localized to a single point of contact, the association between aspiration and tactile sensation could then eventually be generalized to any skin location. However, while the results of Gick & Derrick (2009) and Derrick & Gick (2013) show that air puffs affect VOT perception when the point of contact is the neck or even the ankle, it is suggestive that not just any tactile stimulus produces the effect, as their negative result from the tapping vs. air puff comparison shows. Accordingly, while the pathway to acquiring an association between VOT aspiration and the tactile sensation specific to feeling its effect on the skin is purely speculative, the results from Gick & Derrick (2009) and this confirmatory study indicate that such an association is real. Once available, it joins other potential cues (visual, lexical, etc.) available for exploitation by language

users to disambiguate the speech signal.

2.5 Conclusion

The aim of the current study was to provide solid evidence for audio-tactile integration in the perception of speech. We used voice onset time (VOT) continua to address the critique raised by Massaro (2009) and show that the obtained effect was indeed the result of multimodal perception. Indeed, we found that though there was a shift toward voicelessness in the presence of air puffs, the responses still reflected the expected category boundary, thus showing a true integrative effect, where both the audio and the aero-tactile stimuli were considered. Moreover, the obtained results, a shift in perception towards voicelessness in the presence of air puffs for the two VOT continua but not for the vowel continuum, show that somatosensory information modulates the perception of speech only when it is relevant for the task, that is, only when the somatosensory cues are congruent with those expected in production.

Chapter 2: References

Aiello, J. R., & Jones, S. E. (1971). Field study of the proxemic behavior of young school children in three subcultural groups. *Journal of Personality and social psychology*, *19*(3), 351.

Aiello, J. R., & Aiello, T. D. C. (1974). The development of personal space: Proxemic behavior of children 6 through 16. *Human ecology*, *2*(3), 177-189.

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339-355.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & Oc´onnell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal–hearing and hearing–impaired subjects. *Journal of the Acoustical Society of America*, *90*(6), 2971-2984.

Bicevskis, K. (2015). *Visual-tactile integration and individual differences in speech perception*. (Master's thesis, The University of British Columbia). Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0166756

Burnham, D., & Dodd, B. E. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D. G. Stork & M. E. Hennecke (Eds.), *Speech reading by humans and machines: Models, systems, and applications* (pp. 103–114). Berlin: Springer–Verlag.

Byrd, D. (1993). 54,000 american stops. *UCLA Working Papers in Phonetics*, *83*, 97–116.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of cognitive neuroscience*, *16*(6), 1000-1009.

Cooper, A. M. (1991). *An articulatory account of aspiration in English*. (Unpublished doctoral dissertation). Yale University, New Haven, CT.

DeLeeuw, K. E., & Mayer, R. E. (2008). A comparison of three measures of cognitive load: Evidence for separable measures of intrinsic, extraneous, and germane load. *Journal of Educational Psychology*, *100*(1), 223.

Derrick, D., Anderson, P., Gick, B., & Green, S. (2009). Characteristics of air puffs produced in English "pa": Experiments and simulations. *Journal of the Acoustical Society of America*, *125*(4), 2272-2281.

Derrick, D., & Gick, B. (2013). Aerotactile integration from distal skin stimuli. *Multisensory Research*, *26*, 405–416.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology*, *1*(2), 121-144.

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, *55*, 149-179.

Eagleman, D. M., & Holcombe A. O. (2002). Causality and the perception of time. *Trends in cognitive sciences* 6(8). 323–325.

Eagleman, D. M. (2008). Human time perception and its illusions. *Current*

opinion in neurobiology, *18*(2), 131-136.

Engonopulos, N., Sayeed, A., & Demberg, V. (2013). Language and cognitive load in a dual task environment. *Proceedings of the Annual Meeting of the Cognitive Science Society 35*(35).

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech, Language, and Hearing Research, 24*(1), 127-139.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception & Psychophysics, 36*(4), 359-368.

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America, 99*(3), 3.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross–modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816–828.

Geers, A., & Brenner, C. (1994). Speech perception results: Audition and lipreading enhancement. *Volta Review*, *96*(5), 97–108.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*, 502– 504.

Gick, B., Ikegami, Y., & Derrick, D. (2010). The temporal window of audio-tactile integration in speech perception. *The Journal of the Acoustical Society of America*, *128*(5), EL342-EL346.

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of the Acoustical Society of America*, *123*(4), EL72-EL76.

Goldstein, L., & Fowler, C. A. (2003). Articulatory phonology: A phonology for public language use. In *Phonetics and phonology in language comprehension and production: Differences and similarities*, 159–207.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*(3), 1197-1208.

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In B. E. Dodd & R. Campbell (eds.), *Hearing by eye II: Advances in the psychology of speechreading and audiovisual speec*h, (pp. 3–26). Hove, UK: Psychology Press.

Haggard P., Clark S., & Kalogeras J. (2002). Voluntary action and conscious awareness. *Nature Neuroscience 5(*4), 382–385.Hall, E. T. (1966). *The hidden dimension*. Garden City, NY: Doubleday.

Hardison, D. M. (2007). The visual element in phonological perception and learning. In M. C. Pennington. (ed.), *Phonology in context*, (pp. 135-158). London: Palgrave Macmillan.

Harrar, V., & Harris, L. R. (2008). The effect of exposure to asynchronous audio, visual, and tactile stimulus combinations on the perception of simultaneity. *Experimental brain research*, *186*(4), 517-524.

Isshiki, N., & Ringel, R. (1964). Air flow during the production of selected consonants. *Journal of Speech and Hearing Research, 7*(3), 233-244.

Isshiki, N., & von Leden, H. (1964). Hoarseness: aerodynamic studies. *Archives of otolaryngology, 80*(2), 206-213.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1245-1248.

Jaeger, J. J. (1978). Speech aerodynamics and phonological universals. *Annual Meeting of the Berkeley Linguistics Society 4*, 312-329.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*, 390–404.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Klatt, D. H. (1980). Speech perception: A model of acoustic-phonetic analysis and lexical access. *Perception and production of fluent speech*, 243-288.

Kluender, K. R. (1994). Speech perception as a tractable problem in cognitive science. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics*, (pp. 173-217). San Diego, CA, US: Academic Press.

Kuznetsova A., Brockhoff P. B. and Christensen R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1-26.

Lachs, L., Pisoni, D. B., & Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear and Hearing*, *22*(3), 236-251.

Lefcheck, J. & Sebastian Casallas, J. (2014). *R-squared for generalized linear mixed-effects models.* Retrieved from https://github.com/jslefche/rsquared.glmm

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in english stops. *Language and Speech*, *10*(1), 1-28.

Macleod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.

Massaro, D. W. (1987). Speech perception by ear and eye. In B. E. Dodd & R. Campbell (eds.), *Hearing by eye: The psychology of lip–reading*, (pp. 53–83). Hillsdale, NJ: Lawrence Erlbaum.

Massaro, D. W. (2009). *Caveat emptor: The meaning of perception and integration in speech perception*. Available from Nature Precedings http://hdl. handle. net/10101/npre, 1.

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, *21*(4), 445-478.

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language, 65*(2), 145-160.

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

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, *82*(6), 2145-2147.

Mills, A. E. (1987). The development of phonology in the blind child. In B. E. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip–reading* (pp. 145–161). Hillsdale, NJ: Lawrence Erlbaum.

Miyazaki, M., Yamamoto, S., Uchida, S., & Kitazawa, S. (2006). Bayesian calibration of simultaneity in tactile temporal order judgment. *Nature neuroscience*, 9(7), 875.

Molholm, S., Ritter, W., Murray, M. M., Javitt, D. C., Schroeder, C. E., & Foxe, J. J. (2002). Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive brain research*, *14*(1), 115-128.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & psychophysics*, *58*(3), 351-362.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage*, *59*(1), 781-787.

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

Penfield, W., & Rasmussen, T. (1950). *The cerebral cortex of man: A clinical study of localization of function*. New York, NY: Macmillan.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Randolph, M. (1989) *Syllable-based Constraints on Properties of English Sounds*. (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/4197/RLE-TR-555-21983908.pdf?sequence=1

Reed, C. M., Durlach, N. I., Braida, L. D., & Schultz, M. C. (1989). Analytic study of the Tadoma Method: Effects of hand position on segmental speech perception. *Journal of Speech, Language, and Hearing Research*, *32*, 921–929.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to

understand: A lip– reading advantage with intact auditory stimuli. In B. E. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip–reading* (pp. 97–114). Hillsdale, NJ: Lawrence Erlbaum.

Remez, R. E., Fellowes, J. M., Pisoni, D. B., Goh, W. D. & Rubin, P. E. (1998). Multimodal perceptual organization of speech: Evidence from tone analogs of spoken utterances. *Speech Communication 26*(1). 65–73.

Rosen, S. M., & Howell, P. (1981). Plucks and bows are not categorically perceived. *Perception & Psychophysics*, *30*(2), 156-168.

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 51–78). Malden, MA: Blackwell.

Rosenblum, L. D. & Gordon, M. S. (2001). The generality of specificity: Some lessons from audiovisual speech. *Behavioral and Brain Sciences 24*(02), 239–240.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Attention, Perception & Psychophysics*, *59*(3), 347–357.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*, 1147–1153.

Shimojo, S. & Shams L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current opinion in neurobiology 11*(4), 505–509.

Sparks, D. W., Kuhl, P. K., Edmonds, A. E., & Gray, G. P. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental

features of speech. *Journal of the Acoustical Society of America*, *63*(1), 246-257.

Spence, C., & Bayne, T. (2014). Is consciousness multisensory. In D. Stokes, M. Matthen & S. Biggs (Eds.), *Perception and its modalities*, (pp. 95-132). Oxford: Oxford University Press.

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, *9*(4), 255.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: its organization and maturation. *Hearing research*, *258*(1-2), 4-15.

Stetson C., Cui X., Montague P. R., & Eagleman D. M. (2006). Motor-sensory recalibration leads to an illusory reversal of action and sensation. *Neuron 51*, 651–659.

Stevens, K. N. (1981). Constraints imposed by the auditory system on the properties used to classify speech sounds: Data from phonology, acoustics, and psychoacoustics. In *Advances in psychology* (Vol. 7, pp. 61-74). North-Holland.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics*, *17*, 3-45.

Stevens, K. N. (2000). *Acoustic phonetics*. MIT Press: Lawrence Erlbaum.

Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America*, *111*(4), 1872-1891.

Stoffregen, T. A. & Bardy, B. G. (2001). On specification and the senses. *Behavioral and Brain Sciences 24*(02), 195–213.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio–visual speech perception. In B. E. Dodd, & R. Campbell (eds.), *Hearing by eye: The psychology of lip reading*, (pp. 3–51). Hillsdale, NJ: Lawrence Erlbaum.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607.

Zwicker, E., & Fastl, H. (2006). *Psychoacoustics: Facts and models*. 2nd ed. Berlin: Springer-Verlag.

Chapter 3:

Effects of Aero-Tactile Information on Perception of VOT Continua in Non-Initial Positions

3.1 Introduction

In the previous chapter we demonstrated audio-tactile integration by creating a VOT continua ranging from a CV token with an initial voiceless stop (e.g., /pa/) to a CV token with an initial voiced stop (e.g., /ba/). The participants were asked to identify the sound they heard by pressing a button marked with a corresponding letter. In half of the trials, the participants felt an air puff that was blown on the back of their hand. More voiceless responses were recorded in trials that were accompanied by air puffs than in trials that were not. We concluded that the puffs of air shifted the participants' perception of voicing, showing integration of the auditory and the tactile modalities in the perception of speech. In this chapter we use the same experimental setting to further explore the effect of air puffs on listeners.

One of the main acoustic cues for the voicing contrast for stops is Voice Onset Time (VOT) (Lisker & Abramson, 1964, 1967, 1970; Flege, 1982; Keating, 1984). VOT

refers to the relative timing of vocal fold vibration and the opening of the oral closure in the production of a stop consonant that is followed by a sonorant. The presence or absence of *aspiration* in consonant produced by air pressure from the lungs follows from the value of the VOT. If there is a positive temporal lag between the opening of the oral closure and the vibration of the vocal folds, the speaker produces a puff or air, also known as aspiration. For English, the presence or absence of aspiration is only typically used for disambiguating voicing categories in initial positions, that is, word initially and at the onset of a stressed syllable. These environments were collapsed into one by Kiparsky (1979) who treats both as foot initial. In medial positions aspiration is not a cue for the voicing distinction since it is mostly not part of the physical signal (Lisker 1957, 1984, 2002). Instead, listeners are distinguishing voiced from voiceless sounds in these positions based on one or more of the following: duration of preceding vowel, consonant closure duration, ratio of preceding vowel to consonant closure duration, formants transition at the vowel edge, and voicing during closure (Lisker, 1957, 1986; Kingston & Diehl 1994).

We argue that the puffs of air were interpreted by the participants in the experiment in the previous chapter as aspiration, and therefore as relevant to the task of distinguishing word-initial stops. We predict further that the puffs of air would not be perceived as relevant for the task of distinguishing word-medial stops. In the current study we used the same setup described in the previous chapter to present a continuum of stops in medial positions, ranging from /ˈa.pa/ to /ˈa.ba/, rather than a continuum of stops in initial positions (e.g. from /pa/ to /ba/). Aspiration seldom occurs in this context; thus, we do not expect listeners to interpret the puffs of air as relevant to the task.

Relevance to the task is crucial for integration: information that is not directly related to the sounds being disambiguated is not expected to affect perception. Ito et al. (2009) demonstrated this point, as discussed in the previous chapter.

The kind of direct link between somatosensory stimulus and speech perception that is demonstrated by Ito et al. (2009) is not reflected in Gick & Derrick (2009) and the experiment described in the previous chapter. However, we nonetheless interpret the finding from both studies as suggesting that aero-tactile stimulation is relevant for disambiguating aspirated from non-aspirated sounds. Both studies included a condition where air puffs were not relevant for the disambiguation being made: tapping was found by Gick & Derrick (2009) to be irrelevant, as did aero-tactile stimulation, in an environment where the disambiguated sounds were aspirated to the same degree (the vowel continuum in the previous chapter). We conclude that the aero-tactile stimuli used in these experiments, puffs of air, were interpreted as relevant information. That is, they were congruent with the cues that are expected in production of aspirated sounds.

In Ito et al. (2009), Gick & Derrick (2009), and the experiment conducted in the previous chapter, then, integration occurred when the integrated information was relevant for the task. Since aspiration is not used for distinguishing voicing in medial position, it is predicted to function in a similar way to the rearward deformation in Ito et al. (2009), the tapping in Gick & Derrick (2009), and the vowel condition in the previous chapter. That is, it is not predicted to influence voicing judgments.

3.2 Methods

3.2.1 Participants

37 monolingual native speakers of American English participated in the experiment (26 females; age range 18-54, mean age 32.6, SD = 10.4). The participants were all residents of Southern Connecticut at the time of the experiment but were born and raised in multiple regions of the US, including the northeast, the northwest, the west coast, the Midwest and the south. Their level of education ranges from high school graduates to graduate students. The participants were recruited with flyers and by word of mouth. All were naive to the purpose of the study and had no self-reported speech or hearing defects. They were compensated for their time. All of the participants signed an informed consent form approved by the Yale Human Research Protection Program.

