CROSS-MODAL EFFECTS IN STATISTICAL LEARNING

A Dissertation in Psychology
by
Aaron D. Mitchel

©2011 Aaron D. Mitchel

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

May 2011
The dissertation of Aaron D. Mitchel was reviewed and approved* by the following:

Daniel J. Weiss  
Associate Professor of Psychology and Linguistics  
Dissertation Advisor  
Chair of Committee

Judith F. Kroll  
Distinguished Professor of Psychology, Linguistics, and Women's Studies

Reginald B. Adams  
Assistant Professor of Psychology

Chip Gerfen  
Associate Professor of Spanish and Linguistics

Melvin M. Mark  
Professor of Psychology  
Head of the Department of Psychology

*Signatures are on file in the Graduate School.
ABSTRACT

A central question of research on language acquisition concerns which types of information in the environmental input are available to language learners, as well as the extent to which learners utilize this information. Although research examining this issue has traditionally focused on the nature of the auditory input to language learners, there is a growing body of research indicating that the language learning environment is multimodal and that processes supporting language acquisition exploit cues in the visual domain (e.g. cues in the speaker’s face). In this dissertation, I examined how mechanisms of early language acquisition operate over multimodal input, focusing on processes underlying speech segmentation, an early obstacle faced by language learners. One mechanism believed to support speech segmentation is statistical learning, in which learners use distributional regularities in the input to identify word boundaries. Few studies have examined statistical learning in a multimodal context, despite evidence that perception is fundamentally multisensory. Thus, across a series of three studies, I investigated the interaction of visual and auditory input during statistical learning.

In Chapter II, I tested adults’ ability to use distributional cues to segment streams of tone and shape triplets presented simultaneously. Learners were able to segment each stream so long as the triplet boundaries aligned across streams. When the streams were misaligned, performance dropped to chance. In Chapter III, I used an illusion arising from audiovisual integration (the McGurk effect) to alter the statistical representations of two languages, indicating that learners can integrate audio and visual input during statistical learning. Finally, in Chapter IV, I found that learners were able to use facial cues alone to segment a speech stream. The results of these three studies provide evidence of cross-modal effects on statistical learning,
suggesting that statistical learning does not occur independently across modalities. In addition, these results provide further evidence for the relevance of visual input, in particular cues in the speaker’s face, for processes supporting language acquisition. I conclude the dissertation with a discussion of how these findings inform models of statistical learning, as well as potential broader applications of these results for atypical language development.
TABLE OF CONTENTS

List of Tables........................................................................................................vii
List of Figures........................................................................................................viii
Acknowledgements..............................................................................................x

CHAPTER I. Introduction..........................................................................................1
  Nature of the Input to Learners...........................................................................2
  Modality Specificity of Statistical Learning.........................................................7
  The Present Study: Segmentation in a Multimodal Context...............................15

CHAPTER II. Learning across senses: Cross-modal effects in multisensory statistical learning....................................................................................................18
  Experiment 1a: Tone Sequence Alone.................................................................21
  Experiment 1b: Visual Sequence Alone..............................................................25
  Experiment 2: Correlated, Synchronized Streams..............................................27
  Experiment 3: Uncorrelated, Synchronized Streams..........................................30
  Experiment 4: Uncorrelated, Desynchronized Streams......................................33
  General Discussion.............................................................................................36

CHAPTER III. Cross-modal effects in statistical learning: Evidence from the McGurk Illusion.......................................................................................................42
  The McGurk Effect.............................................................................................43
  Experiment 5.......................................................................................................46
  Experiment 6.......................................................................................................55
List of Tables

Table 1. The artificial languages used in Experiments 5 and 6. Bolded syllables in the audiovisual condition represent the illusory, integrated McGurk syllables. Syllable-to-syllable transitional probabilities are reported for each language in each condition. Transitional probabilities in the audiovisual conditions reflect the statistical structures of the languages if they include the integrated percept.
List of Figures

Figure 1. (a) The shapes used in Fiser & Aslin (2002). The shapes are grouped into the statistically defined triplets. (b) Example temporal sequence of the visual stream in Fiser & Aslin (2002).

Figure 2. Comparison of performance for speech (Saffran et al., 1996b) and tone stimuli (Saffran et al., 1999). Adapted from Saffran et al. (1999).

Figure 3. Artificial grammar used in Conway and Christiansen (2005).

Figure 4. Percent correct identification in a two-alternative forced-choice task across four Experiments. The bar represents chance (50%).

Figure 5. This figure illustrates the correspondence and boundary alignment across modalities in Experiment 2 (panel A), Experiment 3 (panel B), and Experiment 4 (panel C).

Figure 6. Design of the artificial languages used in Experiments 5 and 6. The syllable-to-syllable transitional probabilities are noted for each condition and language. The transitional probabilities in the audiovisual conditions reflect the statistical structure of the languages if participants perceived the McGurk illusion.

Figure 7. Still frame of the video display in Experiments 5 & 6.

Figure 8. Mean d-prime scores for the audio-only and audiovisual display conditions for Language A (panel a) and Language B (panel b) in Experiments 5 and 6, respectively.

Figure 9. Mean endorsement rates (i.e. probability of saying ‘yes’) for the different types of items across conditions for Experiment 5 (left panel) and Experiment 6 (right panel). Significant comparisons (p<.05) are denoted with *, and marginally significant comparisons (p<.10) are denoted with †.

Figure 10. Mean d-prime scores in Experiments 5 and 6, illustrating the crossover interaction of language and display type.

Figure 11. Mean endorsement rates for the two audio-consistent test items across display conditions for Experiment 5 (left panel) and Experiment 6 (right panel).

Figure 12. (a) The words and part-words from the artificial language in Experiment 7. (b) The statistical structure of the artificial language, depicting the transitional probabilities between syllables.

Figure 13. Still frames from the visual stimuli used in Experiments 8 and 9. The aware condition is on the left, and the misinformed condition is on the right.
Figure 14. Results of Experiment 8, visual-only. The results of the aware condition are on the left (blue bar) and the results of the misinformed condition are on the right (green).

Figure 15. Results of Experiment 7 (Audio-only) and Experiment 9 (Aware and Misinformed audiovisual conditions).
ACKNOWLEDGMENTS

Many individuals have helped shape this dissertation and my career, so much so that this section can only begin to adequately express my gratitude for their assistance. First, I would like to acknowledge financial support for this research from the National Institutes of Health Grant R03-HD048996-01 to Dan Weiss, a teaching release award from the Center for Language Science at The Pennsylvania State University, and the Herschel W. Leibowitz Graduate Travel Scholarship from the Department of Psychology at The Pennsylvania State University.

I am thankful for the guidance and support of my wonderful dissertation committee: Dan Weiss, Judy Kroll, Reg Adams, and Chip Gerfen. Dan, thank you for your advice and support throughout my graduate career. You have provided invaluable direction that has greatly shaped my professional development, guiding me to become a better researcher, a more confident speaker, and a more proficient writer. I have truly enjoyed working with you these past five years, and I hope to continue our relationship as colleagues for many years. Judy, thank you for your wisdom and guidance throughout my career. I have always appreciated your advice for my research and for my career. Reg, I have greatly appreciated your suggestions and contributions to my research—you have helped to point me in the right direction as I forge into new domains of research. Finally, Chip, in many ways I have considered you to be a second adviser, and I am extremely thankful for your support over the years—from stimuli creation to how to navigate graduate school. Thank you all for the time and energy you have dedicated to guiding my research and professional development.

This dissertation research would not be possible without the assistance of numerous research assistants over the years: Sarah, Affalter, Pete Alexopolous, Allison Allman, Mark Basile, Beth Buerger, Rebekah Crandall, Nicole Freed, Roseanne Fry, Molly Jamison, Megan
Jordan, Jan Schober, Jane Shin, and Rachel Sweeney. I would like to specifically thank the assistants that allowed me to videorecord their faces to create the stimuli used in this experiment, Brian Giallorenzo and Kevin Weiss. To my fellow members of the Comparative Communication Laboratory, Katie Chapman, Helen Marie Graves, and Tim Poepsel: I thank you for your support both in and out of the lab. I also wish to thank members of the Cognitive Psychology area and the Center for Language Science for their support of my research over the years. In particular, I would like to thank Rich Carlson and David Rosenbaum for their open doors and willingness to answer questions and provide feedback and advice over the years. I would like to acknowledge Morten Christiansen for his collaboration on the research presented in Chapter III, as well as for his feedback and advice on numerous other projects.

To my undergraduate adviser, Brooke Lea: thank you for inspiring me to pursue a degree in Cognitive Psychology. You have been an excellent mentor, colleague, and friend throughout the years.

Thank you to my family for providing unending support and encouragement. Mom and Dad, thank you for cheering me on throughout my many endeavors, and for always being there when I needed an understanding ear and a calming voice. Regina, Dill, and Jess, thank you for welcoming me into your family and for your support and understanding.

Finally, I would not have reached the end without the love and support of my wife, Marissa. I thank you for everything.
CHAPTER I:

Introduction

A fundamental question of language acquisition is how language learners are able to segment a continuous auditory speech stream into discrete units, or words. A growing body of research indicates that the ability to track transitional probabilities across units of speech (hereafter statistical learning) plays a key role in resolving the segmentation problem (e.g., Saffran, Aslin, & Newport, 1996a; Saffran, Newport, & Aslin, 1996b). Despite a wealth of research on statistical learning, there are few studies that have investigated how this mechanism operates over multimodal stimuli, a gap in the literature that is particularly striking given that multisensory integration is a fundamental aspect of perception (Stein & Stanford, 2008).

Language acquisition has long been viewed as an auditory phenomenon, but speech is multimodal, available to both eye and ear (Massaro, 1998; Rosenblum, 2005); thus, here I systematically test how statistical learning mechanisms operate over multimodal stimuli. Pursuant to this objective, the present studies investigate three fundamental questions that will address the role of visual information in speech segmentation. First, is statistical learning in one modality independent of stimuli in another modality? Second, to what extent is information from a visual display of talking faces integrated with an auditory stream during statistical learning? Finally, what are the independent contributions of faces to speech segmentation (i.e. can learners use facial cues alone to segment a continuous speech stream)?

This thesis will begin with an overview of theories on language learning, focusing on one aspect of language acquisition—speech segmentation. I will then describe research that has addressed the question of how statistical learning guides speech segmentation, as well as how statistical learning mechanisms operate over non-speech stimuli from a number of sensory
domains. In chapters 2-4, I will describe a program of research that employed behavioral methods to investigate multisensory integration in statistical learning and whether stimuli in the visual domain (facial cues) alter learning in the auditory domain. Finally, in Chapter 5, I will discuss how the findings from this research inform our understanding of speech segmentation, and, by extension, the course of language acquisition.

**Nature of Input to Language Learners**

Language acquisition is a remarkable feat, as children rapidly acquire a highly complex linguistic system with seemingly little effort. The exceptionality of the human language faculty has led to considerable debate regarding its origin. Nativist approaches assert that infants are innately endowed with at least some linguistically specific knowledge or learning mechanisms (e.g. Crain, Goro, & Thornton, 2006). Conversely, in empiricist or emergentist approaches, no innate linguistic knowledge is assumed; rather, infants induce linguistic structure from environmental input through domain-general learning mechanisms. Much of the debate between these accounts has centered on the amount of linguistic information in the environment. Specifically, nativists often assert that there is insufficient information in the input to support language acquisition (Chomsky, 1980; Crain & Petroski, 2002).

This position, known as the *argument from the poverty of the stimulus* (POS), is typically identified with Chomsky (1980), who argued that our linguistic knowledge surpasses what is evidenced in the language input. POS arguments claim that certain types of linguistic constructions (e.g. negative polarity items, such as “No linguist with any brains admires Chomsky,” Crain & Pietroski, 2002) are too infrequently produced in the input to be learned from the environment, suggesting that language acquisition is at least partially supported by innate knowledge. Furthermore, it is argued that the environmental input does not contain
sufficient *negative* evidence to account for ungrammaticality judgments (for a review, see Pullum & Scholz, 2002). That is, the language input does not contain examples of ungrammatical constructions and yet language learners reliably detect ungrammatical novel sentences. For example, based on constructions available in the language input, children might extract a very general rule that the auxiliary *is* can be reduced to ’s (e.g. *Jim’s happy*); yet, this rule does not apply in all contexts (e.g. the grammatical sentence *Jim is taller than Kim is* cannot be reduced to *Jim’s taller than Kim’s*). According to POS arguments, children do not ungrammatically overgeneralize this rule, even though there is limited evidence in the input that the rule does not apply to every context (Lightfoot, 1998). Nativist accounts claim that the absence of such positive and negative linguistic evidence precludes the possibility of acquiring a language through experience. Empiricists, on the other hand, contend that much of this linguistic evidence can be extracted from systematic properties of the input, suggesting that the language environment is not as impoverished as was previously believed (Behrens, 2006).

Recently, theorists have proposed that the abundance of information in language input about linguistic structure “makes innate stipulations unnecessary” (Behrens, 2006). In particular, a growing body of research indicates that distributional or statistical properties of the input provide insight into the structure of languages (see Saffran, 2003). This view has been supported by evidence that infants and adults are able to use these distributional cues to acquire syntactic categories (Redington, Chater, & Finch, 1998), native phoneme boundaries (Maye, Werker, & Gerken, 2002, Maye, Weiss, & Aslin, 2008), lexical development (Li, Farkas, & MacWhinney, 2004), verb transitivity assignment (Scott & Fisher, 2009), and rudimentary grammar (Gomez & Gerken, 1999). Such findings support the idea that early language development may employ statistical learning mechanisms. These learning mechanisms track the statistical properties of the
input, such as the distribution of words across contexts or the co-occurrence of linguistic features, properties which had largely been excluded from early accounts of language environment (Christiansen, Allen, & Seidenberg, 1998). One aspect of language acquisition in which probabilistic cues have been investigated at length is speech segmentation. The studies in this field illustrate how statistical learning operates in language acquisition.

**The Segmentation Problem.** One of the earliest obstacles confronting infant language learners is how to isolate word boundaries from continuous speech. Speech segmentation is a critical step in early language acquisition, since infants must determine which combinations of sounds represent words prior to the emergence of subsequent aspects of language, such as mapping meaning to expressions and forming syntactic categories. Additionally, since children have no a priori knowledge that, for example, *blanket* is a single word in English and not two words, *blan* and *ket*, word boundaries represent a feature of language that must be acquired through experience.

Speech segmentation is a significant hurdle for language learners because there are no invariant acoustic cues (e.g. stress) that mark word boundaries in all languages (Klatt, 1979). Pauses in speech rarely co-occur with word boundaries, and thus are not reliable cues to word segmentation (Klatt & Stevens, 1973; Reddy, 1976). Moreover, although there are a number of other acoustic cues to word boundary, including stress (Jusczyk, Houston, & Newsome, 1999; Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000), allophonic variation (Jusczyk, Hohne, & Bauman, 1999), and phonotactic cues (Friederici & Wessels, 1993). However, none of these cues are available to all languages, nor do they suffice as independent segmentation strategies. For example, languages vary in which syllable is typically stressed (e.g. word-initial stress in English vs. penultimate stress in Spanish). Thus, infant learners must identify where stress falls in the
language system prior to effectively using it as a segmentation cue. As an alternative to acoustic cues to word boundary, evidence suggests that learners may initially rely on statistical properties inherent in all languages to identify word boundaries, which then provides a scaffold allowing language learners to identify language-specific acoustic cues, such as stress patterns (Thiessen & Saffran, 2003).

**Statistical learning.** A well-documented source of linguistic information is the statistical structure of sound correspondences in the speech that infants hear. In particular, researchers have examined how language learners track the conditional statistic of transitional probability, which is defined as the probability that two sounds co-occur, relative to the overall frequency of occurrence \( P(Y|X) = \frac{\text{frequency of } XY}{\text{frequency of } X} \); Aslin, Saffran, & Newport, 1998). For example, in the often-cited phrase *pretty baby*, the sound *pre* occurs relatively infrequently in English, and in most cases is followed by the sound *ty* (approximately 80% in infant-directed speech; Saffran, 2003). On the other hand, since the sound *ty* occurs at a word boundary, it can be followed by any number of syllables (e.g. *pretty puppy, pretty flower*) and there is only a small likelihood that it will be followed by *ba* (approximately 0.03% in infant-directed speech; Saffran, 2003). Thus, the transitional probability within words (\( pre \rightarrow ty, \ ba \rightarrow by \)) is relatively high when compared with the transitional probability between words (\( ty \rightarrow ba \)). Such local minima of transitional probabilities provide universally available cues to word boundary (Swingley, 1999), and it is possible to use such cues to segment speech.

In a classic series of studies, Saffran and colleagues (1996a, b) provided the initial demonstration of this ability in adults and infants. Saffran et al. (1996b) familiarized adults to a continuous speech stream of an artificial language comprised of six trisyllabic words (*babupu, bupada, dutaba, patubi, pidabu, and tutibu*). The authors stripped the speech stream of all
potential acoustic word boundary cues; consequently, only the statistical structure of the speech stream afforded a means for segmentation, since the transitional probabilities within the word were always higher than the transitional probabilities between words. Participants listened to the synthesized artificial language for 21 minutes, and were then given a two-alternative, forced-choice task between either words and non-words or words and part-words. Part-words consisted of the final syllable of one word and the first two syllables of another word (e.g. *pubupa* from *babupu* and *bupada*), thus part-word foils occurred during familiarization and only differed from statistically defined words by one syllable. Adults were able to accurately discriminate statistically defined words from both part-words and non-words, suggesting that they are able to track of the statistical regularities in speech and then use this information to correctly identify word boundaries (Saffran et al., 1996b).

Using a headturn preference procedure (see Jusczyk & Aslin, 1995), Saffran and colleagues (1996a) extended their previous findings to infants, testing 8-month-olds’ ability to segment speech using statistical learning. In this task, infants listened to a two-minute familiarization passage of a simplified version of the artificial language presented to adults. Following the familiarization stream, infants were tested on four tri-syllabic strings (two words, two part-words), which were repeatedly presented on either the left or right of the infant, accompanied by a red, flashing light. Infants displayed a novelty preference, exhibited by longer looking times for part-words relative to statistically defined words, suggesting that the infants had segmented the speech stream on the basis of transitional probabilities and subsequently became habituated to the statistically-defined words.

The results of these studies demonstrate that a wealth of information exists in the linguistic input to learners, and that language learners may utilize these cues during speech
segmentation, suggesting that statistical learning mechanisms may play a key role in language acquisition. However, it is possible that the linguistic statistical learning mechanism is a component of the human language faculty, or, at a minimum, is subject to language-specific constraints (Yang, 2004). It is therefore necessary to examine the nature of the mechanism or mechanisms that govern statistical learning. Is the mechanism revealed by Saffran and colleagues (1996a,b) unique to language, or is it a domain-general learning mechanism used to track sequential dependencies in linguistic and non-linguistic stimuli? In order to test these competing hypotheses, researchers have examined segmentation of non-linguistic stimuli in different sensory modalities to determine whether stimuli in separate modalities are processed by a single mechanism or by several separate mechanisms.

