THE ROLE OF PRIOR EXPERIENCE IN THE INTEGRATION OF AEROTACTILE SPEECH INFORMATION

by

Megan Keough

B.A. with honors, Brown University, 2014

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Linguistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2019

© Megan Keough, 2019
The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

The Role of Prior Experience in the Integration of Aerotactile Speech Information

submitted by Megan Keough in partial fulfillment of the requirements for
the degree of Doctor of Philosophy
in Linguistics

Examining Committee:

Bryan Gick, Linguistics
Supervisor

Janet F. Werker, Psychology
Supervisory Committee Member

Donald Derrick, New Zealand Institute of Language, Brain, and Behaviour
Supervisory Committee Member

Jessica De Villiers, English Language and Literatures
University Examiner

Anne-Michelle Tessier, Linguistics
University Examiner

Additional Supervisory Committee Members:

Kathleen Hall, Linguistics
Supervisory Committee Member
Abstract

Perceivers receive a constant influx of information about the natural world via their different senses. In recent years, speech researchers have begun to situate speech more firmly within this multisensory experience, moving progressively away from the traditional focus on audition toward a more multisensory approach. In doing so, speech researchers have discovered that, in addition to audition and vision, many somatosenses are all highly relevant modalities for experiencing and/or conveying speech. The current dissertation focuses on the integration of aerotactile somatosensation—the feeling of speech-related airflow on the skin—and whether prior experience with specific speech information modulates aerotactile influence on visual and auditory speech cues to English stops. In Chapter 2, I used a two-alternative forced choice visuo-aerotactile perception task to show that adult English perceivers can integrate aerotactile speech information from a novel visual source. In Chapter 3, I used a two-alternative forced choice audio-aerotactile perception task to demonstrate integration occurs for this population even when the auditory and aerotactile speech cues are presented in a way that does not conform with previous prior experience in the natural world. Finally, in Chapter 4 I used a looking time procedure to test prelinguistic infants on their sensitivity to speech-related airflow during auditory perception and found no evidence that infant stop perception can be influenced by airflow before infants begin babbling. Taken together, these three experiments suggest that while adult perceivers can integrate aerotactile speech information with speech information from other modalities without specific prior experience with the cues, some developmental experience may be required for this ability to emerge.
Lay Summary

While we often think of speech as something we hear, we experience speech through nearly all of our senses. We not only hear our fellow talker when she is speaking, we see her tongue, lips, and jaw move, and may also feel her breath. Moreover, we hear our own speech through bone conduction, feel our tongues and lips move, and feel our own airflow across the insides of our mouths. More and more, research suggests that all of these sensations interact when we process speech. But, how do we connect all these disparate channels of information? One possibility is we begin to connect visual, auditory, and tactile streams of speech information through experience. In this dissertation, I ask how our experience perceiving speech-related airflow—both our own and that of others—affects how and whether we use that airflow to discriminate between sounds.
Preface

This dissertation is my own original work and I was the first author and primary researcher on all collaborative projects.

Chapter 1: Introduction

A version of this chapter has been published in part. Keough, M., Derrick, D., Gick, B. (2019). Cross-modal effects in speech perception. I am the primary author of the chapter with intellectual contributions from Bryan Gick, PhD (supervisor) and Donald Derrick, PhD.

Chapter 2.

I am the primary author of this chapter. The research question and study design were decided in collaboration with Bryan Gick, PhD (supervisor) and Ryan C. Taylor, PhD. The stimuli were created in collaboration with Ryan C. Taylor. I determined and implemented the final study design, supervised the undergraduate research assistants conducting data collection in the Interdisciplinary Speech Research Lab, and wrote the current report. The analyses were done in collaboration with Donald Derrick, PhD. This research was covered under UBC Ethics Certificate: H04-80337. A version of this was presented at the 2019 International Congress of Phonetic Sciences in Melbourne, Australia and will appear in those proceedings.

Chapter 3.

I am the primary author of this chapter. The research question and design were decided in collaboration with Bryan Gick, PhD (supervisor) and Murray Schellenberg, PhD. I determined and implemented the final study design, collected the data myself and through undergraduate research assistants in the Interdisciplinary Speech Research Lab, and wrote the current report. This research was covered under UBC Ethics Certificate: H04-80337.
Chapter 4.

I am the primary author of this chapter. The study design was decided in collaboration with Janet F. Werker, PhD, Bryan Gick, PhD (supervisor), Padmapriya Kandhadai PhD, and Alison Bruderer, PhD. I determined and implemented the final study designs, collected all data, supervised two undergraduates in the coding, conducted all analyses (in collaboration with Donald Derrick, PhD), and wrote the current reports. This research is covered under UBC Ethics Certificate: H95-80023. A version of this work has been presented at the 2018 International Congress of Infant Science.

Chapter 5.

I am the primary author of this chapter, with intellectual contributions and comments from Bryan Gick, PhD (supervisor).
Table of Contents

Abstract............................................................................................................................................. iii
Lay Summary........................................................................................................................................ iv
Preface.................................................................................................................................................. v
Table of Contents ............................................................................................................................ vii
List of Tables ....................................................................................................................................... xi
List of Figures ....................................................................................................................................... xiii
Acknowledgements .......................................................................................................................... xv
Dedication .......................................................................................................................................... xvii

Chapter 1: Introduction ...................................................................................................................... 1

1.1 Speech as a Multisensory Experience .......................................................................................... 2

1.1.1 Audiovisual Speech .................................................................................................................... 5

1.1.1.1 The McGurk Illusion: Fusion or Confusion? ....................................................................... 5

1.1.1.2 Beyond McGurk .................................................................................................................... 8

1.1.1.3 Spatiotemporal Congruence ............................................................................................... 11

1.1.2 Somatosensory Speech ............................................................................................................ 13

1.1.2.1 The Many Somatosenses .................................................................................................... 13

1.1.2.2 The Tadoma Method: A Multimodal Method .................................................................... 18

1.1.2.3 Aerotactile Speech Perception .......................................................................................... 22

1.1.2.4 Spatiotemporal (In-)Congruence and Ecological Validity ............................................... 24

1.2 (Prior) Experience and Speech Perception .................................................................................. 27

1.3 Current Experiments .................................................................................................................... 31
Chapter 2: General vs. Specific Experience ...........................33
  2.1 Methods: Pre-experiment ..............................................37
    2.1.1 Visual Stimuli .....................................................37
    2.1.2 Audiovisual Stimuli .............................................38
    2.1.3 Procedure .........................................................38
    2.1.4 Results .............................................................39
  2.2 Methods: Main Study ..................................................41
    2.2.1 Visual Stimuli .....................................................42
    2.2.2 Aerotactile Stimuli .............................................43
    2.2.3 Procedure .........................................................45
    2.2.4 Analysis .............................................................46
  2.3 Results .................................................................47
    2.3.1 Integration .........................................................47
    2.3.2 Response Time ...................................................52
  2.4 Discussion .............................................................54

Chapter 3: Spatial Congruence ..............................................59
  3.1 Methods .................................................................62
    3.1.1 Stimuli and Airflow Apparatus ...............................62
    3.1.2 Procedure .........................................................63
  3.2 Results .................................................................64
  3.3 Discussion .............................................................67

Chapter 4: Pre-linguistic Infants ..........................................70
  4.1 Multi-sensory Speech Perception in Preverbal Infants ..........73
Chapter 4: Methods

4.2 Methods.................................................................................................................. 79

4.2.1 Participants...................................................................................................... 80

4.2.2 Apparatus and Set up .................................................................................. 80

4.2.3 Stimuli.............................................................................................................. 81

4.2.4 Procedure ...................................................................................................... 84

4.2.5 Analysis and Predictions ............................................................................. 86

4.3 Results............................................................................................................... 90

4.3.1 Trial Type Analysis ...................................................................................... 90

4.3.1.1 Total Looking Time .............................................................................. 90

4.3.1.2 Duration of First Look ......................................................................... 93

4.3.2 Stimulus Stream Analysis .......................................................................... 95

4.3.2.1 Total Look Time .................................................................................. 96

4.3.2.1.1 Comparison 1: ba + paPuff vs pa + paPuff .................................. 100

4.3.2.1.2 Comparison 2: ba + paPuff vs. baPuff + pa .................................. 100

4.3.2.1.3 Comparison 3: ba + baPuff vs. ba + paPuff .................................. 101

4.3.2.2 Duration of First Look ........................................................................... 101

4.3.2.2.1 Comparison 1: ba + paPuff vs pa + paPuff .................................. 105

4.3.2.2.2 Comparison 2: ba + paPuff vs. baPuff + pa .................................. 105

4.3.2.2.3 Comparison 3: baPuff + ba vs. ba + paPuff .................................. 105

4.4 Discussion ......................................................................................................... 106

4.4.1 Study Limitations ....................................................................................... 109

Chapter 5: General Discussion and Conclusion ..................................................... 114

5.1 Summary of Experimental Chapters ............................................................ 114
5.2 Empirical Implications

5.2.1 Perception-Direct or indirect?

5.2.1.1 Ecological or Direct Approach

5.2.1.2 Cognitivist or Indirect Approach

5.3 Future Work and Conclusions

References

Appendix: Additional Methodological Details for Chapter 4

Counterbalancing Details
List of Tables

Table 2.1 Summary of means and standard errors by condition for Audio-only and Audiovisual conditions for the pre-experiment................................................................. 40
Table 2.2 Summary of fixed effects for pre-experiment model......................................................... 40
Table 2.3 Mean and standard errors for percent /pa/ response for the simulated face, human “ba”, and human “pa” experiments by condition (puff vs. no puff). ......................................................... 48
Table 2.4 Summary of fixed effects for the model predicting percent “pa” response.................. 51
Table 2.5 Summary of results for parametric coefficients............................................................... 53
Table 2.6 Approximate significance of smoothing terms............................................................... 54
Table 3.1 Means and standard errors for percent “pa” response with and without airflow for the three stimuli presentation conditions (i.e., Congruent, Incongruent, and Binaural)......................... 65
Table 3.2 Summary of fixed effects................................................................................................. 67
Table 4.1 Mean looking times to trial by stimulus type for the experimental (left) and control (right) groups. ...................................................................................................................... 90
Table 4.2 Summary of fixed effect for the control group................................................................. 92
Table 4.3 Summary of fixed effects for experimental group........................................................... 92
Table 4.4 Means and standard errors for duration of first look.................................................. 93
Table 4.5 Fixed effect for the control group for duration of first look........................................ 95
Table 4.6 Fixed effects for the experimental group for duration of first look............................. 95
Table 4.7 Mean looking time to the checkerboard and standard errors for each stimulus stream for both Experimental and Control groups. ................................................................. 97
Table 4.8 Summary of fixed effects for the experimental group infants for total looking time to trial. ................................................................. 99
Table 4.9 Summary of fixed effects for the control group infants for total looking time to trial. 99
Table 4.10 Mean duration of the first look to the checkerboard and standard errors for each stimulus stream for both experimental and control groups......................................................... 102
Table 4.11 Summary of fixed effects for the experimental group infants for duration of first look to trial. ................................................................. 103
Table 4.12 Summary of fixed effects for the control group infants for duration of first look to trial. ................................................................. 104
List of Figures

Figure 1.1 Research over the last decade has shown that speech occurs in a richly multi-modal space. Auditory, visual, and several of the many somatosenses have been show to interact with one another during speech perception. At a minimum, we know that pressure, proprioception, vibration, and aerotactile somatosensation are all relevant tactile inputs for speech. .......................... 14

Figure 2.1 Still shots from the video stimuli. Top = simulated face, middle = human “ba”, bottom = human “pa”. Left = open mouth prior to plosive production. Middle = compress mouth prior to plosive release. Right = open mouth during vowel production following plosive release. ................................................................. 43

Figure 2.2 Screen capture of stereo track with labels indicating the tone that triggered the airflow release, the duration of the airflow, and the release burst in the waveform. ................................. 45

Figure 2.3 Percent /pa/ response for each visual stimulus type by condition. ......................... 49

Figure 2.4 Loess visualization of mean percent “pa” responses by condition and trial order for all four groups of subjects. Within each group, percent “pa” response was calculated by averaging across all participants for each trial. Condition (puff vs. no puff) was randomized, so for each trial, some participants felt the sensation of airflow while others did not. The lines at 100% and 0% represent individual responses as the response data were numerically coded with 0 for “ba” and 1 for “pa”. ............................................................................................................................................. 50

Figure 3.1 Percent “pa” response by Condition (puff vs. no puff) for Binaural, Congruent, and Incongruent stimulus presentation conditions. In all three conditions, participants gave significantly more “pa” responses during trials with airflow as compared to trials without airflow. ........................................................................................................................................ 66
Figure 4.1 Photographs of the bib worn around the baby’s neck and of the sound-attenuating smock. The vinyl tubing that delivered the airflow was located underneath the fabric and attached to the front of the bib. The tubing was then curved to aim the airflow at the baby’s neck from a distance of 7 cm.

Figure 4.2 Placement of the sine wave relative to the stop burst and vowel onset. The 50ms tone was shifted an additional 20ms earlier to account for system latency.

Figure 4.3 Mean looking times to checkboard by stimulus type and pair for both control group (left) and experimental group (right).

Figure 4.4 Mean duration of first look to the checkerboard by stimulus and pair for the infants in the control group (left) and the infants in the experimental group (right).

Figure 4.5 Total looking times for control group infants (left) and experimental group infants (right) to each stimulus stream. The red diamond in each boxplot signifies the mean.

Figure 4.6 Duration of first look to the checkerboard for control group infants (left) and experimental group infants (right) for each stimulus stream. The mean for each stimulus stream is indicated by a red diamond within each boxplot.
Acknowledgements

First and foremost, I owe the deepest gratitude to my supervisor, Bryan Gick, whose patience and incredible unflappability have been the perfect complement to my high-strung nature. Throughout this process, he kept me focused on the big picture, taught me to seek out knowledge across disciplinary lines, and pushed me to own my successes as readily as my failures.

I must also thank my committee members Janet F. Werker, Donald Derrick, and Kathleen Hall, for their insightful comments during every step of this dissertation. As I fumbled my way through my first infant experiment, Janet challenged me to consider new perspectives (and rethink old ones). She supported me professionally and personally, and it was an honor to have her as a mentor. I owe immeasurable thanks to Donald for many emergency Skype sessions across multiple time zones and for answering all of my statistical questions—both stupid and otherwise. It has been a pleasure to count him as a collaborator. Finally, I am thankful to Kathleen for her incredible attention to detail and precise, logical mind. One’s arguments are seldom as clear as imagined, and Kathleen misses nothing.

I have had the good fortune of being a part of not one, but two labs filled with exceptional people. First, thank you to the members of the Infant Studies Centre for welcoming me without hesitation. It is no exaggeration to say that Chapter 4 could not have been completed without Savannah Nijboer—whose efficiency and managerial expertise are unparalleled—and the incredible volunteers and RAs who booked over a hundred families for my study alone. I owe special thanks to Mary and Matilda for their coding assistance. To Priya, Jen, Erica, Alexis, Sheri, Nicole, Laurie, and so many others: I was inspired every day working with such dedicated scientists and I cannot
thank you enough for what you taught me about research, infants, and comradery. Second, thank you to Murray Schellenberg and the whole Interdisciplinary Speech Research lab for letting me be their “lab mom.” Murray, you are an excellent friend, collaborator, mentor, and human being. I am so grateful to know you. Special thanks to Yik Tung Wong for always going above and beyond as a research assistant. I am thankful I could leave so many critical components in your trusted, capable hands.

Finally, this dissertation would not have been possible without the incredible support of my family and friends. To Elise and Adriana, I’m thankful for the deep friendships we have created, for your honest and generous hearts, and for the many beers we shared along the way. I love you both very much. To Alexis, thank you for listening to me rant my frustrations before reminding me what was really important. To the residents on Oak St, but especially Blake, thank you for the community you gave Julian and me when we needed it most. They say it takes a village to raise a child and together we created one. Lastly, my deepest gratitude to my family for supporting me on every path I have taken in this strange and wonderful life. You taught me to never take myself too seriously, to read read read, and to be open to every experience life has to offer.
For Julian, always. You are my favorite favorite.

And for Sue. Thank you for reminding me that I always have a place.

The world seems a little less exciting without your laughter.

“Whoever you are, no matter how lonely, the world offers itself to your imagination, calls to you like the wild geese, harsh and exciting—over and over announcing your place in the family of things.”

— Mary Oliver
Chapter 1: Introduction

While speech is often described in terms of sound alone, considerable research over the last half-century—and even earlier (e.g., Bell 1867)—supports the notion that speech perception is fundamentally multisensory. Indeed, some natural languages have been found to encode sequences of speech “sounds” without any sound at all (Gick et al. 2012). Gick et al. (2012) reported on ultrasound, video, and acoustic data from native speakers of Oneida and Blackfoot, two languages in which utterance-final vowels are produced soundlessly, even when this silent production results in homophones. The authors found that participants in both languages produced distinct articulatory movements for these phrase-final “sounds” with no acoustic or perceptual difference, suggesting these two languages have phonologically stable segments from an articulatory perspective, even in the absence of an acoustic output. Such results bring into relief the need for more deeply multisensory approaches to speech research. In addition to audition and vision, many somatosenses including proprioception, pressure, vibration, and aerotactile sensation are all highly relevant modalities for experiencing and/or conveying speech. Much of the evidence that has been used in favor of a multisensory view of speech has been derived from laboratory studies of cross-modal effects. Although the field has undoubtedly made progress in understanding how such effects work and how to interpret them, researchers continue to grapple with the most basic questions regarding multisensory perception, i.e.: What sensory inputs are relevant to speech perception? What information or criteria determines in which circumstances we integrate? At what point in processing does integration occur? Obviously, such broad and complex questions cannot be answered in any single body of work. This dissertation focuses on the role of specific prior experience in determining what speech-related sensory information we integrate during
perception. In the experiments detailed in Chapters 2-4, I explore the relationship between prior experience with multisensory cues and speech perception using aero-tactile speech information, a cue related to the airflow that accompanies the production of aspirated stops.

The remainder of this chapter will be split across two topics. First, because the idea of speech as multisensory is so central to this dissertation, I will discuss some of the behavioral and neural evidence demonstrating this perspective. As I mentioned above, much of our understanding of the multisensory nature of speech has derived from studies investigating how speech information from one modality interacts with or influences another during perception. Thus, I will discuss both long-standing cross-modal effects stemming from decades of audiovisual speech research as well as newer findings related to somatosensory effects. Ultimately, I will show that, far from taking place in a one-, two- or even three-dimensional space, speech occupies a highly multidimensional sensory space. Most of these senses have been understood to be relevant only recently, and not all of them are relevant to speech at any given time. But as I will argue in this dissertation, sensory information does not have to be frequently experienced in order to contribute—at least for adult perceivers—and it can derive from unexpected sources as long as the cues relate to a real-world event. Second, I will discuss existing evidence that prior experience modulates speech perception. Finally, I will lay out the specific research questions that drive the experimental chapters.

1.1 Speech as a Multisensory Experience

Until fairly recently, terms such as “cross-modal” and “multisensory” have been almost exclusively limited to two sensory modalities in the speech perception literature—audition and vision—with audition generally seen as the primary or dominant modality; that is, when studies of
speech perception have discussed cross-modal effects, they have generally described influences of specific visual speech information on auditory processing of speech. For example, in their seminal work on audiovisual speech processing, Sumby & Pollack (1954) found that visual speech information can enhance accuracy of auditory speech perception—especially when audio signals are highly degraded. Their groundbreaking work led to a conclusion that now seems obvious: speech has visual characteristics in addition to auditory characteristics. As Bateson and colleagues framed it, “The motor planning and execution associated with producing speech necessarily generates visual information as a by-product.” (Vatikiotis-Bateson et al. 1996, p. 221). Perhaps more than a by-product, the visual information produced during speech may be integrated fundamentally into our representations of speech, as evidenced by the observation that blind and sighted speakers use different lip movements to produce the same speech sounds (Ménard et al. 2016).

Sumby & Pollack’s (1954) enhancement work, while groundbreaking, did not investigate what aspects of the visual signal were useful to perceivers and why, nor did they speculate about what the mechanism might be that allows perceivers to use this information. Nonetheless, their work opened the door to an important shift in thought: researchers could no longer consider audition to be the sole modality for the transmission and processing of speech. Twenty years later, McGurk & McDonald (1976) published their widely known finding that perceivers given incongruent audiovisual stimuli will often choose neither the response consistent with the auditory stimulus nor that which is consistent with the visual stimulus. As will be discussed further below, the decades of scholarly energy, particularly among linguists and psychologists, following McGurk & McDonald (1976) that went into understanding the so-called McGurk effect and audiovisual speech more generally brought us great advances in understanding speech perception from at least
a bimodal perspective; at the same time, however, this focus on audiovisual speech perception diverted attention away from the development of theoretical models that cast speech in a more broadly multisensory space.

While researchers in linguistics and psychology were focused on audiovisual speech, clinical researchers were also investigating cross-modal effects on perception in the form of using tactile speech information as a means of enhancing speech perception for impaired populations (e.g., Alcorn 1932; Sparks et al. 1978). The most well-known example is the Tadoma method (Alcorn 1932; Vivian 1966; Reed et al. 1978), a method most commonly used with deaf and blind individuals who lost their hearing and sight early in childhood (around 18 months of age). In Tadoma, the perceiver places her hand on the face of an interlocutor in such a way that he or she can feel much of the interlocutor’s articulatory movements and their sensory consequences (though see Reed et al. 1989). By placing the thumb at the lips and fanning the fingers across the cheek and neck, the perceiver can feel the movements of the lips and jaw, vibrations at the neck, and airflow at the lips and nose. The success of deaf-blind individuals in using Tadoma to communicate certainly suggests that tactile information is viable for communication. Since most of this work has come from a clinical perspective, the focus has naturally been on the use of tactile information as a means of communication for clinical populations.

Although there has been a great deal of research into how the resulting tactile information can teach impaired perceivers to communicate, there has been less of an attempt to consider how this tactile information—and the fact that it can be used at all—fits into conceptions of how speech perception works for a general population (i.e., outside of purely clinical applications). For example, there is little research in the clinical literature asking what the Tadoma method might tell us about how normally hearing and seeing populations make use of tactile information, and, in
turn, what this might tell us about cognition and communication more broadly. Instead, the assumption has persisted that tactile information can be recruited to support the auditory and visual streams with training if the other streams are unavailable or degraded. This in turn has the effect of reinforcing the assumption that tactile cues require a great deal of training in order to be effective (see, e.g., Bernstein et al. 1991). However, as I will discuss further in later sections, more recent psychological and linguistic research has shown that tactile information can have a modulatory effect on speech perception for even untrained perceivers, suggesting that speech perceivers use signal information from whatever sources they have available, whether audio, visual, or touch.

In the following sections, I first review some of the many important findings in audiovisual speech perception. I will discuss how these findings have shaped our understanding of the what the perceptual system detects and makes use of during audiovisual speech and when this integration occurs. In Section 3, I will focus on identifying the other modalities relevant in speech perception as well as delineating the role of production, with a particular focus on aerotactile somatosensation. Then I will describe how broadening our research to include these additional modalities has enriched our understanding of multi-modal speech perception. Finally, I will end with some important gaps that remain in our knowledge thereby laying the groundwork for the experiments in the following chapters.

1.1.1 Audiovisual Speech

1.1.1.1 The McGurk Illusion: Fusion or Confusion?
Discussions of cross-modal effects in speech perception often begin with the McGurk effect (McGurk & MacDonald 1976), perhaps the most widely known and most-studied cross-modal effect on speech perception. In the original study, the authors presented participants with incongruent auditory and visual speech stimuli (auditory *ba* dubbed over a visual *ga*). The participants were then asked to indicate what they heard. The authors found that participants often responded with neither the syllable that matched the visual token nor with the syllable that matched the auditory token; rather, participants were significantly more likely to respond that the speaker had said *da*, a sequence not present in either modality. These results were compelling in part because they showed that visual information can influence auditory perception even when the acoustic signal is clear and not degraded. In some ways, the McGurk illusion seems largely automatic, suggesting that perceivers cannot avoid integrating the visual information. For example, perceivers report hearing *da* even when explicitly told about the dubbing and instructed to attend to the auditory cue, and after receiving training at attending to the auditory cue (Massaro 1987). Moreover, perceivers appear to be surprisingly insensitive to such factors as gender congruence between face and voice (Green et al. 1991), degradation in the visual signal (Rosenblum & Saldaña 1996), and temporal asynchrony between audio and video signals (though the direction of the asynchrony appears to matter, as discussed in Section 1.1.1.3 below).