3.2.2 Stimuli

3.2.2.1 Acoustic Stimuli

The stimuli were created by recording a male monolingual native speaker of American English. The speaker, a graduate student resident of Southern Connecticut that was born and raised in northern New Jersey and has attended college and earned his BA in Pennsylvania before moving to CT. The speaker produced six tokens of each of the nonce-words /ˈa.pa/ and /ˈa.ba/, with a stress placed on the first syllable. One eight-step consonant closure-duration continuum was created by shortening the closure duration of the voiceless token in log-scaled steps, with the final step matching the duration of the voiced token. Closure durations for each step of the continuum appear in Table 3.1. A

nonlinear (logarithmic) step size was chosen because psycho-acoustic perception tends to follow Weber's law (subjective sensation is proportional to the logarithm of the stimulus intensity); e.g., Zwicker & Fastl (2006). See Rosen and Howell (1981) for results on VOT, and Stevens, (2000, p. 228) for a similar effect on the perception of duration of burst. We chose to vary closure-duration since it has been shown that it provides sufficient information for distinguishing voicing in intervocalic stops by speakers of American English (Lisker, 1957).

Step no.	Closure duration (ms)
1	85
2	55
3	36
4	23
5	15
6	10
7	6
8	4

Table 3.1. Voicing continuum steps showing length of retained silent closure (ms).

A pre-test of the continuum was conducted online as a Mechanical Turk task and was used to assess the quality of the stimuli. The test was run with an independent group of participants that did not take part in the main study (N = 41). They were asked to choose whether they heard an "apa" or an "aba" and rate the goodness of the token on a five step Likert scale. The order of presentation was randomized. The results of the pretest are plotted in Figure 3.1. The category boundary is approximately centered between its

endpoints, that is, its bias (4.3) is close to its midpoint (4.5). The bias was calculated as the 50% crossover point of the psychometric categorization function for the continuum, computed across all listeners. Acuity (a measure of boundary slope) was computed as the difference between the 25% and 75% probabilities for the categorization function. The continuum's acuity is 2.9. Although the responses reflect the expected shape of a categorical distinction function, the goodness ratings are not significantly different along the continuum. This suggests that although closure duration alone is sufficient for distinguishing stops in medial positions, it is not a perfect cue.

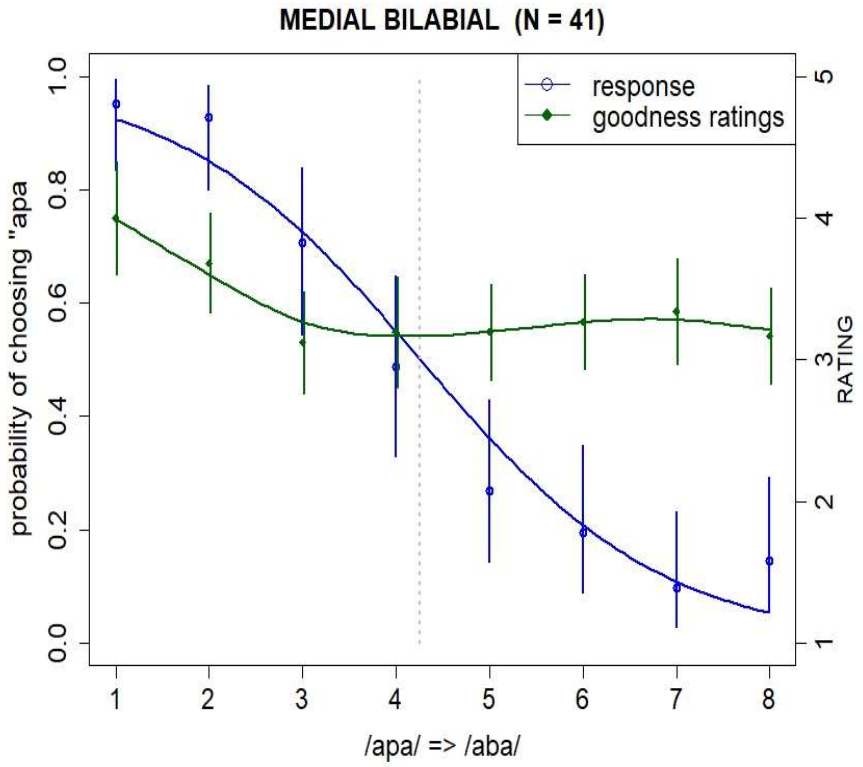

Figure 3.1. Viability test results for the continuum: left scale (blue line) shows probability of choosing voiceless relative to step (dotted vertical line marks 50% crossover point); right scale (green line) shows Likert scale ratings by step. Error bars show 95% confidence intervals.

3.2.2.2 Tactile (Air Puff) Stimuli

The tactile information was delivered as described in section 2.2.2.2. Detectable air turbulence exiting the tube was 87 ms in duration. This timing reflects observed values for voiceless (aspirated) bilabial stops in onset position. This timing was chosen in order to test the hypothesis that aero-tactile stimuli is integrated by listeners only when it is task-relevant. In the previous chapter, where the participants disambiguated bilabial stops in onset position, this timing reflected the aspiration in the voiceless end of the continuum. In the current study, the timing is not appropriate, as the auditory stimuli does not contain aspiration, even in the voiceless end of the continuum.

3.2.3 Procedure

As in the previous chapter, each experimental session included two parts, an initial test to verify that the air puffs were felt but not heard, seen or otherwise perceived, and the main part, which tested participant responses to the auditory stimuli in the presence and absence of air puffs. Stimuli were presented to the participants through ear-enclosing headphones (Sennheiser HD 202 II).

3.2.3.1 Puff Detection Test

The puff detection test was as described in section 2.2.3.1.

3.2.3.2 Perturbed continuum Testing

The perturbed continuum testing was as described in section 2.2.3.2. Five blocks were presented during which sounds drawn from the continuum were tested. Each block included six repetitions of each step of the continuum, for which three instances were accompanied by air puffs and three were not, randomly ordered. In total, the participants were presented with 5 blocks *x* 3 repetitions *x* 2 puff conditions (+/-) *x* 8 continuum steps for a total of 240 separate judgments, with 15 per condition at each continuum step. In each trial, participants were asked to identify the stimulus they heard and to press the corresponding button on a response box: either "apa" or "aba" to indicate the word they heard. The presentation order of the auditory stimuli and the accompanying tactile information (puff present vs. absent) were pseudo-randomized throughout each block. For half the participants, the right button on the response box indicated a syllable with a voiceless consonant (e.g., "apa"). For the other half, the right button indicated a syllable with a voiced consonant.

3.3 Results

3.3.1 Puff Detection Test

In the first block of the puff detection test, when their hand was close to the exit point of the tube, participants correctly discriminated puff/no puff conditions at an average rate of 96.59% (s.d. 2.9), with the worst performer at 90%. An exact binomial test confirms that these recognition percentages were well above chance ($p < 0.01$). In the second block, with their hand positioned away from the tube and everything else the same,

participants were at chance: 50.7% (s.d. 2.6); best performer 58% (binomial test n.s.). These results confirm that the participants felt the puff of air on their hand, but could not hear, see, or otherwise detect it.

3.3.2 Perturbed Continuum Testing

In 337 of the trials (3.8% of the trials) an air puff was requested but not delivered, or not requested but delivered. These trials were excluded from analysis, along with additional 63 trials for which the button-press response time exceeded 7 seconds (~3 s.d.). The data were then modeled with logistic regression in R (R Core Team, 2016) to estimate the effects of puffs on the perceptual boundary. 60% of the participants (N = 22) showed the expected category boundary. The reminder of the participants (N = 15) did not distinguish a category boundary in either +puff or -puff condition. The responses of these participants were skewed towards the voiceless alternative, in both puff conditions. An additional experiment was performed to evaluate the possible priming effect of air puff on some of the listeners who did not distinguish a category boundary (N = 10).

3.3.2.1 Quantifying the effect of puffs on the perceived categories of the participants who showed the expected baseline

Figure 3.2 shows the estimated psychometric functions, pooled across the 22 listeners who showed the expected baseline, in the presence and absence of air puffs. The vertical axis represents the probability of choosing a voiceless token (/ˈa.pa/). The horizontal axis shows the 8 steps along the continuum. The baseline condition, without puff, is

shown in blue lines with circles, and the condition with air puffs is shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25% and 75% probability points along each curve; the distance between these points gives the acuity (a measure of the slope of the boundary). The shift of the bias to the right in the presence of air puffs in the two reflects the fact that there were more voiceless responses in this condition, though not significantly so.

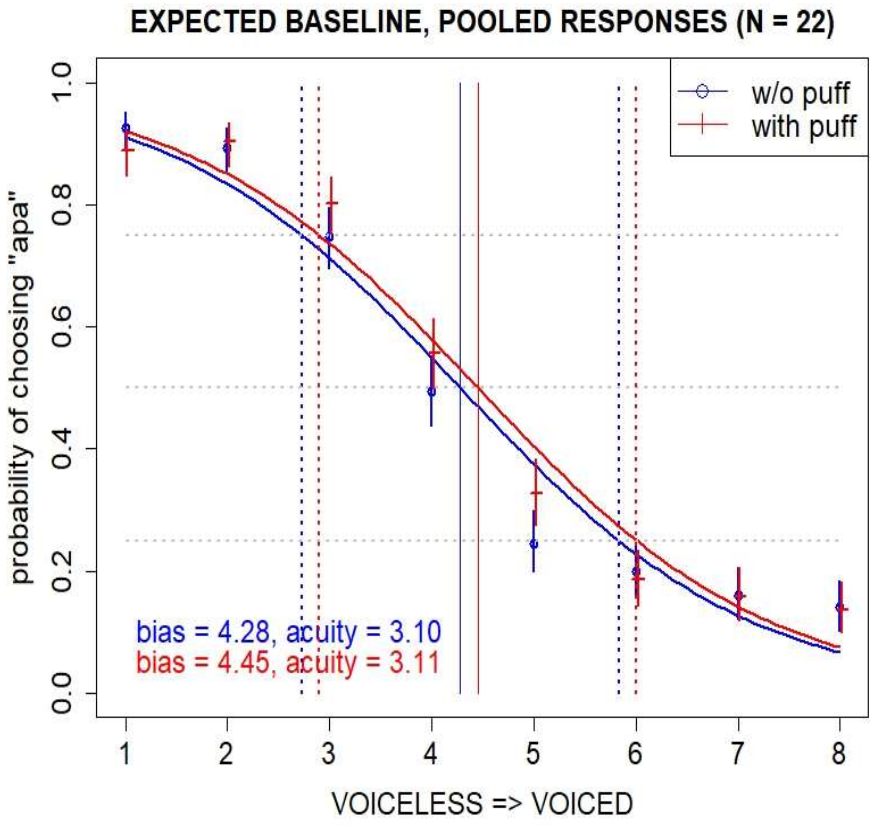

Figure 3.2. Perceived category boundary, pooled across speakers, with (red) and without (blue) an air puff. Vertical lines show the bias (50%) crossover, which is shifted in the direction of voicelessness for +puff trials (though not significantly). 95% confidence intervals are indicated for each pooled response.

A generalized linear mixed-effects model (GLMM) computed with the lme4

package (Bates et al. 2015) was used to assess the significance of the puffs contrast the continuum. In this model[7] the dependent variable (the probability of choosing a voiceless response) was predicted by the fixed effect of PUFF (-/+) and a continuous covariate of STEP, with random intercepts by participant ID, and random slopes for CSTEP by participant. The addition of interaction term for PUFF and CSTEP did not improve the fit of the model ($\chi^2(1) = 0.0141$, $p = 0.906$). The results, summarized in Table 3.2, show no effect of PUFF on the continuum. The effect of CSTEP was significant.

	-Air PUFF (baseline) vs. +Air PUFF		
	Coefficients	z-value	p-value
+PUFF	0.121	1.598	0.11 n.s.
CSTEP	-0.825	-11.383	< 0.001 ***

Table 3.2. Output of the GLMM response model.

An analysis of individual results was conducted to assess the degree to which individual participants were sensitive to the effect of the air puffs. We computed separate logistic regression models for each participant, with response predicted by the fixed effect of PUFF and STEP as a continuous covariate.[8] A little over half of the participants showed small and non-significant shifts towards the voiceless response under +PUFF condition (14/22; binomial test n.s.). One of the participants showed a significant shift (coefficient = 0.721; z = 1.977; p = 0.048). See Table 3.3 for summary statistics.

[7] `glmer(RESP ~ PUFF + CSTEP + (1+CSTEP|ID), family=binomial)`
[8] `glm(RESP ~ PUFF + CSTEP, family=binomial)`

Mean coefficient	0.162
s.d. of coefficient	0.438
Range of Coefficient	-0.5595674 : 1.471482

Table 3.3. Summary of the individual models computed for the participants.

3.3.2.2 Quantifying the effect of puffs on the perceived categories of the participants who did not show the expected baseline

15 participants did not distinguish a category boundary, consistently selecting responses skewed towards the voiceless alternative. Figure 3.3 shows the estimated psychometric functions, pooled across these listeners, in the presence and absence of air puffs. The vertical axis represents the probability of choosing a voiceless token (/ˈa.pa/). The horizontal axis shows the 8 steps along the continuum. The baseline condition, without puff, is shown in blue lines with circles, and the condition with air puffs is shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25% and 75% probability points along each curve; the distance between these points gives the acuity. The bias was not significantly shifted under +PUFF.

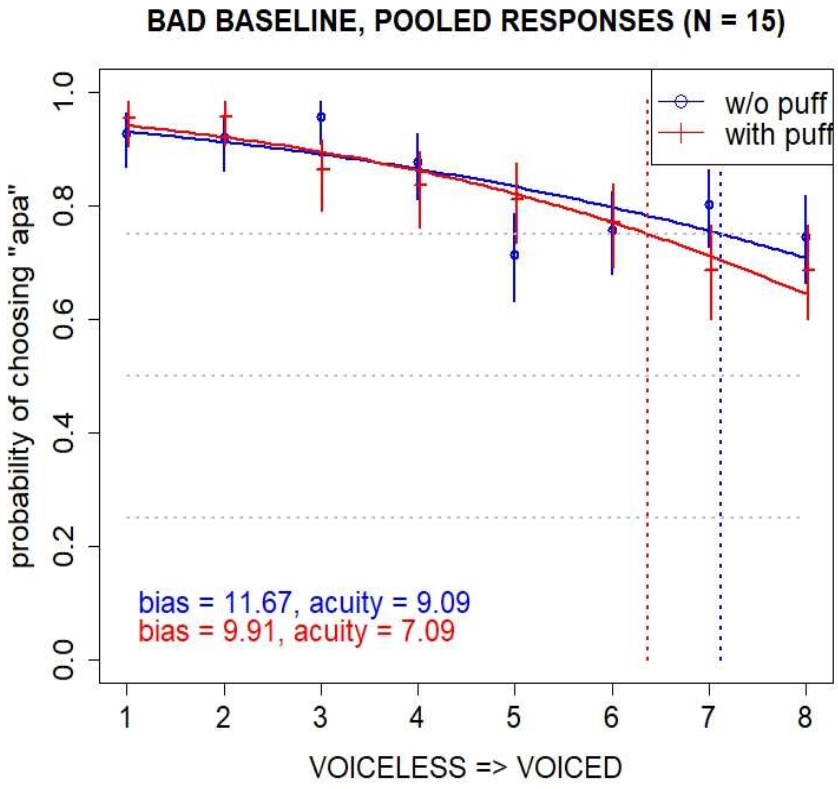

Figure 3.3. Perceived category boundary, pooled across speakers, with (red) and without (blue) an air puff. Vertical lines show the bias (50%) crossover, which is systematically shifted in the direction of voicelessness for +puff trials (though not significantly). 95% confidence intervals are indicated for each pooled response.