**Modality Specificity of Statistical Learning**

To determine whether statistical learning mechanisms extend beyond speech stimuli, Saffran, Johnson, Aslin, and Newport (1999) tested adults’ and infants’ capacity to segment a sequence of musical tones. Adapting their previous speech segmentation study (Saffran et al., 1996a, b), the authors familiarized participants to an input stream comprised of tones grouped into triplets (analogous to the artificial speech stream used previously). The tone sequences were created from pure tones with frequencies corresponding to musical notes (e.g. A, A#, G) within a single octave range, avoiding any standard musical structures (i.e. major/minor chords) or familiar three-tone sequences (e.g. the NBC chimes). The structure of the transitional probabilities was identical to Saffran et al. (1996b and 1996a for adults and infants, respectively), and was the only source of boundary information. In the experiments with adults, participants listened to the familiarization stream for 21 minutes and were then tested with a two-alternative, forced-choice task between tone triplets and tone part-triplets (the 3rd segment of one
tone triplet followed by the 1st and 2nd segment of another tone triplet). In this task, adults reliably identified triplets that were consistent with statistical structure of the familiarization stream. To test 8-month-old infants’ capacity for tone segmentation, the authors familiarized infants to the tone stream for 3 minutes, and then used a head-turn-preference procedure to gauge whether the infants had successfully segmented the tone stream. Infants displayed a significant novelty preference for part-triplets compared with triplets, indicating that infants are able to detect the boundaries between tones using sequential dependencies. The results of this study suggest that the statistical learning mechanism previously demonstrated in a linguistic domain is not specific to speech, since it is activated by non-linguistic stimuli.

A question that emerged from this study is whether this statistical learning mechanism, which computes temporally-ordered dependencies, is specialized for auditory input. Can adults compute sequential, temporal dependencies in the visual domain? Fiser and Aslin (2002) addressed this by adapting the aforementioned statistical learning task to a visual paradigm, familiarizing adults with an input stream comprised of non-canonical, 2D black shapes (see Figure 1a) that moved across the screen in a smooth trajectory. Each segment consisted of one complete movement of a shape emerging from behind a center pillar and travelling to the edge of the window in a horizontal path, and then returning on the same path, with the movement

Figure 1a. The shapes used in Fiser & Aslin (2002). The shapes are grouped into the statistically defined triplets.

Figure 1b. Example temporal sequence of the visual stream in Fiser & Aslin (2002).
concluding behind pillar (see Figure 1b). The subsequent segment in the sequence would then begin as the next shape would emerge on the opposite side of the pillar, move to the edge of the screen and return.

The shapes were grouped into four base triplets (see Figure 1a). These four triplets were then concatenated into a continuous stream. The shape sequence was constructed so that within-triplet transitional dependencies were consistently high (1.0), while transitional probabilities bridging triplet boundaries were consistently low (0.33); thus, transitional probabilities between shape elements provided the only cue to stream segmentation (similar to experiments with tones [Saffran et al., 1999] and speech [Saffran et al., 1996b]). The sequential dependencies here consisted of which shape would appear next in the sequence, rather than where the shape would appear. Thus, in order to segment the base triplets from the continuous stream, participants were required to track temporal-order dependencies in the visual modality. After 6 minutes of familiarization, adults were able to identify shape triplets when tested against statistically inconsistent part-triplets. This ability has also been found in infants as young as two months of age (Kirkham et al., 2002). These results provide evidence that learning mechanisms that extract temporal-order statistics are not specialized for the auditory domain.

Beyond the visual and auditory domains, statistical learning has been demonstrated in the motor domain. Hunt and Aslin (2001) adapted a Serial Reaction Time (SRT) task to be analogous to speech segmentation studies, with fixed within-unit (similar to within-word) structures and variable ordering between units. Participants performed a sequence of movements that were grouped into triplets (i.e. Movement A, Movement B, Movement C would comprise triplet 1) so that the transitional probabilities among movements was high within triplets and low between triplets. Hunt and Aslin found that participants’ reaction time reflected the statistical
structure of the movement sequence, with faster reaction times for within-triplet transitions relative to between triplet transitions, indicating that adults are able to use sequential dependencies to segment movement sequences.

These results demonstrate the ability to use temporal-order adjacent regularities to segment sequential input across a number of sensory domains and stimuli types, underscoring the wide scope of statistical learning. However, though statistical learning is not confined to linguistic input, the specific mechanism or mechanisms supporting this ability may be tailored to language stimuli (Marcus, Johnson, Fernandes, & Slemmer, 2004; Marcus, Fernandes, & Johnson, 2007). According to some nativist perspectives (e.g., Yang, 2004), the mechanism operating on linguistic stimuli (e.g. Saffran et al., 1996a, b) is independent of the mechanisms operating on stimuli in other sensory modalities, and is constrained by parameters of an innate language endowment. In contrast, empiricist frameworks posit that knowledge, across domains, is acquired by a common set of mechanisms and constraints (Seidenberg, 1997). A question that stems from this debate, then, is whether statistical learning is supported by a domain-general mechanism, or by several independent mechanisms. In the next section, I review arguments in favor of a unified, modality-general statistical learning mechanism as well as counter-arguments that propose a distributed, modality-specific account of statistical learning.

Evidence for domain-general statistical learning. Investigations into the stimulus specificity of statistical learning have focused primarily on comparing performance in segmentation tasks across modalities. Proponents of a unified model of statistical learning (e.g. Kirkham et al., 2002) have pointed to cross-modal similarities in the level of performance, the developmental trajectory of statistical learning abilities, and the types of sequential dependencies that can and cannot be learned. First, performance in the statistical learning tasks reported above
has been comparable, regardless of the modality of the stimulus (i.e., 62% for speech [Saffran et al., 1996b], 65% for tones [Saffran et al., 1999], and 66% for shapes [Fiser & Aslin, 2002])\(^1\). In a direct comparison of performance across experiments (see Figure 2), Saffran and colleagues (1999) found no significant difference in the level of performance on tone stimuli from the performance reported in previous studies using speech stimuli with adults (Saffran et al., 1996b) or with infants (Saffran et al., 1996a). The authors interpret this as evidence for similar underlying mechanisms for tones and speech stimuli.

Second, infant studies have revealed that statistical learning comes online at roughly the same age across stimulus modalities. By 8 months old, infants are able to track the temporal-order dependencies in visual (Kirkham et al., 2002), non-linguistic auditory (Saffran et al., 1999) and speech stimuli (Saffran et al., 1996a). The similar developmental trajectory for the ability to compute temporal dependencies suggests that a single, modality general mechanism governs this ability.

Finally, there is evidence that the types of statistical computations that can be performed are similarly constrained across linguistic and non-linguistic stimuli. Newport and Aslin (2004) tested adult learners’ ability to segment speech on the basis of non-adjacent dependencies. Prior work focused on learners’ ability to utilize dependencies between adjacent syllables. Instead,

---

\(^1\) The results of Hunt & Aslin (2001) are not included in this comparison, since the test was not analogous to other statistical learning tasks.
Newport and Aslin familiarized adults to speech streams in which word boundary was marked by dependencies between non-adjacent elements. When the relevant relationships were among non-adjacent segments (i.e. consonant-consonant or vowel-vowel), participants successfully segmented the speech stream. However, when the relevant relationships were among non-adjacent syllables (i.e. CV₁ perfectly predicts CV₃, with a random intervening CV syllable [CV₁ CVₓ CV₃ or CV₁ CVᵧ CV₃]), participants failed to extract this structure and performance was at chance. Interestingly, these results mirror the properties of natural languages, as non-adjacent dependencies at the syllabic level are rare among languages, while non-adjacent dependencies at the segmental level are relatively common (Newport & Aslin).

Newport and Aslin (2004) suggest two possible explanations for this discrepancy between segments and syllables. First, it could be a basic property of speech perception, such that humans tend to favor one type of relationship (phonemic segments) over the other (syllables). Alternatively, statistical learning could be subject to domain general, Gestalt principles of perceptual similarity. That is, the challenge of learning non-adjacent dependencies can be overcome if the non-adjacent elements share some feature that contrasts with the interleaving elements, providing a grouping cue that facilitates the tracking of the relevant relationships. In Newport and Aslin, learners were able to compute non-adjacent dependencies between segments because the statistically relevant elements could be grouped by a shared feature (e.g. all relevant elements were consonants) that was not shared by the irrelevant elements (e.g. all irrelevant elements were vowels).

To test these hypotheses, Creel and colleagues (2004) presented learners with non-linguistic tone stimuli (similar to Saffran et al., 1999) that were statistically patterned by non-adjacent regularities. Each “triplet” was comprised of three tone pairs, totaling six tones per
triplet. In this way, the study mirrored Newport and Aslin (2004), with each tone representing a speech segment (i.e. D₄D₆F₄F₆G₄C₄ in Creel et al. corresponded to CV.CV.CV in Newport & Aslin). When the tones all fell in the same octave range, participants were able to acquire regularities among adjacent tones, but failed to acquire regularities among non-adjacent tones. However, when a perceptual grouping cue was introduced by interleaving tones from two different octave ranges (e.g. D₆D₄F₆F₄G₆C₄), participants successfully acquired non-adjacent regularities while failing to acquire adjacent ones. Thus, regardless of whether the stimuli consisted of tones or speech, learners were only able to segment non-adjacent patterns in the presence of a grouping cue (c.f. Pena, Bonatti, Nespor, & Mehler, 2002; Bonatti, Pena, Nespor, & Mehler, 2005). This pattern of results indicates that statistical learning is globally constrained by general Gestalt perceptual principles of similarity, suggesting that the same statistical learning mechanism operates on speech and non-linguistic auditory stimuli.

Evidence for modality-specific statistical learning. To this point, the studies reviewed above have supported the view that statistical learning is a modality general mechanism by comparing segmentation across modalities. However, this view has recently been challenged by evidence of modality constraints on statistical learning. Conway and Christiansen (2005) argued that disparate methodologies and the lack of controlled procedures and stimuli across segmentation experiments may have masked subtle distinctions in the types of computations made and the overall proficiency of statistical learning across different modalities. Consequently, they tested the ability to segment sequential input in the visual, auditory, and tactile modalities while carefully limiting any differences in methodology across modalities.

In contrast to the aforementioned statistical learning studies, Conway and Christiansen employed an artificial grammar learning task. Artificial grammar tasks (e.g. Reber, 1967) use
simple, finite-state grammars (see Figure 3) to generate sequences of stimuli. Beginning at a
starting position (S1 in Figure 3), strings are created by following allowable paths until reaching
an exit position (S4 or S5 in Figure 3). For example, a legal sequence produced by the
grammar in Figure 3 might be 1-2-1-1-3, generated by moving along the following path: S1-S3-S2-S2-S4. Conway & Christiansen (2005) used this grammar to create a series of audio, visual,
and tactile sequences which were then presented to participants in a between-subjects design (i.e.
participants were only familiarized to one type of
stimuli). The visual stimuli consisted of a black
square that appeared on the screen at one of 5
possible locations in a horizontal array. The
location of the block was constrained by the
grammar (e.g. a legal sequence might be positions
1-2-5-5). The audio stimuli were comprised of
five pure tones that varied in frequency (pitch). Finally, the tactile stimuli were pulses on the
tips of the fingers of one hand. Similar to the visual sequences, the ordering of the tones and the
pulses were constrained by the same artificial grammar used in the visual condition.

After the familiarization phase, participants were tested on novel sequences (using the
same set of stimuli) that were either legal or illegal sequences within the artificial grammar. In
this task, Conway and Christiansen (2005) found quantitative modality constraints on learning,
evidenced by an asymmetry in performance, with an advantage for auditory stimuli (tones)
relative to visual or tactile stimuli. Participants in the auditory condition were more accurate at
categorizing test sequences than participants in the visual or tactile conditions. In addition, the
authors found qualitative differences in the types of computations performed in each modality.
Auditory input resulted in a greater bias toward sequence-final information, whereas there was a bias toward sequence-initial information in the tactile modality. Thus, while participants in the auditory condition were selectively attending to tones that occurred more often at the end of a sequence and using these tones to guide identification of legal sequences, participants in the tactile modality were focusing on pulses that occurred more regularly in the beginning of a sequence to guide identification. Given these quantitative and qualitative differences between modalities, the authors argued for the existence of multiple, parallel statistical learning mechanisms that are directly linked to the sensory modality of the input.

**The Present Study: Segmentation in a Multimodal Context**

Until recently, the issue of modality specificity in statistical learning has been tested by comparing results across individual modalities. Alternatively, an approach to examining this issue that more appropriately reflects the complexity of natural learning environments (see Shams & Seitz, 2008) while avoiding making comparisons across conditions or experiments is to test statistical learning in a multimodal context. To this point, to the best of my knowledge, only a single study has examined statistical learning of multimodal input. Using a multisensory statistical learning task in which learners were asked to simultaneously segment an audio and visual input stream, Seitz and colleagues (2007) found that participants were not only able to simultaneously segment each stream, but they also found no evidence that statistical learning in one modality was influenced by learning in the other modality. That is, participants' ability to segment either stream did not differ when the streams were presented in isolation or when the streams were presented concurrently. From these results, the authors propose that statistical learning is modality independent (i.e., learning in each modality occurs independently, and for that reason is unaffected by stimuli in other modalities). For example, in this view, audiovisual
stimuli are processed as an audio stream and a visual stream, and the ability to learn the audio stream should not be influenced by the visual stream, or *vice versa*. This account, therefore, makes the strong prediction that cross-modal effects should not be observed for multisensory statistical learning.

However, these predictions are not consistent with the growing literature of perceptual studies that support the existence of interactive, integrated perceptual systems (e.g. Stein & Stanford, 2008; Rosenblum, 2005; Ghazanfar & Schroeder, 2006; Calvert et al. 2004; Shimojo & Shams, 2001). Across a number of cognitive domains, there is behavioral evidence for cross-modal interactions in the form of multisensory illusions (for a review, see Shimojo & Shams, 2001) and multisensory advantages in perception (see Massaro, 1998; Shams & Seitz, 2008). In addition, there is neural evidence that perception may be fundamentally multisensory, as regions of the brain that were previously believed to be unimodal sensory cortices (e.g. primary auditory [A1] and primary visual [V1] cortices) have been shown to respond to multisensory input (e.g. Calvert et al., 1997; Ghazanfar & Schroeder, 2006). Given the ubiquity of cross-modal interactions at both the behavioral and neural level, in Chapters II and III, I test the modality independent account of Seitz and colleagues (2007) by examining cross-modal effects in two different multisensory statistical learning paradigms.

In Chapter II, we extend the multisensory paradigm used by Seitz et al. (2007), systematically investigating whether simultaneous segmentation of a tone and shape stream is affected by the cross-modal relationships between the streams. I vary both the correlation between individual elements as well as the alignment of boundary information across streams. If statistical learning is modality independent, then these manipulations should have no impact on segmentation. Alternatively, if statistical learning across modalities is interactive, then these
manipulations should alter segmentation performance.

In Chapter III, I provide a further test of modality independence by incorporating a McGurk illusion, a classic demonstration of audiovisual integration (McGurk & Macdonald, 1976) into a statistical learning task. I present learners with an audiovisual speech stream that contains inconsistencies between the audio and visual input designed to elicit a McGurk illusion (see below). Across two experiments, the McGurk illusion, if perceived, should alter the statistical structure of an artificial language to either remove or enhance transitional probability cues to word boundary. However, if statistical learning is modality independent, then the McGurk illusion should not interact with the statistical representations of the speech streams and therefore should not impact segmentation performance.

Finally, in Chapter IV, I assess whether learners can use facial cues to segment speech. While recent research has examined how facial cues interact with other segmentation cues (Mitchel & Weiss, 2010; Sell & Kaschak, 2009), to the best of my knowledge, no study has explored the independent contribution of facial cues to speech segmentation. If, as I hope to demonstrate in Chapter III, facial cues are integrated with the speech stream at the point of statistical learning, then I might expect these cues to play a role in segmentation. Specifically, visual prosody (i.e. visual cues [e.g. head nodding and lip aperture] to acoustic prosodic structure [e.g. pitch, stress, and rhythm]; Blossom & Morgan, 2006) may signal word boundaries. Learners can use acoustic prosodic cues (pitch and stress) to segment speech (Jusczyk et al., 1999). Furthermore, observers can extract pitch and stress cues from talking face displays (with no accompanying audio signal; Yehia et al., 2002). Thus, in Chapter IV, I examine whether learners can use these visual analogs of prosodic structure (visual prosody) to segment a speech stream.
CHAPTER II

Learning across Senses: Cross-modal Effects in Multisensory Statistical Learning

(Article in Revision at JEP:LMC)

Statistical learning is the process by which learners rapidly acquire structured information from variable environmental inputs in the absence of explicit reward or feedback (Aslin & Newport, 2009). This process has been demonstrated to operate at numerous levels within the domains of language acquisition and visual processing, leading researchers to question whether statistical learning is supported by a singular, modality-general mechanism or by a set of modality-specific mechanisms. After discovering that statistical learning plays a critical role in speech segmentation (Saffran et al., 1996a,b), several studies demonstrated comparable learning with non-speech auditory stimuli (Saffran, Aslin, Newport, & Johnson, 1999), visual shapes (Fiser & Aslin, 2002), and motor tasks (Hunt & Aslin, 2001). Since the levels of performance and types of computations appear to be equivalent across these modalities (e.g. Creel, Newport, & Aslin, 2004), these initial findings were interpreted as empirical support for a modality-general mechanism (Kirkham, Slemmer, & Johnson, 2002). However, this view has been countered by evidence demonstrating modality constraints. Using a different type of statistical learning task, Conway and Christiansen (2005) found a quantitative advantage for the audio modality relative to the visual and tactile modalities, as well as qualitative differences in the types of computations performed in each modality. From this pattern of results, the authors have argued for the existence of multiple, parallel statistical learning mechanisms that are directly linked to the sensory modality of the input, consistent with theories of embodied cognition that
more broadly posit that cognitive representations are rooted in modality-specific sensorimotor systems (e.g., Barsalou et al., 2003).

In addressing whether statistical learning is rooted in a central mechanism or multiple modality-specific mechanisms, researchers initially relied on comparisons across individual modalities. However, the sensory environment is seldom limited to a single modality or input source (Stein & Stanford, 2008); thus, it is likely that statistical learning mechanisms encounter multiple statistical regularities across modalities on a regular basis. Consequently, an alternative approach to examining the modality-specificity of statistical learning is to explore the degree to which multimodal input sources are processed independently. According to a modality-general view, simultaneous multisensory input should be processed by a unitary mechanism, suggesting that statistical learning should not be independent across modalities. Thus, evidence of independence in multisensory statistical learning would provide strong support for modality-specific models of statistical learning.

Sietz and colleagues (2007) investigated this issue by presenting participants with a stream of audiovisual bigrams constructed from distinct two-dimensional shapes and highly contrastive audio stimuli with varied spectrotemporal properties. Familiarization consisted of either unimodal (audio or visual streams in isolation) or multimodal (audiovisual) stimuli. Participants in the multimodal condition were able to correctly identify audio, visual, and audiovisual bigrams that appeared in the familiarization stream when tested against novel bigrams constructed from the same stimuli. This result suggests that statistical learning mechanisms are capable of simultaneously extracting multiple sequential dependencies from multimodal input. Further, Seitz and colleagues (2007) noted that there was no difference in performance in either modality across both familiarization conditions (unimodal and
multimodal), suggesting that simultaneous learning in one modality did not impact learning in the second modality. From this pattern of findings, the authors claim that both streams were processed independently, consistent with a modality-specific account of statistical learning.