While the McGurk illusion has been observed in many studies, there is also considerable evidence that it is not nearly as robust as is often suggested and may even disappear in certain contexts. For example, although it is often noted that the McGurk effect persists when there are incongruities between aspects of the voice producing the auditory stimulus and the face producing the visual stimulus (Green et al. 1991), there are circumstances when an apparent mismatch in the source affects the McGurk effect. Walker et al. (1995) found that the illusion is greatly reduced if
familiar faces and voices are used. In addition, the McGurk effect is sensitive to differences across vowel context (Green et al. 1988) such that while /Ci/ contexts elicit strong McGurk effects, /Ca/ is variable and /Cu/ is quite weak. Further, both coarticulatory cues to the following vowel (Green & Gerdeman 1995) and high attention loads can disrupt the McGurk effect (Alsius et al. 2005; Alsius et al. 2007; Tiippana et al. 2014). Finally, the effect is highly subject-dependent (e.g., Basu Mallick et al. 2015), a fact that was confirmed by a reanalysis of a large corpus of McGurk data (Schwartz 2010).

Perhaps the larger issue, however, is that a key conclusion drawn from this effect seems fallacious: namely the idea that true audiovisual integration equals a fused percept. The concept of fusion stems from the fact that the reported percept matches neither the auditory percept nor the visual one. Proponents argue that the novel percept emerges because the perceiver “fuses” place features from the auditory bilabial [ba] and the visual velar [ga] to arrive at the alveolar [da] (van Wassenhove 2013) a view that is problematic on several levels. First, it implies that [d] is somehow an intermediate segment between bilabial [b] and velar [ɡ]. This further suggests that these stops exist on a continuum and the perceiver is selecting a midpoint between the articulatory features of the incongruent stimuli (e.g., de Gelder et al. 1996). While it is indeed true that the alveolar ridge is positioned between the lips and the soft palate, the muscles and structures involved in these articulations are so different that the idea of alveolar as a kind of bilabial-velar compromise is unlikely. This take on the fusion illusion also assumes that perceivers are correctly identifying the visual stimulus as /ga/, an assumption that is impossible to confirm. More to the point, as noted in Tiippana (2014), the concept of fusion as the true or ideal example of integration ignores the other percepts reported in McGurk-style tasks.
Beyond the above issues with fusion, the idea that only a fused percept indicates true audiovisual integration has in some ways slowed movement toward understanding the more richly multisensory nature of speech. Audiovisual integration happens even when the McGurk effect does not (Brancazio et al. 2002). Once we move past the idea that cross-modal effects on perception always result in a fusion, the landscape of cross-modal speech effects starts to make a bit more sense. Instead, the usual integration response is that having congruous information from more than one modality enhances perception because one sense provides information missing from the signal given to the other sense. In contrast, incongruous cross-modal stimuli generate confusion as the signals provide conflicting information that does not match what normally occurs in real-world speech. That is, processing incongruous cross-modal information (as in studies of the McGurk effect) is, at least in some respects, a fundamentally different kind of task from processing congruous cross-modal information; this is supported by neuroimaging evidence that the brain recruits additional cortical areas when processing incongruent audiovisual speech (Erickson et al. 2014).

1.1.1.2 Beyond McGurk

Beyond the well-known McGurk effect, as described above, movements of the vocal tract involved in the production of speech sounds have well-known visual and auditory consequences (Munhall et al. 2004; Munhall & Vatikiotis-Bateson 2004; Vatikiotis-Bateson et al. 2000; Vatikiotis-Bateson et al. 1996). One of the most interesting discoveries regarding cross-modal audiovisual effects is that these visual speech cues do not just come from observing the oral aperture; rather, perceivers can make use of phonetic information from kinematic movements of
essentially all parts of the face: the lips, jaw, neck, cheeks, eyebrows - and even movements of the whole head (Vatikiotis-Bateson et al. 1996; Yehia et al. 1998); these movements have been shown to convey different types of speech information, and they participate in cross-modal audiovisual effects that operate under a variety of experimental conditions. For example, we have considerable evidence that visual speech information can supplement a degraded or ambiguous acoustic signal (e.g., Sumby & Pollack 1954). Evidence also shows that visual speech information modulates auditory perception even when the auditory signal is clear and unmanipulated. For example, adding visual speech information improves a participant’s tracking rate in speech shadowing tasks (Reisberg et al. 1987) and facilitates listening comprehension for heavily accented or semantically and syntactically complex speech (Arnold & Hill 2001).

Some research in audiovisual perception suggests further that auditory and visual speech cues do not combine in a simple, additive fashion, but rather can be superadditive. Superadditivity refers to the observation that perceivers’ responses to multisensory information are not merely the result of a summation of the unimodal parts; rather, perceivers gain a disproportionate benefit from the addition of cross-modal information. Behavioral studies have shown evidence of superadditivity in audiovisual speech (e.g., McGrath & Summerfield 1985) showing that when perceivers experience audiovisual speech, their responses are more accurate than expected based on the sum of their responses to visual cues and auditory cues in unimodal conditions. While this seems to be true of behavioral results, electrophysiological evidence suggests that audiovisual speech perception may be subadditive at the neural level. Klucharev et al. (2003) reported that when ERP responses to audiovisual stimuli are compared to the sum of unimodal auditory and visual speech stimuli responses, the audiovisual responses were found to be smaller than the sum of the unimodal responses. While the neural findings may at first appear to conflict with the superadditivity shown
in behavioral studies, they may be related to the decreased ambiguity in an audiovisual signal as compared to an auditory only signal. Studies using ambiguous stimuli indicate that, when exposed to congruent cross-modal cues, the brain does not have to work as hard to process the signal as it would if the input were unimodal (Parker & Krug 2003). In addition, a recent meta-analysis drawing on the results of hundreds of EEG studies of audiovisual speech (Baart 2016) reports that audiovisual speech reduces the amplitude of the auditory N1 peak (a negative peak at around 100 ms triggered by a sudden onset of sound) and the subsequent positive P2 peak that occurs around 200 ms compared to audio speech alone, confirming that visual speech production facilitates the processing of auditory information. Similar decreases have been found where lower ERP amplitudes were observed during the easier task of decoding normal speech compared to disordered speech (Theys 2014a, 2014b).

Some of the most revealing cross-modal effects to date in audiovisual speech perception have been seen in observing how brain activation responds to speech, both cross-modal and unimodal. A particularly interesting finding concerns the unimodal perception of visual speech cues during silent speech. Multiple studies have found that silent visual speech information generates activation in the areas of the brain primarily associated with auditory processing (Campbell et al. 2001; Calvert et al. 1997; MacSweeney et al. 2000). For example, Calvert et al. (1997) used fMRI to test the brain activation of normal hearing subjects during a variety of conditions: silent lip reading, heard speech, nonspeech lip/jaw movements, and speech-like lip and jaw movements. The authors reported activation in the primary auditory cortex not only in the condition where participants heard acoustic speech cues, but also during silent lipreading trials and trials with phonetically-plausible lip and jaw movements. Crucially, this activation did not occur during trials in which the participants saw non-speech lip and jaw movements. The authors interpret these
findings as indicating that visual information influences the neural processing of auditory cues long before they are categorized into phonemic categories.

Similarly, incongruent visual speech information modulates responses in the auditory cortex. Sams et al. (1991) used MEG and a McGurk-style task to investigate in which part of the brain the visual information affects auditory processing. The authors note that classical theories assume that the visual information is processed in the occipital cortex before being sent to the angular gyrus for “reorganization into auditory form.” However, they did not find coherent activity in these two regions. Instead, they found responses in the primary auditory cortex. This is further evidence that visually presented articulatory movements modulate responses in a cortical area typically associated with auditory processing. This in turn suggests that the brain processes speech as a multimodal percept rather than as unimodal streams integrated at later stages, a view further supported by evidence that auditory and visual speech information interact in lower-order structures such as the brainstem (Musacchia et al. 2006).

1.1.1.3 Spatiotemporal Congruence

While there has been some debate within the speech perception literature as to the degree to which speech is “special” and not governed by the same requirements as non-speech (Liberman 1984; Fowler 1991; Tuomainen et al. 2005; Vatakis et al. 2008; Vroomen & Stekelenburg 2011), work on audiovisual speech has provided evidence that cross-modal effects are, at least in part, constrained by general properties of the natural world. Many researchers interested in multi-modal perception outside the domain of speech point to the importance that coincidence in space and time plays in governing perceptual integration (e.g., Holmes & Spence 2006; Macaluso & Driver 2005).
Indeed, it appears that relative timing plays an important role in integrating cross-modal speech cues. For example, multiple studies have shown that there exists a temporal window within which integration occurs (Munhall et al. 1996; van Wassenhove et al. 2007). In other words, the onsets of the cross-modal cues must co-occur within a specific window of time in order to be integrated. This window can be as large as 200 ms, suggesting that perceivers are relatively forgiving when determining which incoming information relates to a perceptual event. In addition, the temporal window has been shown to be asymmetrical. Thus, in Munhall et al. (1996), participants showed a significant decline in integration when the visual speech cue (in the form of a video) preceded the auditory cue by more than 180 ms, while in contrast, participants were much less forgiving when the auditory cues preceded the visual: the decline in integration occurred when the cues were offset by just 60 ms. Munhall et al. (1996) suggest that this asymmetry can be explained by facts about how the cues behave in the natural world (i.e., that the speed of light is faster than the speed of sound). Perceivers may thus be sensitive to the natural relationship between the relative speeds of auditory and visual signals, and are more likely to integrate cross-modal cues that agree with basic principles of physics.

The story becomes murkier, however, when spatial congruence is considered. Given that stimuli coming from different directions are unlikely to originate at the same source, spatial incongruence of cross-modal cues may interfere with perceivers’ ability to integrate them. However, though the “spatial rule” appears to hold in multi-modal perception outside of speech (Soto-Faraco et al. 2003), it seems to have little to no effect on audiovisual speech perception (Bertelson et al. 1994; Fisher & Pylyshyn 1994; Jones & Munhall 1997). Studies have shown little evidence that perceivers care about whether cross-modal cues appear to originate from the same spatial location. Indeed, it is this lack of constraint on spatial congruence that enables
ventriloquism, surely the most popularly known of cross-modal speech illusions. Though some studies have tested very small degrees of dislocation (Bertelson et al. 1994; Fisher & Pylyshyn 1994), Jones & Munhall (1997) reported that participants showed a strong McGurk effect even when the auditory and visual stimulus were separated by as much as 90 degrees. Though these findings are somewhat expected given the robustness of the ventriloquism effect, it is surprising that perceivers might rely on synchrony alone. However, it remains possible that their findings are specific to audiovisual processing and not a fact about cross-modal speech perception more generally.

While studies of interactions in audiovisual speech have provided many useful insights, they have left many questions unanswered. The broadest of these questions is whether observations about audiovisual speech are specific to the audiovisual pairing, or whether these observations indicate more general properties that would obtain across any cross-modal (or multi-modal) pairing. Getting at this question requires moving beyond audiovisual speech to compare additional modality pairings. The remainder of this section focuses on studies of cross-modal effects that have attempted to extend the range of modalities.

1.1.2 Somatosensory Speech

1.1.2.1 The Many Somatosenses

Somatosensory space has been identified for decades as a rich yet largely unexplored frontier for speech research (e.g., Abbs & Gracco 1984; Gick et al. 2008; Ghosh et al. 2010; Kelso et al. 1984; Nasir & Ostry 2006; Perkell 2012; Tremblay et al. 2003). While the work in this area has been
vital in establishing somatosensory information as playing an important role in speech, the various
effects referred to in these studies are often attributed to a single modality under the label
“somatosensory”. Although the somatosenses have often been described as a single sense modality
in the speech literature, the term “somatosense” applies to a broad range of different sensory
modalities (Hsiao & Gomez-Ramirez 2011). As Hsiao & Gomez-Ramirez (2011) put it:

“The somatosensory system is best conceptualized as a multi-modal, rather than a
unimodal, processor, comprised of multiple parallel systems carrying information about
numerous aspects of environmental stimuli. To the extent that a given object encountered in the
real world may simultaneously generate multiple tactile impressions, what is remarkable
is that the somatosensory system unites these disparate channels into a unified percept (141).”

Figure 1.1 Research over the last decade has shown that speech occurs in a richly multi-modal space.
Auditory, visual, and several of the many somatosenses have been shown to interact with one another during
speech perception. At a minimum, we know that pressure, proprioception, vibration, and aerotactile
somatosensation are all relevant tactile inputs for speech.
The “disparate channels” described here—the various somatosenses—include senses such as pain, itch, temperature, pressure, vibration, and joint position, among others (Hsiao & Gomez-Ramirez 2011; Wilson et al. 2009), each with its own distinct organs, mechanisms and neural processes, and with some researchers further dividing the numerous somatosenses into subgroups (e.g., tactile, proprioceptive, and vestibular; Anderson and Fairgrieve, 1996). Thus, while perturbation studies have often examined the role of proprioception in speech (e.g., Kelso et al. 1984; Nasir & Ostry 2006), other studies have examined mainly pressure sense (e.g., Ghosh et al. 2010), others the vibrotactile sense (e.g., Bernstein et al., 1991), and still others the aerotactile sense (e.g., Gick & Derrick 2009, Derrick & Gick 2013). The various somatosenses should thus be viewed as being approximately as distinct from one another as they are from those modalities that have been more traditionally recognized in the speech literature (i.e., vision and audition)—and as such, the same kinds of approaches traditionally used to investigate cross-modal processes in speech ought to be brought to bear on each of the different somatosenses. Minimally, in addition to the senses of audition and vision, this should apply to proprioception, pressure, vibration and aerotactile sensation, all of which should be considered highly relevant modalities for experiencing and conveying speech, and each of which merits study both separately and in combination with other modalities in speech. Far from taking place in a one-, two- or even three-dimensional space, speech should thus be viewed as occupying a sensory space that is indeed richly multidimensional (see Figure 1.1).

Many of the above behavioral experiments relating to somatosensory input are focused on speech production rather than perception. Production has at times been left out of discussions of cross-modal effects on speech, or at times treated as an additional modality or sense. However, while there is no “production” sense, the movements of speech production do generate much of
the somatosensory feedback shown to modulate speech perception and are thus important in any
discussion of cross-modal effects. When we speak, we receive continuous sensory input through
multiple modalities: the sense of air flowing across the articulators, the vibration of the vocal folds
in the neck and other structures, pressure and proprioception from facial skin deformation, and of
course, acoustic-auditory feedback, among others. Though we may not be consciously aware of
this sensory feedback, it has been shown that manipulating the somatosensory feedback from our
articulators during speech production causes speakers to alter their production. For example, when
the lips (Abbs & Gracco 1984) or jaw (Kelso et al. 1984) are mechanically perturbed, speakers
adapt their articulatory strategies to compensate for the perturbation. This holds even when the
perturbation does not alter the acoustic signal, suggesting that somatosensory targets themselves
are relevant in the production of speech sounds independent of auditory goals (Tremblay et al.
2003; Nasir & Ostry, 2006).

This cross-modal interaction has been shown to feed into speakers’ productions more
generally. For example, somatosensory acuity, or the keenness of an individual’s perception of
somatosensory input, has been shown to correlate with produced acoustic contrast distance
between /s/ and /ʃ/ (Ghosh et al. 2010). This effect is independent of auditory acuity, which also
suggests that speech sounds have independent somatosensory and auditory perceptual goals. It is
also the case that inputs generated by production can influence and be influenced by auditory
perception. In a novel “skin-stretching” paradigm, Ito et al. (2009) simulate the somatosensory
consequences of lip spreading; they find that stretching the facial tissue of perceivers in such a
way that mimics skin deformation in vowel production influences the perception of the vowel
sound heard by those participants. Strikingly, the reverse also holds: hearing a vowel can shift the
perceived direction of skin stretch (Ito & Ostry 2012). These cross-modal effects hold in spite of
evidence that the lips do not contain muscle spindle proprioceptors to provide input regarding changes in muscle length (Frayne et al. 2016).

When our perception in one modality is affected by our own production, it is unclear whether we are responding to the sensations themselves or to possible sensory consequences simulated by our internal models, known as “efference copy” or “corollary discharge”. It has been suggested that when a motor command is sent to the motor system to initiate an action, a copy of this command signal (the efference copy) is sent to an internal forward model, generating a prediction of the perceptual or sensory consequences of the action (Pickering & Garrod 2013). This prediction is a sensory signal known as corollary discharge (Scott 2012). In this view, auditory-motor adaptation occurs when the incoming sensory signal does not match the prediction. Experimental investigations of corollary discharge in speech production have focused on the auditory consequences of speech production. For example, Tian & Poeppel (2010) showed that the activity of the auditory cortex during articulatory imagery tasks, where participants imagined producing certain sounds, is strikingly similar to its activity during the perception of actual auditory stimuli; this was argued to be a result of predicted auditory feedback (corollary discharge). Further, Curio et al. (2000) and others have shown that responses in the auditory cortex are attenuated during speaking as compared to responses to taped speech, particularly in the left hemisphere. This effect remains even when the tape-recorded speech is that of the participants themselves (Houde et al. 2002). However, Houde et al (2002) showed that when the participant’s auditory feedback is perturbed, the suppression effect disappears. The authors interpret this as evidence that the match between the corollary discharge and the actual acoustic output of the motor action results in reduced neural activity in relevant brain areas. It has further been shown that the corollary discharge involved in auditory imagery constitutes a sufficiently rich and detailed
representation of the predicted sensory information that it can interfere with the perception of actual external auditory signals (Scott 2012). Our immediate responses to sensory feedback during everyday speech production may well be responses to corollary discharge rather than to the physical productions themselves, opening possibilities for novel experimental approaches to understanding cross-modal effects involving speech production.

1.1.2.2 The Tadoma Method: A Multimodal Method

As mentioned above, we have long known from clinical research on the Tadoma method that somatosensory information can, with adequate training, provide a useful aid to communication. As the Tadoma method was initially developed for individuals who have a loss of hearing and vision, it has often been used for populations who no longer have access to speech information from hearing and sight, but who did have access to the visual and auditory cues at one time (Alcorn 1932; Vivian 1966). From a clinical point of view, research has thus focused more on using somatosensory cues as a substitution for missing cross-modal information rather than on the role of somatosensory input as a natural part of everyday speech perception. Nevertheless, it has been observed that individuals who lost their hearing and sight as young as 1.5 years old can become successful communicators with the Tadoma method, with the interpretation that the Tadoma input does not just provide access to the speech stream, but serves to “create a language base” for individuals whose language acquisition was interrupted at a fairly early stage (Reed 1996). Further to this point, experimental work shows that sensorimotor information from the articulators is both available and influential during speech perception for infants as young as six months old (Yeung & Werker 2009; Bruderer et al. 2016; Choi et al. 2019). As I will discuss further in Section 1.2
and Chapter 4, Bruderer et al (2016) and Choi et al (2019) both show that impeding an infant’s articulatory movement—in this case, the curling of the tongue tip—interferes with the infant’s perception of sounds produced with that articulation. These results suggest that the links between articulation and acoustics develop early, well before the infant’s first birthday. This in turn suggests that somatosensory information as it relates to articulatory movement and the resulting speech sound was likely already available to deaf and blind individuals such as those discussed in Reed (1996) in some capacity by the time they lost their hearing and sight. For those born deaf and blind, such links may not develop between the somatosensory information available to them and the visual and auditory information that is unavailable.

As noted above, the focus of the clinical research on the Tadoma method did not generally extend to discussions of speech perception in a non-clinical population. However, it did pave the way for later studies of cross-modal tactile effects in speech, starting with Fowler & Dekle (1991), an important early work in opening up cross-modal speech perception research to modality pairings beyond audition and vision. Perhaps just as importantly, the authors showed that somatosensory speech information is available to even untrained perceivers. In their study, the authors contrasted the influence on auditory perception of somatosensory speech information vs. orthographic cues in a McGurk-like task, asking whether the McGurk effect arises because of cue association in memory or because the cross-modal cues jointly specify the same event in the real world. The authors compared participant responses in two conditions with conflicting cross-modal cues. In one, auditory cues were simultaneously presented with either congruent or incongruent mouth movements. Participants placed a hand over the lips of a speaker and were asked to identify which syllable they had heard, as well as which they had felt. The participants were not able to see the face of the speaker, and thus only had access to tactile and auditory speech information during
the condition. In the second condition, participants saw a congruent or incongruent printed syllable on a computer screen at the same time as they heard a syllable. As in the previous condition, participants responded with which syllable they had heard followed by which they had seen. The authors chose these two situations because in the first, the cross-modal cues jointly specify the real-world event and thus have a causal relationship in the natural word. However, most perceivers have little or no experience feeling the mouth movements of a speaker while listening. In contrast, the second condition offered a situation in which the cross-modal cues are associated only by social convention—spelling—yet the undergraduates tested in the study have experience with the visual and acoustic pairing of spelling and sounds. The authors found that the haptic information, and not the orthographic, influenced categorization, lending support to the idea that haptic speech information can be useful to perceivers without hours of training.

Gick et al. (2008) picked up this line of research, using the Tadoma method to further show that somatosensory information not only influences perception of incongruent syllables but that it can increase accuracy of congruent audio-tactile (AT) and visuo-tactile (VT) speech in a syllable identification task. The authors tested a group of perceivers with no previous training in the Tadoma method on their ability to identify syllables through bimodal pairings of auditory, visual, and tactile speech cues. Accuracy improved by nearly 10% when tactile information was available to perceivers when paired with auditory-only or visual-only speech information. Their findings support the idea that perceivers do not need to have previous experience with information in those specific modality pairings for the cues to enhance perception. This study also highlighted the non-additive nature of multimodal speech perception. The degree to which the Tadoma information augmented auditory or visual speech perception varied considerably between individuals, suggesting that perceivers may use information from one modality more than another. This finding
is in line with the audiovisual results above from Schwartz (2010) that some perceivers are more auditory and others are more visual. Gick et al. (2008) further found that participants whose accuracy increased with the addition of tactile cues in one modality pairing (e.g., AT) tended to benefit less from adding the same cues in the other modality pairing (e.g., VT). Importantly, this negative correlation was unrelated to participants’ base acuity in the individual modalities, demonstrating that the individual differences in relative contribution were not due to different baseline perceptual abilities.

While the Tadoma method has thus played an important role in the development of research into the relevance of somatosensation in speech, its history has been shaped by perhaps its most important characteristic, i.e., that Tadoma is not a unimodal method, but a (highly) multimodal one. That is, if we consider somatosensory space as comprising many distinct modalities, this gives a much more complex view of the Tadoma method than merely adding unimodal (“tactile”) information. On the contrary, the hand position for the Tadoma method is designed to allow the perceiver to pick up, at the very least, vibration from the larynx, air flow from the lips and pressure sensation from the moving articulators. Understanding the multisensory nature of the Tadoma method helps to explain both its power as a communicative tool for clinical populations as well as its challenges as an experimental tool for studying speech perception in a normative population. Naturally, interpreting the Tadoma method as if it conveyed a single sense modality rather than three or more would create confusion in any experimental paradigm, and attempting to create a Tadoma experiment with a single modality as an independent variable would seem certain to fail. Considering this likely confound, while some of the more descriptive aspects of a Tadoma study such as Gick et al. (2008) may remain useful, their results regarding superadditivity merit future study using more easily controlled methods than Tadoma.
1.1.2.3 Aerotactile Speech Perception

Following an aspirated plosive such as [pʰ], air exits the mouth relatively slowly, traveling at a velocity an order of magnitude slower than the speed of sound, with the resulting pressure front dispersing and slowing loglinearly as it advances (Derrick et al. 2009). When this slow-moving pressure front strikes skin, it stimulates mechanoreceptors in the skin and in hair follicles that signal the presence of air flow across the skin, creating an aerotactile sensation (Gick & Derrick 2009); the absence of hair reduces the perceptibility of this sensation (Derrick & Gick 2013). The aerotactile sense is one that presents itself as an excellent candidate for experimental study of cross-modal speech effects. As with audible and visual speech information, the aerotactile signal (air flow) is transmitted externally, it is isolable (i.e., independent of other information such as vibration and pressure), and it is relatively easy to perturb and simulate precisely. Further, air flow can be applied to the skin from any direction and at any body location, and is safe enough to be applied to infants, opening doors for a range of novel kinds of experiments. For these reasons, the aerotactile sense has become the source of a number of novel studies of cross-modal effects.