A generalized linear mixed-effects model (GLMM) computed with the lme4 package (Bates et al. 2015) was used to assess the significance of the puffs contrast the continuum. In this model[9] the dependent variable (the probability of choosing a voiceless response) was predicted by the fixed effect of PUFF (-/+) and a continuous covariate of STEP, with random intercepts by participant ID, and random slopes for CSTEP by participant. The results, summarized in Table 3.4, show no effect of PUFF on the

[9] `glmer(RESP ~ PUFF + CSTEP + (1+CSTEP|ID), family=binomial)`

continuum. The effect of CSTEP was significant. The addition of interaction terms for PUFF and STEP did not improve the fit of the model ($\chi^2(1) = 1.557$, $p = 0.212$).

	-Air PUFF (baseline) vs. +Air PUFF			
	Coefficients	z-value	p-value	Significance
+PUFF	-0.115	-0.95	0.341	
CSTEP	-0.318	-4.37	< 0.0001	***

Table 3.4. Output of the GLMM response model.

3.3.2.3 Assessment of possible priming effect of puffs on the perceived categories of the participants who did not show the expected baseline

In order to assess the possible priming effect of air puffs on the participants that did not show the expected baseline we conducted an additional experiment with 10 of the listeners, who agreed to return and participate. The post-testing was done between 46 and 110 days after the original experiment. It was similar to the original perceptual test but did not introduce any aero-tactile stimuli. That is, it included the perturbed continuum testing, but not the preceding puff detection test. The perturbed continua testing was done with the same procedure as the original experiment, except for the +PUFF condition. The hand of the participants was placed at the same position, close to the exit point of the air-tube, but they were told no air is expected to be blown on it, and indeed, no air was blown. Figure 3.4 shows the estimated psychometric functions, pooled across the 10 listeners who did not show the expected baseline, for the -PUFF trials, in the original experiment and the post-testing. The vertical axis represents the

probability of choosing a voiceless token (/ˈa.pa/). The horizontal axis shows the 8 steps along the continuum. The baseline condition, -PUFF trials from the original experiment, is shown in blue lines with circles, and the trials from the post-testing are shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25% and 75% probability points along each curve; the distance between these points gives the acuity.

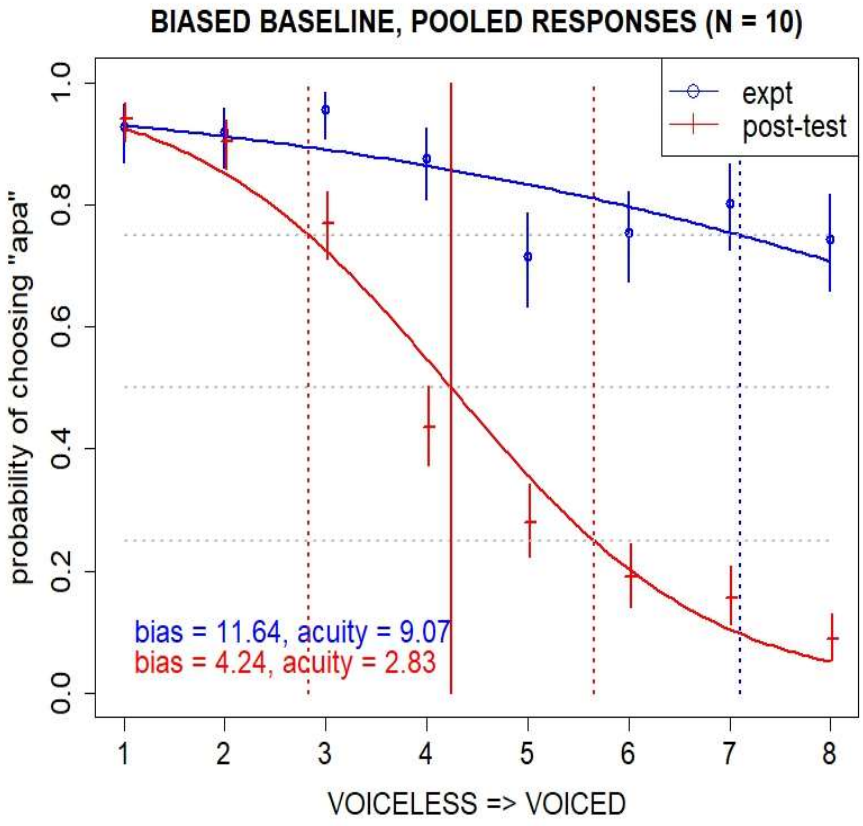

Figure 3.4. Perceived category boundary pooled across speakers for -PUFF trials, in the original experiment (blue) and post-testing (red). Vertical lines show the bias (50%) crossover. 95% confidence intervals are indicated for each pooled response.

A generalized linear mixed-effects model (GLMM) computed with the lme4 package (Bates et al. 2015) was used to compare the -PUFF trials in the original

experiment and the post-testing. In this model[10] the dependent variable (the probability of choosing a voiceless response) was predicted by the fixed effect of CONDition (original experiment/follow-up experiment), a continuous covariate of STEP, and their interaction, with random intercepts by participant ID, and random slopes for CSTEP by participant. These factors were supported by model selection ($\chi^2(1) = 15.598$, $p < 0.001$). The results, summarized in Table 3.5, show a significant effect of condition on the continuum. The positive coefficient indicates decreased likelihood of voiceless response in the later experiment where no air puffs were present. The significant interaction between COND and CSTEP supports the evidently stronger effect of condition at the voiced end of the continuum.

	CONDition: Original experiment (baseline)		
	Coefficients	z-value	p-value
CONDpost	1.186	2.174	0.023 *
CSTEP	-0.208	-1.792	0.073 .
CONDpost:CSTEP	-0.795	-4.777	< 0.001 ***

Table 3.5. Output of the GLMM response model.

3.3.2.4 Assessment of a possible learning effect

An additional analysis was conducted to determine whether there was a learning effect for the population that did not show the expected baseline *in the original experiment*. This analysis was conducted to assess whether the overwhelming tendency to choose a voiceless response was fixed during the perturbed continua testing or learned

[10] `glmer(RESP ~ COND * CSTEP + (1+CSTEP|ID), family=binomial)`

throughout it. First, we added trial number as a factor to the model. The effect of TRIAL was not significant ($z = -1.01$, $p = 0.32$). Then we binned the experiment into an early bin (trials 1-16 out of 48), later bin (trials 17-32 out of 48) and latest bin (trials 33-48 out of 48). The effect of BIN was not significant either ($z = -0.94$, $p = 0.35$). This suggests that the bias in the direction of voicelessness was affected by the puff detection test and already fixed at the beginning of the perturbed continua testing.

3.4 Bayesian Analysis as a Tool for Quantifying the Variable Behavior

3.4.1 Two Patterns of Behavior

Two patterns of behavior were observed in the current study. 60% of the participants in the experiment showed the first pattern, where the expected category boundary was recorded in the baseline condition. 40% of the participants showed the second pattern, where the expected category boundary was not recorded in the baseline condition, and instead the responses were biased towards voicelessness in both the baseline and the experimental condition. For both groups, no shift of the category boundary in the presence of air puffs was found. In the following discussion, the former group of participants will be called “Expected Baseline” and the latter group of participants will be called “Primed”.

The responses of the participants in the Expected Baseline group reflected the expected category boundary in the baseline condition and were not statistically different

in the +PUFF condition. These participants demonstrated the expected behavior, as they disregarded the tactile information as not relevant to the task, or at least as less relevant than it was for the participants in previous studies, where aspiration was found to be relevant for the disambiguation being made (Gick & Derrick, 2009; Derrick & Gick, 2013, Chapter 1). In contrast with the previous studies, the participants in the Expected Baseline group did not integrate the aero-tactile information with the acoustic information. We suggest that no integration occurred since there was nothing to integrate: these participants interpreted the puffs of air as the tactile manifestation of aspiration and did not make use of it since aspiration was not part of the heard signal.

However, the listeners in the current experiment varied in their susceptibility to the effect of air-puffs. The Primed participants neither distinguished the expected category boundary in the baseline condition nor in the +PUFF condition. 10 of the Primed 15 participants agreed to come back for an additional experiment, during which the same acoustic stimuli were used, but no tactile information was presented. The same participants that did not show the expected category boundary in the original experiment, did show it in the post-testing. That is, when no tactile information was provided, they showed the same category boundary as the participants in the Expected Baseline group, suggesting that the overwhelming tendency to choose the voiceless response was affected by the presence of air puffs. Crucially, the effect was due to the presence of the air puffs in the experiment, not in a given trial, since the effect was evident both in trials where there was an exposure to air puffs, and trials where there was no such exposure. That is, this was not an integrative process, where the puffs of air applied to the skin of the participants changed their perception, it was a priming effect. Since no effect of trial

was found, nor an effect in early versus late trials binned together, we conclude that the effect was induced by the exposure to air puffs in the puff detection test. To summarize, the Primed participants did not integrate in the same way participants in the previous studies mentioned above did, as they demonstrated a different behavior, but unlike the participants in the Expected Baseline group, they did react to the presence of air puffs.

3.4.2 Quantifying the Differences between the Groups

One way of understanding the difference between the two groups is in terms of *evidence/signal* and *expectations*. All the participants in the experiment were exposed to the same *evidence* for a category. That is, they heard the same acoustic *signal* and felt the same tactile *signal*. We suggest that the air puffs influenced the Primed participants by changing their *expectations*, and that this change of expectations affected their perceptual behavior. It has been demonstrated that participants' expectations shape their speech perception. Social expectations have been shown to influence word recognition and phoneme categorization (e.g., Niedzielski, 1999; Drager, 2010; Hay & Drager, 2010; McGowan, 2015; Nguyen, 2017). Niedzielski (1999), for example, showed that when American listeners were told they are listening to a Canadian speaker they matched the vowels they heard with vowels that exhibited Canadian Raising, and when they were told they are listening to an American speaker they matched the same vowels to American-Accented vowels. Information-based expectations have been shown to influence perception of acoustic prominence (e.g., Bard & Aylett, 1999; Cole et al., 2010). Cole et al. (2010), for example, showed that naive listeners transcribed highly frequent words (frequent either in the given context or overall in the language) as non-

prominent, even when they were acoustically more prominent than other words in the given context.

We suggest that the Primed participants were primed by the air puffs at the puff detection test to expect voiceless stimuli. We used Bayesian reasoning to quantify the difference in expectations between the two groups. Bayesian models have been used in a wide range of psycholinguistic studies, including studies of categorical perception (e.g., Clayards et al., 2008; Norris & McQueen, 2008; Kleinschmidt & Jaeger, 2015; Norris et al., 2016; Nguyen, 2017). In these models, Bayesian priors are used to model the beliefs, knowledge, or expectations of speakers. Since the participants in the current experiment were in an experimental setting where a binary choice was presented to them, we assume that they expected each of the two choices presented to them (voiceless/voiced) to have a prior probability of 0.5. Other factors that may affect prior probability such as lexical statistics are not relevant in the context of the current study since the experimental tokens we used are non-words. The priming effect that the Primed participants were subjected to was caused by the exposure to air puffs in the puff detection test and reinforced throughout the perturbed continua testing by the continuous exposure to additional air puffs. This priming effect is manifested as a change in expectations, or in Bayesian terms, in the *prior probability* of choosing a voiceless response.

We calculated the difference in prior between the groups as follows, using Bayes's Rule (Lee, 2012): the *posterior probability* in Equation 1, p(voiceless|step+puff), is the probability of choosing a voiceless response given a specific signal (that is, a step along the acoustic continuum either accompanied or not accompanied by a puff of air),

computed across all the participants in a specific group (the Expected Baseline group or the Primed group), in the perturbed continua testing. The *prior probability*, p(voiceless), is the probability of choosing a voiceless response prior to exposure to the data in the perturb continua testing. The *likelihood,* p(step+puff|voiceless), is the probability distribution of a specific signal (acoustic and tactile) for the category voiceless in a given experiment and group of participants, or, in other words, the probability of the signal given the category voiceless. The denominator, p(step+puff) is the overall probability of a specific signal (acoustic and tactile) across both categories (voiceless and voiced) in the perturb continua testing of a specific group of participants.

$$p(voiceless|step + puff) = \frac{p(voiceless)\, p(step + puff|voiceless)}{p(step + puff)}$$

Equation 3.1. Posterior probability of choosing a voiceless response given a certain signal in the perturb continua testing.

For the Expected Baseline group, we computed the posterior probability for each step and puff condition (e.g., STEP 1 +PUFF) from the experimental data, as the proportion of voiceless responses out of all the observations for this specific step and puff condition, across all the participants in a given group of participants. The prior was set at 0.5, as described above. The denominator, the overall probability of the specific step and puff conditions, was computed from the experimental data, as the proportion of the observations for this specific step and puff condition out of all the observations. For example, the proportion of STEP 1 +PUFF for all the step and puff conditions, across all the participants in the group. Then we solved Bayes's rule for the likelihood, as detailed in Equation 2.

$$p(step + puff|voiceless) = \frac{p(voiceless|step + puff)\, p(step + puff)}{p(voiceless)}$$

Equation 3.2. Likelihood: the probability distribution of a specific signal (acoustic and tactile) for the category voiceless in a given experiment and group of participants.

The Primed participants were exposed to the same signal (acoustic and tactile) as the participants in the Expected Baseline group. Therefore, we assume that the likelihoods were the same for the two groups for each specific signal (e.g., STEP 1 +PUFF). Accordingly, we used the likelihoods that were computed for the Expected Baseline group to solve for the priors for the Primed group. We averaged across the resulting priors for the different step and puff conditions to arrive at a single prior. There was no reason to assume the prior was adjusted during the perturb continua testing, since no learning effect was found for this part of the experiment. Consequently, we assume that the change in the expectation to perceive a voiceless sound occurred during the puff detection test and was set before the beginning of the perturbed continua testing. The prior we arrived at after averaging was 0.818, which reflects a much higher probability of choosing a voiceless response for the Primed group than for the Expected Baseline group, whose prior probability of choosing a voiceless response was 0.5.

3.5 Discussion

The current study tested the effect of air puffs on perceptual judgements in the comparison between /ˈa.pa/ and /ˈa.ba/. The predicted result, no shift of the category boundary in the presence of air puffs, was found for the participants who showed the expected category boundary in the baseline condition, where no puffs of air were present

(60% of the participants, the Expected Baseline group). 40% of the participants did not show the expected category boundary in the baseline condition and instead gave responses biased towards voicelessness in all trials, i.e. trials with or without air puffs (the Primed group). A post-test confirmed that without exposure to the apparent priming effect of air-puffs these participants perceived the expected category boundary. The duration of the puffs of air used in the experiment was based on mean observed values for voiceless exemplars in onset position. This was incongruent with the experimental stimuli, which were drawn from stops in medial position, where no aspiration was present. The phonetic dimension that was varied in the stimuli was closure duration. Given that the puffs of air are interpreted by perceivers as aspiration, as argued in the introduction, integration of the air puff with the auditory stimuli was not expected. Indeed, an effect of puff was not found for either the participants who showed the expected baseline or the participants who did not. This is in contrast with the finding from the experiment conducted in the previous chapter, where the participants heard stops in onset position.

In the previous chapter we tested the effect of air puffs on the perception of a continua of sounds ranging from /pa/ to /ba/, from /ka/ to /ga/, and from /hɛd/ to /hɪd/. The current study tested the effect of air puffs on the perception of a continuum of sounds ranging from /ˈa.pa/ to /ˈa.ba/. The previous chapter found that the presence of air puffs significantly increased the likelihood of choosing voiceless responses for the two VOT continua but had no effect on choices for the vowel continuum. In the current experiment no shift of the category boundary in the presence of air puffs was found for either the participants in the Expected Baseline group or the Primed participants. Although the

Primed participants tended to choose voiceless responses overall, no significant difference was found between trials accompanied by air puffs and trials not accompanied by them.