The conclusions of Seitz and colleagues regarding sensory independence must be tempered as their stimuli contained perfect cross-modal correspondence between the input streams. Each audio segment always occurred concurrently with a single visual shape. It is possible that the absence of a significant difference in performance between unimodal and multimodal conditions hinged on the perfect correlation across modalities. Prior research investigating multi-stream visual statistical learning found evidence of distinct learning patterns for perfectly correlated input (Turk-Browne et al., 2008). Further, it is possible that learning could have occurred primarily in one modality and then transferred at test to the second modality. Given these concerns, the goals of the current study are twofold. First, we endeavor to systematically test whether multi-stream statistical learning is processed independently for each modality. Second, we explore the types of cross-modal relationships that influence simultaneous learning of multimodal input streams. We accomplish both of these goals by manipulating the statistical correspondence between individual elements across streams as well as cross-modal boundary alignment (described below). Boundary information may be particularly important for cross-modal learning as demonstrated by Cunillera and colleagues (2010) who found that audio-visual contiguity between a picture presented to learners and dips in transitional probabilities (a statistical cue to word boundaries) facilitated successful performance in a speech segmentation task. Here we ask whether such contiguity might also impact segmentation of both and auditory and visual stream in a cross-modal statistical learning paradigm.
In a series of four experiments, we investigate the influence of cross-modal associations on multimodal statistical learning. In Experiment 1, we collect baseline measures of performance for our input streams presented in isolation. The streams are based on the tone stream from Saffran et al. (1999) and the visual shape stream from Fiser & Aslin (2002). In Experiment 2, we present both streams simultaneously, with perfect predictability between audio and visual elements (similar to Seitz et al., 2007). In Experiment 3, this predictability between streams is removed, as each audio element is equally likely to co-occur with each visual element (and vice versa), although the triplet boundaries are aligned across streams. Finally, in Experiment 4, we further disrupt cross-modal relationships by offsetting the streams such that triplet boundaries are not aligned across streams. If statistical learning is achieved independently across modalities, then disrupting cross-modal relationships should not alter segmentation performance. Alternatively, evidence that statistical learning is sensitive to these manipulations would constrain the types of modality-specific theories to be considered.

EXPERIMENT 1a: Tone sequence alone

Previous statistical learning studies have demonstrated statistical learning with tone streams (Saffran et al., 1999; Creel et al., 2004). Here we test the ability of participants to segment a similar tone stream adapted from the stimuli used in Experiment 3 of Saffran et al. (1999). The goal of Experiment 1a is to replicate this effect and establish a baseline level of performance for the tone stream when presented in isolation.

Methods

Participants. Twenty-six naïve undergraduate introductory psychology students (19 female and 7 male), participating for course credit, were included in the analysis. All participants were monolingual English speakers. Given the role of attention in statistical learning tasks (Toro
et al., 2005; see Weiss, Gerfen, & Mitchel, 2009), we excluded from analysis any participant that gave a self-reported effort level below seven on a ten-point scale (3). By excluding participants on the basis of effort, we adopted a conservative approach to ensure that negative results could not be attributed to inattentiveness.

**Stimuli.** The stimuli in Experiment 1a were modeled after previous studies examining the ability of learners to track statistical dependencies across tone sequences (Saffran, et al., 1999; Creel et al, 2004). We chose tone stimuli rather than speech sounds because it was easier to precisely manipulate duration and produce a more homogenous set. In addition, Seitz and colleagues (2007) used tone stimuli and we wanted to ensure that our results could be directly compared. The tone sequences in this experiment were created from all 12 pure tones within the one-line Western chromatic octave between C₄, or “middle C”, and C₅. Each tone was created using a sine wave generator in Praat (Boersma & Weenink, 2008) based on pitch frequencies set by the Acoustical Society of America, keeping length constant at 1 second per tone. The tones were arranged into 4 groups of three (FGD, G#C#B, CF#D#, and EAA#), forming triplets that avoid any standard musical frame (major/minor chords).

The triplets were concatenated in Praat to form a loop consisting of 24 triplets in a pseudo-random order, with each triplet occurring an equal number of times and no triplet ever following itself. There were no silences between tones, nor were there any other acoustic markers of the triplet boundaries. This loop was repeated four times and concatenated into a 96-triplet block, lasting 4 minutes and 48 seconds. The familiarization stream was concatenated in Praat and encoded in .WAV format with a sampling rate of 44.1 kHz.
The familiarization stream followed an identical structure to statistical learning speech segmentation studies (e.g. Saffran, et al., 1996a) such that individual tones were analogous to syllables and the triplets were analogous to statistically-defined words. Because each tone appears in only one triplet, the transitional probability (the probability of two sounds co-occurring relative to the sounds’ overall frequency of occurrence; see Saffran, 2003) within triplets was 1.00, while the transitional probability between triplets dipped to .33 (since each triplet never followed itself, it could only be followed by one of three other tones). As in the aforementioned segmentation studies, the dips in transitional probabilities provided the only reliable cue to triplet boundaries.

Procedure. Participants were instructed to listen to an audio stream followed by a test that would assess the information learned from the stream. There were no explicit instructions given about the nature of the audio stream, nor were participants informed that it was composed of sequences of triplets. The stream was repeated three times using iTunes software with a one minute silence between each block for a total of 16 minutes.

Following familiarization, participants were given a 16-item, two-alternative, forced choice test, equivalent to the test used in previous segmentation studies (e.g., Saffran, et al., 1996a; Weiss, et al., 2009). In each trial, a statistically-defined triplet (hereafter referred to as a tone-word) was paired with a statistically incompatible triplet (consisting of the last tone from one triplet concatenated with the first two tones of a different triplet; hereafter referred to as a tone-partword). Between test items there was a one second pause, and inter-trial intervals lasted four seconds, during which participants indicated which of the two sequences was most consistent with the audio stream by circling either “1” or “2” on the answer sheet. The test was comprised of the four tone-words, along with 4 tone-partword foils. Each tone-word was paired
with two different tone-partwords twice (counterbalancing for order) yielding 16 total test trials. After completing the test, participants filled out a questionnaire on language background (how many languages spoken, number of years studied, and whether they would label themselves as being bilingual). The questionnaire also included a self-report effort rating (how hard the participants tried in the task). The self-reported effort level is a particularly important metric in subsequent visual and audiovisual conditions, since any visual exposure requires participants to attend to the display.

Results and Discussion

The overall results are presented in Figure 4. The mean test score in Experiment 1a was 11.23 out of 16 (70%), with a standard deviation of 1.95. A one-sample t-test (all tests were 2-tailed) revealed that performance for the tone sequence was significantly above chance (in all experimental conditions chance is defined as 50%, or 8 out of 16), $t(25) = 8.47, p < .001, d = 3.39$. These results represent a successful replication of the findings reported by Saffran et al.

![Figure 4](image-url)
(1999), indicating that statistically defined tone sequences can be segmented into their constituent triplets. Our findings provide a baseline learning rate for comparison in subsequent experiments.

**EXPERIMENT 1b: Visual sequence alone**

In Experiment 1b we familiarize adults to a visual sequence of arbitrary shapes in order to replicate the findings of Fiser & Aslin (2002) and provide a baseline level of performance for the visual stream in isolation.

**Methods**

**Participants.** Twenty-four (17 female and 7 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis any participant that gave a self-reported effort level below seven on a ten-point scale (5).

**Materials.** The stimuli in Experiment 1b were modeled on previous research investigating segmentation of visual-shape sequences (Fiser & Aslin, 2002). We created a movie (Macromedia Director MX 2004) consisting of a sequence of single shapes (the same 12 simple black shapes used in Fiser & Aslin, 2002; see Figure 1a). A 6 cm (5.19°) wide x 10 cm (8.62°) long static black vertical bar was positioned in the center of a 17” LCD Flat Panel monitor at 1024 x 768 pixel resolution. The movie consisted of a single shape moving smoothly, at a constant rate, from the starting position (behind the occluder) out toward the edge of the window (for a distance of 4 cm, 3.47°) in a straight, horizontal path and then returning along the same path, ending behind the occluder. After one shape disappeared behind the occluder, the subsequent shape would emerge from the opposite side and proceed along the same horizontal
plane out toward the opposite side of the screen (see Figure 1b). Each complete movement, from
the occluder to the edge and back, lasted exactly 1 second.

The shapes were grouped into the same four base triplets used in Fiser and Aslin (2002; see Figure 1a). Similar to the tone-words, the base triplets were sequences of three consecutive shapes. The four triplets were concatenated into a continuous movie (Macromedia Director MX 2004) of 24 triplets in a pseudo-random order, with each triplet appearing the same number of times and no triplet following itself. This movie loop was then repeated four times and combined into a movie lasting 4 minutes 48 seconds, consisting of 96 triplets. The movie was then exported as an 800 x 600 Quicktime movie (.MOV) with Sorenson 3 video compression, at a frame rate of 30 frames-per-second.

The statistical structure of the visual sequence was identical to that of the tone sequence in Experiment 1a. The transitional probabilities within triplets were 1.00, while the transitional probabilities between triplets dipped to 0.33. There were no other cues to word boundary other than these transitional probabilities.

Procedure. Participants were instructed to watch a movie followed by a test to assess the information they learned from the movie. Participants then watched the movie clip described above, repeated three times with a one minute pause (during which the screen would turn white) between each block for a total of 16 minutes. The movie clip was presented using iTunes software. Following the clip, participants were given a 16-item, two-alternative, forced choice test, structurally identical to Experiment 1a except that the test items were shape sequences rather than tone sequences. All other aspects of the procedure were identical to Experiment 1a.

Results and Discussion
The mean test score in Experiment 1a was 10.71 out of 16 (67%), with a standard deviation of 2.90 (see Figure 4). A one-sample t-test indicated that performance for the visual shape sequence was significantly above chance, \( t(23) = 4.56, p < .001, d = 1.86 \). An independent samples t-test between performance on the visual and audio tests revealed no significant difference, \( t(48) = 0.75, p = .455, d = 0.22 \). These findings represent a successful replication of the findings reported in Fiser & Aslin (2002). When the visual stream was presented in isolation, the underlying constituent triplets were learnable through their statistical properties. The results provide a baseline comparison for performance on the visual sequence in subsequent experiments in which the shape sequences are presented with accompanying tone sequences. In Experiment 2, we begin to explore whether such audio-visual sequences are learnable, and whether this learning is dependent on the correspondence between the audio and visual stimuli.

**EXPERIMENT 2: Correlated, Synchronized streams**

The goal of Experiment 2 is to determine whether learners can track the statistical structure of two input streams simultaneously in two modalities. Our previous work suggests that learners are capable of tracking multiple sets of sequential statistics within the auditory modality (Weiss, Gerfen, & Mitchel, 2009; Mitchel & Weiss, 2010). Here we test whether this ability extends to sequential statistics presented simultaneously in different modalities, exposing participants to an audio-visual familiarization stream comprised of the audio stream from Experiment 1a and the visual stream from Experiment 1b.

**Methods**

**Participants.** 50 (28 female and 22 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were
monolingual English speakers. We excluded from analysis participants that failed to follow instructions (3) or gave a self-reported effort level below seven on a ten-point scale (10).

**Materials.** The familiarization stream in Experiment 2 was created by combining the tone stream in Experiment 1a with the visual stream in Experiment 1b. Using Adobe Premiere, we synched the audio and visual streams, aligning the onset and offset of each shape and tone (each individual tone and shape presentation lasted one second). The combined, audiovisual stream was then exported as an 800 X 600 Quicktime movie encoded with Sorenson 3 compression, with a total duration of 4 minutes 48 seconds. The statistical structure of each stream was identical to those presented in Experiment 1. The audio and visual streams contained a one-to-one correspondence between each individual tone and a shape element such that tone triplet ABC always coincided with the visual triplet ABC, tone triplet DEF with visual triplet DEF, and so on (see Figure 5a). A pseudo-random ordering was used for presentation, such that no element followed itself and all elements occurred equally often.

![Diagram](image)

**Figure 5.** This figure illustrates the correspondence and boundary alignment across modalities in Experiment 2 (panel A), Experiment 3 (panel B), and Experiment 4 (panel C).
Procedure. All aspects of the procedure were identical to Experiment 1b, except that half of the participants were given the audio test from Experiment 1a, while the other half received the visual test from Experiment 1b. We elected to give participants only one test, rather than both, in order to avoid any potential transfer, interference, or shifts in strategy from one test to the other. Unlike Seitz et al. (2007), we did not use an audio-visual test since they can be solved using multiple strategies (e.g., audio, visual, or audiovisual) and are thus somewhat difficult to interpret. Further, since we manipulate cross-modal relationships in subsequent experimental conditions, it would be difficult to implement an audiovisual test that would be comparable across experiments.

Results and Discussion

The mean test score for the audio test in Experiment 2 was 10.72 out of 16 (67%), with a standard deviation of 2.62 (see Figure 4). A one-sample t-test indicated that performance on the audio test was significantly above chance, \( t(24) = 5.19, p < .001, d = 2.12 \). This was not significantly different from performance in Experiment 1a when the audio stream was presented in isolation, \( t(49) = 0.79, p = .432, d = 0.23 \). The mean test score for the visual test in Experiment 2 was 10.12 out of 16 (63%), with a standard deviation of 2.62 (see Figure 4). A one-sample t-test indicated that performance on the visual test was significantly above chance, \( t(24) = 3.89, p = .001, d = 1.59 \). This was not significantly different from performance in Experiment 1b when the visual stream was presented in isolation, \( t(47) = 0.73, p = .468, d = 0.21 \). An independent samples t-test between performance on the visual and audio tests revealed no significant difference, \( t(48) = 0.79, p = .432, d = 0.23 \).

The results of Experiment 2 indicate that learners are able to track two sets of sequential statistics simultaneously, extending previous work demonstrating this sequentially within a
single modality (Weiss et al., 2009). Using methods more comparable with earlier statistical learning methods (e.g. Saffran et al., 1999; Fiser & Aslin, 2002), Experiment 2 replicates the effect observed by Seitz and colleagues (2007) using a paradigm that tests trigrams structures. Our results provide further evidence that statistical structures in two modalities can be learned simultaneously. As noted above, it is possible that participants only segmented one of the two streams and then transferred knowledge to the other stream during test, a confounding factor that was also present in the Seitz et al. (2007) study. The design of the streams in each of these experiments correlated every audio and visual item with a token in the other modality (e.g. each audio tone was presented simultaneously with the same visual shape throughout familiarization and vice versa). Learners could have transferred knowledge regarding stimulus-specific associations between individual elements across modalities (i.e. tones and shapes) or positional information (Endress & Bonatti, 2007; Endress & Mehler, 2009). For example, learners could have segmented the auditory stream and then noticed that a particular shape always occurred at the onset of an auditory triplet, thereby cueing boundary information. In Experiments 3 and 4 we explore the possibility that multimodal statistical learning is supported by transfer of element-to-element association or positional information, respectively.

**EXPERIMENT 3: Uncorrelated, synchronized streams**

In Experiment 3, the triplets within each stream are ordered such that each element is equally likely to occur with one of four possible elements in the other stream. If the bimodal learning in Experiment 2 was a product of element-to-element transfer of learning, then we would predict that only one stream in Experiment 3 should be learned. If, however, both streams were learned concurrently during familiarization, then here we predict no decrement in performance for either stream.
Methods

Participants. Forty-nine (27 female and 22 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis additional participants that failed to follow instructions (4) or gave a self-reported effort level below seven on a ten-point scale (14), as well as instances of technical failure during the experiment (2).

Materials and Procedure. In Experiment 3 we presented participants with an audiovisual familiarization stream comprised of the audio input stream from Experiment 1a and visual input stream from Experiment 1b. The streams were reordered such that each triplet had an equal probability of co-occurrence with all triplets in the other modality (see Figure 5b). For example, tone triplet ABC was presented an equal number of times with shape triplets ABC, DEF, GHI, and JKL. Thus, unlike Experiment 2, there was no reliable correspondence between particular shape and tone elements. We created a loop of 48 tone triplets with a pseudo-random ordering in Praat and a loop of 48 shape triplets with a distinct pseudo-random ordering in Macromedia Director MX 2004. These loops were then combined using Adobe Premiere. The resulting clip had a duration of 4 minutes 48 seconds and consisted of 96 triplets. The movie was then exported as an 800 X 600 Quicktime Movie with Sorenson 3 compression.

The procedure was identical to Experiment 2. Twenty-five participants completed the audio test, while 24 participants completed the visual test.

Results and Discussion

The mean test score for the audio test in Experiment 3 was 10.28 out of 16 (64%), with a standard deviation of 1.79 (see Figure 4). A one-sample t-test indicated that performance on the audio test was significantly above chance, $t(24) = 6.36, p < .001, d = 2.60$. This was not
significantly different from performance in Experiment 1a when the audio stream was presented in isolation, \( t (49) = 1.81, p = .076, d = 0.52 \). The mean test score for the visual test in Experiment 3 was 11.33 out of 16 (71%), with a standard deviation of 2.55 (see Figure 4). A one-sample t-test indicated that performance on the visual test was significantly above chance, \( t (23) = 6.41, p < .001, d = 2.67 \). This was not significantly different from performance in Experiment 1b when the visual stream was presented in isolation, \( t (46) = -0.79, p = .431, d = -0.23 \). An independent samples t-test between performance on the visual and audio tests revealed no significant difference, \( t (47) = -1.68, p = .100, d = -0.49 \). Further independent samples t-tests revealed no difference in performance in Experiment 2 and Experiment 3 on the audio test (\( t (48) = 0.69, p = .492, d = 0.19 \)) or visual test (\( t (47) = -1.61, p = .115, d = -0.47 \)).

The results of Experiment 3 demonstrate that the findings reported in Experiment 2 were not likely due to element-to-element transfer from one modality to the other. Rather, these results suggest that learners are able to simultaneously extract multiple statistical regularities from audiovisual input. However, since triplet boundaries were aligned across streams, participants that successfully segmented one modality could have used positional information to segment the other stream (see Endress & Bonatti, 2007). In Experiment 4, we offset boundary alignment across streams to remove this source of positional information.

Another explanation for the learning observed in Experiment 3 is that audio and visual pairings may have been represented as unified, bound objects (i.e. AV 1→AV 2→AV 3) as opposed to individual elements (i.e. A1→A2→A3 and V1→V2→V3). If learners perceived the stream in an object-based manner then the transitional probabilities of this AV stream would have still provided consistent word boundary cues (1.0→1.0→0.33). While this account may seem less plausible (due to the infrequent occurrence of each AV triplet), recent evidence
suggests that multi-stream statistical learning may be object-based when the streams are aligned (Turk-Browne et al., 2008). When the streams are partially decoupled, statistical learning switches to a feature-based parsing strategy. Thus, in Experiment 4 we examine whether multimodal statistical learning occurs when the triplet boundaries are temporally decoupled across modalities.

**EXPERIMENT 4: Uncorrelated, desynchronized streams**

In this experiment, we decouple the streams by misaligning word boundaries across the audio and visual streams. Consequently, participants should be unable to transfer positional knowledge of one stream to segment the other. Further, the transitional probabilities of the AV, object-based stream (0.50→1.0→0.50) no longer provide a consistent boundary cue, which should preclude statistical learning on coupled AV stimuli. If learners segment each input stream independently, then we would not expect any change in learning relative to previous conditions. Alternatively, if learning in the two modalities is not entirely independent, then in Experiment 4 we predict that this manipulation should disrupt successful segmentation.

**Method**

**Participants.** Fifty-one (25 female and 26 male) naïve undergraduate introductory psychology students, participating for course credit, were included in the analysis. All participants were monolingual English speakers. We excluded from analysis participants that failed to follow instructions (13) or gave a self-reported effort level below seven on a ten-point scale (17), as well as instances of technical failure during the experiment (1).