While it may seem difficult to imagine that the sensation of air flow on the skin could play a significant role in speech perception, it becomes less so when we remember that much spoken communication happens within the range of distances where air flow can be felt, particularly during language acquisition, and that, as we talk, we simultaneously perceive the sensations created by our own production. Studies of aerotactile speech perception over the last decade have demonstrated that untrained perceivers incorporate aerotactile somatosensation as an informative part of the speech event. Gick & Derrick (2009) showed that aerotactile information can enhance
or interfere with accurate auditory speech perception for native English perceivers in a two-way forced choice task, even for participants who report being unaware of the airflow. When stop-initial syllables (e.g., /pa/, /ba/, /ta/, or /da/) were accompanied by silent puffs of air applied to the participants' neck and hand, participants were significantly more likely to perceive the syllable as aspirated (i.e., /pa/ or /ta/), even in mismatch conditions (i.e., when the air puff occurred with an unaspirated token). In a control study, Gick & Derrick (2009, supplemental materials) compared the effect of applying a different somatosensory stimulus—a tap on the hand—in an otherwise identical experiment; not surprisingly, they observed no effect on auditory perception, highlighting the distinct effects of information conveyed through different somatosenses. The aerotactile effect on auditory perception has been replicated (e.g., Gick et al. 2010; Derrick & Gick 2013) and extended to enhancement of fricative identification (Derrick et al. 2014). Additionally, air flow applied to the skin has also been used to shift speech perception using tokens along a voicing continuum (Goldenberg et al. 2015), showing that the effect of aerotactile input on audible speech is stronger when auditory cues are more ambiguous.

While most aerotactile cross-modal studies have focused on effects on auditory perception, aerotactile cues have also been shown to apply even in the absence of audition, affecting the perception of visual-only speech. For example, Bicevskis et al. (2016) found that aerotactile information influences the perception of silently articulated bilabial stops. Participants were presented with silent videos of a speaker articulating /ba/ or /pa/. During some of the trials, they felt silent puffs of air on their skin. Just as in the above audio-aerotactile findings, the sensation of air-flow on the skin significantly affected the participants categorization such that participants were significantly more likely report that the speaker produced the voiceless aspirated token /pa/ when the silent video was accompanied by air-flow on the skin. The authors argue that these results
indicate that, rather than using the aerotactile information only in support of an ostensibly primary auditory signal, perceivers simply make use of whatever speech information is available. They further propose that this suggests that speech lacks an exclusive primary modality and can be described as “modality neutral”. However, previous research on multimodal speech perception with individuals who have hearing loss suggest this claim merits further investigation. For example, there is some evidence that cochlear-implant users are better at integrating audiovisual speech information if the cross-modal cues are congruent (Rouger et al. 2007), but not when the cues are incongruent, as in a McGurk task (Rouger et al. 2008). This has led some to suggest that without early auditory exposure, typical speech integration processes do not develop (Schorr 2005). If impairment in the auditory modality does interfere with integrating cross-modal cues, then speech may be more dependent on auditory information than the aforementioned results suggest.

1.1.2.4 Spatiotemporal (In-)Congruence and Ecological Validity

As discussed above, previous studies of audiovisual cross-modal effects have revealed that audiovisual speech has some notable spatiotemporal properties. First, temporal congruence of signals is important for integration within an asymmetrical window in a direction consistent with the laws of physics (e.g., Munhall et al. 1996). Second, spatial congruence does not seem to be required for audiovisual integration in speech (e.g., Jones & Munhall 1997). While a certain degree of temporal congruence is important, perceivers do not integrate just any synchronous set of cross-modal cues. For example, Fowler & Dekle (1991) did not find evidence that simultaneously presented orthographic visual cues influenced auditory speech perception. Likewise, as described
in the preceding section, while Gick & Derrick (2009) found a significant effect of a light puff of air on the hand synchronous with auditorily presented plosives, they observed no effect of a light tap on the hand. Finally, in addition to the stop consonant continua, Goldenberg et al. (2015) presented participants with a vowel continuum. As aspiration plays no role in vowel production, the sensation of airflow should not affect vowel perception. As expected, the authors found that the presence of airflow did not affect participant response for vowels, demonstrating that the airflow has no effect on perception when the cue is irrelevant. Together these results suggest that cues must be more than merely synchronous: rather, it seems that cross-modal stimuli must have a lawful, causal relationship in the real world in order to be perceived as integrated. These findings are corroborated by work showing that neural responses to auditory-tactile stimulation are modulated by congruence between the area of the body touched and the bodily origin of the acoustic signal (Shen et al. 2018). The existence of a real-world relationship between two stimuli, however, suggests shared properties along multiple possible axes, such as temporal congruence, spatial congruence and signal relevance.

Previous work on sensorimotor adaptation has shown that temporal congruence is an important predictor of whether perceivers will shift their articulatory strategies in response to perturbation in the feedback. For example, when Max and Maffet (2015) presented their participants with formant-shifted auditory feedback, the participants altered their articulatory strategies to compensate. However, if the auditory feedback was delayed by 100 milliseconds or more, the participants no longer compensated. Such findings suggest that perceivers expect a degree of temporal congruence between speech information from their somatosenses and auditory speech information.

In order to investigate the role of temporal congruence for aerotactile speech information, Gick et al. (2010) tested perception of synchronous and asynchronous presentations of audible /pa/ and
/ba/ plosives in combination with slight, inaudible, cutaneous air puffs to the skin. As with the previous audiovisual findings, results of this audio-aerotactile study showed an asymmetrical window of integration for the enhancement effect (i.e., improved identification of aspirated tokens in trials with airflow), allowing up to 200 ms of asynchrony when the puff followed the audio signal, but only up to 50 ms when the puff preceded the audio signal. Interestingly, however, the same asymmetrical window did not occur for the interference effect (i.e., impaired identification of unaspirated tokens in trials with airflow), consistent with the differential neural processing of congruent and incongruent cross-modal stimuli (Erickson et al. 2014). Bicevskis et al. (2016) observe a similar asymmetrical window in their visual-aerotactile study. It is notable that these asymmetries in both modality pairings occur in the direction that would be expected based on relative signal speed, supporting the view that the perceptual system is built (whether innately or through experience) to accommodate natural differences in physical transmission speed of multimodal signals.

Given perceivers’ constraints on timing, one might expect perceivers to be similarly sensitive to spatial dislocation. There has been no direct manipulation of signal direction in a speech perception task using aerotactile speech information until the experiment detailed here in Chapter 3, but we have some indirect evidence that spatial and temporal factors are not equally important to perceivers in cross-modal pairings beyond audiovisual speech. One way it is possible to observe spatial congruence through tactile stimuli is through body location rather than signal direction. Indeed, one of the more intriguing outcomes of tactile work in speech is the degree to which speech perception is holistic in body space: perceivers integrate aerotactile speech information whether felt proximally, at the neck, or distally, at the hand (Gick & Derrick 2009)—or even at the ankle (Derrick & Gick 2013). It is difficult to imagine that perceivers have significant prior experience
feeling another speaker’s airflow on their ankles. In addition, evidence from audio-tactile studies with non-speech stimuli suggest that perceivers may not attend to spatial alignment when integrating. Sperdin et al. (2010) found no effect of spatial incongruence on response time in a simple detection task with auditory-somatosensoory stimuli: participants responded faster to multisensory trials than to unisensory trials with no significant difference between the congruent and incongruent versions. Moreover, participants in the study were unable to pinpoint the origin of a unimodal auditory or tactile stimulus after being presented with a multimodal trial, even when informed that spatial location would be task-relevant. The foregoing studies on audio- and visual-aerotactile perception suggest that perceivers will integrate air-flow cues without the presence of a plausible real-world source, at least from a spatial point of view.

As I have described at length above, humans experience speech across a wide range of modalities, both alone and in various combinations. An important question that arises from this multi-modal nature is what kinds of experience, if any, perceivers need to be able to integrate these various sensory streams of speech information. In the following section, I will turn to this question.

1.2 (Prior) Experience and Speech Perception

While no research has directly tested the effects of prior experience on a perceiver’s ability to integrate aerotactile speech information, many experiments investigating audio, visual, and sensorimotor speech information have provided evidence for a modulatory role for experience. For example, prior experience with specific audio and visual cues affects the degree to which perceivers integrate incongruent cues during a McGurk task. When Walker et al. (1995) tested native English perceivers on incongruent audiovisual speech, perceivers were significantly less likely to integrate incongruent visual and auditory speech cues if they are familiar with the voice
or face used in the stimuli. Prior experience with a specific language also affects audiovisual speech perception. Werker et al (1992) presented French Canadian monolinguals and English-French bilinguals with incongruent syllables in a McGurk-style task. The French Canadian monolinguals reported hearing significantly fewer instances of the interdental fricative /ð/—a sound not present in their native phonological inventory—than the bilinguals. Instead, the monolinguals responded with the alveolar stop /t/, arguably the closest available sound. As Werker and colleagues note, their results suggest that seeing, hearing, and producing native sounds modulates audiovisual perception for adult perceivers.

Yet while these results and others suggest a modulatory role for experience, the question remains whether experience plays a deterministic role. Any investigation into this question is complicated by just how early in life speech experience begins. Humans are exposed to auditory speech information well before we are born. In utero, fetal hearing matures enough that the fetus can hear low-frequency noises by roughly 22 weeks. By 25-29 weeks gestation, fetuses show consistent responses to non-speech vibroacoustic stimulation (Birnholz & Benacerraf 1983). At term (but before birth), a fetus can discriminate its mother’s voice from another that of another speaker (Kisilevsky et al. 2003), as well as its native language from a rhythmically distinct non-native language (Kisilevsky et al. 2009). Thus, even at birth, neonates have fairly significant experience with auditory and vibrotactile speech information. In addition, an infant gains access to additional cross-modal cues during speech perception almost immediately after birth: visual information certainly, and perhaps aerotactile, as well. This means that neonates tested just hours after birth already have experience with multimodal speech.

Beyond this, fetuses also have significant prenatal non-speech experience that may help form the link between auditory, visual, and self-generated oral-motor cues. Infants frequently
exhibit sucking and swallowing behavior prenatally (Arabin, 2004; Kurjak, Stanojevic, Azumendi, & Carrera, 2005). It has been suggested that this general experience with upper vocal tract movements and the proprioceptive feedback involved may help set infants up to be able to integrate the various speech information without experience producing those specific sounds themselves (Choi et al. 2017). Choi and her co-authors propose that tongue retraction and protrusion assist in the creation of a template onto which visual and auditory speech information about place of articulation can be mapped during perception. By supporting the dorsal speech pathways dedicated to auditory-motor mapping, such non-speech fetal movements help lay the foundation for the integration of audio, visual, and sensorimotor speech information. This, the authors argue, could explain how prelinguistic infants can exhibit sound-specific cross-modal effects for sounds they have experience neither producing nor perceiving (Bruderer et al. 2015; Choi et al. 2019).

General experience may lay the foundation for multimodal perception, but many studies have shown that specific experience affects both unisensory and multisensory perception, though perhaps in an unexpected direction. Infant perceptual systems appear sensitive to a variety of visual and auditory signals they have little to no experience with. Rather than broadening infant perceptual sensitivities, specific experience during development narrows the range of differences their perceptual systems discriminate to characteristics relevant to the infant’s environment. For example, while infants can discriminate both human and non-human primate faces at 6 months, they are only sensitive to differences in human faces by 9 months (Pascalis et al. 2002). Infants also show a decline in their ability to discriminate other-race faces at around 6 months (Kelly et al. 2007). In auditory speech perception, 6-8 month old infants can discriminate many non-native contrasts, but this perceptual sensitivity declines over the first year of life (Werker & Tees 1984; Werker & Lalonde 1988; Kuhl 1998). This effect of specific language experience occurs with
multimodal speech signals, as well. In audiovisual speech perception, infants can detect audiovisual congruence for both native and non-native syllables at 6 and 9 months, but by 11 months are only sensitive to audiovisual congruence for native syllables (Pons et al. 2009). Danielson et al (2017) extended these findings to show a similar shift in infants’ sensitivity to audiovisual phonetic incongruence and provide compelling evidence against a learned relationship between specific sight/sound pairings. The authors presented 6, 9, and 11 month old infants with congruent and incongruent audiovisual syllables using retroflex plosive [ɖ] and dental plosive [d̪] produced by a native Hindi speaker. Crucially, the English monolingual infants in the study had no prior experience with these sounds in their language environment. The authors found that the infants in the incongruent condition were able to detect the phonetic mismatch between visual and auditory speech information for these non-native sounds. Furthermore, the authors found that the ability to detect this audiovisual incongruence shows a decline similar to that seen in infant auditory discrimination of non-native sounds. Danielson and colleagues found that the infants detected phonetic incongruence in non-native audiovisual speech at 6 months and 9 months, but not at 11 months old.

In fact, these aforementioned results from both unisensory and multisensory perception have led some researchers to speculate that perceptual narrowing is pan-sensory (Lewkowicz & Ghazanfar 2006) such that we begin with broad sensitivities across all modalities that narrow through experience with the world. The evidence from the developmental literature that infants integrate audio, visual, and sensorimotor speech information without specific prior experience certainly support this view. However, it is not clear that aerotactile perception ought to behave in the same way. While the mechanism proposed by Choi et al. (2017) described above allows for a mapping from oral motor gestures to place of articulation information in utero, the characteristics
of the uterine environment make it difficult to extend their argument to aerotactile feedback. Because fetuses develop in a fluid-filled amniotic sac, they do not produce airflow—speech-related or not. Therefore, infants cannot have had even general experience with airflow before birth and some additional experience with airflow is likely required. Second, as I will discuss further in Chapter 4, our perceptual experience with speech-related airflow from others is may not be reliable enough to form a link between audio, visual, and aerotactile speech information. Thus, it remains an empirical question whether prior specific experience determines or modulates the integration of aerotactile speech information during perception.

1.3 Current Experiments

The foregoing studies in aerotactile speech perception have demonstrated that adult native English perceivers integrate audio, visual, and aerotactile speech information without prior training. But it remains unknown what if any prior experience enabled the participants to use speech-related airflow during perception. Our perceptual experiences are many and complex due to the richly multisensory nature of our physical environment and there are several perceptual avenues through which perceivers might experience speech-related airflow information. First, perceivers receive somatosensory feedback of their own airflow during production. Second, perceivers may feel the airflow of an interlocutor (e.g., a caregiver speaking closely to an infant, or someone whispering in another person’s ear). Third, perceivers can hear the airflow in the acoustic signal. And finally, perceivers may see the environmental effect of airflow during speech (e.g., a flickering candle as in Mayer et al. 2013). Given these various inputs, the possible ways in which experience modulates aerotactile integration are numerous. To fully understand the role of prior perceptual experience in aerotactile speech perception, considerably more research must be undertaken. This dissertation
aims to chip away at the gap in the literature through three experiments that ask the following research questions:

1) Can adult perceivers integrate multimodal stimuli with which they have prior *general* but not *specific* experience?

2) Can adult perceivers integrate multimodal stimuli that do not conform with their prior experience with the unimodal cues?

3) Do infants need experience producing sounds with aspiration before they can integrate aerotactile speech information during perception?
Chapter 2: General vs. Specific Experience

We commonly encounter digitally rendered faces in our daily lives, whether in animated films, video games, or increasingly in avatar-mediated interactions in virtual reality. The success of these technologies rests upon the notion that perceivers can attribute linguistic sounds coming through speakers or headphones to a simulated face on a screen. For example, moviegoers have no trouble associating an actor’s voice with the speaking face of even an abstract animated character. Many studies in audiovisual (AV) speech perception have made use of this perceptual flexibility and used simulated faces to present visual speech information (e.g., Massaro & Cohen 1990b; Bosseler & Massaro 2003; Massaro & Bosseler 2006; Ouni et al. 2006). It remains unclear, however, whether perceivers process speech from an animated face the same way they do a human face, and thus whether results reported for experiments using simulated interlocutors reflect the same integration processes used in real-world interactions.

One obvious advantage to a digital face over a human one is the increased control available over the details of a visual stimulus. For example, Cohen & Massaro (1990) synthesized visual speech information for each phoneme by varying 11 articulatory parameters controlling points on the model’s underlying grid. This allowed them to create Baldi, a highly programmable animated face with fairly nuanced and precise visual speech information for audiovisual stimuli. When the audiovisual stimuli were used in McGurk-like tasks, the authors found that participants responded with the novel percept, suggesting that perceivers can integrate audio and visual speech information from synthesized faces (Cohen and Massaro 1990). Baldi, and other animated talking heads, have been employed successfully in a variety of audiovisual tasks: to assist hard-of-hearing individuals during telephone conversations (Beskow et al. 2004); as a language tutor for autistic
(Bosseler & Massaro 2003; Massaro & Bosseler 2006) and hard-of-hearing (Massaro & Light 2004) children; and to improve second language learners’ production of non-native consonants (Massaro & Light 2003). Still, speech articulations from even precisely programmed talking heads such as Baldi remain less informative than real faces (Ouni et al. 2006), leaving open the possibility that perceivers make use of synthesized visual speech information differently than natural visual speech information.

Given that most individuals in contemporary society have considerable experience connecting voices with animated or computer-generated characters, it is impossible to say whether the audiovisual findings above demonstrate an automatic integration of cross-modal speech cues. It remains unknown what roles development and experience play. To our knowledge, no research exists examining how infants—a population with arguably less experience perceiving animated talking heads—responds to speech originating from animated faces. However, a study testing uncanny valley effects suggests that young infants are surprisingly insensitive to an avatar’s artificial humanity (Lewkowicz & Ghazanfar 2012). When groups of infants (ages 6, 8, 10, and 12 months) were presented with videos of human actors and realistic avatars silently producing the syllable /ba/, only the 12-month-olds showed an ability to discriminate between the faces. The authors argue that before 12 months, infants may not possess the perceptual expertise to notice the subtle differences between a real face and a realistic computer-generated one. They further propose that our ability to distinguish human from human-like develops over time through experience with human faces. There is ample evidence that infants integrate audio speech information from talking human faces on a video from a very young age (e.g., Kuhl & Meltzoff, 1982; Rosenblum, Schmuckler, & Johnson, 1997; Patterson & Werker 2003; Patterson & Werker 1999; Burnham & Dodd, 2004; Pons et al. 2009; Danielson et al. 2017). Thus, Lewkowicz & Ghazanfar’s finding
that young infants do not discriminate between human faces and avatar faces suggests that infants in their first year of life would likely integrate multimodal speech from a talking animated face, as well. However, the aforementioned study used unimodal stimuli and cannot provide conclusive results. Therefore, it remains possible that the behavioral results from adult experiments with non-human faces reflect a learned association between auditory speech information and a non-human source. To test this, we conducted a series of experiments using a cross-modal cue that participants have no experience perceiving from a simulated face: aerotactile speech information.

While we may not be consciously aware of the tiny bursts of airflow emitted during the production of some sounds (e.g., aspirated stops), airflow is one of many somatosensory inputs our brain receives during speech. Air not only flows across our speech articulators and potentially our extremities, but we may also feel the airflow of others when speaking in close proximity. Previous research has shown that this aerotactile input—the sensation of airflow on the body—can influence the perception of stop consonants. Gick and Derrick (2009) demonstrated that native English speakers are more likely to report hearing an aspirated consonant (pa or ta) when they feel a puff of air during the trial. This is true regardless of whether the consonant present in the auditory stimulus is aspirated. This effect of aspiration occurs when air is applied at the neck, hand, and even the ankle (Gick & Derrick 2009; Derrick & Gick 2013). The discovery of this cross-modal effect provides an excellent method for investigating the role of a human source without the confound of experience; unlike in the case of animated audiovisual speech, we can assume that perceivers have no previous experience with speech-related airflow originating from a computer or movie screen.

While there are currently no studies testing visual-aerotactile integration with an animated face presented on a monitor (the Baldi experiments to date have all employed audiovisual stimuli),
we do know that perceivers integrate speech airflow information when it is paired with a *human* face. Bicevskis et al. (2016) presented participants with video clips of a male face producing silent bilabial articulations. On half of the trials, the visual stimulus was co-presented with a gentle puff of air to the participant’s neck. Participants were significantly more likely to judge a token as aspirated when simultaneously presented with a puff of air on the skin. This raises an important question: how would perceivers treat a synthetic source that is not physically capable of producing airflow—such as a digitally rendered speaker? Unlike a human, a computer’s means of producing sound should not be expected to produce a puff of air in the real world. Aerotactile integration may be automatic enough to occur in the absence of a human source of airflow in the perceiver’s environment. However, Bicevskis et al.’s results may instead show that perceivers are willing to extend physical capabilities to a non-present source when the source is human and therefore physically capable of producing the aerotactile information. If the behavioral evidence from the audiovisual literature reflects an automatic integration process unaffected by the source’s animated nature, perceivers should show no decrease in the aspiration effect described above from the beginning of the experiment. But, if the ecological validity of the source of the visual information matters, two possibilities come to mind: 1) participants may not integrate the aerotactile information at all from a talking animated face, or 2) they may learn to associate it with the face over the course of the experiment.

Bicevskis et al. (2016) demonstrated that perceivers integrate aerotactile speech information co-presented with visual speech information from an ecologically valid though non-present source (a video of a human speaker). Building off this finding, we ask whether aerotactile information presented synchronously with visual speech information from an impossible source—a computer-animated face on a computer monitor—affects perception of consonants the way that
information from a human source does. In light of the audio, visual, and tactile research outlined above, we predict participants will exhibit a “ba” bias for tokens presented without air flow, and a “pa” bias for tokens presented with air flow regardless of whether the visual source is from a natural or artificial face.

2.1 Methods: Pre-experiment

Before we began testing our primary hypothesis, we conducted a pre-experiment to validate our methodology. Specifically, we wanted to be sure that the animated face we had created for the main experiment was sufficiently realistic for the mouth movements to be perceived as speech. We used an audiovisual speech-in-noise task to test this: if the participants perceive the avatar’s facial movement as visible speech information, then that visual speech information should enhance the listeners’ auditory perception resulting in increased accuracy during a speech-in-noise task as human faces are known to do (e.g. Sumby & Pollack 1954; McGrath & Summerfield 1985; Ross et al. 2006).

2.1.1 Visual Stimuli

The simulated face stimulus was generated from a two-dimensional female avatar created using computer animation software (CrazyTalk 8). We then created a single video clip of the avatar producing a bilabial plosive followed by a low back vowel (e.g., /ba/). The software does not offer distinct visemes for /b/ and /p/; there is only a single bilabial articulation available. Then, using the program’s text-to-speech feature, the syllable /ba/ was synthesized and synchronized with the avatar’s articulation of the syllable by aligning the stop burst in the .wav file with the release of the bilabial closure. The resulting clip was then exported to a QuickTime file. A second video clip
of the animated face producing a coronal plosive followed by a low back vowel (e.g., /da/) was also generated using the same method. Again, the software offered only a single viseme for /d/ and /t/. The coronal video was created to make the pre-experiment less monotonous for the participants and was not used in the main study.

2.1.2 Audiovisual Stimuli

To create the audiovisual stimuli used in the Pre-Experiment, the simulated face videos described above (both bilabial and coronal productions) were combined with audio tokens of /ba/, /pa/, /ta/ and /da/ produced by a female native English speaker. The audio stimuli were recorded in a sound-attenuated booth using a head-mounted microphone. The tokens were recorded directly into an iMac computer (2012) using sound-editing software (Audacity 2018) and were combined with the visual stimuli that matched in place of articulation (e.g., visual bilabial and auditory /ba/) using Final Cut Pro (Version 10.1.2). To sync the audio and visual stimuli, the stop burst from the naturally produced token was aligned with the existing stop burst in the TTS-token produced by the software. The TTS-token was subsequently removed resulting in an audiovisual token comprised of a naturally produced audio stimulus and an animated face. This resulted in four audiovisual tokens with the animated face: visual bilabial + auditory /ba/; visual bilabial + auditory /pa/; visual coronal + auditory /ta/; and visual coronal + auditory /da/.

2.1.3 Procedure

16 native English speakers (mean age = 21.58 (SD = 2.87), 8 male, and 8 female) took part in a two-alternative forced-choice task. The participants were tested in a sound-attenuated booth in the Interdisciplinary Speech Research Lab at the University of British Columbia. Participants were
recruited from the University of British Columbia and were compensated $5 for a fifteen minute session. Each participant gave informed consent and completed a language background questionnaire. Participants reported no speech, hearing, or language difficulties.

The stimuli were presented on an iMac (2012) using OpenSesame experimental presentation software (Mathôt et al. 2012) and played through Direct Sound Ex-29 headphones. All participants were presented with four different syllables (/pa/, /ba/, /da/, and /ta/) across two types of trials (Audiovisual, Audio-Only) resulting in 8 items. Each item was presented 24 times for a total of 192 trials. During the Audio-Only trials, participants were presented with a black screen while the audio stimulus played. After each trial, participants were asked to use their mouse to select which syllable they had heard (e.g., “ba” or “da”) from a list on the screen. To avoid ceiling effects, task difficulty was increased by playing multitalker babble for the duration of the experiment. The signal-to-noise ratio was calibrated empirically with a separate pilot group of 16 participants to achieve a success rate of 70%. This resulted in a signal-to-noise ratio of -6 SNR. Finally, both trial order and response order within the response form list were randomized.