The responses of the two groups of participants in the current experiment differ from each other and from those of the group of participants that were tested for the initial bilabial continuum in the previous chapter. In the following discussion, the latter group of participants will be called "Integrators". The responses of the Integrators to the baseline condition, where no puffs of air were presented, was reflective of the expected category boundary. In the +PUFF condition the responses were shifted towards voicelessness, though they still largely reflected the expected category boundary, demonstrating an integrative process, where both the acoustic and tactile stimuli were taken into account.

We have suggested that two factors affected the responses in the current experiment: the first factor is the *signal* that the participants heard and felt, which can be expressed as the *likelihood*, the probability of the signal given the category voiceless. The second factor is the *expectations* the participants had regarding the stimuli, which can be expressed as the *prior* probability of choosing a voiceless response. We suggest further that the Integrators were affected by the same factors. The participants in the current experiment and the Integrators were exposed to different signals (initial position vs. medial position, VOT continuum vs. closure-duration continuum). The aero-tactile information that the Integrators felt was congruent with the auditory information that was provided to them. They were affected by the aero-tactile stimuli, but also took into account the auditory information, as reflected by the fact that the expected category

boundary was observed in both -PUFF and +PUFF conditions, although the probability of choosing a voiceless response was significantly higher in the +PUFF condition. For the two groups of participants in the current experiment the likelihood of choosing a voiceless response was not higher in the +PUFF condition. This result was as expected. We interpret these results as reflecting the fact that the aero-tactile information was integrated by the Integrators, but not by any of the participants in the current experiment. Ito et al. (2009) have demonstrated that somatosensory information can be integrated with auditory information in perception of speech when it is task relevant. We claimed that the tap testing from Gick & Derrick (2009) and the vowel continuum from the previous chapter demonstrate the same point. The current results strengthen our claim: the participants consistently choose voiceless response, in both conditions, not just in the presence of air puffs. That is, there was no integration of somatosensory and auditory information. This is the result of the aero-tactile information not being relevant for the task: aspiration is not one of the cues for distinguishing /ˈa.pa/ from /ˈa.ba/.

The two groups of participants in the current experiment were exposed to the same signal. In Bayesian terms, the *likelihood*, the probability of the signal given the category voiceless, was the same for the two groups. However, the groups differed in the *prior* probability of choosing a voiceless response before any exposure to the data in the perturb continua testing. The participants in both groups took the acoustic information into account, as reflected by the significance of the effect of STEP for both groups, but the Primed participants were primed by the exposure to the puffs of air in the puff test such that they were biased towards choosing a voiceless response. This bias is reflected in the posterior probability of choosing a voiceless response that was observed for this

group, over 0.5 regardless of the STEP and PUFF conditions.

Why were some participants in the current experiment primed by the puffs felt during the puff detection test while others were not? To answer this question, we want first to convince ourselves that the difference between the groups was indeed a difference in the expectations, the prior, and not the likelihood, the probability of the signal given the category voiceless. Had the priors been the same for the Expected Baseline group and the Primed group, the likelihoods must have been different. However, the participants in both groups were exposed to the same signal, under the same conditions. Therefore, it is not likely that the likelihood terms were different. We conclude that the difference between the groups was in their priors. The participants in the Primed group were primed by the aero-tactile stimuli provided during the puff detection test such that they adjusted their expectations and were primed to expect a voiceless response. Why were the members of the Expected Baseline group not affected? It is possible that the participants in this group, most of the participants in the current experiment, were not primed to prefer a voiceless response, but primed to prefer an *aspirated* response. Given an experimental setting where no such response was provided, having given a choice between /ˈa.pa/ and /ˈa.ba/, these participants had no use for the preference for an aspirated response. Their expectations reflected just the experimental setting, where a binary choice was presented, and the distribution of the tokens is expected to be even during the experiment, as it was. Thus, these participants disregarded the air puffs they felt on their hands from time to time during the perturb continua testing as not relevant for the disambiguation being made. The minority of the participants, the Primed ones, had different adjusted expectations. For these participants, aspiration was relevant,

although it was not part of the acoustic signal. The acoustic signal was not aspirated and these participant did not interpret it as such, but they did link aspiration with voicelessness at some level of abstraction, at least enough to justify selecting /ˈa.pa/ based on an expectation for an aspirated sound, or enough to justify expectation for a voiceless sound based on aspiration alone.

What kind of representation of voicing can justify such expectations? This may be the result of a learned association between voicelessness or voiceless stops and aspiration from positions where aspiration is part of the acoustic signal, and generalization of this association at an abstract level. Adopting this explanation would require some mapping rules that associate the sensation of air puff, interpreted as aspiration, with abstract phonological objects (such as distinctive features or categories, depending on the details of the specific proposal). This kind of mapping is often referred to as *phonetic implementation* (Liberman & Pierrehumbert, 1984; Keating, 1985, 1990; Pierrehumbert, 1990. Note that none of these works consider somatosensory information a candidate for such mapping). Browman & Goldstein (1995) criticize models that use rules of phonetic implementation by arguing that they entail a loose relationship between the cognitive and physical level of representation, since the physical representation and the abstract phonological representation can essentially be independent of one another. Adopting phonetic implementation in the current case may be a good demonstration of this point, since it is not clear how a non-aspirated sound may be associated with aspiration.

Another possibility, that does not require phonetic implementation, are exemplar-based approaches (e.g., Goldinger, 1998; Pierrehumbert, 2006; Johnson, 2007). Exemplar based approaches are models of cognitive storage of aggregates of properties

that can include contextual information and fine phonetic detail. That is, the representation of linguistic categories is done in terms of these aggregates, that may include, among other things, sensory information. These models make use of the human capacity to make abstract generalizations, but memory according to these models does not depend on abstract generalizations (Port, 2010). The memory system in these models contains many detailed concrete instances and a set of category labels or another system that organizes the concrete instances in clusters. Abstractions and generalizations can be extracted from this system and are used in learning new phonological categories. In some of the models this learning process is done by estimating a probability distribution over the items that belong to the category. The distribution that is associated with a certain category is learned by observing the specific instances that already exist in the perceiver's perceptual space and applying a label over many instances that share a similar property. That is, clusters of instances in the perceptual space can be used for estimation of probabilistic distributions that are associated with phonological categories (Maye & Gerken, 2000; Maye et al., 2002; Feldman et al., 2013). In this way abstract units can play a role in exemplar models, but they arise in a bottom up fashion, and are not part of the mental representation of phonological units.

Pierrehumbert (2003) suggests that when categories are acquired in a bottom up fashion, they are first learned as positional variants of phonemes and only later refined into context-independent phonemes, using feedback from the community and from the lexicon. The final phonemes/categories in this system are labels over a cognitive map of items on which a metric of proximity is defined. Each such label has a probability distribution associated with it, and the items are represented as clusters of labels. In such

a system, an item might include the labels *aspirated* and *voiceless stop*, but not *foot initial*. This item may be at the tail of the probability distribution for the label *voiceless stop*, but included in the distribution, nonetheless. Such a model can also accommodate the variable behavior of the participants in the current study, as it allows assignment of different representations for different listeners. It is possible that for the majority of the listeners in the current study the space of exemplars that have the label *aspirated* but not the label *foot initial* is not populated, or that its density of population is very low.

In Chapter 5 we will revisit some of the questions that were discussed here in this chapter. Specifically, we will consider the possibility that for some speakers aero-tactile information cues something more abstract than aspiration, such as a [spread glottis] feature. We will argue that the behavior of the Primed participants can be accounted for by some theories of laryngeal phonology but not by others, and moreover, that it might be the case that the Primed participants and the participants in the Expected Baseline group have different phonological representations of stop consonants.

Another question that should be answered is the following. Why was a priming effect found for some of the participants in the current experiment but not for any of the participants in the previous chapter? A pattern where there is an overwhelming tendency to choose a voiceless response was not recorded in the previous chapter. We argue, again, that the puffs are interpreted by speakers as aspiration. In the case of the participants in the previous chapter aspiration was task-relevant, since it was part of the acoustic stimuli and crucial for the disambiguation being made. Therefore, it was factored into the *likelihood*, the probability of the signal given the category voiceless, and not into the *prior*, the expectation, or overall probability of a voiceless response.

3.6 Conclusion

The aim of the current study was to investigate the effect of air puffs on listeners during multisensory integration in speech perception, by utilizing the notion that somatosensory information is integrated with auditory information only when it is task relevant. Since aspiration is not used for distinguishing voicing in medial position, it was not predicted to influence voicing judgments in the comparison between /ˈa.pa/ and /ˈa.ba/. The predicted result, no shift in the category boundary in the presence of air puffs, was found for all the participants. However, 40% of the participants showed a priming effect where a bias towards voicelessness was found for all responses, regardless of the presence of puffs of air. We have argued that this bias is the result of a shift in the expectations of these participants, and modeled it using Bayesian reasoning, as a change in the prior probability of choosing a voiceless response for these participants, but not for the other 60% of the participants.

Chapter 3: References

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339-355.

Bard, E. G., & Aylett, M. P. (1999). The dissociation of deaccenting, givenness, and syntactic role in spontaneous speech. In *Proceedings of the 14th international congress of phonetic sciences (ICPhS)*, Vol. 3, 1753-1756. University of California Berkeley, CA.

Bicevskis, K. (2015). *Visual-tactile integration and individual differences in speech perception*. (Master's thesis, The University of British Columbia). Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0166756

Browman, C. P., & Goldstein, L. (1995). Dynamics and articulatory phonology. In R. F. Port & T. van Gelder (eds.), *Mind as motion*, 175-193. Cambridge, MA: MIT Press.

Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition, 108*(3), 804-809.

Cole, J., Mo, Y., & Hasegawa-Johnson, M. (2010). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology, 1*(2), 425-452.

Derrick, D., Anderson, P., Gick, B., & Green, S. (2009). Characteristics of air puffs produced in English "pa": Experiments and simulations. *Journal of the Acoustical Society of America*, *125*(4), 2272-2281.

Drager, K. (2010). Sociophonetic variation in speech perception. *Language and Linguistics Compass, 4*(7), 473-480.

Feldman, N. H., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological review, 120*(4), 751.

Flege, J. E. (1982). Laryngeal timing and phonation onset in utterance-initial English stops. *Journal of Phonetics 10*(2). 177–192.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*, 502– 504.

Goldenberg, D., Tiede, M. K., & Whalen, D. H. (2015). Aero–tactile influence on speech perception of voicing continua. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*. Glasgow, UK: The University of Glasgow.

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

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics, 48*(4), 865-892.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1245-1248.

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology. In honor of John Ohala*, 25-40.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*, 390–404.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language 60*(2). 286–319.

Keating, P. A. (1985). CV phonology, experimental phonetics, and coarticulation. *UCLA Working Papers in Phonetics, 62*, 1-13.

Keating, P. A. (1990). Phonetic representations in a generative grammar. *Journal of phonetics, 18*(3), 321-334.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Kiparsky, P. (1979). Metrical structure assignment is cyclic. *Linguistic inquiry 10*(3). 421–441.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological review, 122*(2), 148.

Lee, P. M. (2012). *Bayesian statistics: an introduction* (4th ed.). New York: Wiley.

Lefcheck, J. & Sebastian Casallas, J. (2014). *R-squared for generalized linear mixed-effects models.* Retrieved from https://github.com/jslefche/rsquared.glmm

Liberman, M. & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff, R. T. Oehrle, F. Kelley & B. W. Stephens (eds.), *Language sound structure*, 157–233. Cambridge, MA: MIT Press.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language, 33*(1), 42-49.

Lisker, L. (1984). How Is the Aspiration of English/p, t, k/" Predictable"?. *Language and Speech, 27*(4), 391-394.

Lisker, L. (1986). "Voicing" in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech, 29*(1), 3-11.

Lisker, L. (2002). The voiceless unaspirated stops of English. In B. E. Nevin & S. M. Johnson (eds.), *The legacy of Zellig Harris: Language and information into the 21th century*, 233–240. Amsterdam; Philadelphia: John Benjamins.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lisker, L., & Abramson, A. S. (1967). Some effects of context on voice onset time in english stops. *Language and Speech*, *10*(1), 1-28.

Lisker, L. & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th international congress of phonetic sciences (ICPhS)*, vol. 563, 563–567. Academia Prague.

Macleod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.

Maye, J., & Gerken, L. (2000, March). Learning phonemes without minimal pairs. In *Proceedings of the 24th annual Boston university conference on language development*, Vol. 2, 522-533.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101-B111.

McGowan, K. B. (2015). Social expectation improves speech perception in noise. *Language and Speech, 58*(4), 502-521.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

Nguyen, N. (2017). *The Role of Intergroup Attitudes in Speech Perception* (Doctoral dissertation, Western Sydney University (Australia)). Retrieved from https://tinyurl.com/yy3fbkym

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of language and social psychology, 18*(1), 62-85.

Norris, D., & McQueen, J. M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological review, 115*(2), 357.

Norris, D., McQueen, J. M., & Cutler, A. (2016). Prediction, Bayesian inference and feedback in speech recognition. *Language, cognition and neuroscience, 31*(1), 4-18.

Pierrehumbert, J. (1990). Phonological and phonetic representation. *Journal of phonetics, 18*(3), 375-394.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and speech, 46(*2-3), 115-154.

Pierrehumbert, J. B. (2006). The next toolkit. *Journal of phonetics, 4*(34), 516-530.

Port, R. F. (2010). Rich memory and distributed phonology. *Language Sciences, 32*(1), 43-55.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rosen, S. M., & Howell, P. (1981). Plucks and bows are not categorically perceived. *Perception & Psychophysics*, *30*(2), 156-168.

Stevens, K. N. (2000). *Acoustic phonetics*. MIT Press: Lawrence Erlbaum.

Zwicker, E., & Fastl, H. (2006). *Psychoacoustics: Facts and models*. 2nd ed. Berlin: Springer-Verlag.

Chapter 4:

Audio-Tactile Integration in the Perception of Thai

4.1 Introduction

While most of the studies cited in the chapters above, including the two experiments described in these chapters, were performed with speakers of American English, audio-visual integration has also been documented in other languages such as Italian (Bovo et al., 2009), Japanese (Sekiyama & Tohkura, 1991; Massaro et al., 1993; Sekiyama, 1994), Mandarin Chinese (Magnotti et al., 2015), Spanish (Massaro et al., 1993), and bilingual Mandarin Chinese-Dutch (de Gelder, 1992). However, to date no data is available for visuo-tactile or audio-tactile integration in the perception of speech for languages other than English. Employing aero-tactile stimuli in perceptual testing for other languages can be useful in understanding how the tactile stimuli are interpreted by participants during the process of integration. Specifically, in the current study, we ask whether a puff of air applied to the hand of a listener is interpreted as aspiration.

In Chapter 2 we tested the effect of aero-tactile information on perception of VOT continua. In American English voiceless stops are aspirated, with a long lag between the release of the consonant and the onset of voicing, while voiced stops are not aspirated, with a short lag between the release of the consonant and the onset of voicing (Lisker &

Abramson, 1964; Byrd, 1993). The continua we created ranged from long lag to short lag VOT (e.g., /pa/ to /ba/), with end values that reflected the values of a voiceless stop and a voiced stop produced by a native speaker of American English. We found that the presence of aero-tactile information, in the form of puffs of air delivered to the listener's hand, increases the likelihood of choosing a voiceless response. There are several possible explanations for this result. Our focus here is the possibility that the puffs of air were interpreted as the perceptual correlates of aspiration and were thus integrated as a source of information about the heard sound. The current experiment was designed to evaluate this possibility.