**Materials and Procedure.** In Experiment 4 we modified the audiovisual familiarization stream from Experiment 3, such that the cross-modal coherence was disrupted by offsetting
triplet boundaries across streams. This was achieved in Adobe Premiere by moving the initial
segment of the visual stream to the end of the stream. This effectively shifted each visual
segment forward in ordinal position relative to the audio stream, which remained the same as in
Experiment 3. By adjusting the visual stream while keeping the audio stream constant, the triplet
boundaries became misaligned across streams. For example, the beginning of the familiarization
stream consisted of the visual elements BCD (from the triplets ABC and DEF), and the audio
elements ABC. Thus, the triplet boundaries were offset, disrupting the boundary alignment
between the audio and visual streams (see Figure 5c). The correspondence between individual
elements in the audio and visual streams was identical to Experiment 3 (0.25). These two streams
were then combined in Adobe Premiere and exported as an 800 X 600 Quicktime Movie with
Sorenson 3 compression.

All other aspects of the materials and procedure were identical to Experiment 3. Twenty-
six participants completed the audio test, while 25 participants completed the visual test.

Results and Discussion

The mean test score for the audio test in Experiment 4 was 8.96 out of 16 (56%), with a
standard deviation of 3.14 (see Figure 4). A one-sample t-test indicated that performance on the
audio test was not significantly above chance, \( t (25) = 1.56, p = .131, d = 0.62 \). This was
significantly lower than performance in Experiment 1a when the audio stream was presented in
isolation, \( t (50) = -3.13, p = .003, d = 0.89 \). The mean test score for the visual test in Experiment
4 was 8.92 out of 16 (56%), with a standard deviation of 2.48 (see Figure 4). A one-sample t-test
indicated that performance on the visual test was not significantly above chance, \( t (24) = 1.85, p
= .076, d = 0.76 \). This was significantly different from performance in Experiment 1b when the
visual stream was presented in isolation, $t(47) = 2.32, p = .025, d = 0.68$. An independent samples t-test between performance in Experiment 4 on the visual and audio tests revealed no significant difference, $t(49) = .052, p = .959, d = 0.01$.

It is possible that the at-chance performance was a consequence of averaging the scores of subgroups of participants that learned one of the two streams (e.g. half the participants in the auditory test condition successfully learned the auditory stream, and the other half learned the visual stream). To test whether participants in Experiment 4 were learning one or neither of the two streams, we examined the distribution of scores in each test. Separate one-sample Kolmogorov-Smirnov analyses for the auditory and visual tests revealed that each was normally distributed around the group mean (auditory: $K-S = .64, p = .801$; visual: $K-S = .78, p = .582$). This suggests that participants failed to successfully segment either of the two streams at above chance levels.

A one-way ANOVA across all experiments revealed a significant difference in audio test score, $F(3, 98) = 4.12, p = .009$. A second one-way ANOVA revealed a significant difference in visual test scores across all experiments $F(3, 94) = 3.65, p = .015$. Planned contrasts\(^2\) confirmed that, for each test type, performance in Experiment 4 was significantly lower than performance in Experiments 2 and 3 (audio test: $t(98) = -2.61, p = .011, d = -0.53$; visual test: $t(94) = -2.76, p = .007, d = -0.57$)\(^3\).

The results of Experiment 4 demonstrate that when the audio and visual streams were offset, the learning observed in Experiments 2 and 3 was attenuated. Learners require alignment

---

\(^2\) The weights for the contrast analysis were [0, -1, -1, 2] for Experiments 1, 2, 3, and 4, respectively.

\(^3\) Due to the unusually high number of excluded participants in this experiment, we conducted a similar set of analyses that included participants removed on the basis of self-reported effort. This analysis was not substantially different from the filtered analyses. A one-way ANOVA revealed a significant difference in performance between experiments (audio: $F(3, 124) = 3.89, p = .011$; visual: $F(3, 116) = 3.06, p = .031$). Contrasts analyses with identical weights each revealed that performance in Experiment 4 was significantly lower than in Experiments 2 and 3 (audio test: $t(124) = -2.10, p = .038, d = -0.53$; visual test: $t(116) = -2.35, p = .032, d = -0.57$).
of boundaries in order to successfully segment both streams. These results indicate that learning is not independent across modalities.

**General Discussion**

The primary goals of this research were to systematically test whether multi-stream statistical learning is processed independently for each modality and to explore the types of cross-modal relationships that influence simultaneous learning of multimodal input streams. In Experiment 1, we provided baseline levels of performance for segmentation of both a tone and shape input stream presented in isolation. In Experiment 2, we presented learners with an audiovisual stream that consisted of the simultaneous presentation of the audio and visual streams from Experiment 1 and maintained a one-to-one correspondence between elements across modalities. We found that learners were able to successfully segment both the visual and auditory input streams. In Experiment 3, we removed the correspondence between individual elements in the audio and visual streams, a manipulation that did not result in a decrement in performance relative to Experiment 2, as learners were equally successful in learning both streams. In Experiment 4, we further disrupted the relationship between the audio and visual streams by offsetting the triplet boundaries. In this condition, learners were unable to successfully segment either the audio or visual streams at above chance levels.

These findings should inform theoretical accounts of cross-modal statistical learning, particularly regarding the extent to which learning in one modality is achieved independently of learning in a second modality. Previous studies have claimed that during cross-modal statistical learning each stream is segmented independently. According to this view, statistical learning within a particular modality should be unaffected by cross-modal relationships and consequently learners should be capable of simultaneously segmenting two streams in different modalities.
without suffering any decline in performance regardless of any cross-modal associations (Seitz et al., 2007). However, the results presented here in Experiment 4 fail to confirm this theory; learners were unable to segment either stream when presented with multi-modal input in which the boundaries were not aligned. Thus, we conclude that multisensory statistical learning appears to be contingent on some degree of cross-modal coherence and therefore cannot be described as independent (sensu Seitz et al., 2007).

The second goal of this research was to identify the types of cross-modal relationships that affect multisensory statistical learning. As described above, we systematically varied cross-modal relationships across three experimental conditions. In Experiment 2, we presented learners with a multimodal stream characterized by a perfect correspondence between individual elements and triplet boundaries across streams. In subsequent experiments, we removed the element-to-element correspondence (Experiment 3) and boundary alignment (Experiment 4) between streams. Participants successfully segmented both the tone and shape streams when they contained perfect correspondence between elements and boundaries, essentially replicating the results of Seitz et al. (2007) with a different set of stimuli and a different methodology. There was no evidence that removing the correspondence between elements impacted this learning. However, as mentioned, misaligning the boundaries did disrupt learning, suggesting that boundary information is critical for multimodal statistical learning.

Accordingly, we argue that the decrement in learning observed in Experiment 4 was likely due to the importance of boundary information, as opposed to effects arising from transfer of learning or object-based parsing (discussed below). The importance of cross-modal boundary alignment has been further attested in studies of auditory statistical learning. Cunillera and colleagues (2010) found a decline in performance when learners performed a speech
segmentation task in which the onset of a static visual cue was misaligned with the word onsets. Furthermore, word offsets are known to be highly salient events during speech segmentation (Cunillera, Gomilla, & Rodriguez-Fornells, 2008; see also, Echols & Newport, 1992). Together with our results, these findings are consistent with the existence of edge-based positional codes, in which elements occurring at an edge undergo specialized processing (Endress & Mehler, 2010). Learners may extract positional information in conjunction with or in addition to statistical information (see Endress & Mehler, 2009). There may be limits on the number of positional codes that can be tracked at once, and consequently the misalignment of positional information in Experiment 4 may have resulted in the deterioration in performance.

Another possible explanation for the observed pattern of results is that the successful multisensory statistical learning in Experiments 2 and 3 was achieved by learning a single stream and then transferring knowledge of one stream to segment the other. The lack of positional information in Experiment 4 may have hindered transfer and resulted in a decrement in performance. However, our findings are not entirely compatible with this account. In order to transfer knowledge of one stream to the other, we would expect learners to successfully segment one of the two streams, yet neither the aggregated results nor additional Kolmogorov-Smirnov analyses on the distribution of scores in Experiment 4 support this assertion. Rather, learners in Experiment 4 failed to successfully segment either of the two streams. This casts doubt on the notion that the successful learning reported in Experiments 2 and 3 arose as a function of transfer of learning from one stream to the other.

Another possible explanation of our results, as alluded to earlier, is that learners employed an object-based parsing strategy (see Turk-Browne et al., 2008). In Experiments 2 and 3, transitional probabilities between audiovisual objects (i.e. AV→AV→AV) provided cues to
triplet boundaries, whereas they did not cue boundary locations in Experiment 4. However, the large number of possible audiovisual triplets (16) in Experiment 3 and the relative infrequency of each audiovisual triplets (ranging between 12-24 total instances throughout familiarization) make this account less tenable. Moreover, we might have expected a decline in performance in Experiment 3 relative to Experiment 2 due to an increase in computational complexity (in Experiment 2, there were only 4 possible audiovisual triplets and each triplet occurred 108 times throughout familiarization).

Regardless of the underlying source of these findings, our results confirm that adult learners are capable of multisensory statistical learning. This conclusion is consistent with recent demonstrations that statistical learning is sensitive to multi-modal input. While early studies of statistical learning typically focused on learning within a single modality (e.g., Saffran et al., 1996), more recent studies have confirmed that visual input can impact auditory statistical learning (e.g., Mitchel & Weiss, 2010; Sell & Kaschak, 2009; Hollich, Newman, & Jusczyk, 2005; Cunillera et al., 2010). To the best of our knowledge, only one other study (Seitz et al., 2007) has explored how learners segment multiple streams from different modalities.

Our results also provide insight into whether statistical learning has its basis in one central mechanism or in several sensory-specific mechanisms. Had learners been successful in all conditions regardless of our manipulations of cross-modal relationships, then it would have provided compelling evidence for a modality-specific view of statistical learning (e.g. Conway & Christiansen, 2005). However, lack of evidence for modality independence suggests that statistical learning includes at least some modality-general component (i.e., the strictest modality-specific account, as implied by Seitz et al. (2007) does not seem viable). Nevertheless, we do not rule out the possibility of that statistical learning might also include a modality-
specific component. Indeed, a comprehensive model of statistical learning may be comprised of both types of components. Such models emphasizing the mixture of modality-specific and modality-general influences have been posited in other domains, including attention (e.g. Driver & Spence, 1998) and early sensory perception (Besle, Fort, & Giard, 2005). Thus, for example, an interactive activation model (see McClelland and Rumelhart, 1981) with modality-specific processing nodes that are interconnected could account for both modality constraints (Conway & Christiansen, 2005) and cross-modal effects on statistical learning. It may be possible to adjudicate among potential models of multisensory statistical learning by examining the time course of cross-modal effects to determine the point at which information is integrated across senses. Utilizing the temporal precision of electrophysiological measures, future research will endeavor to explore the role of modality-specific and modality-general processes in statistical learning. Previous EEG studies have identified superadditivity (i.e. when the response magnitude to multisensory stimuli supersedes the summed responses to the component unisensory signals) as a signature of multimodal integration (Fort, Delpuech, Pernier, & Giard, 2002; see also, Calvert, Campbell, & Brammer, 2000). We plan to examine superadditive responses in early sensory-perception components (MMN; see Besle et al., 2005) while familiarizing participants to multimodal stimuli in order to detect whether information is integrated immediately or further downstream.

In summary, statistical learning operates within a multidimensional, multimodal sensory environment; thus, we asked whether learners can parse multiple input streams across modality at the same time, and, if so, whether these streams are learned independently. Our experiments demonstrate that learners appear to simultaneously segment two input streams presented in separate modalities. However, this ability demands a certain level of cross-modal coherence, as
disrupting boundary information across modalities attenuates learning, providing evidence against modality independence during multisensory statistical learning.
CHAPTER III

Cross-modal effects in statistical learning: Evidence from the McGurk illusion

The studies reported in Chapter II demonstrated a cross-modal effect on multisensory statistical learning as participants did not successfully segment the audio nor the visual input stream when the positional codes and boundary information were inconsistent across streams. However, a possible limitation of these studies is that the critical finding was the absence of learning in Experiment 4. Such null results can be difficult to interpret since factors outside of the experimental manipulations might have led to lower performance. For example, it has been suggested that the decrement in learning in Experiment 4 could be attributed to out-of-phase neural activity generated by two independent statistical learning mechanisms rather than a direct cross-modal effect of stimuli in one modality on learning in the other modality (M. Christiansen, personal communication, October 30, 2008). Recordings from single neurons in monkey cortex indicate that neural coherence (i.e. in-phase firing of multiple networks of neurons) may enhance learning in a recognition memory task (Jutras, Fries, & Buffalo, 2009), and incoherence in neural spikes inhibits learning (for a review, see Wang, 2010). Because statistical learning produces enhanced neural firing at word boundaries (Sanders, Newport, & Neville, 2002), it is possible in Experiment 4 that the neural responses generated by independent learning mechanisms could have been incoherent, which would have inhibited learning overall. In addition, it is possible that the decrement in performance in Experiment 4 was due to an increased attentional load relative to the other experiments. Evidence from research on attention indicates that observers may switch attention between input streams presented simultaneously in separate modalities (Duncan, Martens, & Ward, 1997). If this is accurate, then learning in Experiments 2 and 3 may have been
achieved by switching attention between tone and shape streams. However, this task of switching attention between streams may have been more difficult in Experiment 4 given the lack of synchrony between positional information across streams. Thus, there may have been greater demands on attentional resources in Experiment 4 than in the other conditions. If statistical learning is constrained by attention (e.g. Turk-Browne et al., 2005), then increased attentional or task demands might lead to lower performance across streams in Experiment 4, rather than a specific cross-modal effect.

Given these potential limitations, an alternative test for modality independence in statistical learning is to explore cross-modal effects in complementary conditions that either impair or facilitate learning. In a series of two experiments, we accomplish this by pairing a visual talking face display with two different artificial languages. If the visual display is integrated with the speech stream, it should produce a McGurk illusion (see below); thus, evidence that the McGurk illusion alters segmentation performance would provide a clear demonstration of a cross-modal effect. Across two experiments, the cross-modal effect (i.e. the McGurk illusion) should either impair (Experiment 5) or facilitate (Experiment 6) learning. Demonstrating contrasting effects of audiovisual integration on statistical learning would provide a stringent test of modality-independent accounts of statistical learning.

**The McGurk Effect**

The McGurk effect is a classic illusion in psychology. First demonstrated by McGurk and MacDonald (1976), the effect is produced when incongruent audio and visual speech signals are integrated to create an illusory percept. McGurk and MacDonald recorded a female speaker producing /ba/, /ga/, /pa/, and /ka/. They then created four dubbed audio-visual recordings, switching the auditory syllables and the lip movements to produce (1) visual /ga/ + audio /ba/,
(2) visual /ba/ + audio /ga/, (3) visual /pa/ + audio /ka/, and (4) visual /ka/ + audio /pa/. When the visual place of articulation was velar and the audio bilabial (1 and 4), participants reported hearing a fused syllable (/da/ and /ta/, respectively). When the visual place of articulation was bilabial and the audio velar (2 and 3), participants reported hearing a combination syllable (/bga/ and /pka/, respectively).

The authors interpreted these results as evidence of audiovisual speech integration. They suggest that the acoustic signal for /ba/ contains some featural overlap with /da/, but not with /ga/. Likewise, the visual signal for /ga/ contains shared features with /da/, but not /ba/. Thus, when the audio signal for /ba/ and visual /ga/ are integrated, there is sufficient perceptual evidence of /da/ to form a fused percept.

This illusion is remarkably robust and has been replicated numerous times in various contexts and languages (e.g. MacDonald & McGurk, 1978; Green, Kuhl, Meltzoff, & Stevens, 1991; Sams, Manninen, Surakka, Helin, & Kättö, 1998; Brancazio & Miller, 2005; Massaro & Cohen, 1996). The McGurk effect is widely regarded as a compelling behavioral index of audiovisual integration (e.g. Massaro, 1998; Green, 1998; Brancazio & Miller, 2005). Demonstrations of the McGurk effect with inverted face stimuli (Massaro & Cohen, 1996; Rosenblum, Yakel, & Green, 2000), point-light displays (Rosenblum & Saldana, 1996), and even incongruities in the gender of the face and voice (Green et al., 1991) all illustrate the robustness of the effect as well as the automaticity of audiovisual integration.

Research on the McGurk illusion demonstrates that audiovisual integration is a fundamental aspect of speech perception (Rosenblum, 2005, 2008; Massaro, 1998). However, it is not clear whether statistical learning mechanisms operate on these integrated percepts. This uncertainty arises from debate concerning the temporal locus of integration in sensory
processing. It has been proposed that integration occurs subsequent to initial modality-independent sensory processing (e.g. Massaro, 1998). In the Fuzzy Logical Model of Perception (FLMP), for example, each sensory input is first evaluated independent of input in other senses, and then these input sources are integrated (Massaro, 1987). In this model, the McGurk illusion arises after initial modality-independent processing of the audio and visual components of the speech input. Since statistical learning is a fast, automatic process (Sanders, Newport, and Neville, 2002), it is possible that statistical learning may occur prior to integration; thus, in this view, the McGurk illusion should not influence statistical learning, consistent with a modality-independent account of statistical learning (Seitz et al., 2007). In the present study, we test this claim by examining the effect of the McGurk illusion on participants' ability to use statistical learning to segment speech.

Across two experiments, we present learners with two miniature artificial languages with contrasting statistical properties paired with a synchronous video of a speaker’s face. In Experiment 5, participants are exposed to a language that provides transitional probability cues to word boundary (Language A) and in Experiment 6, participants are exposed to a language that provides no transitional probability cues to word boundary (Language B). In each experiment, the artificial speech stream is either presented alone (audio-only condition) or paired with a talking face display (audiovisual condition). In the audiovisual condition, inconsistencies between select auditory syllables and visual articulatory gestures are used to elicit a McGurk illusion, which we then use to manipulate the statistical structure of the speech streams (see Figure 6). If, as Seitz et al. suggest, visual and auditory input are processed independently, then we predict that there should be no difference in performance in audiovisual conditions relative to when the speech stream is presented in isolation. If, however, the visual input influences the
processing of the auditory signal (and vice versa), then we would predict that the presence of the McGurk illusion should influence the statistical learning process; thus, the pattern of learning in the audio and audiovisual conditions should differ in predictable directions. Specifically, if participants can integrate visual and auditory input during statistical learning, we would predict to observe an interaction between language (A, B) and display condition (audio-only, audiovisual).

Experiment 5

In Experiment 5, we present learners with a speech stream that contains transitional probability cues to word boundaries. In the audio-only condition, we predict that learners will be able to use these cues to successfully segment the speech stream. In the audiovisual condition, we pair this speech stream with a talking face display that is designed to elicit a McGurk illusion. If participants perceive the McGurk illusion during statistical learning, then this illusion should alter the statistical representation for the speech stream, removing the transition probability cues to word boundary. Thus, we predict that learning should be hindered in the audiovisual condition relative to the audio condition, demonstrating a cross-modal effect on statistical learning.