2.1.4 Results

Accuracy rates increased by roughly 25% when participants were presented with both audio and visual speech information (see Table 2.1) such that participants were better able to identify the initial consonant in the audiovisual condition.
Table 2.1 Summary of means and standard errors by condition for Audio-only and Audiovisual conditions for the pre-experiment.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Mean</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Only</td>
<td>66.72%</td>
<td>1.63%</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>92.36%</td>
<td>1.32%</td>
</tr>
</tbody>
</table>

To test for statistical significance, the data were fit to generalized linear mixed-effects model with Condition as a fixed effect, a random effect of Subject, and a by-subject random slope for Condition. The formula is as follows, with the reference level set to Visual-only:

\[
\text{CorrectResponse} \sim \text{Condition} + (1 + \text{Condition} | \text{Subject})
\]

Where CorrectResponse refers to accurate identification of the syllable and Condition refers to one of Audio-only and Audiovisual. As can be seen in Table 2.2, participants were significantly more accurate during Audiovisual trials as compared to Audio-Only trials. These results demonstrate that participants received a similar enhancement effect from the visual speech information available in the animated face we created as we would expect to see with a human face, validating our use of this particular animated face in the main study.

Table 2.2 Summary of fixed effects for pre-experiment model.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.67</td>
<td>0.02</td>
<td>17.00</td>
<td>38.73</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Audiovisual</td>
<td>0.25</td>
<td>0.01</td>
<td>17.00</td>
<td>13.82</td>
<td>p &lt; 0.001</td>
</tr>
</tbody>
</table>
2.2 Methods: Main Study

For the main study, we used the animated face bilabial video from above as well as two additional human face videos. The human face videos served two functions. First, their inclusion provided necessary controls against which we can compare the simulated face results. Second, the human face stimuli provided an important validation for our simulated face. As mentioned previously, limitations inherent in the animation software meant we were only able to create a single generic bilabial articulation. In contrast, Bicevskis et al. (2016) presented participants with eight tokens each of visual /pa/ and /ba/. More to the point, animation rarely provides the kind of natural variation we see in human speech, making it important to ensure that the results obtained in Bicevskis et al. (2016) using multiple clips of the same talker would replicate in a simplified context. By adding the human face tokens from the aforementioned study we were able to test whether visuo-aerotactile integration persists over multiple presentations of a single token. Unfortunately, the software used to create the animated face did not offer realistic male avatar face. Thus, we were unable to match the gender of the animated face we created to that of the human speaker.

Finally, it is important to note that while we created audiovisual stimuli for the pre-experiment to ensure that the animated face’s articulations were sufficiently realistic, the main experiment was a visual-aerotactile task. Thus, all videos in the main experiment to follow were silent and contained no auditory syllables.
2.2.1 Visual Stimuli

The visual stimuli for the Main Study include 1) Simulated face, 2) human “ba”, and 3) human “pa”. Simulated face is the video of the female avatar from the pre-experiment producing the bilabial articulation, but without the audio signal. The human “pa” and human “ba” stimuli are each a single video of the same human speaker. The “pa” is the speaker producing a voiceless aspirated bilabial stop followed by a low back vowel (/pa/ or [pə]). The “ba” is the same speaker producing a voiceless unaspirated bilabial stop followed by a back vowel (/ba/ or [pa]). Both the human “ba” and the human “pa” video clips were drawn from the no-lag condition of Bicevskis et al. (2016). Still-shots from the visual stimuli, capturing the slightly open mouth at rest, the lip compression prior to stop release, and the open mouth during vowel production can be seen in Figure 2.1.

For all participants, the video stimulus was accompanied by a puff of air synchronized to the release of the obstruent on half of the trials. The following section describes the aerotactile stimulus creation, including the airflow apparatus that delivered the puffs of air.
Figure 2.1 Still shots from the video stimuli. Top = simulated face, middle = human “ba”, bottom = human “pa”. Left = open mouth prior to plosive production. Middle = compress mouth prior to plosive release. Right = open mouth during vowel production following plosive release.

2.2.2 Aerotactile Stimuli

For each of the three videos (simulation, “ba”, and “pa”), we extracted the audio from the video clip. We then split the resulting sound file into a stereo track and inserted a 50 ms 10 kHz sine wave in the left channel using sound-editing software (Audacity, 2018). The voltage from the sine wave triggered the release of air from a compressor (see below for more details). The sine wave
was placed to begin 85 ms, and end 35 ms, before the stop burst to account for system latency. This ensured that the puff of air exiting the tube and the release of the bilabial closure would be synchronous. The left channel of the sound file (with the tone) was then extracted and recombined with the original video clips to create a silent video clip that triggered a simultaneous puff of air. The participants did not hear the tone as the left channel was connected to the airflow apparatus described below and not to a playback system. For the no puff condition, the left channel was left empty so as to avoid activating the air flow system. For all conditions, the right channel was not connected to any playback system, so no audio was produced. Images of the relationship between the underlying acoustic stimuli and the air puff signal are shown in Figure 2.

The puffs of air were generated using a California Air Tools 4610 air compressor located outside of the sound booth. The air compressor was connected to a specially-designed switch box through a ¼” diameter vinyl tube. A second ¼” diameter vinyl tube was connected from the switch box through the access port of the sound booth and attached to a flexible microphone stand. The vinyl tube was adjusted and placed so that the opening of the tube was located 7 cm in front of the participant’s suprasternal notch. As described above, the air release timing was controlled by the 10 kHz sine wave in the left channel of the sound file that accompanied the video clip. The voltage in the sine wave triggered the switch to open a solenoid valve connected to the air compressor. The valve remained open for the duration of the sine wave allowing us to control the duration of the airflow. The pressure in the compressor was set to 7 p.s.i., producing a gentle puff.
2.2.3 Procedure

65 native English speakers, mean age = 21.75 (SD = 4.92), 37 male, and 28 female, participated in the two-alternative forced choice visual-aerotactile perception task. Participants were recruited from the University of British Columbia and were compensated $5 for a fifteen-minute session. Each participant gave informed consent and completed a language background questionnaire. Participants reported no speech, hearing, or language difficulties.

Before the task began, participants were informed that they may feel air during some of the trials but were otherwise given no instructions regarding the air. All participants were tested in a sound-attenuated booth and instructed to keep their head and back against a high-backed chair. Participants were randomly assigned to one of four groups: 1) 16 to simulated face, 2) 17 to human “ba”, 3) 16 to human “pa”, and 4) 16 to no-puff simulated face. In each group, participants were presented with repetitions of one of the visual stimulus types (e.g., a human face silently
articulating “pa”) making visual stimulus type a between-subject manipulation. The Simulated Face no-puff group served as a comparison to the Simulated Face group. In the event that the participants in the Simulated Face group exhibited different response behavior in trials with airflow as compared to participants in the human face groups, the no-puff control would allow us to compare the results to a group of participants who had been presented with the animated visual stimulus without airflow at all.

Following Bicevskis et al. (2016), the video presentation was accompanied by multi-talker babble played through Direct Sound Ex-29 headphones from a second computer located outside the sound booth. The babble served to both mask the sound of the airflow exiting the tube and to make the task more natural. For all participants except those in the no-puff simulated face group, half of the videos were presented with a synchronized puff of air on the neck, and the other half were not, resulting in two within-subject conditions (puff vs. no puff). For each token, participants were asked to indicate on a keyboard which syllable (“pa” or “ba”) they felt the talker in the video had said. Trial order was randomized and the response keys were counterbalanced across participants. The experiments were run on an iMac computer using PsychoPy experimental presentation software (Pierce, 2009).

2.2.4 Analysis

To test our hypothesis, we ran descriptive tests to compare the means and standard deviations of responses for each of the four visual conditions (simulated face, human “ba”, and human “ba”, no-puff simulate face) and each condition (air puff and no air puff). Generalized linear mixed-effects models (GLMM) (Kuznetsova et al. 2017) were then run in R (R Core Team, 2016) on the interaction between visual stimulus type (simulated face, human “ba”, and human “pa”), condition
(air puff or absence of air puff), and the normalized centered trial order. We visualized significant 3-way interactions using ggplot2 (Wickham, 2016) to show linear fit patterns across the three relevant interactions.

Finally, we checked for time course effects on response times by using loess curves to visualize how quickly participants responded based on visual stimulus type. We then used generalized additive mixed-effects models (GAMM) (Wood, 2017) to test for statistically significant differences in response time over the course of the experiment across groups. The non-linear nature of GAMMs allowed us to see differences in response time independent of overall changes by identifying changes in window shape or envelope. In this way, we were able to more effectively look for differences in response times across the groups that may occur during only certain portions of the experiment (e.g., the first half of the trials).

2.3 Results

2.3.1 Integration

A summary of the results of all four experimental groups shows that there is considerable “ba” bias for tokens that were not paired with air puffs, and considerable “pa” bias for those paired with air puffs, as seen in Table 2.2. The percent /pa/ response ranges from 64% to 79% during visual + tactile trials. In contrast, the percent /pa/ response during visual-only trials is consistently below 40%, indicating that the presence or absence of air flow had an impact on behavioral responses.
Table 2.3 Mean and standard errors for percent /pa/ response for the simulated face, human “ba”, and human “pa” experiments by condition (puff vs. no puff).

<table>
<thead>
<tr>
<th>Visual Stimulus Type</th>
<th>Condition</th>
<th>Mean</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulated face</td>
<td>no puff</td>
<td>29.28%</td>
<td>5.49%</td>
</tr>
<tr>
<td></td>
<td>puff</td>
<td>79.16%</td>
<td>4.32%</td>
</tr>
<tr>
<td>Human “ba”</td>
<td>no puff</td>
<td>38.82%</td>
<td>5.49%</td>
</tr>
<tr>
<td></td>
<td>puff</td>
<td>64.87%</td>
<td>5.85%</td>
</tr>
<tr>
<td>Human “pa”</td>
<td>No puff</td>
<td>29.28%</td>
<td>4.9%</td>
</tr>
<tr>
<td></td>
<td>puff</td>
<td>74.64%</td>
<td>5.81%</td>
</tr>
<tr>
<td>No Puff Control</td>
<td>no puff</td>
<td>29.56%</td>
<td>5.63%</td>
</tr>
</tbody>
</table>

Participant responses to visual only (no puff) and visual + tactile (puff) stimuli are presented in Figure 2.3. The figure shows similar response behavior across the visual stimulus groups: for the simulated face and the two human faces, participants exhibit a response bias shift towards “pa” for the visual + tactile condition as compared to the visual only condition. Interestingly, participants in the Human “ba” group show greater inter-subject variability as well as a smaller increase in “pa” responses, as compared to participants in the Simulated Face and Human “pa” groups.
Figure 2.3 Percent /pa/ response for each visual stimulus type by condition.

To test for statistically significant differences in percent “pa” response, the data were fit to a generalized linear mixed-effects model with Condition and Visual Stimulus type as fixed effects, subject as a random effect, and a by-subject random slope for the interaction between Condition and Visual Stimulus Type. The formula is as follows:

\[
\text{Response} \sim \text{Condition} * \text{Visual Stimulus} * \text{Trial} + (1 + \text{Condition} * \text{Visual Stimulus} | \text{Subject})
\]

Where Response is a numerical value, 1 for “pa” and 0 for “ba”, Condition is one of audio or audio + tactile, Visual Stimulus is one of “simulated face”, “human ‘pa’”, or “human ‘ba’”, Trial is the
trial order from 1 to 70, centered and scaled to improve model fit performance, and \((1 + \text{Condition} * \text{Visual Stimulus} \mid \text{Subject})\) represents the random effects term by subject.

Figure 2.4 Loess visualization of mean percent “pa” responses by condition and trial order for all four groups of subjects. Within each group, percent “pa” response was calculated by averaging across all participants for each trial. Condition (puff vs. no puff) was randomized, so for each trial, some participants felt the sensation of airflow while others did not. The lines at 100% and 0% represent individual responses as the response data were numerically coded with 0 for “ba” and 1 for “pa”.

A summary of fixed effects can be seen in Table 2.3. Condition emerged as highly significant \((\beta = 0.51, SE = 0.09, t = 5.65, p < 0.001)\) such that all participants significantly were significantly more likely to respond “pa” on trials with airflow. There were no significant
interactions between Condition and Stimulus Type confirming that participants did not show significantly different response behavior across visual stimulus types. While there appears to be a smaller puff effect for the human “ba” group in Figure 2.3, the interaction between condition and visual stimulus type did not reach significance. Interestingly, the increased “pa” responses over time in trials with airflow for the Human “pa” group seen in Figure 2.4 emerged as significant, as shown by a significant three-way interaction between Trial, Condition, and the human “pa” visual stimulus ($\beta = 0.24$, $SE = 0.12$, $t = 2.10$, $p < 0.05$).

Table 2.4 Summary of fixed effects for the model predicting percent “pa” response.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.31</td>
<td>0.06</td>
<td>5.17</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Human “ba”</td>
<td>0.07</td>
<td>0.09</td>
<td>0.08</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Human “pa”</td>
<td>0.01</td>
<td>0.08</td>
<td>0.17</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ConditionPuff</td>
<td>0.51</td>
<td>0.09</td>
<td>5.65</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Trial</td>
<td>-0.04</td>
<td>0.06</td>
<td>-0.68</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba:ConditionPuff</td>
<td>-0.21</td>
<td>0.14</td>
<td>-1.49</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>pa:ConditionPuff</td>
<td>-0.16</td>
<td>0.12</td>
<td>-1.25</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba:Trial</td>
<td>0.05</td>
<td>0.08</td>
<td>0.57</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>pa:Trial</td>
<td>-0.03</td>
<td>0.08</td>
<td>-0.40</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ConditionPuff:Trial</td>
<td>-0.03</td>
<td>0.08</td>
<td>-0.41</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba:Puff:Trial</td>
<td>-0.05</td>
<td>0.11</td>
<td>-0.41</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>pa:Puff:Trial</td>
<td>0.24</td>
<td>0.12</td>
<td>2.10</td>
<td>p &lt; 0.05</td>
</tr>
</tbody>
</table>
2.3.2 Response Time

A loess-smoothed descriptive visualization of response times in relation to trial order for all three experiments can be seen in Figure 2.5\(^1\). There was little difference in response times across the Simulated Face, Human “ba”, and No Puff control groups. In contrast, the participants in the Human “pa” group were much slower during early trials in the experiment. For all groups, participants responded faster as the experiment wore on and the Human “pa” group response times appear to catch up to the other three groups by roughly a third of the way through the experiment.

A Generalized additive model (GAM) revealed no significant effect of condition (visual vs. visual + tactile) on response times; the model can be seen in the following formula:

\[
\text{Response Time} \sim \text{Visual Stimulus} + s(\text{Trial, } bs = \text{“cr”, } k = 10) + s(\text{Trial, } bs = \text{“cr”, } k = 10, \text{ by } = \text{Visual Stimulus}) + s(\text{Trial, subVisual Stimulus, } bs = \text{“fs”, } k = 10, m = 1)
\]

\(\text{Response Time}\) refers to how long participants took to indicate what they perceived by pressing a key. \(\text{Visual Stimulus}\) is one of “simulated face”, “human ‘pa’”, or “human ‘ba’”. The third term is a cubic regression spline-based smoothing formula for the main effect of Trial. The fourth term is a cubic regression spline-based smoothing formula for the interaction between Trial and Visual Stimulus. The final term is a full factorial smoothing term for Trial by Subject and Visual Stimulus.

---

\(^1\) It is important to note that this figure is for illustration only. The confidence intervals are based on local regressions estimates and are not intended for use as analysis.
Summary of the model results can be seen in Tables 2.4 and 2.5. The adjusted R-squared is 0.257 explaining 28.4% of the deviance.

Figure 2.5 Loess visualization of response time by trial order for simulated face, human “ba”, human “pa”, and no puff control groups. Response times for human “pa” were slower initially as compared to the other three visual stimulus types. In general, however, response times sped up over the course of the experiment.

Table 2.5 Summary of results for parametric coefficients.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.76</td>
<td>0.09</td>
<td>8.21</td>
<td>p &lt; 0.001</td>
</tr>
<tr>
<td>Human “ba”</td>
<td>-0.05</td>
<td>0.13</td>
<td>-0.38</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Human “pa”</td>
<td>0.21</td>
<td>0.13</td>
<td>1.56</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>NoPuff Control</td>
<td>0.04</td>
<td>0.13</td>
<td>0.29</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>
The results confirm that response times sped up throughout the experiment, as seen in the loess visualization in Figure 2.5. In addition, however, while there was no main effect of Visual Stimulus on response time, there was a significant interaction between Trial and Human “pa”. Specifically, participants in the human ‘pa’ group were initially slower at responding than participants in all other groups. This initial slowness was highly significant, even with outlier responses greater than 3 seconds removed from analysis (see Table 2.5). However, response times by trial for the Human “ba” and Control groups did not differ significantly from those of the Simulated Face group, confirming that response times were similar for all three groups across the duration of the experiment.

2.4 Discussion

The current study used a visuo-aerotactile integration task to investigate whether participants distinguish between a non-present yet human speech source and a non-present, non-human speech source during multimodal speech perception. To accomplish this, we tested whether perceivers integrate visual-aerotactile multimodal speech stimuli when the apparent source is not a real human. Our current results clearly demonstrate that perceivers integrate cross-modal speech
information similarly from human and synthesized faces. We predicted that participants would provide more “pa” responses when presented with both visual and aero-tactile stimuli than when presented with only the visual stimulus regardless of the visual source. Indeed, this was true: All participants responded “pa” significantly more often than “ba” in trials where they felt synchronous airflow. This held for both the participants who viewed a human face and those who viewed a simulated face. It is worth noting that, for both groups of participants who viewed the human face, there was an increased percentage of “pa” responses during trials that were accompanied by airflow, just as found in Bicevskis et al. (2016). These results offer an important validation for the animated face stimuli used in the main study because the data suggest that the aspiration effect is stable regardless of how many different articulations perceivers see. As noted above, human speech offers a great deal of natural variation (both visual and acoustic), while animated faces tend to have more simplified and less varied speech articulations. If the aspiration effect found in Bicevskis et al. (2016) had disappeared under the reduced variability in our human face controls, it would be difficult to justify our use of a single animated face. However, the control results demonstrate that normal integration processes are not interrupted by such simplification.

The aspiration effect seen in the animated face group suggests that perceivers do not require previous experience with the specific stimuli pairing in order to integrate; while everyone presumably has general experience integrating visual and aerotactile speech information, participants in the current study did not have prior exposure to the combination of speech-related airflow and an animated talking head. This interpretation is bolstered by the fact that we found no significant difference in response (i.e., “pa” or “ba”) across time between the participants who were presented with the simulated face and those who were presented with a human face. In other words, participants did not change their response strategy as the experiment progressed—
increasing exposure to the stimuli pairing did not affect participant response behavior. Taken together, these results demonstrate that participants integrated the cross-modal cues from the very beginning of the experiment and did not learn to pair the airflow and a simulated human face through experience with the specific stimuli. Such a finding aligns with others in demonstrating that cross-modal speech information need not be frequently experienced to be integrated (Fowler & Dekle 1991; Gick et al. 2008; Derrick & Gick 2013).

In addition, while response times grew faster on the whole as the experiment progressed regardless of condition, there was an interaction between trial and visual stimulus type. Specifically, participants in the human “pa” group were slower to respond at the start of the experiment than participants in any other group. While that difference in response time behavior disappeared roughly a third of the way through the experiment, the result is somewhat surprising. We would have predicted that the participants in the simulated face group would show the slower initial response times had any occurred given their lack of prior experience with the specific cues. The native English perceivers who took part in the current study have arguably the most prior experience with the combination of visual “pa” and airflow, suggesting that this result is not related to experience with the cues. It remains possible that some characteristic of the visual “pa” stimulus makes the articulation more difficult to discriminate than the other visual stimuli. Finally, while cause of this response time difference remains an open question, our finding provides additional support for a visual distinction between /b/ and /p/. While the bilabial stops have often been considered instances of a single viseme (Fisher 1968), more recent work has suggested that perceivers do pick up on subtle visual differences between the articulations (Abel et al. 2011, Bicevskis et al. 2016).
Our results fit well with previous studies suggesting that perceivers may not be overly concerned with having an ecologically-valid, localized source of the multimodal speech information. For example, studies have shown that the source of the visual information in an audio-visual task need not appear to coincide in spatial location with the source of the auditory information (Jones & Munhall 1996; Bertelson et al. 1994; Fisher & Pylyshyn 1994). Jones & Munhall (1996) found no reduction in the McGurk effect when the auditory stimulus was spatially dislocated from the visual stimulus by as much as 90 degrees. Such findings are surprising when one considers that stimuli originating from different locations are unlikely to come from a single natural source. It is further remarkable given that perceivers appear to care about other of aspects of ecology during perception. For example, while perceivers allow a surprising degree of cue asynchrony, the windows over which they will integrate audio, visual, and somatosensory speech information appear to be asymmetrical in directions that are consistent with the relative signal speeds in the physical world (Munhall et al 1996; Gick et al. 2010; Bicevskis et al 2016). In addition, perceivers do not integrate just any synchronous sensory input: cross-modal cues must possess a lawful relationship in the real world in order to be integrated (Fowler & Dekle 1991; Gick & Derrick 2009). But why these ecological factors—temporal congruence and signal relevance—play important roles in integration while spatial congruence and a human source do not remains a question for future research. Most perceivers—and certainly the undergraduate subjects tested in a metropolitan city such as Vancouver—have experienced spatially dislocated audio and visual information (both speech and non-speech) through movie theaters and surround sound speaker systems. Future work should consider how exposure to and experience with different forms of technology modulate perceptual sensitivity to spatial congruence.
Finally, while our findings constitute a valuable addition to the discussion surrounding perceivers’ sensitivity to realism and ecological validity in speech perception, some important gaps remain. First, while we and others have not found behavioral differences in studies comparing real and animated talking heads, there may exist neural or electrophysiological response differences to be discovered. Second, future research should further probe how realistic a source’s human-ness must be for a perceiver to integrate. In the introduction, we speculated that our participants viewing an animated face might not integrate airflow information paired with visual speech information on a monitor even though participants in Bicevskis et al. (2016) did. We posited that, while the participants in Bicevskis may have extended physical capabilities (i.e., speech airflow generation) to a human face, our participants might be less perceptually flexible when presented with a simulated human. As discussed above, we found no evidence of this. Still this does not provide definitive evidence that multisensory integration is automatic enough that perceivers ignore or are unaware of the source’s ecological validity. Our avatar’s very human-like features may have been realistic enough that our participants were willing to extend physical capabilities to it. This raises the question whether there exists a point along a spectrum of human-ness at which perceivers stop integrating multisensory speech information.
Chapter 3: Spatial Congruence

Considerable research outside the speech domain has suggested that both spatial and temporal congruence play important roles in multisensory perception (e.g., Soto-Faraco et al. 2003; Macaluso & Driver 2005). Yet, while the importance of timing in multimodal speech perception has been well demonstrated across different modality pairings (e.g., Munhall et al. 2006; van Wassenhove et al. 2007; Gick & Ikegami 2008; Bicevskis 2016), much less is currently known about the importance of spatial location with respect to speech information. If perceivers are expecting distal source of the cross-modal cues that can be localized within their immediate environment, spatial incongruence may interfere with integration during perception: stimuli coming from different directions are unlikely to originate at the same source. This is supported by evidence showing differences in neural response to multisensory non-speech stimuli coming from the same approximate location in space as compared to stimuli that do not (Macaluso et al., 2005; Teder-Sälejärvi et al. 2005). Such perceptual sensitivity to spatial characteristics of the signal would not be surprising given that cross-modal speech information generally coincides in space and time. In our day to day communication, we typically see and hear the person we are talking to from the same location in space. Moreover, the visual and acoustic information is largely synchronous. In other words, a perceiver’s prior perceptual experience consists primarily of spatially and temporally congruent speech cues.