We used an experimental setting similar to this used in the previous chapters to investigate the effect of puffs of air in Thai, a language chosen for its relevance to the question at hand. Unlike English, Thai has a three-way voicing contrast for labial and alveolar stops. At both places of articulation there are aspirated voiceless stops, unaspirated voiceless stops, and voiced stops (Lisker & Abramson, 1964, 1970; Gandour & Dardarananda, 1982; Gandour, 1985). Thus, Thai speakers make use of aspiration in distinguishing aspirated voiceless stops from the other two stops in the series but not in distinguishing unaspirated voiceless stops from voiced stops. We used bilabial stops, as in the previous chapters. Our primary hypothesis is that air puffs that are felt by listeners are associated in perception with aspiration, thus shifting perception in those environments and contrasts in which aspiration is relevant for the distinction being made. We therefore expect an effect of aero-tactile information in Thai in the contrast between aspirated voiceless stops and unaspirated voiceless stops but not in the contrast between unaspirated voiceless stops and voiced stops. The stimuli for the experiment were

constructed based on productions made by a native speaker of Thai, in order to create continua that reflect accurately the voicing categories of the language. The experiment was conducted in Thailand to minimize influence from other voicing systems.

An effect of air puffs in the comparison between /p^ha/ and /pa/ but not in the comparison between /pa/ and /ba/ would show that the puffs of air are indeed interpreted as aspiration by speakers of Thai. It may seem trivial to assume that a puff of air is being perceived as aspiration, which is essentially a puff of air. However, there is no direct link between the experience a speaker might have from blowing puffs of air through her vocal tract and out of her mouth during the production of speech and a puff of air that is being blown on her hand (or the neck, or the ankle, cf. Gick & Derrick, 2009; Derrick & Gick, 2013). The lack of a direct connection between production and perception as in the skin deformation experiments of Ito et al. (2009) and in audiovisual perception differentiates the current case from past research in this area.

An argument for a direct connection between production and perception can also be made for audio-visual integration. Many researchers have demonstrated an effect of visual information about the shape of lips during the production of speech and speech perception (e.g., Reisberg et al., 1987; Macleod & Summerfield, 1990; Arnold & Hill, 2001). Humans have rich experience both with hearing speech sounds that are produced by themselves as they move their lips, and with hearing the same speech sounds that other speakers produce while moving their own mouth. Thus, although the connection in this case is not direct, it is robust enough to explain why this phenomenon might occur. However, adults do not typically have lots of opportunities to associate speech produced by people other than themselves with aero-tactile sensation. Even the aero-tactile

consequences of speech produced by a person herself is not expected to be felt by her very often in places such as her hand or ankle. Children are typically found in close proximity to caretakers and other children, thus a connection between certain speech sounds and a puff of air may be formed during early childhood and generalized into any point of contact on the skin. However, the connection is less direct and such an experience is less robust than it is in other cases discussed above.

An effect of air puffs in the comparison between /p^{h}a/ and /pa/ but not in the comparison between /pa/ and /ba/ would provide additional evidence for the link between the somatosensory stimulus, speech perception, and speech production, and would additionally serve as a control to the results from medial positions in English (See Chapter 3). It has been established that in medial positions, that is, intervocalic or post-tonic positions, in English and other related languages, aspiration is typically not part of the physical signal, or if it exists, it is less prominent than aspiration in initial positions (Lisker, 1957, 1984, 1986, 2002; Kingston & Diehl, 1994). Crucially, aspiration is not required for distinguishing voiceless from voiced stops in this position. In the previous chapter we tested continua where stops occur in medial rather than initial positions (e.g. /ˈa.pa/ to /ˈa.ba/ rather than /pa/ to /ba/). The medial continuum was built by manipulating closure duration rather than VOT. The hypothesis was that aero-tactile information is associated with aspiration and concomitant long positive VOT and is thus not expected to shift perception toward voicelessness in the case of medial positions in English. Indeed, although the exposure to aero-tactile stimuli had a priming effect on some of the participants, it did not shift the perception of participants towards voicelessness in comparison to their baseline preferences. These results contrast with results from previous

studies, including Gick & Derrick (2009), Derrick & Gick (2013) and the experiment conducted in Chapter 2, who limited their scope to word-initial position in American English, where VOT is a primary cue to stop voicing contrasts (Lisker & Abramson, 1964, 1967, 1970; Flege, 1982; Keating, 1984). These three studies did find an effect of puffs of air on speech perception.

A minority of the participants (40%) in Chapter 3 were affected by the aero-tactile information such that it biased their perception towards voicelessness. However, this result was recorded for all the trials, whether accompanied by puffs of air, or not. That is, the effect was present both in the baseline and in the experimental condition, and perception was biased towards a voiceless response in all trials, whether a puff of air was felt or not. We concluded that this was the result of a priming effect, caused by exposure to air puffs during a validation test of the aero-tactile stimuli before the main part of the experiment. In other words, there was an effect of air puffs for this group of participants, indicating an association between the puff and the voiceless category, even in medial position. However, this was a priming effect and not an instance of multisensory integration. The acoustic signal was not aspirated, and these participants did not interpret it as such, but they did link aspiration with voicelessness at some level of abstraction.

Such a link between aspiration and voicelessness is not predicted to be formed for speakers of Thai. In English, voiceless stops are aspirated in some positions, thus speakers may include aspiration in their abstract representation of voicelessness, which may affect the way the speakers perceive voiceless sounds, even in an environment in which these sounds are not aspirated. In contrast, in Thai the lack of aspiration in /pa/ plays an important role in distinguishing it from /p^ha/. Therefore, the structure of the voicing categories in Thai

makes it significantly harder for listeners to associate the phoneme /pa/ with aspiration. Accordingly, we predict an effect of air puffs in the comparison between /p^{h}a/ and /pa/ but not in the comparison between /pa/ and /ba/. Such results would show that speakers of Thai interpret aero-tactile stimuli as aspiration, and moreover, that the interpretation of tactile stimuli may vary across languages and depends, among other factors, on the structure of the voicing categories in the language.

4.2 Methods

4.2.1 Participants

42 monolingual native speakers of Thai participated in the experiment (17 females; age range 18-30, mean age 22.63, SD = 2.33). The participants were all residents of Bangkok at the time of the experiment but were born and raised in multiple regions of Thailand. The participants were recruited with flyers and by word of mouth. All were naive to the purpose of the study and had no self-reported speech or hearing defects. They were compensated for their time. All of the participants signed an informed consent form approved by the Yale Human Research Protection Program and Chulalongkorn University.

4.2.2 Stimuli

4.2.2.1 Acoustic Stimuli

The stimuli for the medial continua were created by recording a male native speaker of

the Bangkok dialect of Thai, who was also proficient in English. The speaker was not exposed to a significant amount of English before he was 10 and does not communicate in English daily. The speaker produced six tokens of each of the syllables /p^{h}a:/, /pa:/, and /ba:/. Two eight-step continua were created. An aspirated voiceless to unaspirated voiceless continuum was created by removing the initial burst from the aspirated voiceless token and then shortening the aspiration in log-scaled steps, with the 8th step matching the duration of the unaspirated voiceless token. The 8th step was then replaced by the actual unaspirated voiceless token. An unaspirated voiceless to voiced continuum was created by shortening the pre-voicing from the fully voiced token in log-scaled step, with the 8th step being the fully voiced token itself. The first step was the unaspirated voiceless token (the same one that is used as step 8 in the other continuum). Table 4.1 summarizes the durations of the aspiration for the aspirated voiceless to unaspirated voiceless continuum and the durations of the pre-voicing for the unaspirated voiceless to voiced continuum. A nonlinear (logarithmic) step size was chosen because psycho-acoustic perception tends to follow Weber's law (subjective sensation is proportional to the logarithm of the stimulus intensity); e.g., Zwicker & Fastl (2006). See Rosen and Howell (1981) for results on VOT, and Stevens, (2000, p. 228) for a similar effect on the perception of duration of burst.

/pha:/ - /pa:/	
Step no.	Length of aspiration (ms)
1	129.17
2	88.3
3	62.57
4	44.4
5	33.3
6	26.74
7	21.19
8	0

/pa:/ - /ba:/	
Step no.	Length of pre-voicing (ms)
1	0
2	08.58
3	13.12
4	20.18
5	31.28
6	47.43
7	73.16
8	113.02

Table 4.1. Voicing continuum steps showing length of aspiration (ms) for the aspirated voiceless to unaspirated voiceless continuum and length of pre-voicing (ms) for the unaspirated voiceless to voiced continuum.

A pre-test of the continua was conducted online as a Mechanical Turk task and was used to assess the quality of the stimuli. The test was run with an independent group of participants that did not take part in the main study (N = 41). The participants were native speakers of Thai and participated online from Thailand. They were asked to choose whether they heard a /p^{h}a:/ or a /pa:/ (when sounds from the first continuum were presented), or /pa:/ or /ba:/ (when sounds from the second continuum were presented), and rate the goodness of the token on a five step Likert scale. The sounds were all presented in the same test. The instructions were in Thai. The instruction and target tokens were written in Thai orthography. The results of the pretest are plotted in Figure 4.1.

The bias was calculated as the 50% crossover point of the psychometric categorization function for the continuum, computed across all listeners. Acuity (a measure of boundary slope) was computed as the difference between the 25% and 75% probabilities for the categorization function. The category boundary for the aspirated voiceless to unaspirated voiceless continuum (left panel) is approximately centered between its endpoints, that is, its bias (4.3) is close to its midpoint (4.5). The category boundary for the unaspirated voiceless to voiced continuum (right panel) is not as centralized and is skewed towards the voiced end (bias = 5.2), and its acuity (3.9) is shallower than that of the other continuum (1.8). These responses reflect the expected shapes of categorical distinction functions, but the goodness ratings are not significantly different along the continua.

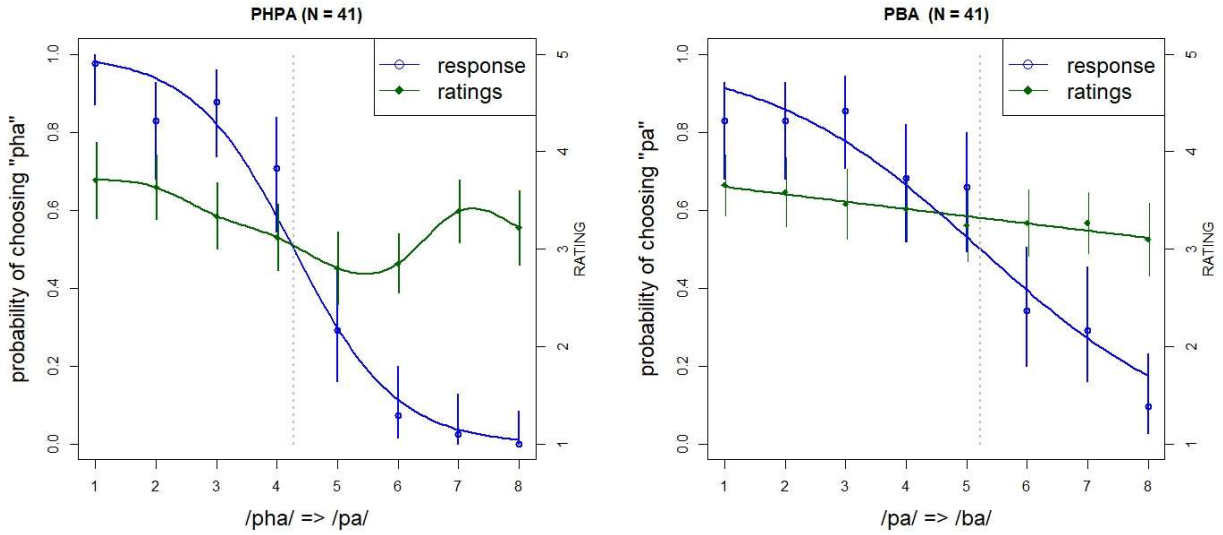

Figure 4.1. Viability test results for the continuum: left scale (blue line) shows probability of choosing voiceless relative to step (dotted vertical line marks 50% crossover point); right scale (green line) shows Likert scale ratings by step. Error bars show 95% confidence intervals.

4.2.2.2 Tactile (Air Puff) Stimuli

The tactile information was delivered as described in section 2.2.2.2. Detectable air turbulence exiting the tube was 100 ms in duration. These timing reflect observed values for aspirated voiceless bilabial stop in onset position in Thai (Lisker & Abramson, 1964, 1970; Gandour & Dardarananda, 1982).

4.2.3 Procedure

As in the previous chapter, each experimental session included two parts, an initial test to verify that the air puffs were felt but not heard, seen or otherwise perceived, and the main part, which tested participant responses to the auditory stimuli in the presence and absence of air puffs. Stimuli were presented to the participants through ear-enclosing headphones (Beyerdynamic DT 770 Pro 80 ohm). The experiment was conducted at the Linguistics Department in Chulalongkorn University in Bangkok. The consent forms, payment forms and any additional materials were written in Thai. The instructions were read to the participants in Thai by a native speaker. During the experiment the participants interacted with a native speaker of Thai, in Thai. No other languages were spoken during the experiment.

4.2.3.1 Puff Detection Test

The puff detection test was as described in section 2.2.3.1.

4.2.3.2 Perturbed continuum Testing

The perturbed continuum testing was as described in section 2.2.3.2. Five blocks were presented during which sounds drawn from both continua were tested. Sounds from both continua were presented together, and the participants were asked to make a 3-way choice. The unaspirated voiceless token, that was identical in both continua, was not presented twice. That is, 15 steps were presented, not 16. Each block included six repetitions of each of the 15 steps, for which three instances were accompanied by air puffs and three were not, randomly ordered. In total, the participants were presented with 5 blocks *x* 3 repetitions *x* 2 puff conditions (+/-) *x* 15 continuum steps for a total of 450 separate judgments, with 15 per condition at each continuum step. In each trial, participants were asked to identify the stimulus they heard and to press the corresponding button on a response box: either "พา" (/pha:/), "ปา" (/pa:/), or "บา" (/ba:/) to indicate the word they heard. The presentation order of the auditory stimuli and the accompanying tactile information (puff present vs. absent) were pseudo-randomized throughout each block. For one sixth of the participants, the left button on the response box indicated a syllable with an aspirated voiceless consonant, the middle button on the response box indicated a syllable with an unaspirated voiceless consonant, and the right button on the response box indicated a syllable with a voiced consonant. The other five possible combinations were presented each for roughly one sixth of the participants (5 of the combinations were presented to 8 participants, one of the combinations was presented to 7 participants).

4.3 Results

4.3.1 Puff Detection Test

In the first block of the detection test, when their hand was close to the exit point of the tube, participants correctly discriminated puff/no puff conditions at an average rate of 95.38% (s.d. 5.04), with the worst performer at 89%. An exact binomial test confirms that these recognition percentages were well above chance ($p < 0.01$). In the second block, with their hand positioned away from the tube and everything else the same, participants were at chance: 49.95% (s.d. 0.3); best performer 50% (binomial test n.s.). These results confirm that the participants felt the puff of air on their hand, but could not hear, see, or otherwise detect it.