Methods

Participants. 96 (70 female, 26 male) English-dominant participants were included in the analyses. Participants were recruited from the Introductory Psychology subject pool at The Pennsylvania State University, and were given course credit for participating. Participants were randomly assigned to one of two familiarization conditions (audio-only and audiovisual, 48 participants in each condition). Participants were excluded from analysis if they failed to follow instructions (9) or if there was a technical error during the experiment (2).
Stimuli and Materials. The auditory stimuli consisted of an artificial language with four tri-syllabic (CV.CV.CV) words (see Table 1). Six consonants and six vowels were combined to form a total of six CV syllables. Each CV syllable was created in a similar manner as previous statistical learning experiments (see Weiss, Gerfen, & Mitchel, 2009, 2010; Mitchel & Weiss, 2010) by recording a male speaker producing CVC syllables, with the final consonant being one of three possible places of articulation (bilabial, alveolar, or velar). Coda consonants were recorded to preserve the co-articulatory vowel-to-consonant transitions when the CV syllables were later concatenated into trisyllabic words. Each CVC syllable was then hand-edited in Praat, removing the coda consonants and equating vowel duration. The syllables were synthesized in Praat, overlaying the same pitch (f0) contour onto each syllable in order to remove any pitch or stress cues to segmentation and then concatenated to form the words.

![Figure 6](image.png)

Figure 6. Design of the artificial languages used in Experiments 5 and 6. The syllable-to-syllable transitional probabilities are noted for each condition and language. The transitional probabilities in the audiovisual conditions reflect the statistical structure of the languages if participants perceived the McGurk illusion.

The four words were concatenated into a continuous stream in a pseudo-random order, such that each word appeared an equal number of times and no word ever followed itself. In addition, the order of words in the stream was constrained such that words 1 and 2 were only
followed by words 3 and 4, and vice versa. This order constraint allowed us to use the McGurk illusion (if perceived) to alter the statistical structure of the entire language even though we only manipulated 2 word-final syllables. That is, given the order constraint, if participants perceived 4 word final syllables (as in Language B), then each between-word transition was 50% (each word-final syllable could precede only 2 possible syllables). However, if participants perceived only 2 word final syllables (as in Language A), then the ordering constraint created a scenario in which the word-final transitional probability was 25% (each word-final syllable could precede all 4 possible syllables; see Table 1). The continuous speech stream was comprised of three four-minute blocks, each block containing 288 words, for a total familiarization of 12 minutes and 864 words. In between each block there was a one minute silence during which the screen turned white. There were no acoustic cues to word boundary. The artificial language (Language

<table>
<thead>
<tr>
<th>Display Condition</th>
<th>Audio-only</th>
<th>Audiovisual (if integrated)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language A</td>
<td>so bæ pã</td>
<td>so bæ pã</td>
</tr>
<tr>
<td>(Experiment 5)</td>
<td>je lu mi</td>
<td>je lu ni</td>
</tr>
<tr>
<td></td>
<td>bæ je pã</td>
<td>bæ je ðã</td>
</tr>
<tr>
<td></td>
<td>lu so mi</td>
<td>lu so mi</td>
</tr>
<tr>
<td>Transitional probabilities</td>
<td>.5 → .5 → .25</td>
<td>.5 → .5 → .5</td>
</tr>
</tbody>
</table>

| Language B                 | so bæ ðã    | so bæ ðã                    |
| (Experiment 6)             | je lu mi    | je lu ni                    |
|                            | bæ je pã    | bæ je ðã                    |
|                            | lu so ni    | lu so ni                    |
| Transitional probabilities | .5 → .5 → .5 | .5 → .5 → .25               |

Table 1. The artificial languages used in Experiments 5 and 6. Bolded syllables in the audiovisual condition represent the illusory, integrated McGurk syllables. Syllable-to-syllable transitional probabilities are reported for each language in each condition. Transitional probabilities in the audiovisual conditions reflect the statistical structures of the languages if they include the integrated percept.
A) had consistently higher transitional probabilities within words (0.50) than between words (0.25); thus, in the audio-only condition, transitional probabilities provided cues to word boundaries (see Table 1).

The visual display was created using methods from a previous study (Mitchel & Weiss, 2010). A Sony Handicam was used to video-record research assistants lip-synching to an audio-stream while reading from a list of words mounted behind the camera. The video was then hand-edited in Adobe Premiere v.3 to ensure that the audio stream and video display were synchronous, aligning them such that the articulatory gestures of the lips coincided with the corresponding auditory event. The video was cropped to only display the lip movements (see Figure 7), and then exported as an 800 X 600 Quicktime movie with Sorensen 3 video compression.

In the audiovisual condition, the visual stream differed from the audio stream in two word-final syllables (audio: /mi/ and /pa/, visual: /gi/ and /ka/, respectively). These syllables were selected because they consistently (~90% of the time) produce a McGurk fusion effect (MacDonald & McGurk, 1978). If integrated, the participants should perceive /ni/ and /ta/. If perceived in this manner, the new transitional probabilities of the audio stream would have been the same within and between words (0.50; see Table 1).

**Test Materials.** Learning of the statistically defined words was tested using a 24 item word-identification task. The test was the same for each condition and consisted of 6 words, 3 part-words, and 3 non-words. The six words were sub-divided into three classes. During familiarization, two of the words in

![Figure 7. Still frame of the video display in Experiments 5 & 6.](image-url)
the language were consistent across audio and visual input (/so bae pa/, /je lu mi/). At test, these items were considered to be AV words. The other two words in the language had inconsistencies between the audio and visual input (producing the McGurk illusion). At test, we tested participants' identification of audio-only words, which were taken from the audio stream (/je lu mi/, /bae je pa/) and McGurk words, which were the auditory equivalent of what participants should have heard if the McGurk illusion produced a fused, integrated percept (/je lu ni/, /bae je ta/). Non-words were combinations of syllables that did not occur together during familiarization. In addition, each syllable in a non-word was in the same position (initial, medial, final) that it occurred in the words used for familiarization (e.g. the words ABC and DEF might be combined to form non-words AEF or DBC). Part-words were formed by combining the third syllable of one word with the first and second syllables of another word (e.g. ABC and DEF might be combined to form part-words ADE and FAB). Each item was presented twice in a random order. There was no video display during test. Participants judged whether each item was a word by pressing the keys marked “yes” or “no” on a keyboard.

**Analysis.** Using signal detection theory, d’ (hit rate – false alarm rate) was calculated to determine participants’ sensitivity to detecting words. Because d’ incorporates both participants' ability to identify words and their ability to reject non- and part-words, it is a sensitive measure of participants' knowledge of word boundaries. In the audio-only condition, hit rate was measured as the average endorsement rate (the proportion of trials participants selected ‘yes’) of AV and Audio-only word items and false alarm rate was the average endorsement rate of McGurk words, non-words, and part-words. In the audiovisual condition, hit rate was the average probability of endorsing AV and McGurk words and false alarm rate was the average endorsement rate of audio-only words, non-words, and part-words.
In this task, a d’ of 0 represents chance performance (participants were equally likely to endorse words, non-words, and part-words), while a d’ significantly above 0 represents learning (participants were more likely to endorse words than part-words or non-words). A measure of response bias, c, was calculated to determine if participants favored one response (“yes” or “no”) over the other. Response bias was measured by summing the hit rate and false alarm rate and then multiplying that total by -0.5.

**Procedure.** In the audio-only condition, participants were instructed to listen to an audio stream and informed they would be tested on knowledge acquired from this familiarization. Participants were not informed that the audio stream was an artificial language. The familiarization stream and test were presented using E-prime software. Following the test, participants completed a questionnaire that assessed language background and effort level (using a ten-point scale).

In the audiovisual condition, participants were instructed to view a short movie and informed that they would be tested following the movie. There were no explicit instructions given about the nature of the movie, nor were participants informed that the audio stream was composed of an artificial language. Familiarization streams were presented using iTunes (version 7.0) software. Following familiarization, participants completed the identification test, presented using E-Prime software. After the test phase, participants completed the same questionnaire as in the audio-only condition.

**Results**

In the d’ analysis, the mean d’ score in the audio-only condition was 1.27 (SD=1.13), and was significantly greater than chance (defined as 0), $t(47) = 7.79, p < .001, d = 2.27$ (see panel a, Figure 8). All tests were two-tailed. The mean d’ score in the audiovisual condition was 0.10.
(SD=.68), and was not significantly different from chance, \( t(47) = 1.03, p = .310, d = 0.30 \).

Performance in the audio-only condition was significantly higher in the audio-only condition than in the audiovisual condition, \( t(94) = 6.15, p < .001, d = 1.27 \). The response biases, \( c \), for the audio-only and audiovisual conditions were -0.47 (SD=.77) and -0.27 (SD=.44), respectively, reflecting a liberal response bias.

![Figure 8](image)

Figure 8. Mean d-prime scores for the audio-only and audiovisual display conditions for Language A (panel a) and Language B (panel b) in Experiments 5 and 6, respectively.

In addition to calculating \( d' \), we also report here the endorsement rates (the probability that a participant would chose “yes” to an item) for each type of test item (see Figure 9). The endorsement rate for the McGurk consistent words was significantly higher in the audiovisual condition (49%) than in the audio-only condition (23%), \( t(94) = 5.39, p<.001, d=1.11 \). There were no other significant differences between conditions in endorsement rates for the four other item types (\( p’s>.05 \)).
The goal of Experiment 5 was to examine the effect of an incongruent visual display on the segmentation of an auditory artificial language. When the speech stream (Language A, which contained transitional probability cues to word boundary) was presented without an accompanying visual display, segmentation performance was significantly above chance. When paired with the incongruous visual display, there was a significant drop in performance, as measured by $d'$. These results suggest that auditory and visual signals are integrated during statistical learning, as the pattern of learning across the audio-only and audiovisual conditions indicates that the McGurk illusion altered the statistical representation of the auditory language, influencing segmentation patterns. This is inconsistent with modality-independent accounts of statistical learning (e.g. Seitz et al., 2007). However, disrupting performance with a McGurk illusion is not sufficient evidence to discount modality independence. As in Chapter II, it is
possible that this decrement in performance was the by-product of a general effect on processing. For example, in the audiovisual condition, participants were required to monitor a stream of information in the auditory and visual modality simultaneously, whereas participants in the audio-only condition were only required to monitor a single input source. In addition, the presence of the visual display (which consisted of moving lips) may have distracted participants in the audiovisual condition. These task constraints between conditions may have led to the observed decrement in learning, rather than the McGurk illusion.

Finally, one potential limitation of Experiment 5 is that the d' analysis may have underestimated learning in the audiovisual condition, since the method of calculating d’ differed across display conditions. In the audio-only condition, endorsement of an audio-consistent word was coded as a correct identification. In the audiovisual condition, endorsement of these items was coded as a false alarm. Consequently, the high endorsement rate for audio consistent words in the audiovisual condition contributed to the low d’. Indeed, other than the McGurk items, there was no significant difference in participants’ endorsement rates, suggesting that the visual display may not have significantly affected segmentation performance. However, as we discuss at greater length in Experiment 6, it is also possible that the high rate of endorsement for audio-consistent words is driven by a preference bias for one of the two items (/je lu mi/), which was strongly endorsed regardless of the condition or experiment (see below).

To address these concerns, Experiment 6 is designed to provide a contrast to Experiment 5, testing whether it is possible to use the McGurk illusion to facilitate learning. Thus, participants are exposed to an artificial language that contains no word-boundary cues in the audio-only condition. In the audiovisual condition, if the participants integrate the visual and auditory streams, then the McGurk illusion should alter the transitional probabilities between
adjacent syllables to provide segmentation cues.

**Experiment 6**

In the previous experiment, incorporating a McGurk illusion that altered the underlying statistics of an input stream resulted in a decrement in learning. In Experiment 6, participants are presented with a speech stream that has flat transitional probabilities (i.e. transitional probabilities are equivalent between and within words) and hence no cues to word boundaries. In the audio-only condition, learning is predicted to be at chance. In the audiovisual condition, if participants integrate the audio and visual streams, the resulting McGurk illusion should alter the statistical structure of the language. The illusory syllables will make the word-final transitional probabilities dip below the within-word transitional probabilities, thereby cuing the location of word boundaries. If participants perceive the McGurk illusion in the audiovisual condition, then we predict Language B should be segmented at above chance levels. While the McGurk illusion impaired performance in Experiment 5, the McGurk illusion, if perceived, should facilitate performance in Experiment 6. Thus, the complimentary predictions of Experiments 5 and 6 should provide a strong test of whether multisensory integration occurs during statistical learning.

**Methods**

*Participants.* 96 (63 female, 33 male) English-dominant participants were included in the analyses. Participants were recruited from the Introduction to Psychology subject pool at The Pennsylvania State University. The participants were randomly assigned to one of the two familiarization conditions (audio-only and audiovisual), with 48 participants in each condition. Participants received course credit for their participation. Participants were excluded from analysis for failing to follow directions (7) or if there was a technical error during the experiment.
Stimuli. The stimuli for Experiment 6 were created in an identical manner to those in Experiment 5. In Experiment 6, the artificial language (Language B) had flat transitional probabilities within and between words (0.50→0.50→0.50; see Table 1). The statistical structure of Language B was the same as the McGurk-altered structure of Language A in the audiovisual condition of Experiment 5. Without statistical cues to word boundary, it was predicted that this language should not be learned in the audio-only condition. As in Experiment 5, the incongruent visual display differed from the audio stream in two word-final syllables (see Table 1). These two visual syllables, if integrated with the audio syllables, produced a McGurk effect, yielding two fused syllables. If participants perceived these fused syllables, then the transitional probabilities for the McGurk-altered sound inventory should have been higher within words (0.50) than between words (0.25), providing a statistical segmentation cue (see Table 1).

The test was identical to that used in Experiment 5, except for the two AV-consistent words (/so bae ta/ and /lu so ni/ in Experiment 6) which were different than the AV consistent words in Experiment 5 (/so bae pa/ and /lu so mi/ in Experiment 5). All other aspects of the procedure and analysis were identical to Experiment 5.

Results

In Experiment 6, the mean d’ score in the audio-only condition was 0.28 (SD=0.53), a level of performance that was significantly above chance, \( t(47) = 3.71, p = .001, d = 1.08 \). The mean d’ score in the audiovisual condition was 0.53 (SD=0.73), which was significantly above chance, \( t(47) = 5.03, p < .001, d =1.47 \) (see Figure 8). There was a liberal response bias in both the audio-only condition (\( c = -0.12, SD=.28 \)) and the audiovisual condition (\( c = -0.31, SD=0.47 \)). An independent samples comparison revealed a marginally significant increase in learning in the
audiovisual condition over the audio-only condition, \( t(94) = 1.90, p =.060, d = 0.39 \).

Furthermore, there was a significant difference in performance between Language A and Language B in both the audio-only and audiovisual conditions. In the audio-only condition, participants had higher d’ scores for Language A than for Language B, \( t(66.53) = 5.49, p <.001, d = 1.35 \). In the audiovisual condition, however, d’ scores were higher for Language B than for Language A, \( t(94) = -2.98, p =.004, d = -0.61 \).

Across Experiments 5 and 6, a 2 (Language A, Language B) X 2 (audio-only display, audiovisual display) factorial ANOVA revealed a main effect of language \( (F(1, 188) = 5.93, p=.016, \eta^2=0.02) \) and a main effect of display type \( (F(1, 188) = 16.08, p<.001, \eta^2=0.06) \). In addition, there was a significant crossover interaction between language and display, \( F(1, 188) = 37.83, p<.001, \eta^2=0.15 \) (see Figure 10). 

![Figure 10. Mean d-prime scores in Experiments 5 and 6, illustrating the crossover interaction of language and display type.](image)

The mean endorsement rate for audiovisual consistent words (see Figure 9) was significantly greater in the audiovisual condition (70%) than in the audio-only condition (59%), \( t(94) = 2.05, p = .043, d=.42 \). The mean endorsement rate for the McGurk consistent test items was marginally significantly greater in the audiovisual condition (64%) than in the audio-only condition (55%), \( t(94) = 1.91, p = .059, d=.39 \). There were no other significant differences between conditions in the endorsement rates for the other item types \( (p’s>.05) \). Across experiment 5 and 6, a repeated measures ANOVA with condition (audio-only and audiovisual for both experiments, totaling 4 conditions) as a between-subjects factor revealed a significant
difference in endorsement rate between the two audio-consistent test items (/je lu mi/ and /so bae pa/; see Figure 11), $F(1, 188) = 7.32, p<.001, \eta^2=0.24$, as well as a significant interaction between the difference in endorsement rate for these two items and condition, $F(3, 188) = 5.34, p=.001, \eta^2=0.06$. A one-way ANOVA revealed that endorsement rates for the first audio-consistent word (/je lu mi/) did not differ across experiments, $F(3, 188) = 0.11, p=.956, \eta^2<.01$. A second one-way ANOVA revealed that endorsement rates for the second audio-consistent word (/so bae ta/) significantly differed across experiments, $F(3, 188) = 9.33, p<.001, \eta^2=0.13$.

**Discussion**

Experiment 6 was designed to provide a contrast to Experiment 5. We presented participants with an artificial language whose transitional probabilities between adjacent syllables were identical both between and within words, thereby providing no transitional probability cues to word boundary. In the audiovisual condition, we induced a McGurk illusion that altered the syllable inventory such that the transitional probabilities between words were lower than the transitions within words, thus cuing the location of word boundaries. Consequently, we predicted that the McGurk illusion in the audiovisual condition should facilitate participants’ ability to use statistical cues to segment a continuous speech stream. The
findings from Experiment 6 support this prediction. In the audio-only condition, learning was not robust (though above chance, see below). However, in the audiovisual condition, learning was well above chance and performance was significantly greater than in the audio-only condition. Moreover, the results of Experiment 6 undermine the argument that the results of Experiment 5 were due greater demands on attention in the audiovisual condition, as these same demands should have resulted in poorer performance in the audiovisual condition of Experiment 6. In addition, the pattern of endorsement rates in Experiment 6 support our conclusions from the d’ analysis, as participants were significantly more likely to endorse the audiovisual consistent words (a measure of learning that is unbiased by their interpretation of the McGurked items). Thus, while the d’ analyses in Experiment 5 may have underestimated learning in the audiovisual condition, these analyses appear to accurately reflect a significant increase in segmentation performance in the audiovisual condition of Experiment 6.

It is important to note that performance in the audio-only condition in Experiment 2 was significantly above chance. Given that the transitional probability structure of this language affords no boundary cues, this above-chance performance likely reflects participants’ sensitivity to frequency cues to word boundary (words necessarily occurred more frequently than part-words, while non-words and McGurk words never occurred). Despite being above chance, learning of Language B in the audio-only condition was significantly lower than Language A, audio-only, suggesting that segmentation performance was hindered in the absence of transitional probability cues to word-boundary. This is consistent with the findings of Aslin and colleagues (1998), who demonstrated that transitional probability cues guide segmentation when frequency cues are controlled for. In the present study, although there were frequency cues to word boundary in all conditions, segmentation performance was greatest when transitional
probability cues were also available (audio-only condition of Experiment 5, audiovisual condition of Experiment 6).

It is also worth noting that while performance in the audiovisual condition of Experiment 6 (Language B) was not equivalent to learning in the audio-only condition of Experiment 5 (Language A), this is not surprising given that the McGurk illusion is contingent on sustained attention to the video display throughout familiarization. If participants failed to attend to the video stream in the audiovisual condition, they would not have perceived the McGurk illusion each time it occurred; instead, participants would have occasionally perceived the audio-consistent words, raising the level of statistical noise and, on average, lowering the level of performance relative to the audio-only condition of Experiment 5 in which transitional probability cues in the speech stream reliably cued word boundary.