Given this natural spatial and temporal congruence, we might expect the synchrony and spatial location of cues to influence integration in speech perception tasks. As I discussed in Chapter 1, temporal congruence has proven to meet this expectation for audiovisual, audio-aerotactile, and visual-aerotactile speech perception (Munhall et al. 1996; Gick, Ikegami, &
Derrick 2010; Bicevskis et al. 2016). The aforementioned studies have demonstrated that cross-modal cues must be co-presented within a certain temporal window in order for perceivers to integrate them into a unified percept. In contrast, several audiovisual speech perception studies have suggested that while perceivers are sensitive to temporal congruence, they are surprisingly insensitive to mismatches in spatial congruence (Jones & Munhall 1996; Bertelson et al. 1994; Fisher & Pylyshyn 1994), a point I will return to below. For example, Jones and Munhall (1997) tested participants on a McGurk task in which the auditory and visual stimuli were presented from different directions. The visual stimulus was presented on a single monitor directly in front of participants while the audio played from speakers located at 30-degree intervals in azimuth from the source of the visual cues. The authors found no significant effect of spatial dislocation on consonant identification in the audiovisual condition: participants showed a strong McGurk effect even when the auditory and visual stimuli were separated by as much as 90 degrees. On the basis of their findings, the authors conclude that “spatial incongruencies do not substantially influence the multimodal integration of speech signals.” (Jones & Munhall, 1997, p.13).

However, it remains possible that the results reported in Jones & Munhall (1997) are only true of audiovisual speech perception and do not pertain to multimodal speech more generally. Previous work on spatial congruence in multimodal speech perception (including Jones & Munhall 1997) has been limited to audiovisual paradigms in which the auditory stimulus direction is shifted and the visual stimulus remains constant. Humans are known to possess a bias toward the visual stimulus during sound localization such that they perceive the visual stimulus as the origin of the auditory information (e.g., Bertelson & Aschersleben 1998). This makes sense given visual information provides a more reliable cue to the spatial location of the signal’s origin than auditory information. This ventriloquism effect, as it is often called, results in a mislocation of sounds to
the apparent visual origin and could be driving the apparent spatial insensitivity found in Jones & Munhall’s results. Thus, it remains an open question whether we would find the same insensitivity to spatial information during a task with a non-visual modality pairing. Fortunately, the fact that aerotactile (airflow) cues influence speech perception (e.g., Gick & Derrick 2009) provides an alternative modality pairing—audio-aerotactile—that lends itself particularly well to the current question. Not only does the audio-aerotactile pairing allow us to avoid the potential confound of visual bias, but both the audio signal and the puff can be presented laterally, allowing for 180-degree dislocation of the cross-modal cues.

There is some evidence to suggest that aerotactile and auditory cues may not need to be spatially congruent for perceivers to integrate them during speech perception. For example, perceivers have been shown to integrate audio-aerotactile cross-modal cues when the puffs of air are felt on the hand, neck, and even ankle (Gick & Derrick 2009; Derrick & Gick 2013). However, it is important to note that while shifting the tactile stimulus location on the body away from where the auditory stimulus was presented did not eliminate the aspiration effect, it did affect integration: Participants were less influenced by air felt at the ankle as compared to the hand and neck (Derrick & Gick 2013). This effect difference could be interpreted as support for the hypothesis that localization does matter at least to some degree. However, receptor density varies across body locations making it difficult to know whether the effect difference found in the aforementioned study truly arose from dislocating the tactile and auditory stimuli in space or from physiological differences in perceptual sensitivity. Thus, in the current study we manipulated signal direction—a more controlled method than body location for investigating the effects of spatial incongruence on speech perception.
The audiovisual results discussed above lead toward the perhaps surprising prediction that perceivers do not attend to spatial alignment during speech perception, regardless of the modality pairing. Given this, we wanted to test the most extreme version of spatial dislocation possible (i.e., lateralization). The current study tests directly whether auditory and aerotactile speech cues presented co-directionally are more easily integrated than those presented contra-directionally by presenting audio and aerotactile speech information form opposing sides. We predict that participants will show the aspiration effect found in Gick & Derrick (2009) during trials with spatially incongruent stimuli just as in trials with spatially congruent stimuli. In other words, we predict no significant difference in response behavior across conditions.

3.1 Methods

3.1.1 Stimuli and Airflow Apparatus

Stimuli were created using eight tokens each of /ba/ and /pa/ (phonetically [pa] and [pʰa]) produced by a male native speaker of English; audio tokens were those used in the original Gick and Derrick (2009) study. Each token was modified for directionality using the Pan effect plugin in Audacity (Audacity, 2018) to create 3 acoustic signal conditions: Right, Left, and Binaural presentation. This consisted of shifting the acoustic signal into either just the left or the right track in the .wav file. The end result was a sound file that played the syllable in either the left ear or the right. For the binaural condition, the original stimuli were not modified.

To create the airflow stimulus, a 50-millisecond 23 kHz sine wave inserted into the left channel of the stereo track triggered the release of air from the compressor through a solenoid valve. The sine wave began 70 ms before the consonant burst such that the onset of the airflow and the release of the consonant closure were synchronous. In previous experiments (e.g., Gick &
Derrick 2009) participants heard only the right channel in both ears (i.e., the channel without the sine wave). However, in the current study, both channels were required to present the stimuli in one ear or the other. Thus, participants were necessarily presented with the channel containing the sine wave. To confirm that the 23 kHz sine wave was inaudible, a control group of 12 participants were presented with the audio stimuli and were asked after each trial whether the token had contained a tone. The participants were consistently unable to identify the tokens that contained a sine wave, confirming the viability of the stimuli for the main study.

The airflow apparatus consisted of a California Air Tools 4610 air compressor connected to a switch box via a ¼-inch diameter vinyl tube (all of which was located outside the sound booth). A second ¼-inch vinyl tube passed from the switch box through the access port of the sound booth and was attached to a flexible boom arm fitted to a microphone stand (see Chapter 2 for further details regarding the airflow set up). The open end of the tube was positioned ~7 cm away from the side of the participant’s neck. The position of the airflow tube (to the left or right of the neck) remained constant for each participant but was counterbalanced across participants.

3.1.2 Procedure

Twenty-four native English speakers (12 female, 12 male) were recruited through word of mouth from outside the Linguistics department and were compensated at the rate of $10/hr. for their participation. All participants reported no hearing or speaking difficulties. Participants were tested in a high-backed chair in the sound booth of the Interdisciplinary Speech Research laboratory at the University of British Columbia. They were instructed to keep their back against the chair during the study and were informed that they might feel airflow during some of the trials. Participants took part in a two-alternative forced choice task largely following Gick & Derrick (2009) followed
by a language background questionnaire. During the task, participants were presented with audio tokens of the syllables /ba/ and /pa/ (acoustically [pa] and [pʰa]) over headphones. To increase task difficulty, pink noise was played for the duration of the experiment. The signal-to-noise ratio was calibrated empirically with a separate control group of participants such that auditory discrimination accuracy reached about 70% on average. This resulted in a signal-to-noise ratio of -6 SNR.

As noted above, the participants heard each token three ways over the course of the experiment: in the left ear only, in the right ear only, or binaurally. On half of the trials, participants felt a gentle puff of air at the side of the neck. Of the trials with airflow, the airflow direction either matched the sound direction (“Congruent”), did not match (“Incongruent”), or occurred with the sound presented in both ears (“Binaural”). Participants were presented with each item four times resulting in 192 total trials. Immediately following each trial, participants were asked to indicate which syllable they had heard using buttons on their keyboard. The button response code was counterbalanced across participants.

3.2 Results
For all stimuli presentation conditions, we found a considerable “pa” bias for trials with airflow and a considerable “ba” bias on trials without airflow. The means and standard deviations of responses for each of the three stimuli presentation types (binaural, congruent, and incongruent) and each airflow condition (air puff and no air puff) are summarized in Table 1. Participants responded with “pa” 25-30% more on average during trials with airflow. This demonstrates that the presence of the aerotactile information greatly influenced response behavior as predicted. More to the point, this increase in “pa” responses occurs across all three stimuli presentation conditions:
participants were influence by the airflow regardless of whether the auditory and aerotactile stimuli were spatially congruent, spatially incongruent, or presented binaurally. As Figure 3.1 shows, the participants exhibited a smaller increase in “pa” responses in the incongruent condition, which would be in keeping with the decreased ankle effect in Derrick & Gick (2013).

Table 3.1 Means and standard errors for percent “pa” response with and without airflow for the three stimuli presentation conditions (i.e., Congruent, Incongruent, and Binaural).

<table>
<thead>
<tr>
<th>Stimuli Presentation Condition</th>
<th>Airflow Condition</th>
<th>Mean % “pa” response</th>
<th>Std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Congruent</td>
<td>Puff</td>
<td>57%</td>
<td>1.8%</td>
</tr>
<tr>
<td></td>
<td>No Puff</td>
<td>32%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Incongruent</td>
<td>Puff</td>
<td>54%</td>
<td>1.8%</td>
</tr>
<tr>
<td></td>
<td>No Puff</td>
<td>29%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Binaural</td>
<td>Puff</td>
<td>57%</td>
<td>1.5%</td>
</tr>
<tr>
<td></td>
<td>No Puff</td>
<td>26%</td>
<td>1.8%</td>
</tr>
</tbody>
</table>
Figure 3.1 Percent “pa” response by Condition (puff vs. no puff) for Binaural, Congruent, and Incongruent stimulus presentation conditions. In all three conditions, participants gave significantly more “pa” responses during trials with airflow as compared to trials without airflow.

To test for statistically significance differences in percent “pa” response across conditions, a generalized linear mixed-effects model was run with stimulus presentation (i.e., binaural, congruent, and incongruent) and airflow condition (i.e., air puff and no air puff) as fixed effects, a random effect of subject and a by-subject random slope for stimulus presentation and condition. The formula is as follows:

\[
\text{Response} \sim \text{Condition} + \text{Stimulus Presentation} + (1 + \text{Stimulus Presentation} + \text{Condition} | \text{subject})
\]
Where \( \text{Response} \) is 0 for “ba” and 1 for “pa; \( \text{Stimulus Presentation} \) is one of the following three trial types “Congruent”, “Incongruent”, and “Binaural”; \( \text{Condition} \) is “Puff” and “No Puff”, and \( (1 + \text{Stimulus Presentation} + \text{Condition} \mid \text{subject}) \) represents the random effects term by subject. Condition emerged as the only significant factor (\( \beta = 1.43, \ SE = 0.22, \ t = 6.42, \ p < 0.001 \)). We found no effect of Stimulus Presentation and no interaction between Stimulus Presentation and Condition, as can be seen in the summary of fixed effects in Table 3.2. Thus, while the mean “pa” response decreased when the aerotactile and auditory cues were incongruent, this difference did not turn out to be statistically significant.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-1.01</td>
<td>0.23</td>
<td>-4.37</td>
<td>( p &lt; 0.001 ) ***</td>
</tr>
<tr>
<td>Incongruent</td>
<td>-0.13</td>
<td>0.11</td>
<td>-1.18</td>
<td>( p &gt; 0.1 )</td>
</tr>
<tr>
<td>Binaural</td>
<td>-0.08</td>
<td>0.13</td>
<td>-0.66</td>
<td>( p &gt; 0.1 )</td>
</tr>
<tr>
<td>ConditionPuff</td>
<td>1.34</td>
<td>0.24</td>
<td>5.54</td>
<td>( p &lt; 0.001 ) ***</td>
</tr>
</tbody>
</table>

### 3.3 Discussion

The current study tested whether audio-aerotactile integration of speech occurs with spatially incongruent stimuli. As predicted, participants were significantly more likely to report having heard /pa/ when airflow was co-presented. We did not find evidence that spatial congruence significantly interferes with the integration of audio and aerotactile speech information. Participants showed statistically similar response behavior across conditions: syllables accompanied by synchronous cutaneous airflow were significantly more likely to be judged as
aspirated than those without airflow, regardless of whether the cues coincided in source direction. These findings extend the results from the audiovisual literature detailed above (Jones & Munhall 1996; Bertelson et al. 1994; Fisher & Pylyshyn 1994) to a non-visual modality pairing. During speech perception, participants do not appear to use spatial cues to determine if information from different modalities belong to the same event.

While it remains possible that spatial incongruence could interfere with integration if the spatial information were task-relevant, research on auditory-somatosensory integration outside speech perception does not support this. In Sperdin, Cappe, and Murray (2010), participants exhibited faster response times during both incongruent and congruent multisensory trials, but they struggled to identify the spatial origin of a given unisensory (audio or tactile) cue following a multisensory trial. This effect persisted even when participants were explicitly informed that they would have to answer questions about which direction the information had come from following some trials. Thus, task-relevance did not affect participants’ sensitivity to the individual origins of a cross-modal cues. The authors interpret their findings to suggest that spatial information about unisensory cues is not merely irrelevant to perceivers following multisensory trials, but unavailable.

Finally, our findings offer more than a confirmation of Jones and Munhall’s assertion that spatially dislocating speech signals does not interfere with integration. Our results, when considered in conjunction with the findings that perceivers integrate speech-related airflow felt at a variety of locations across the body (Gick & Derrick 2009; Derrick & 2013), point to a more holistic perception of speech-related sensory information with no specific real-world experience required. It is unlikely that perceivers have significant experience with speech airflow felt at the ankle. More to the point, and noted in the introduction, it is highly unlikely that perceivers have
experience with co-occurring speech information coming from opposing directions. Given this, it seems unlikely that perceivers could have learned to integrate spatially dislocated cues through experience with specific stimuli. It thus appears that perceivers integrate cross-modal speech information regardless of specific perceptual experience with the cue combination, even when aspects of the incoming stimuli diverge from the natural structure of events in the perceiver’s world.
Chapter 4: Pre-linguistic Infants

The experiment in this chapter investigates the role of prior experience with one’s own production in the ability to integrate aerotactile information during perception. To accomplish this, I test pre-linguistic infants in an audio-aerotactile task. In the previous chapter, I showed that adult perceivers integrate cross-modal cues without previous experience with the specific stimuli. However, because our perceptual abilities change across development, it remains possible that the ability to integrate novel pairings as an adult perceiver is a result of early specific experience. One possible avenue through which infants could become capable of integrating speech-related airflow would be experience as producers of a language featuring contrastive aspiration.

Though no direct evidence exists to suggest that production experience plays a central role in the emergence of aerotactile integration, it makes some intuitive sense that feeling one’s airflow while simultaneously feeling one’s articulators and hearing the acoustic output could provide a way to build a multimodal representation of sounds. In fact, some computational models of speech production and acquisition (e.g., the DIVA model Guenther, 1994; Guenther, 1995; Tourville & Guenther 2011) predict that production experience during an initial babbling stage could be where infants learn initial mappings between articulatory movements and their acoustic and sensory consequences. The DIVA model, an adaptive neural network that describes the sensorimotor interactions inherent in articulatory control, undergoes a “babbling phase” designed to roughly mirror the linguistic experience of a typically-developing infant before the model begins producing speech. During this phase, the model’s control system generates pseudo-random articulations through the simulated vocal tract that provide auditory, tactile, and
proprioceptive feedback signals. These feedback signals allow the system to create two mappings: one is phonetic-to-orosensory, and the other is orosensory-to-articulatory. The first mapping can be thought to link a sound with the vocal tract target that produces it, while the second mapping links the vocal tract movement with the articulation needed to produce that sound. In this model, then, production experience plays a central role in linking sensory outputs and the vocal tract configurations or movements that generate them. What does this mean for non-simulated language learners? This would suggest that an infant’s experience with self-generated feedback during speech production may allow them to integrate tactile and proprioceptive speech cues with auditory cues. More to the point, given that airflow is an integral part of speech generation, speech-related aerotactile somatosensation should arguably be mapped during the babbling stage as well. If so, infants without prior babbling experience may not have the capability to integrate the acoustic and airflow outputs produced during an aspirated stop.

Research on speech perception in disordered populations offers additional evidence of a possible causal link between production experience and our perception of sounds. In children with phonological disorders, for example, there is some evidence that difficulty producing a sound may negatively affect the ability to perceive it. In other words, if children do not have experience accurately producing a sound, they may have a harder time identifying it. As explained in Byun (2012), the misarticulations of a child who has a phonological disorder become a large part of the child’s input. If the child accepts these incorrect productions as instances of a target phoneme, it could shift the boundaries of that phoneme and thus make it more difficult for the child to perceive. Furthermore, there is some evidence that production training can improve perceptual sensitivity. Shuster (1998) found that for children with disordered production and perception of /ɹ/, therapy targeting their production of the sound...
resulted in significantly improved performance in a sound judgement task. Researchers have also showed that even relatively short periods of motor-training can facilitate speech sound recognition in articulator-specific ways for typically-developing adults (Glenberg et al. 2008; Sato et al 2011).

Finally, atypical production experience may interfere with typical audiovisual integration. Dejardins et al. (1997) found that preschoolers who exhibit production errors are less influenced by the visual information in an audiovisual speech perception task. The results, as the authors argue, suggest young children with speech production issues may not have as strongly developed visual speech representations because of their difficulties with production. This may also be true about links between sensorimotor and acoustic speech information. Seidl et al (2018) tested the effects of sensorimotor feedback during perception on future productions of a novel word for two groups of 4-year-old English-speaking children: children with speech sound disorders and typically-developing children. The authors found that only the typically-developing group altered their subsequent productions in response to articulatory restriction. As the authors note, this suggests that children with speech sound disorders may be impaired in their ability to integrate somatosensory feedback during perception, potentially leading to difficulties linking perceptual representations and articulatory configurations.

Yet, infants also appear to integrate auditory, visual, and some sensorimotor input before they become language users, leaving open the possibility that infants can also integrate aerotactile speech information. In the following section, I will turn to recent findings in infant multi-sensory speech perception to argue that whether infants can integrate all speech-related sensory information before becoming language-users remains an open question.
4.1 Multi-sensory Speech Perception in Preverbal Infants

There is considerable evidence in the literature that infants can integrate sensory cues from different modalities in many contexts well before they begin speaking. More and more, research points to speech perception as a multisensory process from birth. There is even some evidence that neonates are able to match a voice to a talking face when presented with either a neutral expression or an image of a face making speech sounds (Aldridge, Braga, Waltong & Bower 1999). Bahrick, Licklitter, and Flom (2004) propose that amodal information—that is, information that is redundantly present across modalities—may recruit infant selective attention and make bi-modally specified perceptual events easier for infants to process and learn from. In other words, early sensitivity to amodal stimulus attributes such as rhythm, synchrony, co-location, and rate of motion may support infants in detecting meaningful, unified events out of moving articulators and the concomitant sound of speech.

Though it remains unclear to what extent neonates are able to match stimuli on the basis of phonetic information, infants are attuned to the congruence of modality-specific phonetic information by around 18 weeks. For example, infants begin matching visual and auditory phonetic information at 2-4 months, looking longer to the talking face that is consistent with the sound they are hearing than to the talking face producing a different sound (Kuhl & Meltzoff, 1982; Patterson & Werker 2003; Patterson & Werker 1999). A few studies have reported McGurk-like effects in infants as young as five months (Rosenblum, Schmuckler, & Johnson, 1997; Burnham & Dodd, 2004), though the effect seems to be weaker than in adults (Dejardins
& Werker, 2004). This may provide evidence that some visual experience with speech may be necessary to move beyond synchrony detection to integrating phonetic cues across modalities. However, it is also important to note that prelinguistic infants have been shown to detect phonetic incongruence in audiovisual speech for sounds they have no experience producing or perceiving. This sensitivity occurs even when the cross-modal information is synchronized and infants cannot rely on amodal attributes to match the cross-modal information. As noted in Chapter 1, Danielson et al (2017) used a preferential looking paradigm to test 6, 9, and 11 month-old infants on their sensitivity to phonetic incongruence in non-native audiovisual speech. They found that the 6- and 9-month old infants (but not the 11 month-olds who had already undergone perceptual attunement to their native language) detected the incompatibility between the visual speech articulations and the acoustic signal during incongruent trials. Because the authors used non-native sounds, the infants could not have been using previous experience with the specific sound-sight pairing during the task and their results provide a strong argument against an explicit learning of associations between specific sound-sight pairings.

The literature suggests that infants as young as 2 months can make use of sensorimotor information during speech perception such that they are influenced by feedback from their own oral tract while they are listening to speech sounds (e.g., Yeung & Werker, 2013; Bruderer et al., 2015; Choi, Bruderer, & Werker, 2019; Coulon et al 2013; Kuhl & Meltzoff 1982, 1984). For example, Yeung & Werker found that manipulating an infant’s lip configuration can influence her audiovisual perception. The authors presented 4.5-month-old infants with two simultaneous videos of a talking face producing a vowel (either /i/ or /u/) and an audio track playing one of the preceding vowels. During the test trials, the caregiver held either a soother (to mimic the lip rounding present in /u/) or a flat teether (to mimic the lip spreading in /i/) in the infant’s mouth.
The authors found a contrast effect: the infants whose lip movements matched the heard vowel looked preferentially at the video of a talking face whose articulation did not match the heard vowel.

Sensorimotor information also modulates speech perception in the absence of visual speech information. Bruderer et al (2015) demonstrated that interfering with an infant’s ability to make a sound’s articulatory movements affects the infant’s ability to discriminate that sound. In the task, 6 month old infants were presented with a minimal pair of non-native stops that differed only in the tongue shape necessary to produce them. Infants at this age are typically able to discriminate these sounds (dental [d] vs. retroflex [ɖ]) even though the sounds are not present in their language environment. In one condition, the caregiver held a flat teether in the infant’s mouth; the teether depressed the infant’s tongue such that the curled tongue posture used to articulate retroflex [ɖ] was impossible to achieve. In the control condition, the caregiver used a gummy teether that did not interfere with the tongue position. Unlike the control group infants, the infants in the flat teether group whose tongue movements were impeded were unable to discriminate the non-native contrast. Moreover, the interference occurs even when infants have no prior auditory experience with the sounds. In a follow-up study, Choi, Bruderer and Werker (2019) replicated the findings and further showed that sound-relevant motor perturbation has a similar effect on familiar speech sounds. Using the same gummy teether from Bruderer et al. (2015) that did not interfere with dental-retroflex discrimination, the authors tested English-learning babies on a phonetic distinction present in their native language: bilabial /b/ and alveolar /d/. The shape of the gummy teether prevented the infants from making a bilabial closure—much like the flat teether in Bruderer et al. prevented the infants from making the curled tongue posture. The authors found, as predicted, that interfering with the bilabial closure interfered with
bilabial discrimination. The combined results demonstrate that specific, sound-relevant motor perturbation interferes with perception independent of experience producing or perceiving that sound.

Taken together, much of the forgoing evidence suggests that even for pre-linguistic infants, visual, auditory, and sensorimotor feedback interact and can modulate speech perception without specific experience with the unimodal cues. In a recent commentary, Choi and colleagues make a compelling argument that prenatal and early postnatal experience with the oral tract could account for this surprising ability. Well before birth, fetuses spontaneously produce a variety of oral motor articulations, including tongue protrusion and retraction, and it has been suggested that these movements play an important role in both the development of the aerodigestive system and the initial organization of the somatosensory and motor cortices (Keven & Akins 2017). Choi et al. (2017) argue that these early movements provide information about the boundaries and capabilities of the upper vocal tract (e.g., size, shape, and possible configurations) and create a mapping onto which infants can project auditory and visual speech information during perception. At first, their proposal would appear to support the possibility of audio-aerotactile integration before infants become producers of their native language. If prelinguistic infants are indeed influenced by feedback from their articulators, it is not a great leap to predict that they would also be influenced by the airflow those articulators produced. Yet, we have no a priori reason to assume that the influence of aerotactile speech information emerges at the same time or through the same mechanisms. More to the point, the prenatal oral-motor movements that underlie Choi et al.’s proposal are lingual and do not involve the glottis. Second, these movements do not produce airflow. In fact, aerotactile somatosensation is a sensory input that infants have no experience with at birth because of the very nature of their environment in utero.
Thus, we cannot assume that the ability to integrate most speech-related sensory information without specific experience means an ability to integrate all speech-related sensory information.

To date, no research has focused on aero-tactile integration in infant speech perception. In the current chapter, I investigate the role of prior experience with self-generated multisensory speech cues—specifically, speech-related airflow generated during the production of aspirated stops. Production is of course not the only avenue through which infants and adult perceivers can experience speech-related airflow. Infants may be learning from the speech of others. This is a plausible explanation given that many infants spend much of their early lives being closely held by their caregivers, who coo and sing and talk to their child. The infant, then, is likely experiencing speech-related airflow that is temporally and spatially aligned with the co-occurring auditory and visual speech information emanating from her caregivers’ mouths. However, it is also likely that the infant only sometimes (and not always) feels a puff of air when they hear an aspirated sound. Thus, another speaker’s speech may not the most reliable method of learning the relationship between the signals. One’s own speech, if one produces aspiration, might offer a more dependable option.