4.3.2 Perturbed Continuum Testing

In 268 of the trials (1.4% of the trials) an air puff was requested but not delivered, or not requested but delivered. These trials were excluded from analysis, along with additional 49 trials for which the button-press response time exceeded 5 seconds (~3 s.d.). For the analysis the data was divided into two separate sets, one set containing steps 1 to 8, that will be referred to as the voiceless continuum (aspirated voiceless to unaspirated voiceless), and a second set containing steps 8 to 15, that will be referred to as the unaspirated continuum (unaspirated voiceless to voiced). The responses were converted from ternary to binary in the following way: in the voiceless continuum all the voiced responses were binned together with the unaspirated voiceless responses and contrasted with the aspirated voiceless responses. In the unaspirated continuum all the aspirated

voiceless responses were binned together with the unaspirated voiceless responses and contrasted with the voiced responses. There were 5.38% voiced responses in the voiceless continuum (405 voiceless responses out of total of 7529 responses), and 0.6% aspirated voiceless responses in the voiceless continuum (45 voiceless responses out of total of 7529 responses). No significant difference was found in either of the continua between the occurrences of these responses in +PUFF and -PUFF conditions. We prepared an additional dataset where the voiced responses were discarded from the voiceless continuum, and the aspirated voiceless responses were discarded from the unaspirated continuum. The results for model selection, direction of results, and levels of significance were the same as the results reported here. The data reported here was modeled with logistic regression in R (R Core Team, 2016) to estimate the effects of puffs on the perceptual boundary. The analysis was conducted separately for each of the continua since they differ in step size and type (the voiceless continuum was created by manipulating aspiration duration, whereas the unaspirated continuum was created by manipulating pre-voicing duration). Figure 4.2 shows the estimated psychometric functions, pooled across speakers, in the presence and absence of air puffs. The vertical axis in the left panel represents the probability of choosing an aspirated voiceless token. The vertical axis in the right panel represents the probability of choosing an unaspirated voiceless token. The horizontal axis shows the 8 steps along the continuum. The baseline condition, without puff, is shown in blue lines with circles, and the condition with air puffs is shown in red lines with crosses. Vertical solid lines show the bias (50% crossover point), and vertical dotted lines mark the 25% and 75% probability points along each curve; the distance between these points gives the acuity (a measure of the slope of the

boundary). The shift of the bias to the right in the presence of air puffs for the voiceless continuum reflects the fact that there were more aspirated voiceless responses in this condition. This contrasts with the unaspirated continuum which shows no shift in bias under puffs.

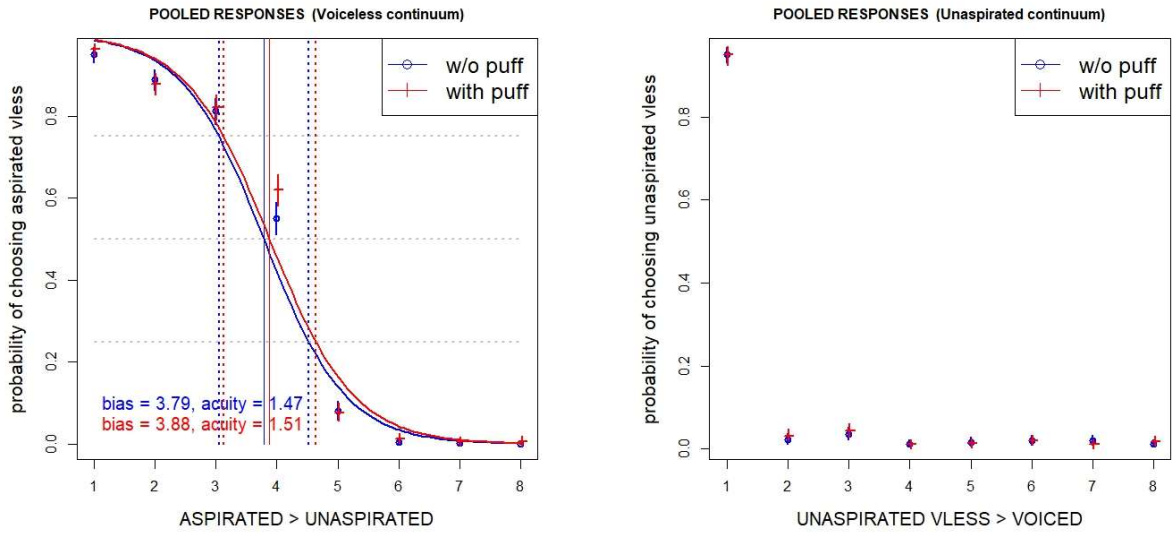

Figure 4.2. Perceived category boundaries pooled across speakers, with (red) and without (blue) an air puff. Vertical lines show the bias (50%) crossover, which is systematically shifted in the direction of aspirated voiceless for +puff trials in the voiceless continuum (left). There is no significant shift in the unaspirated continuum (right). 95% confidence intervals are indicated for each pooled response. The fitted lines were removed from the right panel since they obscured the sharp shift between steps 1 and 2.

4.3.2.1 Quantifying the effect of puffs on perceived categories

A generalized linear mixed-effects model (GLMM) computed with the lme4 package (Bates et al. 2015) was used to assess the significance of the puffs contrast for each of

the continua separately as discussed above. In these models[11] the dependent variable (the probability of choosing an aspirated voiceless response in the voiceless continuum, the probability of choosing an unaspirated voiceless response in the unaspirated continuum) was predicted by the fixed effect of PUFF (-/+) and a continuous covariate of STEP, with random slopes of CSTEP by participant ID (random slopes were supported by model comparisons, $\chi^2(2) = 634.03$, $p < 0.001$ for the voiceless continuum, $\chi^2(2) = 218.37$, $p < 0.001$ for the unaspirated continuum). The addition of an interaction term for PUFF and CSTEP did not improve the fit of the models ($\chi^2(1) = 1.607$, $p = 0.205$ for the voiceless continuum, $\chi^2(1) = 0.893$, $p < 0.345$ for the unaspirated continuum). The results, summarized in Table 4.2, show a significant shift under +PUFF on the voiceless continuum in the direction of aspirated voiceless, and no effect of PUFF on the unaspirated continuum. Marginal R^2 for these models (a measure of effect size), representing the proportion of variance explained by fixed factors alone, was computed using the method of Nakagawa & Schielzeth (2013), as implemented by Lefcheck & Casallas (2014). The effect of CSTEP was significant for both continua.

[11] `glmer(RESP ~ PUFF + CSTEP + (1+CSTEP|ID), family=binomial)`

Continuum	-Air PUFF (baseline) vs. +Air PUFF			
	Coefficient	z-value	p-value	Marginal R^2
Voiceless (/p^{h}a/ to /pa/)	0.232	2.66	0.008 **	0.701
Unaspirated (/pa/ to /ba/)	-0.108	-1.15	0.249 n.s.	0.776

Table 4.2. Output of the GLMM response model for each continuum. For the voiceless continuum the effect of +PUFF was to increase the likelihood of an aspirated voiceless response; the unaspirated continuum was unaffected. R^2 shows the proportion of variance explained by the fixed factors alone.

The responses to the unaspirated continuum were not significantly different in the presence vs. absence of air puffs. However, the responses did not reflect the expected category boundary that was recorded in the pre-test (see Figure 4.1). The responses for the first step, at the unaspirated voiceless end, reflect the expected choice, unaspirated voiceless. The responses for the other steps of the continuum are overwhelmingly voiced, with no noticeable difference between steps or participants.

4.3.2.2 Analysis of Individual Results

To assess the degree to which individual participants were sensitive to the air puff effect we computed separate logistic regression models for each, with response predicted by the fixed effect of PUFF and STEP as a continuous covariate.[12] About 70% of the participants showed a shift towards aspirated voiceless responses under +puff in the voiceless continuum (30/42; binomial test $p < 0.01$). About a third of the participants showed small and non-significant shifts towards voiced responses under +puff in the

[12] `glm(RESP ~ PUFF + CSTEP, family=binomial)`

unaspirated continuum (14/42; n.s.). See Table 4.3 for summary statistics.

Continuum	mean coefficient	s.d. of coefficient	range of coefficient
Voiceless (/p^{h}a/ to /pa/)	0.26766	0.479	-0.87388 : 1.66863
Unaspirated (/pa/ to /ba/)	-0.00845	-0.548	-0.99308 :1.02929

Table 4.3. Summary of the individual models computed for the participants

4.4 Discussion

The current study expanded the set of languages in which audio-tactile integration has been investigated. The Thai language was chosen for its relevance to the question of how participants interpret tactile stimuli during the process of multisensory integration. Specifically, the association between puffs of air and voiceless sounds, observed in a priming effect for some of the English-speaking participants in Chapter 3, is not expected for speakers of Thai. In Thai, aspiration is the basis for the contrast between aspirated and unaspirated voiceless stops and thus cannot be associated with unaspirated voiceless stops. The main hypothesis was that the puffs of air are interpreted by Thai speakers as the perceptual correlate of aspiration. Testing initial continua ranging from /p^{h}a/ to /pa/ and from /pa/ to /ba/ provided an opportunity for comparison between a case where aero-tactile information is predicted to affect speech perception, and a case where it is not predicted to have such an effect. We found that, as predicted, participants were affected by puffs of air in the comparison between /p^{h}a/ and /pa/ such that the presence of the puffs significantly increased the likelihood of choosing /p^{h}a/, but not affected by puffs of air in the comparison between /pa/ and /ba/. These results show that the puffs of air are

interpreted by speakers of Thai as aspiration. This clarifies the connection between aero-tactile stimuli and speech perception: the puffs of air are perceived as aspiration, and are thus available for listeners as one of the potential cues for aspirated phonemes.

The responses for the baseline (-PUFF) in the voiceless continuum (from /p^{h}a/ to /pa/) mirrored the responses in the pre-test (see Figures 4.1 and 4.2). In both cases, the results reflected the expected discrimination function. The responses for the experimental condition (+PUFF) were still reflective of the auditory stimuli but were shifted in the direction of /p^{h}a/, thus demonstrating integrative process, where both the auditory and the tactile stimuli are being taken into consideration by the participants. The responses for both the baseline and the experimental condition in the unaspirated continuum (/pa/ to /ba/) significantly diverge from the responses in the pre-test (see Figures 4.1 and 4.2). While the responses to the pre-test largely reflected the expected category discrimination function, the responses in the experiment did not reflect the expected baseline. The first step, the unaspirated voiceless end of the continuum, was categorized as /pa/, as expected. All the other steps were unexpectedly categorized as /ba/, by all the speakers. This was done regardless of the presence of puffs of air or its absence.

There is a shared pattern between the results in the current study, the results of Chapter 2 and the results of Chapter 3. The voiceless continuum in the current study and the initial continua in Chapter 2 (ranging from /pa/ to /ba/ in American English) are both contexts in which aspiration was part of the acoustic stimuli. In both contexts, the puffs of air increased the probability of choosing the response that the participants are typically producing with aspiration (/p^{h}a/ in Thai, /pa/ in American English). In both cases the responses to the baseline condition (-PUFF) were as expected, reflective of the categorical

distinction function. The unaspirated continuum in the current study and the medial continuum in Chapter 3 (ranging from /ˈa.pa/ to /ˈa.ba/ in American English) are both contexts in which aspiration was not part of the acoustic stimuli. In both cases there was no significant difference between trials that were accompanied by puffs or air and trials that were not. In both cases at least some of the participants (100% in the current case, 40% in the case of the experiment conducted in Chapter 3) did not show the expected categorical distinction function in the baseline condition (-PUFF).

In the previous chapter we used a Bayesian model to explain the variable behavior found in their study. The same reasoning can be used to account for the unexpected results in the current study. In the previous chapter we argued that the participants that did not show the expected baseline were primed to prefer a voiceless response by the puffs of air to which they were exposed during the puff test phase, prior to the perturb continua testing. We modeled this priming effect as part of the prior. Such a model could work based on the assumption that the representation of the voiceless phoneme /pa/ can include aspiration at some level of abstraction. This assumption cannot be made in the current case, since aspiration is the basis for distinction between /p^ha/ and /pa/ in Thai and therefore cannot be a part of the representation of /pa/. The prior, then, cannot be affected in this fashion in the unaspirated continuum in the current study. However, the prior *is* different between the current study and the previous chpater. In both experiments the participants were in a similar setting, but while the participants in the previous chapter were presented with a binary choice, the participants in the current study were presented with a ternary choice. This is also a main difference between the pre-testing, where a binary choice was presented, and the testing conducted during the study. In a

non-biased experimental setting, the participants are assumed to expect an even distribution of tokens and consequently to have a prior of 0.5 for a binary choice and a prior of 0.33 for a ternary choice. In a Bayesian model the response made by the participants, the posterior probability, depends both on the prior and the likelihood. As the prior decreases, the weight of the likelihood increases (see Equation 3). Crucially, in the current case the prior was greater in the pre-testing than in the testing conducted during the study. This means that the weight of the likelihood, the signal given the category, was greater. In other words, the participants were more affected by the signal in the testing than in the pre-testing.

Two factors are of importance here: first, steps 2-8 in the unaspirated continuum were pre-voiced, whereas step 1 was not (see section 2.2.1). Second, the puffs of air did not affect the responses for the unaspirated continuum. The participants disregarded the puffs of air when exposed to stimuli drawn from this continuum, as expected, since aero-tactile information was not relevant for the task. That is, participants were particularly attentive to the signal, because of the ternary choice that had reduced the weight of the prior. Crucially, the signal they were particularly attentive to included only the acoustic information, since the tactile information was disregarded as not relevant. As a result, the participants categorize all the tokens that were pre-voiced as voiced, even when the pre-voicing was considerably shorter than in typical voiced stops in Thai. In contrast, in the pre-testing, where the prior was greater, participants showed the expected category boundary, where tokens with shorter pre-voicing were categorized at times as unaspirated voiceless, with lower frequency of categorizing as voiced for stimuli with shorter periods of pre-voicing.

4.5 Conclusion

The aims of the current study were first, to expand the set of languages in which audio-tactile integration in speech perception has been shown to operate, and second, to show that aero-tactile information is being interpreted by listeners as aspiration during the process of integration. Under the assumption that the puffs of air used in the experiment are indeed interpreted as aspiration, we predicted an effect of air puffs on speakers of Thai in the comparison between /p^{h}a/ and /pa/ but not in the comparison between /pa/ and /ba/. Indeed, these results were obtained, showing both that speakers of Thai are subject to audio-tactile integration and that they interpret aero-tactile stimuli as aspiration.

Chapter 4: References

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology*, *92*(2), 339-355.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bernstein, L. E., Demorest, M. E., Coulter, D. C., & Oc´onnell, M. P. (1991). Lipreading sentences with vibrotactile vocoders: Performance of normal–hearing and hearing–impaired subjects. *Journal of the Acoustical Society of America*, *90*(6), 2971-2984.

Bicevskis, K. (2015). *Visual-tactile integration and individual differences in speech perception*. (Master's thesis, The University of British Columbia). Retrieved from https://open.library.ubc.ca/collections/ubctheses/24/items/1.0166756

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica, 29(*4), 203.

Burnham, D., & Dodd, B. E. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. In D. G. Stork & M. E. Hennecke (Eds.), *Speech reading by humans and machines: Models, systems, and applications* (pp. 103–114). Berlin: Springer–Verlag.

Byrd, D. (1993). 54,000 American stops. *UCLA Working Papers in Phonetics*, *83*, 97–116.

de Gelder, B., & Vroomen, J. (1992). Auditory and visual speech perception in alphabetic and non-alphabetic Chinese-Dutch bilinguals. *Advances in psychology (83)*. 413-426.

Derrick, D., Anderson, P., Gick, B., & Green, S. (2009). Characteristics of air puffs produced in English "pa": Experiments and simulations. *Journal of the Acoustical Society of America*, *125*(4), 2272-2281.