Finally, there is evidence that, across experiments, one of the audio-consistent test items (/je lu mi/) was endorsed at a high rate regardless of the condition. It is possible that this word may be inherently more ‘word-like’ to native English speakers, and thus participants may have been likely to identify this item as a word irrespective of familiarization. The other audio-consistent item (/so bae pa/), on the other hand, varied as a function of the experimental condition (and trended in the predicted direction). Thus, this item may more accurately represent the effect of our manipulation on how learners segmented the artificial languages. In addition, because the d’ value in Experiment 6 included both audio-consistent items, the endorsement bias for /je lu mi/ may lead to an underestimate of participants’ successful segmentation in the audiovisual condition.

**General Discussion**

The goal of the present study was to test whether the computational mechanism or
mechanisms underlying statistical learning parse input from multiple modalities separately, processing each input source completely independent of input from another modality (see Seitz, et al., 2007). In order to test this account, we examined whether input from multiple modalities could be integrated during the statistical learning process, utilizing the McGurk effect to manipulate the perceived statistical structure of a speech stream. In Experiment 5, we presented learners with a speech stream that contained transitional probability cues to word boundary (Language A). The stream was either presented in isolation (audio-only condition) or synchronized with a visual display (audiovisual condition) that elicited a McGurk illusion in two word-final syllables. Perceiving the McGurk illusion altered the statistical structure of Language A so that it no longer contained conditional probability cues to word boundary. In Experiment 6, we designed an artificial language (Language B) in which word boundaries were not cued by transitional probabilities; however, if participants perceived the McGurk illusion in the audiovisual condition, then the resulting statistical structure provided boundary information. If statistical learning mechanisms operated on independent, modality-specific representations prior to integration, then the McGurk illusion should have had a minimal effect on how the speech stream was segmented. Contrary to this prediction, we found that the McGurk illusion altered the pattern of learning in a manner consistent with audiovisual integration. In Experiment 5, participants successfully segmented Language A when presented in isolation, but performance declined in the audiovisual condition (as measured by d’). In Experiment 6, the presence of the visual display facilitated learning.

The complimentary pattern of learning across experiments was consistent with what would have been expected if conditional statistics were calculated over a sound inventory containing the illusory, integrated syllables (/ni/ and /ta/), providing evidence that learners can
integrate audio and visual input during a statistical learning task. Although modality-independent accounts of statistical learning posit that statistical learning in one modality should be independent of stimuli in other modalities (Seitz et al., 2007), they do not preclude multisensory integration or the perception of the McGurk illusion. For example, it is possible that statistical learning occurs prior to multisensory integration, consistent with the time-course of integration proposed by some models of multisensory perception (e.g. Massaro, 1998). However, in order for statistical learning to be modality independent, the mechanisms underlying this process should not operate over integrated percepts. In this account, participants in the audiovisual conditions of the present study should have calculated transitional probabilities across a sound inventory that only included stimuli from the auditory input. Thus, statistical learning in the audio-only and audiovisual conditions should have been equivalent. However, in both Experiment 5 and 6, we found that the McGurk illusion altered segmentation performance in the audiovisual condition, suggesting that statistical learning was operating over the integrated percepts.

In addition to demonstrating audiovisual integration during statistical learning, the results of the studies reported here provide evidence of a specific cross-modal effect, rather than a global decrement in performance. That is, our results conformed to a priori predictions that a visual display would hinder learning of Language A and enhance learning of Language B. Because the visual display influenced learning in both directions, the results of Experiments 5 and 6, unlike the results of Experiment 4 from Chapter II, cannot be attributed to a general effect of incorporating a talking face display. Our results therefore challenge the claims of a modality-independent account of statistical learning, suggesting instead that the mechanisms supporting statistical learning do not process multimodal input completely independently.
While the present study provides evidence of audiovisual integration during statistical learning, further research is necessary to evaluate whether this process is supported by a modality-general (Kirkham et al., 2002) or a modality-specific mechanism (Conway & Christiansen, 2005), since it is possible that modality-specific mechanisms might operate over integrated percepts if integration occurs through association. While there is debate over the nature of integration (reviewed in Chapter V), theories of association (e.g. Bernstein et al., 2004; Bernstein, 2005) propose that multisensory input is processed as separate, modality-specific representations that become associated further upstream. If integration occurs through association, then the audio and visual input are never fused into a single audiovisual perceptual representation. Thus, in terms of the present study, the association of the audio and visual input would elicit the McGurk effect and alter segmentation, even though statistical learning was supported by separate, modality-specific mechanisms. As we discuss in Chapter V, this framework may implicate statistical learning as a mechanism with both modality-general and modality-specific components.

If, as the results reported here suggest, facial information is integrated during speech segmentation, then it is possible that this information may have a role in the segmentation process. A growing number of studies are beginning to highlight the role faces may play in a variety of language acquisition tasks (e.g. Weikum et al., 2007), guided by evidence that the speakers' face conveys a considerable amount of linguistic information (Kuhl & Meltzoff, 1984; Patterson & Werker, 2003; Yehia et al., 2002). Chapter IV extends this line of research into the domain of speech segmentation by asking whether learners can use facial cues, alone, to segment speech.
CHAPTER IV:
Visual Speech Segmentation: Using Facial Cues to Locate Word Boundaries in Continuous Speech

The input to language learners is not restricted to the auditory modality (Massaro, 1998). In particular, the speaker's face is both highly salient to infants (e.g. Morton & Johnson, 1991) and is linguistically informative (e.g. Patterson & Werker, 2003; see below). Thus, it is not surprising that a growing number of studies have begun to demonstrate the importance of faces for language acquisition processes (e.g. Weikum et al., 2007; Soto-Faraco et al., 2007; Patterson & Werker, 2003; Hollich et al., 2005). Extending this line of inquiry, the goal of Chapter IV is to examine the contribution of cues provided by the speakers’ face to speech segmentation, an early component of language acquisition. Few studies have examined the role of faces in speech segmentation, despite evidence that faces convey linguistic structure that may facilitate speech segmentation. In particular, faces contain cues to the prosodic structure (e.g. rhythm, stress, and pitch) of accompanying speech (e.g. head nodding and lip aperture; Yehia, et al. 2002; Graf, Cosatto, Strom, & Huang, 2002) that may help learners identify word boundaries. Here we investigate whether adults are able to use visual cues to acoustic prosody (e.g. hereafter visual prosody) to segment continuous speech in the absence of additional segmentation cues. Before describing the experiments, we address the types of information conveyed in the face of the speaker that learners might use to segment a speech stream.

What’s in a Face?

Faces provide a particularly salient visual cue (see Nelson, 2001) and a potentially rich source of information for segmentation. An abundance of linguistic content is conveyed to listeners by viewing a talking face (known as visemic information). Visemic information includes both prosodic and phonetic cues (Kuhl & Meltzoff, 1982). For example, infants as
young as 4 months of age are able to match auditory syllables with the corresponding visual syllable using only the spectral information (i.e. formant structure) in the vowels (Kuhl & Meltzoff, 1984; Patterson & Werker, 1999). In addition, visual speech displays have been shown to facilitate speech perception (Sumby & Pollack, 1954), suggesting that not only is this information available in the visual input, but that observers are able to utilize visemic information to facilitate the acquisition of auditory information.

Of particular relevance to the present study, Yehia and colleagues (2002) found that head movements convey information about the pitch, lexical stress, and syntactic boundaries of the speech stream. Specifically, visual prosody is cued by x-axis rotation of the head (head nodding) and lip aperture (how far apart the lips are; Yehia et al., 2002; Graf et al., 2002). Since prosody (e.g. pitch and stress) in the auditory stream is a prominent cue to word boundary that learners can use to segment speech (see Jusczyk et al., 1999), if a talking face display contains visual prosodic cues, it is reasonable to predict that such cues might be used to segment a synchronous speech stream.

As mentioned above, there is a paucity of research dedicated to visual contributions to speech segmentation. Although no study, to the best of my knowledge, has assessed the independent contribution of visual speech in speech segmentation, two studies have examined whether facial cues from a talking face display can enhance statistical learning of audiovisual speech (Mitchel & Weiss, 2010; Sell & Kaschak, 2009). In Mitchel & Weiss, a single artificial language was paired with a synchronous talking face display to determine whether the addition of a facial display would facilitate statistical learning. The authors found no change in learning in the audiovisual condition relative to when the artificial language was presented in isolation. Likewise, Sell & Kaschak (2009) tested the effect of visual speech on statistical learning by
presenting adults with an artificial language in three conditions: audio-only (speech stream in isolation), an audiovisual condition (paired with a talking face display), or in a visual-only condition (participants watched the facial display with no corresponding audio). While participants were able to use visual cues to successfully segment the visual speech stream in the visual-only condition, there was no difference between the audio-only and the audiovisual condition, indicating that visual speech neither facilitates nor hinders statistical learning. However, in both studies, the presence of statistical cues may have obviated the need to rely on visual speech cues provided by the talking face display. Given that the benefit of facial information is most noticeable in contexts in which participants cannot rely entirely on the auditory input, such as in noisy environments (Sumby & Pollack, 1954; Grant & Seitz, 2000; Hollich et al., 2005), visual speech may play a greater role in speech segmentation if the transitional probability cues to word boundary are reduced.

In the following experiments, we examine whether adults can extract prosodic information in talking faces (consistent with Yehia et al., 2002) and use this information to segment a speech stream in the absence of alternative segmentation cues. In Experiment 7, we test participants' baseline ability to segment an artificial language with minimal transitional probability cues to word boundary (i.e. low predictability of two sounds co-occurring within a word). We then create two visual speech streams, one in which the assistant creating the visual display is aware of word boundaries, and one created by a second assistant that is misinformed of the word boundaries (see below). We predict that if the assistant is aware of word boundaries, then he will impart visual prosodic cues while recording the talking face display. If the participant is misinformed of word boundaries, then he will not impart prosodic cues. In Experiment 8, we test participants' ability to segment the aware and misinformed visual speech
streams in isolation (no auditory stimuli during familiarization or test). If the aware assistant imparts prosodic cues while recording the visual display, then participants should successfully segment the visual display. Conversely, performance on the visual display created from the misinformed assistant should not be above chance. Finally, in Experiment 9 we pair the auditory speech stream with each visual display, familiarizing participants to an aware audiovisual display and a misinformed audiovisual display. We then test participants on the same audio-only test from Experiment 7. If participants are able to use visual speech cues alone to segment speech, then segmentation performance for the aware audiovisual condition should be significantly greater than performance in the audio-only condition (Experiment 7) and performance for the misinformed audiovisual condition.

**Experiment 7: Audio-only**

**Methods**

*Participants.* 30 (16 female, 14 male) undergraduate Introductory Psychology students at The Pennsylvania State University participated in this study for course credit and were included in the analysis. All participants were monolingual English speakers. Participants were excluded from analysis if their self-rated effort was 5 or below on a scale of 10 (1), if they failed to follow instructions (2), or if there was technical failure during the experiment (2).

*Stimuli.* Participants were familiarized to an artificial language comprised of six trisyllabic words (see Figure 12a). The auditory stimuli were created in an identical manner as those in Experiment 5.

Each syllable was used three times and occurred in every possible word position (i.e. the 1st, 2nd, or 3rd syllable in a word). In addition, there were ordering constraints ensuring that transitions within words did not also occur at word boundaries, nor was any syllable followed by
itself. For example, the syllable /bo/ precedes the syllables /ke/ and /pu/ within words; thus, the word /pu taj bo/ could only be followed by /da pu gi/, /gi bo pu/ and /taj da ke/. Consequently, within-word transitional probabilities were not highly predictive (.33), and were equivalent to the between-word transitional probabilities of artificial languages used in previous studies (e.g. Saffran et al., 1996a; Weiss, Gerfen, & Mitchel, 2009, 2010; Mitchel & Weiss, 2010). While the transitional probability between-words (.11) was consistently lower than within-word transitional probabilities, thereby affording cues to word boundary, we believe the lowered predictability of within-word transitional probabilities reduced the availability of these cues as a segmentation strategy. Thus, participants should not robustly learn this artificial language. The words were then concatenated into a 15 second clip of 18 words. This clip was then looped 16 times to create a four minute block, consisting of 288 words. The familiarization stream consisted of three four-minute blocks, with a one-minute break between blocks, for a total of 12 minutes of familiarization (864 words).

The test stimuli consisted of the 6 words and 6 part-words. The part-words were formed by concatenating the third syllable of one word with the first and second syllables of another word (i.e. 3-1-2). Thus, part-words were heard during the familiarization stream, albeit less often than words. All of the test words and part-words were created in an identical manner as the

<table>
<thead>
<tr>
<th>Words</th>
<th>Part-words</th>
</tr>
</thead>
<tbody>
<tr>
<td>bo ke taj</td>
<td>taj ke gi</td>
</tr>
<tr>
<td>pu taj bo</td>
<td>bo da pu</td>
</tr>
<tr>
<td>ke gi da</td>
<td>da gi bo</td>
</tr>
<tr>
<td>da pu gi</td>
<td>gi taj da</td>
</tr>
<tr>
<td>gi bo pu</td>
<td>pu bo ke</td>
</tr>
<tr>
<td>taj da ke</td>
<td>ke pu taj</td>
</tr>
</tbody>
</table>

Figure 12. (a) The words and part-words from the artificial language in Experiment 7. (b) The statistical structure of the artificial language, depicting the transitional probabilities between syllables.
familiarization stream and the test items in Experiment 1.

**Procedure.** The familiarization stream and test were presented using E-Prime software. Participants wore noise-cancelling headphones and were instructed not to remove the headphones during the experiment. Participants were instructed to listen to an audio stream and were informed that they would be tested on the knowledge extracted from the stream. Both written and verbal instructions were given, with the verbal instructions read from a script to ensure reliability between experimenters. The experimenter was present throughout the experiment to monitor participants and ensure that they followed instructions.

In order to test for successful segmentation, participants were asked to discriminate words from part-words in a two-alternative, forced-choice test presented using E-Prime software. The test was auditory, with no accompanying visual display. In each test trial, participants were presented with a word and a part-word, separated by a one second pause. Participants responded on a keyboard, identifying which test item corresponded to a word in the familiarization stream by pressing a key to indicate the first or second test item. Each test word was paired with every test part-word, resulting in 36 test trials. The order of presentation was counterbalanced, so that words and part-words appeared in the first or second position an equal number of times. Participants’ responses were recorded in E-Prime. After testing, participants were given a questionnaire that assessed their self-reported level of effort.

**Results and Discussion**

The mean percent of words chosen in Experiment 7 was 52.96% (SD=8.44). This level of performance was not significantly above chance (50%): \( t(29) = 1.92, \ p = .064 \). This provides a baseline level of performance for segmentation of the auditory speech stream. Consistent with our predictions, performance was not above chance when transitional probability cues to word
boundary were reduced. While performance was nearly above chance, it was not robust, and this study provides a baseline level of performance to compare with later audiovisual conditions.

**Experiment 8: Visual-only**

In Experiment 8 we test adults’ ability to segment visual speech without any accompanying auditory stimuli. We test participants on two different visual speech streams, the first recorded by an assistant that is aware of word boundaries (the *aware* condition), and the second recorded by an assistant that is misinformed of word boundaries (the *misinformed* condition). We predict that only the assistant that is aware of word boundaries will impart visual prosody onto the visual speech stream. When the visual speech stream contains visual prosodic cues (the aware condition), participants should successfully segment the visual speech stream. This prediction is consistent with previous findings that learners are able to use facial cues to segment a visual speech stream (Sell & Kaschak, 2009). However, when the visual speech stream does not contain consistent prosodic cues (the misinformed condition), then performance should not be above chance.

**Methods**

**Participants.** 59 (48 female, 11 male) Introductory Psychology students at The Pennsylvania State University participated in the study for course credit and were included in analysis. 29 participants were assigned to the aware condition, and the other 30 participants were assigned to the misinformed condition. Participants were all monolingual English speakers. We excluded from analysis any participants that reported an effort level at or below 5 on a 10-point scale (5) and any instance of technical malfunction (2). We also removed one participant who reported being deaf for the first 5 years of life.

**Stimuli.** The visual stimuli were created by digitally video recording two different male
research assistants lip-synching the artificial language (see Figure 13).

During recording, the familiarization audio stream from Experiment 7 was played on a nearby computer while the assistant read from a list of the items comprising the stream. In the aware condition, the assistant read from a list of words. Since the words were separated on the list, the assistant was aware of word boundaries. It was hypothesized that the assistant’s knowledge of the word boundaries resulted in visual prosodic cues during recording that could potentially provide segmentation cues during familiarization. In the misinformed condition, a second assistant read from a list of part-words. For example, the assistant would see the part-words /pu bo ke/, /taj ke gi/, /da gi bo/, which correspond to the words /bo ke taj/, /ke gi da/, and /gi bo pu/.

Consequently, any visual markers to acoustic cues (e.g. stress, pitch) that are evident in the visual display should have indicated segmentation at a different location than frequency cues.

In the aware condition, the list of words the assistant read from only contained six unique words that were repeated. Because the list in the misinformed condition contained part-words, there were 18 unique items, making it more difficult to maintain accuracy while lip synching to a full speed video. To resolve this issue, the video was recorded at half-speed. During video-recording, the audio stream that the assistant was lip-synching to was slowed to 50% of the original speed using Praat software. After recording, the movie was restored to the original speed of the audio stream by digitally speeding the video using iMovie. This allowed the assistant to lip-synch with the movie at a manageable rate while still maintaining the same rate as

Figure 13. Still frames from the visual stimuli used in Experiments 8 and 9. The aware condition is on the left, and the misinformed condition is on the right.
the audio stream, allowing us to synch the video-recording with the original audio stream that had an unaltered speed. Finally, the list the misinformed assistant read from began at a different point in the audio stream than the aware assistant. The list in the misinformed condition, because it consisted of 3-1-2 partwords, had an extra syllable at the beginning of the stream (starting on the 3rd syllable of the preceding word) and two extra syllables at the end of the stream (ending on the 1st and 2nd syllables of the following word). Thus, to make the misinformed and aware videos compatible, the first and last two syllables of the misinformed stream were removed.

In both conditions, assistants were asked to minimize their head movements by keeping the inion of their occipital protuberance (i.e. the highest point of the bump on the back of their heads) affixed to a point (the back of a thumb-tack) on the wall behind them. While we were interested in head movements produced during lip-synching, we wanted to reduce the range of those movements for several reasons. First, a wide range of movements would produce large jerks of the head when concatenating the loops of the videos (see below). Second, we sought to minimize any differences in the size of head movements, on average, that might naturally exist between the aware and misinformed actors. Finally, this constraint would allow subtle cues to prosody (e.g. lip aperture, small nods of the head; Graf et al., 2002; Blossom & Morgan, 2006) to emerge, though perhaps reduced, while preventing large artifacts in movement.

These movies formed 15 second clips of 18 words. The clips were imported into Adobe Premiere© and were faded in from black over a period of one second at the beginning of the clip and then faded out at the end of the clip. The clips were faded in order to remove any jerky head movements that resulted from looping the clips to form the familiarization stream. The clips were looped 16 times to create a four minute block, consisting of 288 words. Familiarization consisted of three four-minute blocks, with a one-minute break between blocks, for a total of 12
minutes of familiarization (864 words), which was the same duration and number of words as presented in Experiment 7.

In Experiment 8, the test stimuli consisted of 6 frequency defined visual words and 6 frequency defined visual part-words (using the same items as Experiment 7). Test items were created by extracting video segments from the two visual streams (aware and misinformed) by hand using Adobe Premiere© software. Because the test-items were extracted from the familiarization stream, all items (word and part-word) were seen during familiarization. However, as in Experiment 7, words occurred more frequently than part-words during familiarization. The test stimuli for the misinformed and aware conditions were created in the same manner. The words and part-words that were used, as well as how these items were paired together at test, were the same across conditions.