The current study seeks to address the question regarding the developmental trajectory of audio-aerotactile integration with an experiment testing 6.5-8-month-old English-acquiring infants on the ability to use aero-tactile cues to distinguish between /pa/ and /ba/ in difficult listening conditions. Building on the fact that infants have been shown to distinguish between /p/ and /b/ in normal listening conditions at 6-8 months of age (e.g., Eimas et al., 1971; though see Burns et al. 2007 for evidence that infants may not discriminate the English boundary until after 8 months), the current study aims to see whether infants can integrate an additional cue (in this case, gentle puffs of air on the neck) to help them discriminate ambiguous stimuli without having
had previous experience with producing the cue. While infants at this age may have begun producing bilabial stops during babbling, research suggests that these stops are mostly voiceless unaspirated sounds with a short voicing lag even for infants as old as 12 months (Whalen, Levitt, & Goldstein 2009; Enstrom 1982). In fact, there is evidence to suggest that VOT may not be produced categorically until closer to two or even three years of age (Hitchcock & Koenig, 2013). Thus, it is unlikely that the 6-8-month-old infants in the current study would have experience feeling airflow across their own lips while producing [pʰ].

Given this developmental trajectory, 6-8-month-old infants may lack the relevant aspiration experience needed to integrate aerotactile speech cues. A hypothesis in which integration emerges through productive experience predicts that preverbal infants lack access to airflow as a speech cue during perception. If such a hypothesis is true, the infants in the current study should not be influenced by speech-related cutaneous airflow during the task and respond similarly to the control group. In contrast, if the hypothesis is false, preverbal infants should show an effect of aspiration similar to that of the adult participants from Gick & Derrick (2009). That is, the infants in the current study should treat unaspirated tokens (i.e., /ba/) accompanied by a puff of air as more like an aspirated token (i.e., /pa/).

For the aspirated tokens, the prediction is less clear. The airflow on syllables with aspirated onsets could theoretically serve as a redundant cue to information already present in the acoustic signal (i.e., the aperiodic noise of the aspiration). Indeed, for adults, the airflow increased accurate identification in Gick & Derrick (2009). Based on this, we would predict that the infants would treat a /pa/ accompanied by an air puff as roughly equivalent to a plain /pa/. However, given that infants in this age range are still in the process of narrowing their native phonetic categories, they may not be judging the tokens on the basis of language-specific phonetic
distinctions. If this is the case, then the infants may treat a /pa/ accompanied by a puff of air as a separate category altogether (i.e., something like an extra-aspirated /p/).

Regardless, if our results reveal that the infants’ perception is not influenced by the airflow, this will provide evidence that production experience plays a role in audio-aerotactile integration. It would also raise the question of why some multisensory integration in speech perception requires production experience while others do not (e.g., audio-visual speech perception). On the other hand, if the infants do treat the unaspirated tokens as aspirated when they feel the puff of air, a production-based hypothesis would not be supported and additional mechanisms must be investigated.

4.2 Methods

In the current study, the infants took part in a modified alternating/non-alternating sound presentation task (Best & Jones 1998; Yeung & Werker 2009). In this type of paradigm, infants are exposed to two different types of trials: an alternating trial, in which repetitions of two different sounds are presented (e.g., /ba/ and /pa/), and a non-alternating trial, in which repetitions of identical sounds are heard (e.g., /pa/ and /pa/). Often this paradigm employs a familiarization phase. However, the aim of the study was to test the infant's baseline ability to use aero-tactile information to discriminate between two sounds, rather than their ability to learn to use the cue. This is crucial if the question at hand concerns whether the infants are currently able to integrate aero-tactile information during speech perception and not whether they can be taught to use it. Because of this, the choice was made not to employ a familiarization phase. Instead, the infants only experienced a series of test trials, as in Bruderer et al. (2015). In alternating/non-alternating paradigms, infants are assumed to have discriminated if they look
longer to one type of trial than the other. Generally, in experiments without a familiarization phase, infants will look longer to an alternating stimulus. Therefore, in the current study, the infants are predicted to look longer to trials that they experience as alternating. As will be discussed further in Section 2.3 below, which trials they experience as alternating will depend on whether the infants are integrating the aero-tactile information.

4.2.1 Participants

48 English-acquiring infants (24 female; mean age = 7 m, 11 d; age range = 6 m, 15 d - 8 m, 4 d) were recruited from a database of families who had been approached at a local maternity hospital shortly after birth and had indicated their interest in participating in studies. As measured through parent reporting, all infants were exposed to a minimum of 80% English and had not been diagnosed with any developmental disorders. Data from an additional 34 infants were not included due to fussiness (n=25), not meeting the language criteria (n=3), and equipment error (n=6). Before beginning the session, caregivers were informed about the study procedure and gave written consent for participation. After the task, parents completed a language background questionnaire. At the end of the session, infants were given a t-shirt and a certificate as a token of appreciation for participating.

4.2.2 Apparatus and Set up

Following Gick and Derrick (2009), an air compressor attached to a solenoid valve in a switchbox comprised the airflow device. The air puffs were delivered at ~ 6 p.s.i. through 1/4-inch vinyl tubing that passed through a cable port from the observation room to the study room. The tube then attached to the front of a flexible plastic bib around the infant's neck. This kept the
mouth of the tube a constant 7 cm from the infant's neck and ensured that the airflow hit the infant's neck each time. The bib and tube were covered with fabric to keep the infant from grabbing or move the tubing (see Figure 4.1 below). In addition, a custom sound-attenuating cloak attached to the high chair ran from the floor to just under the infant's chin. In effect, this created a separate acoustic space in which the airflow occurred, thus ensuring that the infants only experienced the airflow as a tactile sensation. Infants were excluded from analysis if at any point during the experiment the bib and tube came out from under the cloak as it could no longer be guaranteed that the infant was not hearing the air puff.

Figure 4.1 Photographs of the bib worn around the baby’s neck and of the sound-attenuating smock. The vinyl tubing that delivered the airflow was located underneath the fabric and attached to the front of the bib. The tubing was then curved to aim the airflow at the baby’s neck from a distance of 7 cm.

4.2.3 Stimuli

The auditory stimuli for each trial were created in a sound editing program (Audacity Team, 2016) by concatenating naturally produced tokens of /ba/ and /pa/ (six tokens each) that had been produced by a male native English speaker for the original Gick and Derrick (2009) study. The
/ba/ tokens were phonetically voiceless unaspirated stops (i.e., [p]) with VOTs less than 10 ms. The /pa/ tokens were phonetically voiceless aspirated stops (i.e., [pʰ]) with an average VOT of 60 ms. Twelve 20-second stimuli streams were created: six non-alternating (NonAlt) stimuli streams contained 12 presentations each of either /pa/ or /ba/ tokens at an ISI of 750 ms, and six alternating (Alt) stimuli streams contained six presentations each of /pa/ and /ba/ tokens at the same ISI. The tokens were placed in the right channel of a stereo track which was subsequently embedded in pink noise at +2 SNR. The noise was included both to reduce the ability to discriminate a native contrast they have been shown to be able to discriminate and thus reduce the risk of ceiling effects, and to mask in part the noise of the airflow. In the left channel, 50-ms sine waves generated at a frequency of 10kHz triggered the release of the airflow. The waves were time-aligned with the syllables such that, after adjusting for system latency, the air puff exited the tube at the same time as the stop burst 50-ms (see Figure 4.2). This resulted in around 65 ms of aspiration and was done to mimic the natural timing of aspiration during English stops.
Figure 4.2 Placement of the sine wave relative to the stop burst and vowel onset. The 50ms tone was shifted an additional 20ms earlier to account for system latency.

In total, four stimuli stream types were created. Every other syllable was accompanied by a puff such that, as far as the tactile modality is concerned, all trials were alternating. As discussed above, the crucial difference across trials, then, was whether the tactile stimulus reinforced an existing phonological distinction or interfered with it, thereby influencing the infant's perception of whether the trial stimulus alternated. Table 1 below shows the four types of stimuli streams created, the trial type (Alt or NonAlt) predicted if the infants are only processing the acoustic signal, and the trial type (Alt or NonAlt) predicted if the infants are integrating the aerotactile stimulus as part of the speech event.
Table 1. The table below describes the four types of stimuli streams that the infants were presented with, as well as the predicted percept (i.e., an alternating or a non-alternating trial) depending on whether the infants integrate the aerotactile stimulus.

<table>
<thead>
<tr>
<th>Stimuli stream</th>
<th>Predicted Trial Type</th>
<th>Predicted Trial Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Without Integration</td>
<td>With Integration</td>
</tr>
<tr>
<td>$paPuff + pa$</td>
<td>NonAlt</td>
<td>NonAlt</td>
</tr>
<tr>
<td>$paPuff + ba$</td>
<td>Alt</td>
<td>Alt</td>
</tr>
<tr>
<td>$pa + baPuff'$</td>
<td>Alt</td>
<td>NonAlt</td>
</tr>
<tr>
<td>$ba + baPuff'$</td>
<td>NonAlt</td>
<td>Alt</td>
</tr>
</tbody>
</table>

4.2.4 Procedure

The infants were randomly assigned to one of two groups: the experimental group and the control group ($n = 24$ for each). The infants in the experimental group were presented with synchronous puffs of air during the trials, as described above. In contrast, the infants in the control group never felt airflow during the experiment; the end of the tube that delivered the airflow was covered such that the air released into a fabric pocket, rather than against the infant’s skin. Otherwise, the infants in the two groups performed identical tasks.

The infants were tested in a quiet, dimly lit room while seated in a high chair in front of a computer monitor that was positioned in the center of a black curtain. Caregivers, who were seated in a chair next to the infant, were asked not to point or talk during the session. Non-verbal reassurance, such as nodding or smiling, was encouraged to keep the infants calm. In addition, caregivers listened to music over headphones during the session to ensure they didn't
unconsciously influence their child's reactions. A closed-circuit camera recorded the infant's face through a slit in the curtain directly below the computer monitor. From another room, an experimenter monitored the infant's face through the video display and controlled the stimulus presentation using computer software (Cohen et al., 1993). The auditory stimuli were presented free field at ~ 65dB over a speaker located behind the black curtain.

The study began with a silent, colorful animation to attract the infant's attention. This was followed by a silent checkerboard trial to give the infants an opportunity to look at the novel stimulus before the test trials begin. Before the onset of each trial, the infant's attention was drawn to the monitor by a spinning waterwheel. Once the infant was looking at the screen, the test trial began and a red and black checkerboard and the stimuli stream were presented simultaneously. When the trial finished, the same waterwheel animation reappeared to return the infant's attention to the screen. The study consisted of 16 trials over two blocks with a 10 second animated video as a break in between blocks. The two blocks were identical except in the ordering of stimulus presentation. In each block, the infants were presented with four Alt trials and four NonAlt trials (in which the Alt or NonAlt designation reflects the predicted trial type if the infants are integrating the aero-tactile information as outlined in Table 1 above), with every other trial being alternating. The order of the first stimulus was counterbalanced across infants, such that half of the infants experienced an alternating stimuli stream first and half experienced a non-alternating stream. All infants experienced all stimulus stream types. The order of presentation for the aerotactile stimulus was also counterbalanced. As mentioned previously, an air puff was present on every other syllable in each trial. To control for potential order effects, half of the infants were presented with the air puff on the odd (first, third, fifth, etc) syllables, while the other half were presented with the air puff on the even (second, fourth, sixth, etc)
syllables. In both counterbalancing cases, the order of presentation was then reversed for the second block (see Appendix A for further details).

The video recordings were converted to QuickTime movies. Total looking time and duration of the first look to the checkerboard served as the dependent measures and looking time to trials was coded offline frame by frame by the author and two undergraduate research assistants. The author and the first undergraduate assistant each coded half of the experimental trials, with the undergraduate coding an additional 20% of the trials coded by the author. The two undergraduates each coded half of the experimental group trials with the first undergraduate coding an additional 20% of the trials coded by the second undergraduate. The trials were coded with the audio muted so that the coders were blind to which type of trial the infant was experiencing. Infants who were unable to complete at least 8 of the 16 trials (e.g., did not encounter each stimulus stream at least two times) were excluded from analysis.

4.2.5 Analysis and Predictions

Total looking time to the checkerboard and duration of first look were analyzed across 4 trials of each stimulus stream for a total of 16 possible trials. Both measures of infant attention are widely used in the infant literature (see Aslin 2009 for an excellent discussion of looking time measures as a technique for investigating infant cognition), though it has been suggested that an infant’s first look to the stimulus provides a more sensitive measure of infant attention across longer time periods as it gives an indication of their initial interest in the stimulus before they become distracted (McCall 1971). Some previous studies using similar paradigms have used total looking time to checkerboard (Yeung & Werker 2009; Bruderer et al. 2015), while other studies have opted to analyze First Look (Hayes & Slater 2008; Jusczyk et al 1999). Given the novel
procedure I have used here and the exploratory nature of the study, I have included both measures of infant attention.

The data were analyzed in two ways. First, following previous literature (Yeung & Werker 2008; Bruderer et al. 2015), the trials were analyzed by Trial Type (e.g., Alt vs. NonAlt). This is the traditional way of analyzing looking time data collected using an alternating/non-alternating sound presentation task because we are measuring changes in looking behavior from one type of trial to the next. However, it is important to keep in mind that the trials in the current study only followed an alternating/non-alternating sequence on the basis of an integrated percept, not on the basis of the acoustic signal. Thus, analyzing by trial type here requires collapsing the four stimulus streams into alternating and non-alternating trial types in a manner that presupposes integration.

Integration was inferred if the infants showed significantly longer looking times to one type of trial (alternating or non-alternating). However, previous studies have taken longer looking time to either type of trial as an indication that the infants are discriminating the two sequences and thus, showing increased interest in one type (see Best & Jones 1998; Bruderer et al. 2015 for longer looks to alternating trials, and Maye, Werker, & Gerken 2002; Yeung & Werker 2009; Choi et al. 2019 for longer looks to non-alternating trials). Generally, in Alt/NonAlt paradigms without a familiarization, infants look preferentially to the alternating streams. Given that the trial types assume an integrated percept, I predicted that only the infants in the experimental group would show a difference in looking behavior between trial types. The infants in the control group, who did not feel the airflow during the trials, should not perceive the trials along these lines and thus should not show a difference in looking behavior between trial types. With this in mind, I make the following predictions regarding the infants in this study:
1) Infants in the experimental group will show longer total looking times and longer initial looks to the monitor for Alt trials (i.e., ba + paPuff, baPuff + ba) as compared to NonAlt trials (i.e., baPuff + pa, pa + paPuff)

2) Infants in the control group will show no difference in either total looking times or initial look duration between trial types

Second, because the current paradigm is more complex than that used in previous work (e.g., Bruderer et al 2015), an additional analysis was required. Recall that while the infants were presented with two predicted types of trials (Alt vs. NonAlt), there were four distinct stimulus streams. Therefore, the assumptions in the first analysis, namely that the infants were treating both predicted alternating (or non-alternating) streams equivalently, could be problematic. Thus, I also analyzed the data using stimulus stream as an independent variable instead of collapsing the streams into predicted types. In this analysis, there are three important comparisons to make. Below, I list them with the associated predictions.

1) **ba + paPuff vs pa + paPuff**: The streams in this first comparison should be treated as alternating and non-alternating, respectively, by both groups of infants regardless of integration. Thus, for both experimental and control groups, I predict that infants will look longer to ba + paPuff.

2) **ba + paPuff vs. baPuff + pa**: This comparison is important because the streams are predicted to be different only if the infants in the experimental group are integrating the airflow; the streams are acoustically identical, but the location of the puff of air (if
integrated) is predicted to shift the percept toward an aspirated token for the experimental group. Therefore, if the infants in the experimental group do integrate the airflow, then I predict that the infants will look longer to ba + paPuff than to baPuff + pa. In contrast, I predict that the infants in the control group who are not feeling the airflow will not show differences in looking behavior between the two. However, if the experimental group infants do not integrate, then both groups are predicted to treat the two stimulus streams as roughly equivalent (i.e., both /ba/ + /pa/) and thus show no significant difference in looking times between stimulus streams.

3) **ba + baPuff vs. ba + paPuff** : These stimulus streams offer the third important comparison because the predicted looking behavior between trials differs only if the experimental group infants are integrating the airflow. Under integration, the infants should treat both streams as alternating because the co-presentation of a puff of air with the unaspirated /ba/ token renders that token more like an aspirated /pa/.

Therefore, if the infants are sensitive to aerotactile somatosensation, I predict that the infants in the experimental group will exhibit similar looking behaviors across the two stimulus streams, while the control group infants will look longer to ba + paPuff. If the infants do not integrate the airflow, both the experimental group and the control group will show similar looking behaviors to the two stimulus streams (i.e., look longer to ba + paPuff).
4.3 Results

4.3.1 Trial Type Analysis

In the following sections, I describe the results obtained by comparing alternating and non-alternating trials. First, I test for effects of Trial Type on an infant’s total looking time to the monitor during the trial. Second, I test for effects of Trial Type on the duration of an infant’s first look to the monitor.

4.3.1.1 Total Looking Time

Within groups, mean looking times were similar to both alternating and non-alternating trials. Overall means and standard errors for total looking time to Type for each group can be seen in Table 4.1. However, as Figure 4.3 demonstrates, infant looking times decreased over the course of the experiment for both groups, as expected. Looking time generally decreases as infant attention wanes during the study. Response behavior was quite variable across the pairs (and across infants) for each Type, but it is clear from Figure 4.3 that neither the infants in the experimental group nor those in the control exhibited consistently longer looking times to alternating trials as predicted with integration.

Table 4.1 Mean looking times to trial by stimulus type for the experimental (left) and control (right) groups.

<table>
<thead>
<tr>
<th>Trial Type</th>
<th>Experimental Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (sec)</td>
<td>Standard Error</td>
</tr>
<tr>
<td><em>Alt</em></td>
<td>9.82</td>
<td>0.31</td>
</tr>
<tr>
<td><em>NonAlt</em></td>
<td>9.85</td>
<td>0.32</td>
</tr>
</tbody>
</table>
Figure 4.3 Mean looking times to checkboard by stimulus type and pair for both control group (left) and experimental group (right).

To test for statistical significance in looking time differences, a linear mixed-effects model was computed for each of the infant groups. The formula was as follows:

\[ \text{Total.Look} \sim \text{TrialType} \times \text{Pair} + (1 + \text{TrialType} | \text{Participant}) \]

Where Total.Look refers to the total amount of time the infant was coded as looking at the monitor during each trial; Type refers to whether the trial type was alternating or non-alternating; Pair refers to the pair order from 1 to 8; and \((1 + \text{TrialType} | \text{Participant})\) represents the random effects term by participant.

The fixed-effects for the model for each group can be seen in Tables 4.2 and 4.3. Pair emerged as a significant predictor of model fit for both the experimental group \(\beta = -0.64, SE = \)

\[ \]
0.10, \( t = -6.23, p < 0.001 \) and the control group (\( \beta = -0.74, SE = 0.11, t = -6.88, p < 0.001 \)), confirming that the infants significantly decreased their looking to the stimulus over the course of the experiment regardless of whether or not they felt a puff of air. This is expected given that infants tire easily and generally lose interest as an experiment progresses. Moreover, the task was a difficult one: the infants were not only seated in a highchair apart from the caregiver, they were also wearing the sound- attenuating cloak and the bib.

### Table 4.2 Summary of fixed effect for the control group.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.45</td>
<td>0.7</td>
<td>73.67</td>
<td>20.59</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt</td>
<td>0.09</td>
<td>0.7</td>
<td>314.87</td>
<td>0.14</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Pair</td>
<td>-0.74</td>
<td>0.11</td>
<td>316.87</td>
<td>-6.88</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt:Pair</td>
<td>-0.01</td>
<td>0.15</td>
<td>314.28</td>
<td>-0.05</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

### Table 4.3 Summary of fixed effects for experimental group.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.35</td>
<td>0.72</td>
<td>56.2</td>
<td>17.15</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt</td>
<td>0.07</td>
<td>0.66</td>
<td>292.91</td>
<td>0.11</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Pair</td>
<td>-0.64</td>
<td>0.10</td>
<td>295.91</td>
<td>-6.23</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt:Pair</td>
<td>-0.03</td>
<td>0.14</td>
<td>292.05</td>
<td>-0.12</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

As predicted, there was no effect of Trial Type and no interaction between Type and Pair for the infants in the control group, as expected since the infants were not feeling airflow during
the task (see Table 4.2). However, there was also no effect of Trial Type or interaction between Trial Type and Pair for the experimental group (see Table 4.3). In sum, there was no significant difference in total looking time between Alt and Non-Alt trials for the experimental group, suggesting that the infants were not integrating the aerotactile cue.

4.3.1.2 Duration of First Look

Overall means and standard errors for duration of first look to Type for each group can be seen in Table 4.4 below. As above, there is very little difference within group for each type of trial in terms of the means. Unsurprisingly, first look duration also decreased over the course of the trial for both groups (see Figure 4.4). While the infants in the experimental group do not exhibit longer first looks to alternating trials across the board, they do appear to be looking longer to alternating trials in Pair 1, Pair 4, and Pair 7. Given the ordering of the individual stimulus streams (see Appendix A), this likely indicates that the infants were treating the individual stimulus streams differently. This will be addressed in the analysis in Section 4.3.2.

Table 4.4 Means and standard errors for duration of first look.

<table>
<thead>
<tr>
<th>Trial Type</th>
<th>Mean</th>
<th>Standard Error</th>
<th>Mean</th>
<th>Standard Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alt</td>
<td>2.46</td>
<td>0.18</td>
<td>3.63</td>
<td>0.2</td>
</tr>
<tr>
<td>NonAlt</td>
<td>2.22</td>
<td>0.17</td>
<td>3.69</td>
<td>0.23</td>
</tr>
</tbody>
</table>
To test for statistical differences in first look duration, a linear mixed-effects model was computed for each of the infant groups. The formula was as follows:

$$\text{First.Look} \sim \text{TrialType} \times \text{Pair} + (1 + \text{TrialType} | \text{Participant})$$

Where First.Look refers to the duration of time from when the infant was first coded as looking at the monitor to the first look away from the monitor for more than one second; Type refers to whether the trial type was alternating or non-alternating; Pair refers to the pair order from 1 to 8; and \((1 + \text{TrialType} | \text{Participant})\) represents the random effects term by participant. Summaries of the fixed-effects for each group can be seen in Tables 4.5 and 4.6. As is evident in Figure 4.4, the duration of the first look significantly decreased over the course of the experiment. Again, this is an expected result given the realities of infant attention. As predicted, trial type was not a
significant predictor of model fit for the infants in the control group, showing that they did not exhibit longer initial fixations during one type of trial. In contrast, the infants in the experimental group showed a trend toward significantly longer first looks during alternating trials as predicted, though this did not reach significance ($\beta = -0.71$, $SE = 0.04$, $t = -1.74$, $p = 0.08$).

**Table 4.5 Fixed effect for the control group for duration of first look.**

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.27</td>
<td>0.48</td>
<td>132.69</td>
<td>11.04</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt</td>
<td>0.42</td>
<td>0.56</td>
<td>314.17</td>
<td>0.72</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Pair</td>
<td>-0.4</td>
<td>0.08</td>
<td>319.04</td>
<td>-4.69</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt :Pair</td>
<td>-0.08</td>
<td>0.12</td>
<td>314.28</td>
<td>-0.72</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

**Table 4.6 Fixed effects for the experimental group for duration of first look.**

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.64</td>
<td>0.39</td>
<td>81.48</td>
<td>9.36</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt</td>
<td>-0.71</td>
<td>0.04</td>
<td>291.85</td>
<td>-1.74</td>
<td>p = 0.08</td>
</tr>
<tr>
<td>Pair</td>
<td>-0.28</td>
<td>0.06</td>
<td>297.9</td>
<td>-4.38</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>TypeNonAlt :Pair</td>
<td>-0.11</td>
<td>0.08</td>
<td>291.89</td>
<td>1.29</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

**4.3.2 Stimulus Stream Analysis**

In the following sections, I describe the results obtained by comparing the individual stimulus streams rather than collapsing the streams in alternating and non-alternating trial types. As above, I first test for effects of Stimulus on an infant’s total looking time to the screen during the
trial. Second, I test for effects of Stimulus on the duration of an infant’s first look to the checkerboard.

4.3.2.1 Total Look Time

Infants in both groups showed small differences in looking behavior between some of the four stimulus streams, highlighting the importance of this second analysis. Table 4.7 shows the condition means and standard errors for both the experimental and control groups for each stimulus stream. Infants in the experimental group show slightly longer looking times during paPuff + pa trials and paPuff + ba trials. In contrast, the infants in the control group showed the longest looking times to paPuff + pa and ba + baPuff, the two non-alternating streams. This is somewhat unexpected given that infants tend to show preferential looking to alternating trials when there is no familiarization (e.g., Bruderer et al. 2015). However, as noted above, longer looking time to either type of trial has been taken as evidence of discrimination in previous work. Moreover, as is evident from both Table 4.7 and Figure 4.5, most of these looking time differences are quite small (less than 600 ms). However, the stimulus streams paPuff + pa and ba + baPuff as compared to pa + baPuff for the control group seem to be exceptions.
Figure 4.5 Total looking times for control group infants (left) and experimental group infants (right) to each stimulus stream. The red diamond in each boxplot signifies the mean.