Flege, J. E. (1982). Laryngeal timing and phonation onset in utterance-initial English stops. *Journal of Phonetics 10*(2). 177–192.

Fowler, C. A., & Dekle, D. J. (1991). Listening with eye and hand: Cross–modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 816–828.

Gandour, J. (1985). A voiced onset time analysis of word-initial stops in Thai. *Linguistics of the Tibeto-Burman Area, 8*(2), 68-80.

Gandour, J., & Dardarananda, R. (1982). Voice onset time in aphasia: Thai. I. Perception. *Brain and language, 17*(1), 24-33.

Geers, A., & Brenner, C. (1994). Speech perception results: Audition and lipreading enhancement. *Volta Review*, *96*(5), 97–108.

Gick, B., & Derrick, D. (2009). Aero-tactile integration in speech perception. *Nature*, *462*, 502– 504.

Gick, B., Jóhannsdóttir, K. M., Gibraiel, D., & Mühlbauer, J. (2008). Tactile enhancement of auditory and visual speech perception in untrained perceivers. *Journal of the Acoustical Society of America*, *123*(4), EL72-EL76.

Goldenberg, D., Tiede, M. K., & Whalen, D. H. (2015). Aero–tactile influence on speech perception of voicing continua. In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*. Glasgow, UK: The University of Glasgow.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America*, *108*(3), 1197-1208.

Ito, T., Tiede, M., & Ostry, D. J. (2009). Somatosensory function in speech perception. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(4), 1245-1248.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research*, *46*, 390–404.

Keating, P. A. (1984). Phonetic and phonological representation of stop consonant voicing. *Language 60*(2). 286–319.

Lachs, L., Pisoni, D. B., & Kirk, K. I. (2001). Use of audiovisual information in speech perception by prelingually deaf children with cochlear implants: A first report. *Ear and Hearing*, *22*(3), 236-251.

Lefcheck, J. & Sebastian Casallas, J. (2014). *R-squared for generalized linear mixed-effects models.* Retrieved from https://github.com/jslefche/rsquared.glmm

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language, 33*(1), 42-49.

Lisker, L. (1984). How Is the Aspiration of English/p, t, k/" Predictable"?. *Language and Speech, 27*(4), 391-394.

Lisker, L. (1986). “Voicing” in English: A catalogue of acoustic features signaling/b/versus/p/in trochees. *Language and speech, 29*(1), 3-11.

Lisker, L. (2002). The voiceless unaspirated stops of English. In B. E. Nevin & S. M. Johnson (eds.), *The legacy of Zellig Harris: Language and information into the 21th century*, 233–240. Amsterdam; Philadelphia: John Benjamins.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lisker, L. & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the 6th international congress of phonetic sciences (ICPhS)*, vol. 563, 563–567. Academia Prague.

Macleod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43.

Magnotti, J. F., Mallick, D. B., Feng, G., Zhou, B., Zhou, W., & Beauchamp, M. S. (2015). Similar frequency of the McGurk effect in large samples of native Mandarin Chinese and American English speakers. *Experimental brain research, 233(*9), 2581-2586.

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, *21*(4), 445-478.

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

Mills, A. E. (1987). The development of phonology in the blind child. In B. E. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip–reading* (pp. 145–161). Hillsdale, NJ: Lawrence Erlbaum.

Nakagawa, S., & Schielzeth, H. (2013). A general and simple method for obtaining R^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133-142.

R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Reed, C. M., Durlach, N. I., Braida, L. D., & Schultz, M. C. (1989). Analytic study of the Tadoma Method: Effects of hand position on segmental speech perception. *Journal of Speech, Language, and Hearing Research*, *32*, 921–929.

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip– reading advantage with intact auditory stimuli. In B. E. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip–reading* (pp. 97–114). Hillsdale, NJ: Lawrence Erlbaum.

Rosen, S. M., & Howell, P. (1981). Plucks and bows are not categorically perceived. *Perception & Psychophysics*, *30*(2), 156-168.

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Attention, Perception & Psychophysics*, *59*(3), 347–357.

Sekiyama, K., & Tohkura, Y. I. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America, 90*(4), 1797-1805.

Sekiyama, K. (1994). Differences in auditory-visual speech perception between Japanese and Americans: McGurk effect as a function of incompatibility. *Journal of the Acoustical Society of Japan (E), 15(*3), 143-158

Sparks, D. W., Kuhl, P. K., Edmonds, A. E., & Gray, G. P. (1978). Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech. *Journal of the Acoustical Society of America*, *63*(1), 246-257.

Stevens, K. N. (2000). *Acoustic phonetics*. MIT Press: Lawrence Erlbaum.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*(2), 212-215.

Zwicker, E., & Fastl, H. (2006). *Psychoacoustics: Facts and models*. 2nd ed. Berlin: Springer-Verlag.

Chapter 5:

Aero-Tactile Integration in Speech Perception and the Phonological Representation of Voicing

5.1 Aero-Tactile Stimuli as Phonologically Relevant Information

In the experiment discussed in Chapter 2 we tested the effect of aero-tactile information on perception of VOT continua in American English. We found that the presence of aero-tactile information, in the form of a puff delivered to the listener's hand, increased the likelihood of hearing a voiceless sound. There are several possible explanations for this result, some discussed in the chapters 2 and 3. Of primary interest in the following discussion is the possibility that the puffs of air were perceived as phonologically relevant information. The experiment detailed in chapters 3 can serve to evaluate this possibility. This experiment assessed the effect of aero-tactile information on the perception of medial stops in American English.

In medial stops in American English aspiration is not used by listeners in distinguishing voiceless from voiced sounds. We hypothesized that aero-tactile information is associated with aspiration, and thus predicted that it will not shift perception toward voicelessness in this context. The predicted result was found for all the

participants. However, 40% of them, whom we labeled the Primed group, showed a priming effect where a bias towards voicelessness was found for all responses, regardless of the presence of puffs of air. Crucially aspiration was relevant to voicing in that it positively primed voicelessness for the Primed participants. In this experiment, the acoustic signal was not aspirated, and these participants did not interpret the puff as aspiration, but they did link the puff with voicelessness at some level of representation, at least enough to justify selecting /ˈa.pa/ more often than /ˈa.ba/. At the end of Chapter 3 we asked the question what kind of representation of voicing can justify such expectations and considered two options. We suggested, first, this might be the result of a learned between voicelessness or voiceless stops and aspiration from positions where aspiration is part of the acoustic signal, and generalization of this association at an abstract level. Then we considered exemplar-based approaches as a way of accounting for the variation observed in the chapter. The question at the heart of this debate is why do some speakers associate puffs of air with voiceless stops in all positions, while others associate puffs of air with voiceless stops only if they appear in a position where aspiration is expected. This chapter has two goals. The first is to try to answer this question by considering a different angle: the possibility that for some speakers aero-tactile information cues something more abstract than aspiration, such as a [spread glottis] feature.

The second goal of this chapter is to investigate the closeness of the link between the phonetic signal and abstract phonological structure. The results obtained in the three experiments described in this dissertation allow a deeper probe into how aero-tactile information relates to phonological representations than has been available in previous work. To date, no work in speech perception has made use of the somatosensory

dimension of speech to inform conclusions about phonological representations. However, there are open debates in phonology, particularly in laryngeal phonology, that could be informed by such novel evidence. Several theoretical approaches to phonological representation, particularly approaches to laryngeal phonology, have been advocated for since the middle of the last century. The following discussion will focus on three prominent ones, *the Standard Feature-Based Approach* (the Standard Approach henceforth, e.g., Lisker & Abramson, 1964; Keating, 1984; Lombardi, 1991, *Laryngeal Realism*, which is also a feature-based approach (e.g., Iverson & Salmons, 1995, 1999, 2003b), and *Articulatory Phonology*, a gesture-based approach (Browman & Goldstein, 1986, 1989, 1992). We argue that the behavior of the Primed participants can be accounted for by Laryngeal Realism or Articulatory Phonology but not by the Standard Approach. Moreover, we suggest that it might be the case that the Primed participants and the other group of participants in the same experiment have different phonological representations of stop consonants. If this is indeed the case, a full account of the observed variation may require adopting more than one theoretical approach.

5.2 Theoretical Approaches to Phonological Representation

5.2.1 Feature-Based Approaches

Honeybone (2005) follows Hall (2001) in dividing the models of feature-based phonological representation into two main groups, split by their approach to how segments are characterized in terms of laryngeal specifications in different languages. Specifically,

the two approaches differ in how they view the distinction between voiced and voiceless stop consonants. This point can be illustrated by considering a large group of languages that support a two-way voicing distinction. In some of these languages (e.g., English, German) the so-called voiced stop consonants are phonetically unaspirated voiceless stops (/p/, /t/, /k/) and the voiceless stop consonants are phonetically aspirated voiceless stops (/p^h/, /t^h/, /k^h/). We will call these languages *Aspirated*. In the other languages in this group (e.g. French, Russian) the voiced stop consonants are phonetically voiced stops (/b/, /d/, /g/) and the voiceless stop consonants are phonetically unaspirated voiceless stops (/p/, /t/, /k/). We will call these languages *Voiced*. While both approaches recognized these facts, they account for them using different phonological processes.

5.2.1.1 The Standard Approach

The Standard Approach, defined by Hall (2001, p. 32) as "broad interpretation of the feature [voice]", maintains that the underlying laryngeal contrast is the same in Aspirated and Voiced languages. The allophonic/surface output forms are argued to be derived by a set of phonological processes. Hall (2001) traces this approach back to Lisker & Abramson (1964). Keating (1984) and Lombardi (1991) explicitly argue for this approach. Honeybone (2005) notes that, as suggested by its name, most standard language descriptions adopt this approach. He lists Macpherson (1975), Booij (1995), Wiese (1996), and Hammond (1999) as examples (for Spanish, Dutch, German and English, respectively). Other prominent accounts in this approach are Keating (1990) and Kingston & Diehl (1994).

5.2.1.2 Laryngeal Realism

Laryngeal realism, defined by Hall (2001, p. 32) as "narrow interpretation of the feature [voice]" and labeled by Kager et al. (2007) "the multiple feature hypothesis", holds that the underlying laryngeal specifications is fundamentally different in Aspirated and Voiced languages. The phonologically voiced stops in the Voiced languages are represented with a feature such as [voice]. The phonologically voiceless stops in the Aspirated languages are marked with a feature such as [spread glottis]. The phonetically unaspirated voiceless stops (phonologically voiced in the Aspirated languages, phonologically voiceless in the Voiced languages) are argued to be laryngeally neutral, or unspecified. Crucially, the representation of the stops in each category in this approach is uniform across contexts and surface manifestations. In contrast, in the Standard Approach voiceless stops in a language such as American English are represented as aspirated in initial positions (e.g., [+spr gl], Keating, 1990), and as unaspirated in medial positions (e.g., [-spr gl], Keating, 1990). Hall (2001) traces this approach back to Jakobson (1949). Honeybone (2005) mentions Anderson & Ewen (1987) as relatively early advocates of this approach, Harris (1994) who argues for the approach on independent phonological grounds, Iverson & Salmons (1995, 1999, 2003b) and Iverson & Ahn (2007), whose accounts for this approach has probably been the most influential, Jessen (1998), who picked up on Jakobson's work independently from Iverson & Salmons but later adopted their terminology (Jessen & Ringen, 2002), and Petrova (2002) and Honeybone (2002) who independently applied the approach to historical processes and language change. Other accounts in this approach are Spencer (1996), Avery & Idsardi (2001), Iverson & Salmons (2003a, 2006), and Beckman et al. (2013). Laryngeal Realism has also been applied in

various subfields of linguistic theory such as historical linguistics (Calabrese & Halle, 1998), language typology (Kehrein & Golston, 2004), psycholinguistics (Brown, 2004), and language acquisition (Kager et al, 2007).

5.2.2 Articulatory Phonology: A Gesture-Based Approach

In Articulatory Phonology (Browman & Goldstein, 1986, 1989, 1992), gestures are the basic units of the phonological structure. They are understood as linguistically relevant control parameters of the vocal tract. There are three types of gestures in this theory: constriction gestures, tonal gestures, and modulation gestures. Stop consonants are represented by constriction gestures. These are spatiotemporal in nature and defined in terms of phonologically relevant tasks, such as lip closure. Constriction gestures are computationally modeled by the Task Dynamics model of sensorimotor control and coordination (e.g., Saltzman, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989). This model implements the phonological units in the speech production system as dynamical systems and makes predictions about the change in vocal tract constriction over the course of controlled movements. It attempts to account for the connection between surface (phonetic) variability and underlying phonological invariance. Gestural scores provide input to the Task Dynamics model. The gestural score in Figure 5.1 demonstrates the temporal intervals of the velum (VEL), tongue body (TB), lips aperture (LIPS), and glottis (GLO) gestures in the English word *palm*. Timing is represented by the horizontal axis, and temporal intervals during which the gestures are active are represented by the boxes. The degree and the location of the gestures are given in the labels contained in the boxes. For example, the tongue body (TB) gesture for producing the vowel /a/ is specified

for a narrow constriction degree at the glottis. Diagram A is the abstract gestural score. The horizontal lines in diagram B represent a possible spatiotemporal realization of the movements of the articulators given in the abstract score in diagram A. In the Task Dynamics model, constriction gestures are realized by coordinated actions of the articulators. These gestures are associated with planning oscillators (clocks) that determine their onset (Goldstein et al., 2006; Nam et al., 2009). The oscillators are coordinated to each other to form more complex units, such as syllables, words and phrases. These coordination relations are temporal relations: the oscillators are coupled to each other, that is, they are timed with respect to each other, either synchronously (in-phase) or sequentially (anti-phase).

Voiceless stops are represented in Articulatory Phonology by two gestures: an oral constriction gesture and a glottal closing-and-opening gesture (see Figure 5.1). Voiced stops, on the other hand, are represented by a single oral constriction gesture (Browman & Goldstein, 1986; Goldstein & Browman, 1986). That is, the voicing contrast can be characterized in Articulatory Phonology as the presence or absence of the relevant glottal gesture. Aspiration is represented by the relative timing of the peak glottal opening and the release of the stop gesture and aerodynamic conditions (Browman & Goldstein, 1986; Goldstein & Browman, 1986; Browman & Goldstein, 1992).

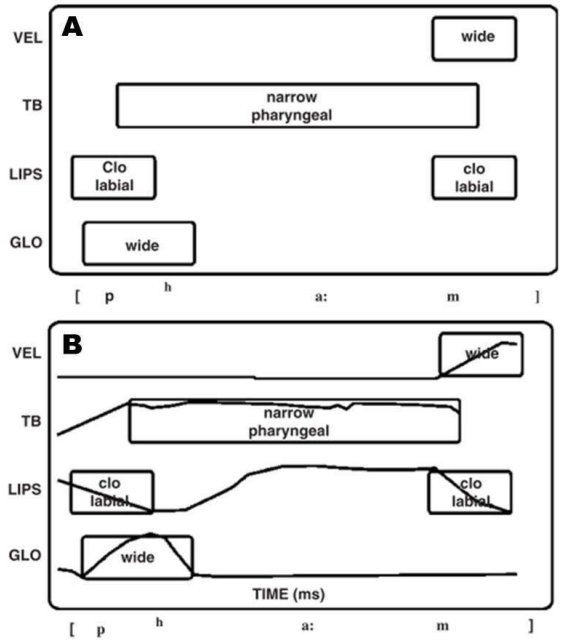

Figure 5.1. Gestural Score of the English word palm. Adapted from Moen (2006).