Procedure. Both conditions used the same procedure. The familiarization stream was presented using iTunes software. Participants were instructed to watch a 15 minute movie and that they would be tested on what they learned immediately following familiarization. They were also instructed several times that the movie did not have any sound. In addition, we asked participants to keep their headphones on to reduce ambient noise. The experimenter remained in the room during the experiment to monitor participants and ensure that they remained focused on the video display and followed directions.

Participants' knowledge of word boundaries was tested using a 2afc task between visual words and visual part-words. The test was presented using E-Prime 2 software. For each test trial, a visual word and visual part-word were presented, separated by a one-second pause. Participants were instructed to choose which item, 1 or 2, was more likely to be a word from the movie. They made their choices by pressing the corresponding key on a keyboard. The test was
self-paced—after each trial, participants moved to the next trial by pressing the spacebar. Unlike the auditory test in Experiment 7, the visual test did not exhaust every possible pairing of words and part-words. Words and part-words were paired so that the word and part-word were extracted from different points in the familiarization stream (i.e. no word and part-word test pair could have been taken from the same time window in the visual stream). Because of this constraint, each word was tested against 2 part-word foils, with each test pair presented twice in counterbalanced order resulting in a total of 24 test trials.

Results and Discussion

The mean percent of words chosen in the aware condition was 58.48% (SD=8.87%; see Figure 14). This level of performance was significantly above chance (50%), \( t(28) = 5.15, p < .001 \), all tests two-tailed. The mean percent of words chosen in the misinformed condition was 49.44% (SD=10.54%), which was not significantly different from chance, \( t(29) = -0.29, p = .775 \). An independent samples t-test revealed that performance in the aware condition was significantly greater than performance in the misinformed condition, \( t(57) = 3.56, p = .001 \).

Consistent with our predictions, participants were able to successfully segment the aware visual display and this level of performance was significantly greater than in the misinformed condition.

The results of this experiment have two key implications. The first is that the above chance performance for the aware face, as well as the lack of learning in the

![Figure 14. Results of Experiment 8, visual-only. The results of the aware condition are on the left (blue bar) and the results of the misinformed condition are on the right (green).](image)
misinformed condition, suggests that there were facial cues to word boundary in the aware stream, but not in the misinformed stream. This provides indirect evidence that the aware assistant encoded visual prosodic structure while recording the video display. In the general discussion we consider which facial feature(s) might have cued prosodic structure.

The second key contribution of this experiment is that, irrespective of which specific cues were utilized, participants were able to use information in the speaker's face to segment a visual speech stream. This is consistent with a previous study that similarly found that adults were able to identify visual words over visual part-words after familiarization to a visual speech stream (Sell & Kaschak, 2009). In addition, in the present study we were able to rule out a possible frequency-based explanation of these findings. As in the auditory speech stream in Experiment 7 (and virtually all statistical learning studies), words occur more often than part-words during familiarization to the visual speech stream. Thus, a simple explanation of our results (and of Sell & Kaschak's results) is that participants chose the words more often than part-words because they were more familiar with word test items. However, if this were the case, then we would have expected performance to be equivalent for the misinformed and aware conditions, since the same frequency imbalance between words and part-words was present in both conditions. Frequency cues were equally available in the misinformed and aware conditions, yet performance in the aware condition was significantly greater; thus, the results of Experiment 8 suggest that adults are able to use facial cues specific to word boundary to segment a visual speech stream.

A final consideration is that we might have expected performance in misinformed condition to be below chance. Since the misinformed assistant read from a list of part-words, it is reasonable to predict that he would have imparted prosodic cues that marked the location of
part-words. If perceived, we would expect participants to segment the stream in locations consistent with part-words rather than words. Since the 2afc test was between words and part-words, if participants had used facial cues to segment the stream at part-word boundaries, then they should have consistently selected part-words over words, which would have been reflected as performance below-chance. However, participants in the misinformed condition did not chose part-words more often than words (performance was at chance), suggesting that the misinformed assistant did not impart visual prosodic cues at part-word boundaries. We believe that this likely reflects the fact that the misinformed assistant did not encode the items on the list as lexical items (i.e. words) to the same degree as the aware assistant. As mentioned, the misinformed assistant read from a list of 18 unique items, whereas the aware assistant read from a list of 6 items that were repeated three times during the list; thus, it is possible that it was more difficult to encode each item in memory and thereby less likely that each part-word was treated as a lexical word. This notion is supported by anecdotal evidence from the assistants; the aware assistant was able to recall the words from the list two weeks after recording the video, suggesting that he had encoded the words into long-term memory, whereas the misinformed assistant was unable to recall most of the items on the list shortly after making the video. If the misinformed assistant did not view each item as a word, then he may have been less likely to impart prosodic cues.

It should be noted that there are two potential confounds in this experiment (also applicable to Experiment 9). First, there was a different actor used across conditions. This was necessary to prevent the assistant used to record the language in Experiment 8 from having his knowledge of word boundaries influence his production of prosodic cues in Experiment 9. Since this assistant remembered the location of word boundaries, it seemed potentially detrimental to
subsequently read from the same list using different boundaries. However, it is possible that participants may have preferred one face to the other, and therefore not have attended to the face of the misinformed actor to the same degree as the aware actor. Further, it is possible that one actor had stronger facial cues or visual prosody. The second difference concerns the method of creating the videos. As mentioned, the misinformed video was produced at a slower rate and then digitally sped, whereas the aware video was recorded in real-time. It is therefore possible that the difference in performance across conditions was the result of using this different recording method. Digitally speeding the video may have made the misinformed video less naturalistic, which consequently distracted participants in the misinformed condition relative to the aware condition. Since this potential confound is also present in Experiment 9, we return to this issue in the General Discussion and remark on planned studies that aim to address this concern.

**Experiment 9: Audiovisual familiarization, audio-only test**

The previous experiment demonstrated that participants are able to extract word boundary cues from a talking face display. In Experiment 9, we test whether adults can use these cues to segment the audio stream from Experiment 7 when presented synchronously with the visual streams from Experiment 8. If participants are able to use facial cues to segment speech, then we predict that pairing the audio stream with the aware visual display will facilitate segmentation performance, while pairing the audio stream with the misinformed visual display should not facilitate performance.

**Methods**

**Participants.** 60 (35 female, 25 male) undergraduate Introductory Psychology students participated in this experiment for course credit and were included in the analysis. Participants
were divided into two conditions, with half in the aware condition and half in the misinformed condition. All participants were monolingual English speakers. We excluded from analysis any participant whose self-reported effort level was at or below 5 on a 10-point scale (5), anyone who failed to follow directions (2) In addition, one participant’s test score was excluded because it was a statistical outlier⁴.

**Stimuli.** The familiarization streams for the aware and misinformed conditions were created in an identical manner. Each 15 second video clip from Experiment 8 was overdubbed with the 15 second audio clip from Experiment 7, creating an aware audiovisual clip and a misinformed audiovisual clip. Each video clip was then hand edited in Adobe Premiere© to synch to onset of the lip movements with the onset of the syllables in the audio stream. Similar to Experiment 8, the audio and visual streams were faded in over a period of one second at the beginning of the clip and then faded out at the end of the clip to remove head movement artifacts between clips. The clip was then looped 16 times to create a four-minute audiovisual block, that was repeated three times during familiarization (with a one second pause between blocks) for a total familiarization of 12 minutes.

The test was identical to the auditory-only test from Experiment 7.

**Procedure.** As in Experiment 8, the familiarization stream was presented using iTunes software. Participants in the aware condition were presented with the aware audiovisual stream, while participants in the misinformed condition were presented with the misinformed audiovisual stream. Participants in each condition were instructed to watch a 15 minute movie and informed they would be tested on what they learned. Participants were instructed to keep their headphones on throughout the experiment.

⁴ Outliers were defined as any data point that fell outside the range specified the following formula (see Cohen & Lea, 2004): lower bound, Quartile 1 - 1.5(Quartile 3-Quartile 1); upper bound, Quartile 3 + 1.5(Quartile 3-Quartile 1). The same formula was applied in each experiment, but only identified an outlier in Experiment 9.
The procedure for the audio test was identical to Experiment 7, and was the same for both the misinformed and aware conditions.

**Results and Discussion**

The mean percent of items correctly identified in the aware audiovisual condition was 59.72%, which was significantly greater than chance (50%), \( t(29) = 5.97, p < .001 \) (see Figure 15). The mean percent of items correctly identified in the misinformed audiovisual condition was 54.17%, which was significantly above chance, \( t(29) = 2.52, p = .018 \). Independent samples comparisons revealed that performance in the aware audiovisual condition was significantly greater than in both the misinformed audiovisual condition \( (t(58)=2.39, p=.020, d=0.63) \) and the audio-only condition from Experiment 7 \( (t(58) = 3.01, p = .004, d= .77) \). Performance in the misinformed audiovisual condition did not differ from performance in the audio-only condition \( (t(58)= -0.53, p=.597, d = -0.14) \). Similarly, a one-way ANOVA revealed an overall difference in performance on the audio test across the three conditions in Experiments 7 and 9 (audio-only,
aware audiovisual, and misinformed audiovisual), \( F(2, 87) = 5.02, p=.009, \eta^2=0.10 \). A Tukey HSD post-hoc analysis confirmed the results of the independent t-tests, revealing significant differences between the aware audiovisual condition (Experiment 9) and the audio-only (Experiment 7) and misinformed audiovisual (Experiment 9) conditions (\( p \)'s <.05), but no difference between the misinformed audiovisual and audio-only conditions (\( p >.05 \)). A post-hoc contrast analysis (the weights for the contrast analysis were \([-0.5, 1, -0.5]\) for the audio-only, aware audiovisual, and misinformed audiovisual conditions, respectively, which were selected in order to compare performance in the aware audiovisual condition with performance in each of the other two conditions) indicated that performance in the consistent audiovisual condition was significantly greater than in the other two conditions, \( t(87)=3.12, p=.002 \).

In Experiment 9, the presence of visual cues to word boundary in a talking face display (aware audiovisual condition) facilitated the segmentation of a synchronous speech stream. No facilitation was observed when the talking face display did not contain visual cues to word boundary (misinformed audiovisual condition). This suggests that participants in the aware condition were able to first extract visual prosodic cues (as in Experiment 8) and then apply this knowledge to segment the concurrent audio stream. Since there was no visual display at test, the above chance performance in the aware audiovisual condition could not be a function of visual cues to word boundary being available during test. Instead, our results suggest that participants used visual cues during familiarization to mark word boundaries in the auditory stream, suggesting that the visual input was integrated with the audio stream during learning, consistent with the results from Chapter III.

Furthermore, the facilitation in segmentation performance in the aware audiovisual condition does not appear to be simply a function of increased attention to a talking face display. Faces inherently attract attentional resources (Valenza et al., 1996); thus, it is possible that
participants had elevated levels of attention in the aware audiovisual condition relative to the audio-only condition and, since higher levels of attention are related to successful statistical learning (see Turk-Browne et al., 2005), this could have facilitated segmentation via transitional probability cues in the aware audiovisual condition. However, transitional probability cues were available in the misinformed audiovisual condition, as well as the enhanced attention brought on by a talking face display; thus, had the gains in segmentation performance in the aware condition been due to increased attention, then we would have anticipated to observe similar gains in performance in the misinformed audiovisual condition. Since there was no significant difference in performance between the audio-only and misinformed audiovisual conditions (and performance in the misinformed audiovisual condition was significantly poorer than in the aware audiovisual condition), this suggests that the facilitation observed in the aware condition was not a by-product of increased attention. However, because of the differences in the actors used and manner of creating each video, we cannot rule out attention as a possible explanation for our results. That is, if participants did not attend to the misinformed face to the same degree, then this discrepancy in attention may account for the difference in segmentation performance across conditions.

**General Discussion**

Across three experiments, we examined whether adult participants could utilize facial cues to word boundary to segment a continuous speech stream. In Experiment 7, we tested participants' ability to segment an artificial language that contained no acoustic segmentation cues (e.g. stress cues, Jusczyk et al., 1999) and reduced transitional probability cues to word boundary (e.g. Saffran et al., 1996a). Participants failed to segment the artificial speech stream when presented in isolation. In Experiment 8, we tested participants' ability to segment two
different visual speech streams, created by two different assistants while lip-synching to the auditory stream from Experiment 7. We manipulated whether the assistants who created the visual displays were aware or misinformed of word boundary during recording, with the prediction that when aware of word boundaries, assistants would likely impart visual prosodic contours in their facial movements if they were aware of word boundaries, but not if they were misinformed of word boundaries. Consistent with this prediction, participants were able to successfully segment the aware visual speech stream (without any accompanying audio signal) but failed to segment the misinformed visual speech stream. In Experiment 9, we asked whether participants could use these facial cues to segment the auditory stream from Experiment 7. We paired the audio stream with the aware and misinformed visual speech streams from Experiment 8 and then presented participants with the auditory-only test from Experiment 7. While pairing the audio stream with the aware visual speech stream significantly facilitated audio speech segmentation, pairing the audio stream with the misinformed visual speech stream did not improve segmentation performance.

The pattern of results from these three experiments provide the first evidence that learners may use cues from the speaker's face to resolve the segmentation problem. To date, research into speech segmentation has focused almost exclusively on the auditory input to language learners, identifying a number of auditory-based segmentation strategies including acoustic (Jusczyk et al., 1999; Friederici & Wessels, 1993; Houston et al., 2000) and distributional mechanisms (Saffran et al., 1996a,b; Brent & Siskend, 2001; Christensen, Allen, & Seidenberg, 1998). However, no strategy for speech segmentation has been proposed that incorporates visual cues from talking faces. The results of this study therefore provide a novel contribution to the speech segmentation literature by identifying an additional source of
information that language learners can exploit during speech segmentation.

The results of this study also demonstrate that facial cues are integrated with the speech stream during speech segmentation, supporting the results of Experiments 5 and 6 in Chapter III. In the present study, facial cues, which were used to segment a visual speech stream, were generalized to the audio stream. In the aware audiovisual condition of Experiment 9, participants used cues in the visual input to segment an auditory speech stream, providing evidence of a cross-modal effect on speech segmentation. Thus, mechanisms supporting speech segmentation do not appear to be modality independent, consistent with the conclusions from Chapters II and III.

A potential limitation of the current study is that there is only indirect evidence of the presence of prosodic cues. The present study relied on the logical inference that if prosody is cued in the speaker’s face (Yehia et al., 2002), and the actors were aware of word boundaries, then they would instinctively inflect a prosodic structure when recording the visual displays that would then be perceived by the participants. However, it is unclear which facial features cued the location of word boundaries. To address this concern, we attempted to measure a number of facial features to determine if any cues were correlated with word boundary. Using Matlab, we measured the vertical (y) and horizontal (x) position of various facial features (eyebrows, pupils, nose, mouth, and chin) of the aware facial display on a frame-by-frame basis. These measurements allowed us to calculate a number of facial cues that are known to be linguistically informative (see Yehia et al., 2002), including eyebrow height, eye gaze, lip aperture (vertical distance between lips), lip rounding (horizontal distance between the corners of the lips), and rotation of the head around the x, y, and z axes. We then measured the correlation of each of these cues with the corresponding syllable (1st, 2nd, or 3rd) and segmental (i.e. each individual
consonant and vowel, 1-6) positions of the speech stream. However, in this preliminary analysis, no individual facial cue was significantly correlated with word boundary or word position. Nonetheless, it is possible to speculate about which cues may have marked word boundary, since two cues did emerge as relatively more consistent with word boundaries than the others. Lip aperture tended to be greatest for the vowel in the first syllable of words than in the second or third syllables. In addition, x-axis rotation (nodding up and down) tended to align with word boundaries, as the face was tilted up for the first syllable and then gradually lowered throughout the word, tilting down on the final syllable (i.e. the assistant tended to nod at the offset of each word). This is consistent with previous analyses of visual speech that suggest x-axis rotation (nodding) and lip aperture are facial features that correlate with the prosodic structure of the accompanying speech signal (Graf et al., 2002). Again, it should be stressed that this is purely speculative, since our analysis did not reveal any significant correlations between facial features and word boundary. In the future, more fine-grained analyses with greater spatial and temporal specificity (e.g. using the specialized Image and Video Processing Toolkit for Matlab software) may be able to identify the cues to word boundary available in the visual stream. Alternatively, future work may be able to use eye-tracking methods to determine which facial features learners attend to during segmentation. This latter method would also potentially provide insight into how feature detection strategies might develop during the time-course of segmentation (i.e. whether learners moderate which facial features they attend to across familiarization).

A second potential limitation of the present study concerns a possible methodological confound for our results. As mentioned earlier, it is possible that any difference in performance between the aware and misinformed conditions in Experiments 8 and 9 was an artifact of using different actors and different video-recording methods for the aware and misinformed visual
streams. While the critical finding across experiments was that the aware facial display facilitated segmentation performance relative to the audio-only condition, these potential confounds prevent us from ruling out an alternative accounts for this finding. Specifically, the goal of the misinformed condition in Experiment 9 was to preclude the possibility that increased attention to a talking face display was responsible for the facilitation in the aware audiovisual condition. However, since participants may have paid less attention to the face in the misinformed condition (due to properties of the face itself, or the manner in which the video was created), we cannot rule out attention as an alternative account of our findings. Planned studies will address this concern by creating a new 'aware' visual display using the same assistant and same video-recording methods as were used to create the misinformed display in the present study. If this new aware visual display facilitates segmentation, then this will alleviate concerns about the stimuli used here and bolster our claims that participants can use facial cues to segment speech.

It is important to note that we found a benefit of facial information despite restricting the range of head movements in the present study. Prior studies investigating visual prosodic cues did not constrain the types of head movements of the actors, instead opting for a more naturalistic approach (Graf et al., 2002; Yehia et al., 2002; Munhall et al., 2004; Blossom & Morgan, 2006; Sell & Kaschak, 2009). In particular, Yehia and colleagues measured head movement and the movement of facial articulators as two actors recited simple sentences. The actors were free to move their heads in any direction and as wide a range as was natural. In contrast, actors in the present study were asked to restrict their global head movements. Since head movements during natural discourse are likely much larger and more fluid, the results of the present study may actually underestimate the role of facial cues in speech segmentation, as it is
possible that visual cues will be more accessible if the head movements are not restricted. Thus, future research will attempt to make videos in a more naturalistic manner and allow actors to use a wider range of head motion. More pronounced head movements may also help us more accurately identify which features mark word boundaries in speech.

Finally, given that we are ultimately interested in how facial cues are incorporated during language acquisition, it will important to assess the development of this ability in infants. Given evidence that newborns have an innate preference for attending to faces (Goren, Sarty, & Wu, 1975; Morton & Johnson, 1991; Valenza, Simion, Macchi Cassia, & Umiltà, 1996; Simion, Macchi Cassia, Turati, & Valenzia, 2001; Turati, 2004), facial cues may be highly salient for infants acquiring a language. Future work will assess whether infants at varying points in development can use facial cues to guide speech segmentation. In addition, this work will examine whether the use of facial cues changes across development. There is recent evidence that visual speech perception, like auditory speech perception (see Werker & Tees, 1984), becomes tailored to the native language, as infants gradually become less sensitive to non-native visual speech cues during the first year of life (Weikum et al., 2007). These findings suggest that, if infants are able to use facial cues to segment language, their sensitivity to these cues may become honed to the native language during development; thus, it will be important to assess infants from early (6-8 months of age) to later (12-14 months) stages of development.
CHAPTER V

General Discussion

Although research into language acquisition has primarily considered it to be an auditory phenomenon, the language learning environment is typically multimodal (Massaro, 1998; Shams & Seitz, 2008); thus, the goal of the present studies was to investigate how mechanisms of early language acquisition operate over multimodal input. To this end, I examined the interaction of visual and auditory input during statistical learning, a mechanism believed to support early speech segmentation (Saffran et al., 1996a,b).