Table 4.7 Mean looking time to the checkerboard and standard errors for each stimulus stream for both experimental and control groups.

<table>
<thead>
<tr>
<th>Stimuli stream</th>
<th>Experimental Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean (sec)</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$paPuff + pa$</td>
<td>10.14</td>
<td>0.68</td>
</tr>
<tr>
<td>$paPuff + ba$</td>
<td>10.13</td>
<td>0.52</td>
</tr>
<tr>
<td>$pa + baPuff$</td>
<td>9.73</td>
<td>0.67</td>
</tr>
<tr>
<td>$ba + baPuff$</td>
<td>9.44</td>
<td>0.55</td>
</tr>
</tbody>
</table>
As discussed in Section 4.2.5, there are three important stimulus stream comparisons to test in order to determine if the infants in the experimental group were integrating the airflow during the task. To test for statistical significance in total looking time differences, a linear mixed-effects model was computed for each of the infant groups. The formula was as follows:

1) \[ \text{Total.Look} \sim \text{Stimulus} \times \text{Trial.Number} + (1 + \text{Stimulus} + \text{Trial.Number} | \text{Participant}) \]

Where \text{Total.Look} refers to the total amount of time the infant was coded as looking at the monitor during each trial; \text{Stimulus} is one of “paPuff + pa”, “paPuff + ba”, “pa + baPuff”, “ba + baPuff”; \text{Trial.Number} reflects the trial order from 1 to 16, centered and scaled to improve model fit performance; and \((1 + \text{Stimulus} + \text{Trial.Number} | \text{Participant})\) represents the random effects term by participant. The reference level was set to the stimulus stream “ba + paPuff” because that particular stimulus stream is predicted to be alternating for both groups regardless of whether the experimental group integrates the airflow.

Summaries of the fixed effects for the experimental group and the control group can be found in Tables 4.8 and 4.9, respectively. Only Trial Number emerged as a significant predictor of model fit, but was highly significant for both the experimental group \((\beta = -0.40, SE = 0.10, t = -4.13, p < 0.001)\) and the control group \((\beta = -0.37, SE = 0.09, t = -4.36, p < 0.001)\) such that total looking time to checkerboard decreased for the infants over the course of the study. As noted in the Trial Type analysis above, overall reductions in looking time are to be expected in infant work as the infants grow tired or bored.
Table 4.8 Summary of fixed effects for the experimental group infants for total looking time to trial.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>12.38</td>
<td>0.83</td>
<td>30.96</td>
<td>15.27</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff</td>
<td>0.30</td>
<td>0.73</td>
<td>182.57</td>
<td>-0.41</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff</td>
<td>-0.21</td>
<td>0.83</td>
<td>226.66</td>
<td>-0.25</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>baPuff + pa</td>
<td>-0.38</td>
<td>0.79</td>
<td>199.58</td>
<td>-0.48</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Trial.No</td>
<td>-0.40</td>
<td>0.10</td>
<td>20.49</td>
<td>-4.13</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff: Trial.No</td>
<td>-0.04</td>
<td>0.08</td>
<td>247.27</td>
<td>-0.57</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff: Trial.No</td>
<td>-0.01</td>
<td>0.09</td>
<td>235.63</td>
<td>-0.15</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>baPuff + pa: Trial.No</td>
<td>-0.10</td>
<td>0.09</td>
<td>243.77</td>
<td>-0.11</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

Table 4.9 Summary of fixed effects for the control group infants for total looking time to trial.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>14.22</td>
<td>0.93</td>
<td>31.05</td>
<td>15.37</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff</td>
<td>0.83</td>
<td>0.80</td>
<td>175.45</td>
<td>1.05</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff</td>
<td>-0.50</td>
<td>0.93</td>
<td>189.01</td>
<td>-0.54</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>baPuff + pa</td>
<td>-1.17</td>
<td>0.87</td>
<td>185.29</td>
<td>-1.34</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Trial.No</td>
<td>-0.37</td>
<td>0.09</td>
<td>39.58</td>
<td>-4.36</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff: Trial.No</td>
<td>-0.03</td>
<td>0.09</td>
<td>269.62</td>
<td>-0.29</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff: Trial.No</td>
<td>0.11</td>
<td>0.10</td>
<td>276.55</td>
<td>1.06</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>baPuff + pa: Trial.No</td>
<td>0.13</td>
<td>0.09</td>
<td>269.57</td>
<td>1.38</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

In the following sections, I will turn to the three comparisons laid out in Section 4.2.5.
4.3.2.1.1  **Comparison 1: ba + paPuff vs pa + paPuff**

For this stimulus stream comparison, recall that infants in both the control group and the experimental group were predicted to look longer to ba + paPuff than pa + paPuff regardless of integration. This is because the stimulus streams are predicted to be alternating and non-alternating, respectively, both with and without the integration of the puff of air. However, neither group exhibited longer total looking times to the checkerboard for either stimulus stream (see Tables 4.8 and 4.9). This result is surprising given that the predicted looking behavior did not differ depending on integration. These null results do not provide evidence for or against the experimental group infants integrating the airflow during the task, but instead suggest that neither group of infants successfully discriminated these syllable streams during the study.

4.3.2.1.2  **Comparison 2: ba + paPuff vs. baPuff + pa**

For this stimulus stream comparison, the predicted looking behavior changes if the experimental group infants are integrating the airflow during the task. Recall that, without the puff of air, these two streams were identical. Thus, if there was no integration, the infants in both groups should show similar looking behaviors for the two stimulus streams. If the experimental infants did integrate the airflow, however, they were predicted to look longer to ba + paPuff, the alternating stimulus stream. As evident in Tables 4.8 and 4.9, neither group showed significantly longer total looking time to either ba + paPuff or baPuff + pa, in this case matching the prediction under no integration. The null results from this comparison could be interpreted as evidence that the infants in the experimental group were not influenced by the airflow during the trials.
4.3.2.1.3  **Comparison 3: ba + baPuff vs. ba + paPuff**

As above, integration of the puff was predicted to shift looking behavior for the experimental group in this final stimulus stream comparison. In this comparison, however, the infants in the experimental group should only look longer to one of the stimulus streams if they are *not integrating* the airflow. In contrast, if the infants *do integrate*, the presence of the puff on the /ba/ should push their perception of baPuff + ba toward an alternating sequence similar to ba + paPuff. Just as in Comparisons 1 and 2 above, neither the experimental nor the control group infants showed signs of longer total looks to trial for either stimulus stream. These null results do not match the predictions for either integration or no integration; in both sets of predictions, at a minimum the control group was predicted to look longer to ba + paPuff.

4.3.2.2  **Duration of First Look**

In the following models, duration of first look to checkerboard replaced total looking time as the dependent measure. As noted in Section 4.2.5, first look to the stimulus can be a more sensitive measure of infant preference because infant attention wanders over the course of the trial. Means and standard errors for first look duration can be found in Table 4.10. Infants in both groups exhibited only small differences in mean first look durations across stimulus streams. Both groups show the longest first look to ba + paPuff trials, arguably the most clearly alternating stream. However, both groups also show the shortest initial look to ba + baPuff, which the experimental groups were predicted to treat as alternating (and therefore look longer to) if integrating the puff.
Table 4.10 Mean duration of the first look to the checkerboard and standard errors for each stimulus stream for both experimental and control groups.

<table>
<thead>
<tr>
<th>Stimuli stream</th>
<th>Experimental Group</th>
<th>Control Group</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Standard Error</td>
</tr>
<tr>
<td>$paPuff + pa$</td>
<td>2.35</td>
<td>0.68</td>
</tr>
<tr>
<td>$paPuff + ba$</td>
<td>2.73</td>
<td>0.33</td>
</tr>
<tr>
<td>$pa + baPuff$</td>
<td>2.37</td>
<td>0.38</td>
</tr>
<tr>
<td>$ba + baPuff$</td>
<td>2.30</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Figure 4.6 Duration of first look to the checkerboard for control group infants (left) and experimental group infants (right) for each stimulus stream. The mean for each stimulus stream is indicated by a red diamond within each boxplot.
To test for statistical significance, a linear mixed-effects model was computed for each of the infant groups. The formula was as follows:

1) \( \text{First.Look} \sim \text{Stimulus} \ast \text{Trial.Number} + (1 + \text{Stimulus} + \text{Trial.Number} | \text{Participant}) \)

Where \( \text{First.Look} \) refers to the duration of time from when the infant was first coded as looking at the monitor to the first look away from the monitor for more than one second during each trial; \( \text{Stimulus} \) is one of “paPuff + pa”, “paPuff + ba”, “pa + baPuff”, “ba + baPuff”; \( \text{Trial.Number} \) reflects the trial order from 1 to 16, centered and scaled to improve model fit performance; and \( (1 + \text{Stimulus} + \text{Trial.Number} | \text{Participant}) \) represents the random effects term by participant. Again, the reference level was set to the stimulus stream “ba + paPuff” because that particular stimulus stream was predicted to be alternating for both groups regardless of whether the experimental group integrates the airflow.

Table 4.11 Summary of fixed effects for the experimental group infants for duration of first look to trial.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>3.82</td>
<td>0.51</td>
<td>27.41</td>
<td>7.52</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff</td>
<td>-1.19</td>
<td>0.47</td>
<td>57.89</td>
<td>-2.54</td>
<td>p &lt; 0.05 *</td>
</tr>
<tr>
<td>ba + baPuff</td>
<td>-1.50</td>
<td>0.50</td>
<td>152.34</td>
<td>-3.01</td>
<td>p &lt; 0.01 **</td>
</tr>
<tr>
<td>baPuff + pa</td>
<td>-0.76</td>
<td>0.46</td>
<td>132.81</td>
<td>-1.65</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Trial.No</td>
<td>-0.29</td>
<td>0.05</td>
<td>17.99</td>
<td>-4.04</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff: Trial.No</td>
<td>0.011</td>
<td>0.05</td>
<td>219.15</td>
<td>2.37</td>
<td>p &lt; 0.05 *</td>
</tr>
<tr>
<td>ba + baPuff: Trial.No</td>
<td>0.15</td>
<td>0.05</td>
<td>188.96</td>
<td>2.81</td>
<td>p &lt; 0.01 **</td>
</tr>
<tr>
<td>baPuff + pa: Trial.No</td>
<td>0.06</td>
<td>0.05</td>
<td>205.87</td>
<td>1.20</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>
Table 4.12 Summary of fixed effects for the control group infants for duration of first look to trial.

<table>
<thead>
<tr>
<th>Fixed effect</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>df</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>5.75</td>
<td>0.75</td>
<td>31.68</td>
<td>7.62</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff</td>
<td>0.30</td>
<td>0.62</td>
<td>262.73</td>
<td>0.49</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff</td>
<td>-1.76</td>
<td>0.74</td>
<td>230.59</td>
<td>-2.39</td>
<td>p &lt; 0.05 *</td>
</tr>
<tr>
<td>baPuff + pa</td>
<td>-1.01</td>
<td>0.69</td>
<td>214.59</td>
<td>-1.47</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>Trial.No</td>
<td>-0.24</td>
<td>0.06</td>
<td>51.71</td>
<td>-3.77</td>
<td>p &lt; 0.001 ***</td>
</tr>
<tr>
<td>pa + paPuff:Trail.No</td>
<td>-0.06</td>
<td>0.07</td>
<td>269.40</td>
<td>-0.86</td>
<td>p &gt; 0.1</td>
</tr>
<tr>
<td>ba + baPuff:Trail.No</td>
<td>0.15</td>
<td>0.08</td>
<td>271.54</td>
<td>1.89</td>
<td>p = 0.059</td>
</tr>
<tr>
<td>baPuff + pa:Trail.No</td>
<td>0.09</td>
<td>0.07</td>
<td>286.96</td>
<td>1.18</td>
<td>p &gt; 0.1</td>
</tr>
</tbody>
</table>

As above, Trial Number emerged as a significant predictor of model fit and was highly significant for both the experimental group ($\beta = -0.29, SE = 0.05, t = -4.04, p < 0.001$) and the control group ($\beta = -0.24, SE = 0.06, t = -3.77, p < 0.001$) such that looking decreased for the infants over the course of the study. For the experimental group, there were also significant interactions between Trial Number and two of the stimulus streams such that first look durations to pa + paPuff and ba + baPuff decreased less over the course of the experiment than first look durations to ba + paPuff (see Table 4.11). For the control group, there was a similar trend for ba + baPuff trials, but the interaction did not reach standard levels of significance ($\beta = 0.15, SE = 0.08, t = 1.89, p = 0.059$).

In the following sections, I will turn again to the three comparisons laid out in Section 4.2.5.
4.3.2.2.1 **Comparison 1: ba + paPuff vs pa + paPuff**

The control group did not exhibit significantly different looking behavior across the two stimulus streams, contrary to the prediction that both groups of infants would show longer first look durations to ba + paPuff in this comparison regardless of integration. However, the infants in the experimental group did show evidence of significantly shorter initial looks to the checkerboard during pa + paPuff trials ($\beta = -1.18$, $SE = 0.48$, $t = -2.44$, $p < 0.05$), suggesting that at least the experimental group infants were discriminating the two stimulus streams.

4.3.2.2.2 **Comparison 2: ba + paPuff vs. baPuff + pa**

As I noted above, this comparison offers an important test of integration because the streams are acoustically identical. As predicted, the infants in the control group (who were not presented with airflow at the neck) did not look significantly longer to either stream. Crucially, the infants in the experimental group also did not exhibit longer first looks to either stimulus stream, contrary to the prediction under integration. Recall that if the infants in the experimental group were integrating, they were predicted to show shorter first looks to trials with the stimulus stream baPuff + pa; the co-presentation of the airflow with the /ba/ should have pushed their percept toward an aspirated /pa/. Instead, the experimental infants showed no significant difference in first look duration to either stimulus stream, suggesting that the infants were not integrating the airflow.

4.3.2.2.3 **Comparison 3: baPuff + ba vs. ba + paPuff**

In this final comparison, the experimental infants should only exhibit similar first look durations to the two stimulus streams if they are integrating the airflow. Otherwise, the experimental
infants, like the infants in the control group, were predicted to treat baPuff + ba and ba + paPuff as non-alternating and alternating, respectively. As can be seen in Table 4.12, the infants in the control group exhibited significantly shorter first looks during baPuff + ba trials ($\beta = -1.75, SE = 0.71, t = -2.61, p < 0.05$), as predicted. In addition, the infants in the experimental group showed significantly shorter first looks during baPuff + ba trials, as compared to ba + paPuff trials ($\beta = -1.81, SE = 0.05, t = -3.15, p < 0.01$). Crucially, this looking behavior matches the predictions for no integration outlined in Section 4.2.5, providing additional evidence that the infants in the experimental group were not integrating the puff of air during the task.

4.4 Discussion

In this study, I asked whether perceivers require speech production experience to be able to integrate auditory and aero-tactile cues during speech perception by testing prelinguistic infants in an audio-aerotactile perception task. I predicted that, if integrating, the infants in the experimental group would look longer to the alternating trials because the aerotactile information in the stimulus would push their perception toward an aspirated token. Recall that the overall prediction was that the presence of the aerotactile cue, or puff, would cause the infants to treat /ba/ syllables as more /pa/-like because they would incorporate the airflow during perception and show an aspiration effect like that seen in Gick and Derrick (2009). Given this, the presence of the puff of air was predicted to shift the perceived nature of the trial (i.e., alternating or non-alternating) depending on whether the puff accompanied a sound that naturally produces a burst of air. In other words, the puff of air would interfere with the infant’s ability to discriminate between aspirated and unaspirated tokens when the airflow occurred with a /ba/ syllable. The
results presented here however, do not provide evidence of that prelingual infants integrate aerotactile somatosensation during speech perception.

In this study, I conducted two analyses (Trial Type and Stimulus Stream) with two dependent measures (Total Look and First Look) to test for evidence of integration. While the infant’s total looking time to trial offered little insight in either analysis (the null results across the board are difficult to interpret as they support neither prediction clearly), the more sensitive measure of first look duration offers some evidence that prelingual infants do not integrate speech-related airflow during perception. In the results from the Trial Type analysis, the infants in the control group (for whom airflow was present but not felt during the trials) showed no preference for one type of trial over the other, as predicted. The experimental group showed a non-significant trend toward a longer first look to alternating trials, which is consistent with the integration predictions. However, the results from the stimulus stream analysis suggest that this trend was likely driven by longer looking times to one of the “alternating” stimulus streams: ba + paPuff. When I compared first look duration for the two “alternating” streams, the infants in the experimental group showed significantly longer first looks to ba + paPuff than to baPuff + ba. More to the point, in both Comparisons 2 and 3 the experimental group infants patterned with the control group infants; their first look durations to the relevant stimulus streams matched the predictions for no integration. Thus, the findings reported here offer some evidence that 6–8-month-old infants do not integrate aerotactile somatosensation during the perception of stop consonants.

These results are remarkable given the evidence outlined in the introduction that infants integrate other multi-sensory speech cues well before they begin speaking (e.g., audiovisual speech). Infants are sensitive to cross-modal speech information from very early in life—both
visual and sensorimotor speech cues modulate perception in prelinguistic infants. However, as I discussed in the introduction, Choi et al.’s (2017) mechanism for linking these cross-modal cues is not easily extended to speech-related airflow given the nature of the fetal environment. As I have argued in Chapter 2 of this dissertation, proprioception and (aero)tactile somatosensation should be considered separate modalities for the purposes of speech perception research. Unlike proprioceptive feedback, the aerotactile cue involves a physiological response to a secondary sensory consequence of a movement. It may be that perceivers need at least some experience generating and perceiving that sensory consequence before the cue is relevant during a perception task. This is especially true considering the unreliable relationship between aerotactile cues and other speech information in another’s speech. In our common experience as English perceivers, visual “ba” is produced with auditory “ba” and little or no turbulent airflow during the stop release. Visual “pa” is produced with an auditory “pa”, and considerable turbulent airflow during the stop release. Importantly, this turbulent airflow sometimes, but not always, hits your skin when you encounter a “pa” from a person speaking close to you, but almost never when you encounter a “ba” even from someone very close to you. It would not be surprising then if infants develop the ability to integrate auditory and visual speech information more quickly than speech-related airflow. Such results are in keeping with what has been proposed in some computational models of speech production (for example, see Guenther, 1994; Guenther, 1995; Tourville & Guenther 2011) offering additional evidence that infants may learn some sensory consequences of articulatory movements through productive experience.
4.4.1 Study Limitations

While the above results offer tantalizing evidence that preverbal English-acquiring infants are not influenced by speech-related airflow cues, the findings fall short of offering conclusive evidence that perceivers require experience feeling their own airflow during production to integrate auditory and aerotactile information. Several important limitations make the current findings preliminary.

First, the null Trial Type results are tricky to interpret because most of the sequences presented could only be perceived as alternating/non-alternating if the infants were integrating the airflow during the trial. Thus, with the puff removed from the perceptual event, the trials no longer follow an alternating/non-alternating pattern (see Appendix A for additional information about the trial order) and the study design offers no baseline measure of infant perception of this contrast in noise. To test whether /pa/ and /ba/ are generally discriminable in noise at this age without a puff, the infants must be presented with a different sequence in which the trial type is based on the acoustics alone. Unfortunately, the control group of infants in the current study do not provide this comparison because they performed the exact same task with the exception of the sensation of the airflow. Thus, for control group infants who had no aerotactile input to integrate with the acoustics, the trials did not follow an Alt/NonAlt sequence. Previous research has demonstrated that infants can successfully discriminate place of articulation differences such as /ba/ vs. /ga/ in noise (Nozza et al 1991). However, voicing distinctions may be more difficult as one of the main cues to the distinction (turbulent airflow noise) is at least partially masked. It is possible that the noise levels were simply too high here for the infants to discriminate this subtle cue consistently. Infants have been shown in previous work to have a significant disadvantage in both detecting and identifying sounds in noise compared to adults (Nozza et al.
1991, Nozza 1990). The work by Nozza and colleagues suggests that infants may be unable to discriminate consonants in signal-to-noise ratios below +8. Previous aerotactile perception tasks with adults have used a noise level of -6 SNR (e.g., Gick & Derrick 2009 and Bicvskis et al 2016, among others) to degrade the signal. In the current task, I used a signal-to-noise ratio of +2 to mask the sound of the air exiting the tube. While I did not include the noise to purposely make the task harder for the infants, the noise level needed to mask the sound of the air may have made the task too hard nonetheless. In the work by Nozza and colleagues, 7-11-month-old infants could not discriminate auditory /ba/ and /ga/ in trials with an SNR lower than + 8. And while some evidence suggests that visual information facilitates infant perception of auditory speech in noise, the facilitative effect appears linked to domain-general sensitivities to temporal synchrony in audio-visual stimuli and not dependent on phonetic information (Hollich et al. 2005).

Second, the more interpretable Stimulus Stream analysis results did not align with the no integration predictions in all comparisons. Specifically, it is unclear why the infants in the experimental group showed longer first looks to the alternating sequence in Comparison 1 but the control group did not. Recall that the stimulus streams in this comparison (ba + paPuff vs. pa + paPuff) were predicted to be alternating and non-alternating (respectively) regardless of whether there was integration. Thus, I predicted that both groups would look longer to ba + paPuff, which the results did not support. Moreover, the effect of stimulus in Comparison 3 was smaller for the control group as compared to the experimental group, leaving open the possibility that the puff had some non-linguistic effect on the experimental group. For example, the synchronous puff stimulus may have increased the experimental group infants’ attention to the stimulus thereby improving discrimination, though further investigation would be necessary to
determine this. This would be in keeping with the findings from Hollich et al (2005) that
synchrony between cross-modal cues facilitates infant audiovisual discrimination in noise.

Finally, and perhaps most importantly, the effect sizes are quite small across the board
and it remains possible that the current sample size is too small to find an effect of airflow with
the current paradigm. While the aspiration effect is largely reliable in adults across studies, the
effect is more subtle than a categorical shift from an unaspirated to an aspirated percept. For
example, while the adult participants in Chapter 3’s spatial congruence study showed a
significant increase in “pa” responses when the syllable was accompanied by airflow, their
average percent “pa” response did not exceed 60%. For most of the adults, then, a perceptual
shift towards an aspirated token occurs some but not all of the time when they feel airflow.
Infants, whose perceptual systems are still developing, would likely show an even smaller and
less robust an effect. This would be in line with infant McGurk results noted above suggesting
infants integrate the incongruent audio and visual cues but to a lesser degree than adults
(Desjardins & Werker 2004). Given the likelihood of a small effect, assuming there is one to be
found, looking time behavior may not be a sensitive enough measure to determine whether
infants can integrate these cues. One possible alternative is pupillometry. Several adult studies
provide evidence for a positive relation between increased attention and pupil size (Goldwater,
1972, Laeng et al., 2012, Sirois & Brisson, 2014) and researchers have begun employing the
measure in infant work as well. Unlike looking time (which is a behavioral measure),
pupillometry is a measure of a physiological response—changes in pupil dilation—that
constitutes an involuntary marker of activity in the nervous system. Moreover, pupil response
does not decrease with fatigue the way global looking time does. Thus, measuring pupil dilation
may offer a more sensitive assessment of infant response to novel or interesting stimulus and enable us to discover whether the airflow sensation shifts the infant’s perception of the syllable.

In sum, the findings described here leave open the question of production experience in emergence of aerotactile speech perception. Had the experimental infants shown evidence of integration, it would have been possible to rule out speech production as the experience that makes the cue available in perception. However, I did not find evidence that the sensation of airflow shifted infant perception toward an aspirated token. Yet, even if the results I obtained reflect an inability to integrate, we are still left to consider a few possible mechanisms other than production experience. For one, the ability to integrate airflow may simply emerge later than for other cues. Speech-related airflow is not reliably felt from others’ speech and thus infants may need extended exposure to it. In our common experience as English perceivers, visual “ba” is produced with auditory “ba” and little or no turbulent airflow during the stop release. Visual “pa” is produced with an auditory “pa”, and considerable turbulent airflow during the stop release. Importantly, this turbulent airflow sometimes, but not always, hits you when you encounter a “pa” from a person speaking close to you, but almost never when you encounter a “ba” even from someone very close to you. Given this, it would be unsurprising if infants develop the ability to integrate speech-related airflow more slowly than auditory and visual speech information.