5.3 Aero-Tactile Integration and the Phonological Representation of Voicing

We argue that the aero-tactile stimuli used in our experiments, the puffs of air, was associated in perception with aspiration. In Laryngeal Realism aspiration is represented by the feature [spread glottis] (e.g., Iverson & Salmons, 2006). In the Standard Approach aspiration has been analyzed as a phonetic category that maps to the more abstract phonological feature [-voice] (e.g., Keating, 1984; Kingston & Diehl, 1994). In Articulatory Phonology aspiration is represented by the relative timing of the peak glottal opening and the release of the stop gesture and aerodynamic conditions (Browman & Goldstein, 1986; Goldstein & Browman, 1986; 1992). This is similar to a privative [spread

glottis] feature but differs in that timing is incorporated explicitly. The debate between more abstract phonological features, such as [voice], and features that capture Laryngeal Realism echoes a broader issue in phonological theory over the degree of abstractness of phonological representations, i.e., how faithfully phonological representations reflect the phonetics.

We designed our first experiment (detailed in chapter 2) to demonstrate that aero-tactile information plays a role in perception in a position where aspiration serves as the basis for the voicing distinction. By looking at cases such as the second and the third experiment (detailed in chapters 3 and 4, respectively), where aspiration does not play a role in distinguishing voicing, it became possible to obtain a better understanding of the structure of laryngeal contrasts. This is of particular theoretical interest because whether or not an aero-tactile effect is expected in positions such as non-foot-initial stops in American English depends on the particular theory of laryngeal phonology. An effect of aero-tactile information such that it shifts perception towards voicelessness, had it been obtained in medial positions in American English, could have been interpreted as the result of a learned association between voiceless stops and aspiration generalized from initial positions. We did not obtain such a result. However, the behavior of the Primed participants suggests that at least some speakers have formed a general association between voiceless stops and air puffs. The phonological representation advocated by theories such as Laryngeal Realism or Articulatory Phonology enables this kind of association. For example, such a result is consistent with Laryngeal Realism that represent both initial (aspirated) and medial (unaspirated) voiceless stops in languages such as American English with the same feature, e.g., [spread glottis] (Iverson & Salmons, 2006).

However, this result poses a challenge for other representational approaches. In Keating (1990), for instance, voiceless stops in initial and medial position in English are treated as separate phonetic categories, aspirated ([+spr gl]) and unaspirated ([-spr gl]). On this account, the priming observed is unexpected.

Testing initial continua in Thai (as detailed in chapter 4) provided an opportunity for direct comparison between a case where aspiration is a cue for the voicing distinction, and a case where it is not. We argued that aero-tactile stimuli play the role of the perceptual correlate of aspiration, thus we expected speakers of Thai to be affected by it in the comparison between /p^{h}a/ and /pa/ but not in the comparison between /pa/ and /ba/. In our first experiment (chapter 2), with speakers of American English, we found that aero-tactile stimuli yield more voiceless responses, that is, more /pa/ than /ba/ responses. Such an effect would have been unexpected in the same comparison in Thai, under any theory discussed above. None of the feature-based approaches assign an aspiration-related feature (e.g., [spread glottis]) to either /pa/ or /ba/ in Thai. /ba/ is assigned the feature [voice] and /pa/ is viewed laryngeally neutral, thus it is not assigned a relevant feature. Features such as [spread glottis] are reserved in languages such as Thai for the representation of /p^{h}a/. In Articulatory Phonology /pa/ in Thai is represented by the presence of a closing-and-opening glottal gesture. This is the same gesture that is used to represent /pa/ in English and it is independent of the representation of aspiration. Aspiration is represented by the relative timing of the peak glottal opening and the release of the stop gesture together with specific aerodynamic conditions. In Thai, but not in English, a specific relative timing is also part of the representation of /pa/. The relative timing is different for /pa/ and /p^{h}a/,

determining that the latter is aspirated, and the former is not. Thus, aspiration is not part of the representation of /pa/ in Thai in this theory as well.

To sum up, the three experiments we have conducted establish the effect of aero-tactile information on perception and show that this information is perceived as phonologically relevant. Our second experiment, testing medial positions in American English, can serve as a test-case to evaluate different theories of laryngeal phonology. If in the comparison between /pa/ and /ba/ speakers would have shown similar behavior to the behavior observed in initial positions, this could have suggested that the same logic applies to both cases - that is, in both aspiration is one of the cues associated with voicelessness, regardless of the existence of the cue in the physical signal. While we did not obtain this result, we found that some of the participants are primed by the exposure to aero-tactile stimuli, at least enough to justify selecting /ˈa.pa/ based on an expectation for an aspirated sound, or enough to justify expectation for a voiceless sound based on aspiration alone. In either case, for these participants, the representation of voicelessness includes aspiration, or a puff of air, at some level. The theories that allow such a representation maintain that voiceless sounds in languages such as English are associated with aspiration (represented as a certain glottal configuration) at the phonological level of representation, not at the phonetic level. This can be done, for instance, by assuming that voiceless stops in these languages have the feature [spread glottis], as Laryngeal Realism does, or by associating them with a glottal gesture, as Articulatory Phonology does. Specifically, in Articulatory Phonology, the fact that in medial positions the temporal conditions that are required for an actual production of aspiration are not met does not interfere with the association of aero-tactile stimuli and voicelessness. It is nonetheless

part of the relevant information used by speakers to recover the gestures composing the phoneme /p/. This might mean that the participants in the second experiment who did not seem to have formed a general association between voicelessness and aspiration have a different phonological representation of stop consonants than the Primed participants. A version of the Standard Approach, for instance, where voiceless consonants are associated with aspiration only in certain contexts by allophonic rules, could account better for the behavior of these participants than Laryngeal Realism. It remains an open question what are the factors that determine which phonological representation will be maintained by a speaker, as well as the questions whether and how this phonological representation changes during the speaker's lifetime.

Chapter 5: References

Anderson, J. M., & Ewen, C. J. (1987). *Principles of dependency phonology*. Cambridge: Cambridge University Press.

Avery, P. & Idsardi, W. (2001). Laryngeal dimensions, completion and enhancement. In T. A. Hall (Ed.), *Distinctive Feature Theory* (pp. 41-70). Berlin/New York: Mouton de Gruyter.

Beckman, J., Jessen, M., & Ringen, C. (2013). Empirical evidence for laryngeal features: Aspirating vs. true voice languages. *Journal of Linguistics*, *49*(2), 259-284.

Booij, G. (1995). *The phonology of Dutch*. Oxford: Oxford University Press.

Browman, C. P., & Goldstein, L. M. (1986). Towards an articulatory phonology. *Phonology*, *3*, 219-252.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, *6*(2), 201-251.

Browman, C. P., & Goldstein, L. (1992). Articulatory phonology: An overview. *Phonetica*, *49*(3-4), 155-180.

Brown, J. C. (2004). Eliminating the segmental tier: evidence from speech errors. *Journal of psycholinguistic research*, *33*(2), 97-101.

Calabrese, A. & Halle, M. (1998). Grimm's and Vemer's Laws: a new perspective. In J. Jasanoff, H. C. Melchert & L. Oliver (Eds.), *Mfr curad· studies in honor of Calvert Watkins* (pp. 47-62). Innsbruck: Institut fiir Sprachwissenschaft, University of Innsbruck.

Goldstein, L. & Browman, C. P. (1986). Representation of voicing contrasts using articulatory gestures. *Journal of Phonetics 14(*2). 339–342.

Goldstein, L., Byrd, D., & Saltzman, E. (2006). The role of vocal tract gestural action units in understanding the evolution of phonology. In M. A. Arbib (Ed.), *From action to language: the mirror neuron system* (pp. 215–249). Cambridge: Cambridge University Press.

Hall, T. A. (2001). Introduction: Phonological representations and phonetic implementation of distinctive features. In T. A. Hall (Ed.), *Distinctive Feature Theory* (pp. 1-40). Berlin/New York: Mouton de Gruyter.

Hammond, M. (1999). *The Phonology of English: A Prosodic Optimality-Theoretic Approach*. Oxford: Oxford University Press.

Harris, J. (1994). *English sound structure*. Oxford: Blackwell.

Honeybone, P. G. (2002). *Germanic obstruent lenition: some mutual implications of theoretical and historical phonology*. (Doctoral dissertation, Newcastle University).

Honeybone, P. (2005). Diachronic evidence in segmental phonology: the case of laryngeal specifications. In M. van Oostendorp & J. van de Weijer (Ed.), *The internal organization of phonological segments* (pp. 319-354). Berlin/New York: Mouton de Gruyter.

Iverson, G. K., & Ahn, S. C. (2007). English voicing in dimensional theory. *Language Sciences*, *29*(2-3), 247-269.

Iverson, G. K. & Salmons, J. C. (1995). Aspiration and laryngeal representation in Germanic. *Phonology*, *12*(3), 369-396.

Iverson, G. K. & Salmons, J. C. (1999). Glottal spreading bias in Germanic. *Linguistische Berichte*, 135-151.

Iverson, G. K. & Salmons J. C. (2003a). Laryngeal enhancement in early Germanic, *Phonology 20*. 43-74.

Iverson, G. K., & Salmons, J. C. (2006). On the typology of final laryngeal neutralization: Evolutionary Phonology and laryngeal realism. *Theoretical Linguistics*, *32*(2), 205-216.

Iverson, G. K. & J. Salmons, J. C. (2003b). Legacy specification in the phonology of Dutch. *Journal of Germanic Linguistics 15*, 1-26.

Jakobson, R. (1949). *On the identification of phonemic entities*. The Hagues: Mouton.

Jessen, M. (1998). *Phonetics and phonology of tense and lax obstruents in German*. Amsterdam: John Benjamins.

Jessen, M., & Ringen, C. (2002). Laryngeal features in German. *Phonology*, *19*(2), 189-218.

Kager, R., van der Feest S., Fikkert, P., Kerkhoff, A., & Zamuner T. S. (2007). Representations of [voice]: evidence from acquisition. In E. J. van der Torre & J. van de Weijer (Eds.), *Voicing in Dutch: (De)voicing – phonology, phonetics, and psycholinguistics* (pp. 41-80). Amsterdam & Philadelphia: John Benjamins.

Keating, Patricia A. 1984. Phonetic and phonological representation of stop consonant voicing. *Language 60(*2). 286–319.

Keating, Patricia A. 1990. Phonetic representations in a generative grammar. *Journal of phonetics 18(*3). 321–334.

Kehrein, W., & Golston, C. (2004). A prosodic theory of laryngeal contrasts. *Phonology*, *21*(3), 325-357.

Kingston, J., & Diehl, R. L. (1994). Phonetic knowledge. *Language*, *70*(3), 419-454.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384-422.

Lombardi, L. (1991). *Laryngeal features and laryngeal neutralization*. (Doctoral Dissertation, University of Massachusetts, Amherst). Published by Garland, New York, 1994.

Macpherson, I. R. (1975). *Spanish phonology: descriptive and historical*. Manchester: Manchester University Press.

Nam, H., Goldstein L., & Saltzman, E. (2009). Self-organization of syllable structure: a coupled oscillator model. In F. Pellegrino, I. Chitoran, E. Marsico & C. Coupé (Eds.), *Approaches to Phonological Complexity* (pp. 299–328). Berlin/New York: Mouton de Gruyter.

Petrova, O. B. (2002). *The evolution of the English obstruent inventory: An optimality theoretic account*. (Doctoral Dissertation, University of Ohio).

Saltzman, E. (1986). Task dynamic co–ordination of the speech articulators: A preliminary model. In H. Heuer & C. Fromm (Eds.), *Generation and modulation of action patterns* (pp. 129–144). Berlin: Springer-Verlag.

Saltzman, E. & Kelso, J. A. (1987). Skilled actions: A task–dynamic approach. *Psychological Review 94(*1). 84–106.

Saltzman, E. & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology 1(*4). 333–382.

Spencer, A. (1996). *Phonology: Theory and Description.* Oxford: Blackwell.

Wiese, R. (1996). *The phonology of German*. Oxford: Oxford University Press.

Chapter 6:

Conclusion

This dissertation had four aims. The first aim was to provide a solid evidence for audio-tactile integration in the perception of speech, using aero-tactile stimuli. The results of the experiment outlined in chapter 2 satisfy this aim. In this experiment we evaluated the effect of air puffs on gradations of VOT along a continuum. Three continua were tested: bilabial, velar, and a vowel continuum used as a control. The presence of air puffs was found to significantly increase the likelihood of choosing voiceless responses for the two VOT continua but had no effect on choices for the vowel continuum. At the same time, the responses to the VOT continua were reflective of the distinction function expected according to the acoustic stimuli. This indicates that during the decision-making process, both auditory and aero-tactile inputs were taken into consideration, suggesting that this is indeed an example of multisensory integration. Moreover, analysis of response times revealed that the presence of air puffs lengthened responses for intermediate (ambiguous) stimuli and shortened them for endpoint (non-ambiguous) stimuli. The slowest response times were observed for the intermediate steps for all three continua, but for the bilabial continuum this effect interacted with the presence of air puffs: responses were slower in the presence of air puffs, and faster in their absence. This suggests that during integration auditory and aero-tactile inputs are weighted differently by the perceptual system, with

the latter exerting greater influence in those cases where the auditory cues for voicing are ambiguous.

The second aim of this dissertation was to investigate the effect of the air puffs on listeners during the process of integration, by utilizing the notion that somatosensory information is integrated with auditory information only when it is task relevant. The experiment detailed in chapter 3 satisfied this aim by assessing the effect of aero-tactile information on the perception of medial stops in American English. This case study was chosen because VOT differences are not typically used for disambiguating stop voicing contrasts in this context. We hypothesized that aero-tactile information is associated with aspiration and concomitant long positive VOT, and thus predicted that it is not expected to shift perception toward voicelessness in the case of medial positions in English. The notion of task-relevance is crucial for this prediction: aspiration is relevant for the disambiguation made in the first experiment (chapter 2), but not for the disambiguation made in the second experiment (chapter 3). Thus, a shift in perception driven by exposure to puffs of air was expected in the former case, but not in the latter. Indeed, while a shift in perception was found for the VOT continua in the presence of air puffs in the first experiment, no shift was found for any the participants in the second experiment. However, 40% of the participants in the second experiment showed a priming effect where a bias towards voicelessness was found for all responses, regardless of the presence of puffs of air. We have argued that this bias is the result of a shift in the expectations of these participants, and modeled it using a Bayesian reasoning, as a change in the prior probability of choosing a voiceless response for these participants, but not for the other 60% of the participants.

The last two aims of this dissertation were to show that aero-tactile information is indeed being interpreted by listeners as aspiration during the process of integration, and to expand the set of languages in which audio-tactile integration in speech perception is shown to operate. These aims were satisfied by the experiment discussed in chapter 4. This experiment evaluated the effect of aero-tactile information on perception of an initial VOT continua in Thai. Thai exhibits a three-way voicing contrast, with aspirated voiceless stops, unaspirated voiceless stops, and voiced stops. If the aero-tactile stimuli are perceived as aspiration, they are predicted to shift the perception of voicelessness in Thai only in the case where aspiration is a cue for the voicing distinction. That is, in the comparison between aspirated voiceless stops and unaspirated voiceless stops, but not in the comparison between unaspirated voiceless stops and voiced stops. Indeed, we found that speakers of Thai were affected by the air puffs in the comparison between /p^{h}a/ and /pa/ but not in the comparison between /pa/ and /ba/. These results present a strong case for the claim that aero-tactile stimuli is being interpreted as aspiration during the process of integration.