Second, aerotactile integration may stem from general experience manipulating the vocal tract rather than specific experience with stop production. If so, it may perhaps emerge as infants gain conscious control over their glottis. Infant glottal control develops relatively late compared to labial and lingual control. Initially, this may seem at odds with Choi et al.’s proposal that prenatal experience with early emerging stereotypies such as suckling and tongue protrusion
behaviors may help form the link between perception and production (Choi et al. 2017), a view supported by the clear evidence for sensorimotor influences described in the introduction (e.g., Bruderer et al. 2015). However, as I noted in the introduction, there is no a priori reason to assume that the developmental timing of sensorimotor influences and aerotactile influences is the same. First, unlike proprioceptive feedback, aerotactile cues (and perhaps vibratory cues, as well) involve the perception of a secondary sensory consequence of an articulatory movement, rather than the perception of that movement itself. It may be that perceivers need at least some general experience generating and perceiving that sensory consequence before the cue is relevant during a perception task. Second, airflow is one sensory cue that infants cannot have experience with in utero. These differences coupled with the lower reliability of the cue as described above may delay the ability to use the cue in perception. Future work should investigate a broader range of ages to isolate the relevant mechanism or mechanisms that contribute to its emergence.
Chapter 5: General Discussion and Conclusion

Speech research during recent years has moved progressively away from its traditional focus on audition toward a more multisensory approach. Until fairly recently, the bulk of the research has focused on interactions between auditory and visual speech information. However, in addition to audition and vision, many somatosenses including proprioception, pressure, vibration and aerotactile sensation are all highly relevant modalities for experiencing and/or conveying speech. Far from occurring in a one-, two- or even three-dimensional space, speech occupies a highly multidimensional sensory space. The considerable research described in this dissertation has shown that these modalities can interact during both speech perception and production, raising important questions about what, if any, kinds of experience affect not only when and how perceivers integrate, but the emergence and development of multisensory integration. In this dissertation, I have sought to deepen our understanding of the relationship between specific prior perceptual experience and multisensory integration through three sets of behavioral experiments using auditory, visual, and aerotactile speech information. The following section summarizes the findings in Chapters 2-4.

5.1 Summary of Experimental Chapters

In Chapter 2, I demonstrated that perceivers do not require prior experiences with specific visual and aerotactile speech stimuli in order to integrate them during speech perception. It can be difficult to identify valid methods for testing the necessity of prior specific experience with adult perceivers given our copious experience with speech in different environments and from different
sources. In this study, I created a computer-generated face for the visual stimulus. Then, I presented the participants with speech information they have no experience associating with a simulated face: aerotactile somatosensation. While studies in audiovisual speech perception often employ digitally-rendered faces, it remained unknown whether a virtual face is treated as if it were identical to a human face for the purposes of integration or if the integration of speech cues from a simulated face must be learned at the time of exposure. The use of aerotactile speech information enabled me to test whether participants could integrate airflow from a novel source. Participants were randomly assigned to groups in which they viewed either a simulated face or a human face silently articulating a bilabial stop consonant. In half of the trials, the video was accompanied by a puff of air. The results show perceivers treat computer-generated faces and human faces in a similar fashion: for all groups, the sensation of airflow biased participants toward an aspirated token from the beginning of the experiment. Computer-generated talking faces do not produce airflow during speech. Thus, while perceivers have general experience with animated faces and auditory speech, they cannot have experience with animated faces and aerotactile speech cues. Yet, the participants in the experiments in Chapter 2 integrated airflow paired with a simulated face from the beginning of the study. This suggests that general experience with human faces and speech-related airflow is sufficient for perceivers to integrate airflow co-presented with an animated face.

In Chapter 3, I showed that perceivers integrate spatially incongruent audio and aerotactile speech cues, suggesting we are able to integrate multimodal stimuli that do not conform with our previous experience. Participants took part in a two-alternative forced choice task in which they were presented with /ba/ and /pa/ syllables in noise. The syllables were presented one of three ways: in the left ear only, in the right ear only, or binaurally. In half of the trials, participants felt synchronous puffs of air at the side of the neck. For the trials that included both audio and
aerotactile speech information, there were two conditions: one in which the cross-modal stimuli came from opposing directions and one in which the cross-modal stimuli came from the same direction. Following each trial, participants were asked to indicate which syllable they had heard ("pa" or "ba"). During trials with airflow, participants provided significantly more "pa" responses as compared to trials with no airflow regardless of whether the auditory stimulus was /ba/ or /pa/.

Crucially, participant response was not affected by manipulations of spatial congruence: the participants integrated the airflow regardless of whether the cues appeared to originate from the same spatial location. These results offer additional evidence supporting the findings in Chapter 2 that perceivers can integrate cross-modal cues with which they do not have specific prior experience. In addition, the results in Chapters 3 add more generally to the multisensory speech literature. As I discussed in Chapter 1 and Chapter 3, previous audiovisual speech perception work has suggested that perceivers are remarkably tolerant of spatial incongruence when the cross-modal cues are auditory and visual (Jones and Munhall 1997; Bertelson et al. 1994; Fisher and Pylyshyn 1994). The results reported in this chapter add crucial support to Jones and Munhall (1997)’s claim that multimodal speech perception is unaffected by spatial dislocation by extending questions about spatiotemporal congruence beyond audiovisual interactions.

Lastly, I showed in Chapter 4 that while adult perceivers appear to integrate aerotactile speech information without specific prior experience with the cue pairing, some type of experience may be required for this ability to emerge during development. I reported on looking time data from 48 English-acquiring infants collected using a modified alternating/non-alternating paradigm. The infants were presented with stimulus streams containing 20-second long sequences of /ba/ and or /pa/ syllables that either alternated between syllables (i.e., ba – pa – ba – pa) or did not (i.e., ba – ba – ba – ba). Infants in the experimental group felt synchronous, gentle puffs of air
at the suprasternal notch during some of the tokens and it was hypothesized that the presence of the airflow on the unaspirated /ba/ tokens would cause the infants to treat the tokens as more /pa/-like. This would in turn influence their perception of the stimulus streams as alternating or non-alternating. The results data indicated that the infants did not integrate the multisensory cues. While it appears infants are able to integrate articulatory-motor information without experience generating or even hearing the sounds, the results from Chapter 4 leave open the possibility that specific experience is required for certain types of sensory information. This further underscores the need to consider the many somatosenses separately in future research.

Now I will return to the three research questions I laid out in the final section of Chapter 1 and briefly answer them using the findings from the aforementioned experiments.

1) **Can adult perceivers integrate multimodal stimuli with which they have prior general but not specific experience? If so, can they do this without learning an association between the cues at the time of exposure?**

The answer seems to be yes. Considering the findings from the experiments in Chapters 2 and 3, perceivers appear able to integrate cross-modal cues from new or familiar but altered stimulus pairings when they have had significant prior experience with those cues more generally. For example, the participants in Chapter 2 were presented with aerotactile cues and a computer-animated face. As mentioned previously, the participants likely did not have experience with speech-related airflow from a computer or from an animated face. However, they would have had significant prior experience with speech-related airflow and human talking faces, as well as auditory speech information from animated faces. Yet, the participants in the simulated face group
did not show significant differences in integration from the participants in the human face groups. It would seem that their general prior experience with these cues enabled them to integrate speech information from a novel specific cue pairing.

In addition, the participants exhibited similar response patterns across the experiment demonstrating that they did not learn to associate the simulated talking face and the aerotactile cues through exposure during the trials. These findings are in line with previous work showing that perceivers untrained in the Tadoma Method can integrate tactile speech information with other sensory cues (Fowler & Deckle 1991; Gick et al. 2008). As discussed in Chapter 1, it was long assumed that tactile speech information could only be useful during perception after significant training. However, both of the foregoing studies found that tactile speech information interfered with and facilitated that perception of auditory and visual speech information by untrained perceivers.

The results from Chapter 3 offer additional support for this interpretation. The participants in the spatial congruence study were presented with aerotactile and auditory cues—two cues that the adults tested should have plenty of prior experience with. However, by lateralizing these familiar cues I created a multimodal speech event that should be outside of their common experience: rarely if ever do we experience cross-modal speech information that originates from a common source yet arrives from opposing directions. In this study, the participants were unaffected by the spatial dislocation of the cues, integrating the auditory and aerotactile cues similarly across conditions. This suggests that their lack of experience with the cues presented in this specific manner did not interfere with their ability to integrate the speech information during perception. As above, general experience with the cues seems to have been sufficient for the perceivers to integrate in this novel situation.
2) *Can adult perceivers integrate multimodal stimuli that do not conform with their prior experience with the cues?*

As far as the spatial location of the cues is concerned, the answer appears to be yes. As noted above, perceivers have arguably little to no experience with synchronous lateralized speech cues originating from a single source. More to the point, perceivers *do* have significant experience with synchronous cues that originate from the same location in space. Thus, the lateralized cues employed in Chapter 3 are not only presented in a way perceivers have yet to experience, they are presented in a way that may conflict with what perceivers have experienced previously in the natural world. Yet the participants in Chapter 3 showed similar levels of integration across the stimulus direction conditions as measured by their percent “pa” responses. Dislocating cross-modal speech information does not interfere with integration in a perception task, offering additional evidence that certain aspects of the multimodal stimulus in a laboratory perception task can violate the way that the cross-modal cues typically occur in the natural world and perceivers continue to integrate them.

3) *Can infants integrate aerotactile cues before they have experience producing them?*

Possibly not, though further research is necessary to confirm this. The results in Chapter 4 do not offer evidence that pre-linguistic infants are able to integrate aerotactile speech information. However, as I discussed at length in Chapter 4, it remains possible that the paradigm used was not sensitive enough to detect such a small effect or that aerotactile integration simply takes longer
developmentally to emerge. Ongoing work in the Interdisciplinary Speech Research lab suggests that consistent experience feeling one’s own airflow during aspirated stop production may not be required for adult native English perceivers to integrate auditory and aerotactile information (Keough et al. 2018). We tested congenitally hard-of-hearing (HoH) perceivers in a visual-aerotactile paradigm following the methods used in Chapter 2. Congenitally hard-of-hearing (HoH) individuals have been shown to have difficulty producing voicing distinctions and exhibit greater inter- and intra-speaker variability in aspiration production (Lane & Perkell 2005; Khouw & Ciocca 2007). For example, Lane and Perkell (2005) found that prelingually deaf adults showed a reduced voice onset time (VOT), usually producing the voiceless unaspirated stop regardless of whether the target should be aspirated. Furthermore, Khouw and Ciocca (2007) found no significant difference in VOT between voiced and voiceless initial stops for their hearing-impaired participants. Their resulting inconsistent aspiration during stop production may affect their association of aerotactile cues with aspiration and in turn affect their ability to use aerotactile cues in perception. While we are currently still recruiting participants, preliminary results show no effect of hearing loss—the HoH perceivers were significantly more likely to respond /pa/ when they felt airflow, just like the control group participants who had normal hearing. That our HoH participants showed influences of air flow on speech reading despite not having contrastive aspiration in their own productions suggests a role for either additional perceptual exposure from others’ speech or for additional general experience with the vocal tract. I will return to the question of future research in Section 5.3 below.
5.2 Empirical Implications

The results described in this dissertation have implications for theoretical issues in speech perception and sensory integration research. One such issue—namely, direct and indirect theories of speech perception—will be discussed in the following sections.

5.2.1 Perception-Direct or indirect?

While it is clear from decades of research that speech information from different modalities gets integrated during speech perception, theories diverge on whether this integration is dependent upon experience—either general or specific—with the cross-modal cues. Understanding the role of experience in establishing links between different sensory streams could provide a key to understanding how these cross-modal links develop. Historically, there has been an opposition in ways of looking at how we integrate cross-modal speech cues (e.g., Massaro 1987, 1998; Massaro & Chen 2008; Fowler 1986, 1996, 2010). While new theoretical positions have evolved in recent years (e.g., Hickok & Poeppel 2004, 2007), the dichotomy has continued to influence many assumptions about cross-modal effects and their interpretation. More to the point, these different approaches make different predictions about the role of experience in our ability to integrate. Given their longstanding prominence, I would be remiss in not considering the implications of the current findings for the theoretical debate. In the following sections, I will situate the results from Chapters 2-4 with respect to both approaches: the ecological (or direct), as well as the cognitivist (or indirect).
5.2.1.1 Ecological or Direct Approach

On one side of the theoretical fence we see “direct” or ecological approaches to speech perception, such as those of Fowler (1986, 1991, 1995), Bregman (1990), and Best (1995), modeled to varying degrees on Gibson (1966, 1979/1986). In this view, perception does not involve the detection of sensory input at the receptors and top-down interpretation. Instead, proponents argue that because perception is the means through which perceivers have evolved to interact with their environment, perceivers directly experience the causal source of the sensations. Thus, in such a view, you could argue that perceivers should be able to integrate multisensory information without prior experience provided that the multisensory information is ecologically valid. Yet, the participants in the studies in Chapters 2 and 3 exhibited a remarkable insensitivity to the ecological validity of the cues, both in terms of the potential source of the cues and in terms of how the speech event is structured. In Chapter 2, the participants who experienced speech-related airflow from a computer-generated source did not integrate that airflow differently from the participants who experienced airflow paired with a human face, thus appearing not to distinguish between ecologically possible sources of speech-related airflow (i.e., the talking human face used in Bicevskis et al. 2016) and impossible ones (i.e., the computer-animated talking face used here). Additionally, the participants in the study in Chapter 3 showed no evidence of being affected by spatial dislocation even though a speech event with lateralized cues arguably does not conform to how speech events are typically structured in the natural world. Thus, the findings from Chapter 3 suggest that perceivers can integrate ecologically possible and ecologically impossible speech events similarly, at least with respect to spatial congruence. In general, the results reported here seem a tough fit for an ecological perspective. If we directly perceive the source of the cues, how is it that we integrate things that are not ecological? But perhaps most importantly, why are infants unable to integrate aerotactile
and audio speech information? As noted above, in the ecological view, perception occurs without a mediating cognitive process that draws on acquired associations between the unisensory streams. If perceivers need not learn to associate the information carried by the disparate sensory streams, it is unclear then how to interpret the fact that the infants in Chapter 4 appear not to have integrated the aerotactile speech information.

5.2.1.2 Cognitivist or Indirect Approach

On the other side is a Helmholtzian-based cognitivist approach in which perception is mediated by some higher-order cognitive process (e.g., Massaro 1987; Schwartz 2010). From this perspective, the perceptual system detects and processes sensory cues from each modality separately; these cues are subsequently integrated into a percept through top-down processing. In one such model, Massaro’s fuzzy logic model of speech perception (FLMP) theory (Massaro 1987, 1998), perceivers evaluate the features independently present in the auditory and visual speech signals, integrate the unisensory streams derived from commonly experienced information in the real world to decide upon the most appropriate percept by consulting prototype representations of the syllables in memory. According to Massaro & Cohen (1993, p. 128):

“The integration process combines the auditory and visual modalities as independent sources of evidence for the occurrence of syllable prototypes. An identification decision is made on the basis of the relative goodness of match of the stimulus information with relevant prototype descriptions.”
Given the central role for memory representations in theories such as FLMP, prior experience would be predicted to play an important role: what we have—*and haven’t*—experienced, should affect what we integrate during speech perception. While the results reported here do not speak to whether perceivers consider the sensory streams independently, the findings are, on the whole, consistent with a view in which perceivers identify tokens based on a relative goodness of fit with previously experienced tokens. For example, the adult participants in Chapter 2 were able to integrate aerotactile speech information from a novel source with which they did not have specific prior experience, likely because the animated talking face used here was sufficiently like the human faces in their general experience. In this case, the best possible match in the participant’s memory for a silent animated bilabial articulation co-presented with synchronous airflow would be /p/. Similarly, we might expect perceivers to integrate the spatially dislocated audio-aerotactile cues in Chapter 3 because sensory information is judged on its *relative goodness* as a match rather than needing to match prior experience perfectly in a cognitivist approach. While the multisensory cues in Chapter 3 did not match prior experience in terms of spatial congruence, the cues conformed in two other important respects: temporal synchrony and signal relevance. This “good enough” fit seems to have allowed the participants to integrate the audio-aerotactile cues. Finally, the infant results in Chapter 4 are not unexpected given the roles of memory and prototypes in a cognitivist approach. As mentioned previously, prototypical representations are built up through commonly experienced information. Pre-linguistic infants may not have had enough experience with speech-related airflow to form prototypes that include it and thus, the cue would not factor into their identification of a syllable.
5.3 Future Work and Conclusions

When the evidence from visuo-aerotactile and audio-aerotactile perception is combined with the audiovisual literature we can see that, while cross-modal effects in speech perception are constrained by temporal congruence and by signal relevance, they appear surprisingly unconstrained by spatial congruence (either in terms of direction or location). In short, if two signals are roughly synchronous and are of an appropriate type to fit with a common source, we are likely to try to assign them to a single environmental cause, regardless of our previous experience with the cues. But while this generalization holds true for adult perceivers, further research is required to understand the mechanisms through which they arrive at that state. It remains possible that early experience with speech-related airflow plays a crucial role in forming the links between the cross-modal cues that allow for such a robust pattern of integration. Considerably more research lies ahead of us regarding experience and the developmental trajectory of aerotactile integration.

One important remaining question concerns how specific language experience affects aerotactile integration. One way to address this question is to ask how native speakers of languages without aspiration, such as Spanish, would behave in the task. The voicing contrast for Spanish stops consists of pre-voiced stops and voiceless unaspirated stops. Average VOTs for the two types are around -100 milliseconds for prevoiced stops and around 15-20 milliseconds\(^2\) for the voiceless unaspirated. If experience with aspiration in the language environment is key to developing adult-like aerotactile integration behavior, then native speakers of Spanish would not be predicted to

---

\(^2\) Some variation in VOT has been reported across Spanish dialects, but the differences appear quite small (e.g., with 5 milliseconds). For example, one cross-dialect of Central American Spanish reports a voicing lag for /p/ of 9.8 ms in Guatemalan Spanish and 14 ms Peruvian Spanish (Williams 1977).
integrate airflow as a cue to aspiration in the manner that adult native English speakers do. At present, all published aerotactile integration studies have focused on English. Fortunately, ongoing work testing native speakers of Thai and of Mandarin will shed more light on how language specific the effect is.

A second, and related, step is to investigate how experience with a specific language and its perceptual attributes affects aerotactile integration across the first year of life. It has long been established that experience with a specific language during infancy affects which speech cues we are sensitive to, narrowing our perceptual sensitivity to sounds and contrasts relevant to our native language(s) (Kuhl 1988; Werker & Tees 1984; also see Werker & Gervain 2013 for a review). Significant research provides evidence for this perceptual narrowing and the important role of experience within the first year of life for both unisensory and multisensory perception. In auditory speech perception, for example, 6-8-month-old infants can discriminate many non-native contrasts, but this perceptual sensitivity declines over the first year of life (Werker & Tees 1984; Werker & Lalonde 1988; Kuhl 1998). In audiovisual speech perception, infants can detect audiovisual congruence for both native and non-native syllables at 6 and 9 months, but by 11 months are only sensitive to audiovisual congruence for native syllables (Pons et al. 2009). These examples from both unisensory and multisensory perception have led some researchers to speculate that perceptual narrowing is pan-sensory (Lewkovicz & Ghazanfar 2006). However, it remains unknown whether audio- and visuo-aerotactile perception show early broad multisensory perceptual tuning to both native and non-native speech attributes as seen in audiovisual speech. This is another place where researchers must look to learners of a language without contrastive aspiration. If, contrary to audio, visual, and audiovisual speech perception, aerotactile speech perception does not exhibit an initial broad tuning that narrows through
experience, it would suggest that at least some cross-modal cue integration requires specific language experience to emerge.

Third, it remains to be seen whether aerotactile speech information affects speech beyond the syllable level. In all three studies, I tested the participants on a single pair of CV syllables, a speech perception experience that bears little resemblance to natural communication. Future work must investigate whether aerotactile speech information remains relevant in more ecological tasks, such as with words, phrases or in running speech.

Finally, as compared to audiovisual interactions, research into the tactile dimensions of speech perception is still in its infancy. However, the potential in this field of research is vast; hopefully, it will one day be as widespread as research into audiovisual speech perception is today. This research has already led to innovation in nasalance, turbulent speech airflow recording (Derrick et al. 2014; Derrick et al. 2015) and artificial air flow production techniques (Derrick & De Rybel 2015), and has helped uncover answers to questions about how the skin responds to speech air flow compared to other types of air flow. But most importantly, expanding our focus to include these additional modalities will usher in a truly multisensory approach to speech perception research. The findings reported in this dissertation reveal that different kinds of sensory information may interact with speech and experience in different ways. Future research in cross-modal effects should expand to consider each of these modalities—audition, vision, proprioception, aerotactile somatosensation, and more—both separately and in combination with other modalities in speech. In this way, continued research into cross-modal effects in speech perception will bring a deeper understanding of how humans sense and interact with the world—and, just as importantly, a deeper understanding of how our interactions with the world shape our perception of speech.
References


https://doi.org/10.1111/1467-7687.00052


https://doi.org/10.1016/j.cub.2005.03.046


https://doi.org/10.3109/14722240410001700258


Bell, A. M. (1867). *Visible speech: The science of universal alphabetics or, self-interpreting physiological letters, for the writing of all languages in one Alphabet*. Simpkin, Marshall & Company.


[https://doi.org/10.1121/1.4778348](https://doi.org/10.1121/1.4778348)


Fowler, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *The Journal of the Acoustical Society of America, 89*(6), 2910–2915. https://doi.org/10.1121/1.400729


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


https://doi.org/10.1111/j.1467-8624.2005.00866.x


https://doi.org/10.1017/S0140525X16000911


https://doi.org/10.1080/02699200500195696


https://doi.org/10.1111/1467-9280.02435


https://doi.org/10.1038/264746a0

https://doi.org/10.1016/j.cub.2007.08.064

https://doi.org/10.3109/02699206.2015.1079247


https://doi.org/10.1007/s00221-005-0071-5


https://doi.org/10.1017/S0140525X12001495


Appendix: Additional Methodological Details for Chapter 4

This appendix provides additional details regarding the stimuli and trial ordering used in the infant experiment in Chapter 4.

<table>
<thead>
<tr>
<th>Stimulus Number</th>
<th>Stimulus</th>
<th>Trial Type</th>
<th>Motivation for prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>paPuff + ba</td>
<td>Alternating</td>
<td>Aero-tactile information reinforces existing distinction</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Aero-tactile information influences perception so /ba/ is /pa/-like</td>
</tr>
<tr>
<td>2</td>
<td>baPuff + ba</td>
<td>Alternating</td>
<td>Aero-tactile information influences perception so /ba/ is /pa/-like</td>
</tr>
<tr>
<td>3</td>
<td>baPuff + pa</td>
<td>Non-Alternating</td>
<td>Aero-tactile information influences perception so /ba/ is /pa/-like</td>
</tr>
<tr>
<td>4</td>
<td>paPuff + pa</td>
<td>Non-Alternating</td>
<td>Aero-tactile information reinforces existing distinction</td>
</tr>
</tbody>
</table>

Counterbalancing Details

There are four conditions in order to counterbalance the order of the puff (puff begins on first syllable or puff begins on second) and the order of the stimulus type (alternating initial, non-alternating initial). The four orders are described below and the numbers in the blocks refer to the stimulus sets listed in the table. If a participant is in Order 1 or 2, the first block will be alternating initial and the second block will be non-alternating initial (with the reverse order for participants in Orders 3 and 4). The counterbalancing for the puff occurs similarly across blocks: for participants in Orders 1 and 3, the first block has the order puff/no puff in the stimulus, while the second block has the order no puff/puff. Finally, for all groups, the first block always begins with stimulus streams 1 and 4, because those are where we expect to see the strongest effect.
**Condition 1:** (n=6, 3 female):

Block 1: Alternating Initial, puff initial

1-4-2-3-1-3-2-4

Block 2: Non-Alternating Initial, puff final

4-2-3-1-3-2-4-1

**Condition 2:** (n=6, 3 female):

Block 1: Alternating Initial, puff final

1-4-2-3-1-3-2-4

Block 2: Non-Alternating Initial, puff initial

4-2-3-1-3-2-4-1

**Condition 3:** (n=6, 3 female):

Block 1: Non-Alternating Initial, puff initial

4-1-3-2-3-1-4-2

Block 2: Alternating Initial, puff final

2-4-1-3-2-3-1-4
**Condition 4:** (n=6, 3 female):

Block 1: Non-Alternating Initial, puff final

4-1-3-2-3-1-4-2

Block 2: Alternating Initial, puff initial

2-4-1-3-2-3-1-4