In Chapter II, I tested whether multi-stream, multimodal statistical learning is achieved independently for each modality. In Experiment 1, I measured participants' baseline ability to segment a tone and shape stream when presented in isolation, replicating previous findings (Saffran et al., 1999; Fiser & Aslin, 2002). In Experiment 2, I paired these tone and shape streams, presenting learners with an audiovisual stream that maintained a perfect correlation between individual elements across streams. In this condition, participants were able to segment each stream, a finding that is consistent with prior research in multisensory statistical learning (Seitz et al. 2007). In Experiment 3, I removed the perfect correspondence between tone and shape elements. Performance in this condition remained above chance for each stream, and was equivalent to Experiment 2. In Experiment 4, we offset the alignment of the triplet boundaries across streams, which led to a decrement in performance, as learners were unable to successfully segment either the audio or visual streams. These results demonstrate a cross-modal effect on statistical learning (see below), as the relationship of boundary or positional information between streams influenced participants' ability to segment each stream.
In Chapter III, I further investigated the claims of modality independence by examining whether learners can integrate audio and visual input during statistical learning. To test for audiovisual integration, I exploited a classic audiovisual illusion, the McGurk effect, in order to manipulate the statistical properties of artificial languages. In Experiment 5, baseline measurements of Language A revealed that participants were readily able to segment this language on the basis of transitional probabilities. However, when the same language was paired with a face display, the perceived McGurk illusion altered the statistical structure of the language so that adjacent regularities no longer provided cues to word boundary, obstructing participants’ ability to segment the speech stream. In contrast, in Experiment 6, the McGurk illusion in the audiovisual presentation altered the statistical structure of the language (Language B) to facilitate segmentation performance compared to when the language was presented in the auditory modality alone. The complementary pattern of results in Experiment 5 and Experiment 6 suggest that participants are able to integrate audio and visual information during statistical learning. In addition, consistent with the results of Chapter II, the findings of Experiments 5 and 6 provide evidence of specific cross-modal effects on statistical learning, as information in the visual modality altered participants segmentation of an auditory speech stream.

Finally, the goal of the studies in Chapter IV was to examine whether visual information, and specifically facial cues, might have functional benefits for early language acquisition. To address this question, talking face displays were paired with a statistically uninformative language (i.e. the sound correspondences between adjacent sounds did not robustly cue word boundaries). In Experiment 7 I tested participants’ ability to segment this audio stream. When presented in isolation, the language was not learned significantly above chance. In Experiment 8, I presented learners with two different visual speech streams, one produced by an aware
assistant (the assistant read from a list of words) and one produced by a *misinformed* assistant (the assistant read from a list of part-words). At test, participants in the aware condition were able to discriminate between visual words and visual part-words, whereas participants in the misinformed condition were at chance. In Experiment 9, I paired each of the two visual displays from Experiment 8 with the audio stream from Experiment 7. After familiarization, participants completed the same audio-only test from Experiment 7. Segmentation performance in the aware audiovisual condition was significantly greater than performance in the audio-only and misinformed audiovisual conditions. This indicates that participants were able to use facial cues to segment an auditory speech stream. In addition, performance in the misinformed audiovisual condition did not differ from the audio-only condition, suggesting that the facilitation observed in the aware audiovisual condition was not a general property (e.g. enhanced attention) of incorporating face displays in segmentation tasks.

The findings from the experiments reported here provide insight into how statistical learning and speech segmentation mechanisms operate over multimodal stimuli. Consequently, these results have several implications both for theories of statistical learning as well as for how visual input, specifically information from the speakers’ face, may help guide early mechanisms of language acquisition. In the remainder of this chapter, I will first discuss how these results inform theories of modality independence and modality specificity in statistical learning, and subsequently will consider how visual input may, more broadly, contribute to language acquisition.

**Modality Independent Statistical Learning**
One of the principle goals of the research presented in Chapters II and III was to examine whether statistical learning in one modality is independent of stimuli in another modality. Previous research in multimodal statistical learning found that learning in each modality did not appear to be influenced by the simultaneous acquisition of a stream in the other modality (Seitz et al., 2007). As such, the authors concluded that statistical learning is modality independent. In this view, learning in one modality should not be influenced by stimuli in other modalities; that is, there should be minimal cross-modal effects on multisensory statistical learning.

In contrast to the predictions of Seitz and colleagues, I demonstrated cross-modal effects on statistical learning in two different multimodal contexts. In a multi-stream, multisensory segmentation task similar to the one employed by Seitz et al., I found that participants were able to simultaneously segment both a tone and shape stream so long as triplet boundaries were aligned across streams (Chapter II). In comparison, when the boundaries were misaligned, performance dropped to chance for both streams. The only distinction between Experiment 3 (in which participants successfully segmented each stream) and Experiment 4 (in which participants failed to segment either language), was the relative position of triplet boundaries across streams. If each stream was processed independently, however, then the position of triplet boundaries in one modality should have had no bearing on participants’ ability to segment the stream in the other modality. Instead, the pattern of learning across experiments indicates that multisensory statistical learning is sensitive to the coherence and alignment of boundary information across modalities. Thus, these results, which provide evidence of cross-modal effects on multisensory statistical learning, are inconsistent with a modality independent account of statistical learning.

The results of Experiments 3 and 4 also provide insight into the types of cross-modal relationships that may impact multisensory statistical learning. In each case, I disrupted one type
of cross-modal relationship. While offsetting the alignment of boundary information (Experiment 4) had deleterious effects on performance, removing the one-to-one mapping of individual tone and shape segments did not impact performance, as performance in this condition (Experiment 3) was equivalent to an earlier condition that preserved this correspondence (Experiment 2). The juxtaposition of successful segmentation in Experiment 3 with the inability to segment either stream in Experiment 4 suggests that not all types of cross-modal relationships equally affect multisensory statistical learning. In particular, the results presented in Chapter II provide evidence that boundary information may be particularly critical for multimodal statistical learning (see also, Cunillera et al., 2010; 2008).

As previously noted, while the results from Chapter II provide evidence of a cross-modal effect on statistical learning, the source of this effect is not clear, as the decrement in performance in Experiment 4 could potentially be attributed to global effects on processing (e.g. increased attentional or task demands). A more stringent test of modality independence, then, would be to demonstrate that information from two modalities was integrated during statistical learning, predictably altering segmentation performance in contrasting patterns. To this end, in Chapter III, I demonstrated audiovisual integration during statistical learning, utilizing the McGurk illusion to both hinder (Experiment 5) and facilitate (Experiment 6) statistically-guided speech segmentation. Across experiments, I observed a crossover interaction in segmentation performance between languages and display type, such that the presentation of a talking face display (eliciting a McGurk illusion) during familiarization differentially affected segmentation performance of the two languages in an auditory-only test. This pattern of results is consistent with what would have been expected if the sound inventories of each language included the integrated percept produced by the McGurk illusion (i.e. transitional probabilities were
calculated over /je lu ni/ and /so bae ta/ rather than /je lu mi/ and /so bae pa/). Thus, the findings from the experiments reported in Chapter III provide behavioral evidence that learners can integrate audio and visual input during statistical learning. Further, these results demonstrate specific cross-modal effects (i.e. the effect was dependent on the context, rather than an overall decrement or facilitation across contexts) on statistical learning, since the effect of visual articulatory cues on auditory statistical learning was contrastive across the two audiovisual conditions (Language A and Language B). This suggests that the cross-modal effect observed in Chapter III was not due to global effects on processing, such as increased attentional demands in the audiovisual condition, as these global effects were equivalent across audiovisual conditions.

The results of the first six experiments provide evidence of cross-modal effects on statistical learning. Such cross-modal effects are inconsistent with modality independent accounts of statistical learning (e.g. Seitz et al., 2007). Instead, these results establish that statistical learning mechanisms, which are believed to broadly support language acquisition processes, incorporate input from multiple modalities. If these mechanisms inclusively process stimuli across modalities, what are the consequences of such modality inclusivity for statistical learning?

One possibility is that statistical learning may benefit from the availability of multisensory input. Because the perceptual environment is multimodal, perceptual learning systems most likely evolved to operate optimally in multisensory environments and, consequently, learning may be more efficient and effective in multisensory contexts (Shams & Seitz, 2008). Consistent with this, there is evidence that multisensory training facilitates unisensory learning. For example, Seitz and colleagues (2006) asked participants to visually detect and discriminate between motion patterns of dots on a screen (i.e. are the dots moving to
the left or right?). While the test was always only in the visual modality, training was either visual only or audiovisual. During audiovisual training, the dots were presented with an auditory stimulus (white noise) that appeared to move in the same direction as the dots (auditory movement was induced by linearly changing amplitude between left and right speakers). Audiovisual training resulted in a 60% increase in the rate of learning (i.e. amount of training required to reach asymptote) as well as greater maximal levels of performance (Seitz, et al., 2006). Furthermore, auditory recognition of voices is enhanced by the arbitrary pairing of the voices with pictures of faces during training (von Kriegstein and Giraud, 2006). Likewise, Lehman and Murray (2005) found that memory for visual objects was improved if the objects were presented with semantically congruent auditory stimuli (e.g. a picture of a dog and the sound of a dog barking). Enhanced multisensory learning has also been observed across a range of species, including fruit flies (Guo & Guo, 2005) and quail embryos (Lickliter, Bahrick, & Honeycutt, 2002), suggesting that sensitivity to cross-modal relationships is not limited to primates, but is instead a general property of perceptual systems.

Although there are a number of potential explanations for enhanced multisensory learning (see Shams & Seitz, 2008), one possibility that has received considerable empirical support is an increase in access and attention to information that is redundant across modalities in multimodal rather than unimodal stimuli (see Bahrick, Flom, & Lickliter, 2004). Stimuli consist of both modality-specific information (e.g. pitch or shape) and amodal information, or information that is not specific to a given modality and which is redundant across multiple modalities (e.g. rate, rhythm, intensity; Gibson, 1966). For example, a bouncing ball conveys visual information (e.g. color, texture, and shape) and auditory information (e.g. pitch). In addition, the auditory and visual stimuli are synchronous, share an origin, share a spatial
location, and have a common pattern of rate, rhythm, and intensity (Bahrick et al., 2004). According to the *Intersensory Redundancy Hypothesis* (Bahrick & Lickliter, 2000), redundancies in this amodal information increases the salience of temporal and spatial regularities, guiding learners to attend to shared properties of the input and enhancing perceptual capabilities (for a review, see Bahrick & Hollich, 2008). Since these redundancies are not available in a single modality, perception is enhanced in multimodal conditions relative to unimodal conditions. For example, infants' ability to detect changes in the rhythm of a swinging hammer was greater for bimodal (audio and visual) presentation than for either audio or visual stimuli alone (Bahrick & Lickliter, 2002). If, as I suggest here, statistical learning mechanisms incorporate information from multiple modalities, then the availability of redundant amodal information in multimodal contexts may increase attention to temporal and spatial regularities, thereby facilitating statistical learning.

This claim is bolstered by recent evidence that rule learning may be facilitated by redundant multimodal information (Frank, Slemmer, Marcus, & Johnson, 2009). Frank and colleagues presented 5-month-old infants with synchronous audiovisual strings of colored shapes and speech syllables that followed either an ABA or ABB sequencing rule. In this bimodal condition (in which information was presented in auditory and visual modalities), infants were able to successfully learn abstract rules, discriminating between novel strings that either followed or violated the rule. However, in unimodal control conditions, 5-month-old infants failed to acquire the rule structures of the shape or speech strings in isolation. The ability to extract rules from unimodal speech stimuli is known to be available by 7 months of age (Marcus et al., 1999). Thus, the results from Frank et al. (2009) indicate that redundant multimodal information facilitates the extraction of rule structures, allowing infants to acquire these structures at an
earlier age than rule learning in unimodal stimuli. Likewise, statistical learning mechanisms guiding speech segmentation may not only be facilitated by multimodal stimuli, but may actually be available to learners at an earlier point in development than has been indicated by previous unimodal studies (e.g. Saffran et al., 1996a; Thiessen & Saffran, 2003).

Given evidence that perceptual processes, including perceptual learning, are augmented by the availability of information from multiple modalities, it is likely that there are gains in statistical learning in multimodal contexts. Specifically, redundant information across modalities may enhance learners' ability to extract temporal regularities early in the language acquisition process. As mentioned, statistical learning has been investigated primarily in unimodal contexts. Thus, the studies reported here in Chapters II and III provide a critical starting point for investigating statistical learning in multimodal environments by demonstrating that statistical learning mechanisms are not modality independent. Future work is needed to assess the degree to which statistical learning may be enhanced by the availability of redundant information. For example, since the attentional and perceptual benefits of inter-sensory redundancy are greatest in early infancy (see Bahrick & Hollich, 2008), future research with infants may indicate that, as in the rule-learning study reported above (Frank et al., 2009), statistical learning abilities may be apparent at an earlier age when tested under multimodal learning conditions.

**Modality Specific Statistical Learning**

In addition to providing evidence against modality independence, the findings from Chapters II and III may inform modality-specific theories of statistical learning (e.g. Conway & Christiansen, 2005), which posit that input from different modalities is processed by separate statistical mechanisms. While the findings reported here do not rule out such theories, they do
provide two pieces of evidence that challenge the claims of modality-specific accounts of statistical learning. First, Seitz et al. (2007) suggested that the absence of cross-modal effects in their multisensory paradigm was consistent with modality-specific statistical learning. However, the results from Chapter II temper this assertion, as I observed cross-modal effects on multisensory statistical. Second, in Chapter III, I demonstrated that learners can integrate audio and visual information during statistical learning, and this integration can alter the statistical representation of artificial languages. In the strictest version (as suggested by Seitz et al., 2007), modality-specificity would predict that statistical learning should not operate over integrated audio-visual percepts, though it may be possible to account for multisensory integration within other modality-specific frameworks (see below). The results from Chapters II and III provide evidence that, at some level, statistical learning includes a modality-general component. In contrast, Conway & Christiansen (2005) found quantitative and qualitative differences in statistical learning of auditory, visual, and tactile sequences, providing evidence for modality-specific constraints on statistical learning. How, then, might one reconcile the modality-general effects observed here with evidence of modality-specific constraints on statistical learning?

**Task constraints.** The first possibility is that the modality-specific effects observed in Conway & Christiansen (2005) may reflect the type of tasks used in their paradigm. Their study employed an artificial grammar learning (AGL) task, which has three key distinctions from the statistical learning tasks presented here, as well as those used in prior studies that found similarities in statistical learning across modalities (e.g. Saffran et al., 1999; Fiser & Aslin, 2002; Newport & Aslin, 2004; Creel et al., 2004). First, the tests used in AGL and traditional statistical learning paradigms require learners to perform fundamentally different tasks (Saffran, Pollack, Seibel, & Shkolnik, 2007). In the test phase of the AGL task used by Conway & Christiansen
(2005), participants were presented with novel strings that were either legal or illegal sequences based on the artificial grammar used to create the familiarization strings. Thus, to successfully identify legal test sequences, participants were required to generalize knowledge to sequences that had not occurred during familiarization. Statistical learning tasks, on the other hand, typically test participants on sequences that had occurred during familiarization, assessing learners’ ability to track associations between specific stimuli. Thus, properties of the stimulus (i.e. modality) may constrain the generalizability of learning, which would affect AGL tasks but not statistical learning tasks (see Saffran et al., 2007).

Second, there may have been differences between the types of regularities available in the three AGL tasks in Conway and Christiansen (2005). In the visual and tactile tasks, participants tracked sequences of images or pulses that varied in both spatial location and temporal order; thus, these tasks required participants to extract structure from spatio-temporal dependencies. In the audio task, the auditory stimuli did not vary in spatial location and, as such, participants in the audio task were acquiring structure from purely temporal regularities. It is possible that these differences in the visual/tactile and audio tasks contributed to the quantitative and qualitative modality constraints Conway and Christiansen (2005) reported.

Finally, there is debate whether AGL tasks and statistical learning tasks are supported by entirely different mechanisms. It has been proposed that AGL is guided by abstract, algebraic-like rule learning and statistical learning is guided by associative mechanisms (Pena et al., 2002), though this proposition is not universally accepted (see Perruchet & Pacton, 2006). Given these significant distinctions between AGL and statistical learning tasks, it is possible that the discrepancies between the findings of the present studies and those of Conway & Christiansen (2005) could be the result of using fundamentally different tasks.
A Mixture Model. A second, and potentially more interesting possibility is that statistical learning mechanisms may reflect both modality-specific and modality-general components. There is precedent for the mixture of modality-specific and modality-general, or cross-modal effects in both attention (e.g. Driver & Spence, 1998) and early sensory perception (e.g. Besle et al., 2005). For example, in attention research, modality-specific and cross-modal effects have been observed within the same task (see Driver & Spence, 1998). When participants were presented with a stream of either auditory or visual stimuli in rapid succession, their ability to identify targets declined if the target was presented less than 500 ms after the previous stimulus (Duncan, Martens, Ward, 1997). This is known as the attentional blink effect. When the stimuli stream alternated between audio and visual stimuli, no attentional blink effect was observed, suggesting that the visual and auditory stimuli were monitored by modality specific attentional resources. However, in the same task, overall performance was poorer in the bimodal condition relative to the unimodal conditions. Target monitoring in each modality was influenced by the occurrence of stimuli in another modality, demonstrating a cross-modal effect on the allocation of attentional resources (Driver & Spence, 1998). Thus, in the same attentional blink task, researchers observed effects that were consistent with modality-specific and modality-general accounts of attention.

Similarly, Besle and colleagues (2005) found evidence of early audiovisual integration in basic sensory processing, even though the ultimate representations in sensory memory did not encode a trace of the integrated percept. The authors used an audiovisual oddball detection task, in which, for the majority of trials, participants are presented with the same stimulus. Interspersed throughout these standard trials, participants are presented with a 'deviant' stimulus that is slightly different from the standard stimulus along some dimension. If participants
perceive this change (typically implicitly), then it results in an increased Mismatched Negativity (MMN) ERP component. In the audiovisual condition of Besle et al., participants were presented with the same pairing of audio and visual stimuli for 76% of trials. In the remaining trials, deviations from this standard pairing were presented and MMNs were measured. The authors compared MMNs in this condition with MMNs elicited in unisensory conditions (participants were presented with either audio or visual stimuli), finding a significant difference in the spatio-temporal properties of the MMNs across conditions. The authors argue that this difference in MMNs reflects early cross-modal interaction of the audio and visual components of the audiovisual stimuli. In contrast, topographical analyses of the distribution of waveforms across the scalp indicate that, in the audiovisual condition, the three different types of changes (changes to the audio stimulus, the visual stimulus, or both) were generated in separate, sensory-specific cortical areas. On the basis of this evidence, the authors claim that the MMN processes largely operate at a modality-specific level, with no encoding of the integrated audiovisual percept in sensory memory. The combination of these two findings suggest that MMN processes to audiovisual stimuli likely reflect the interaction of separate audio and visual sensory representations.

The results of Besle and colleagues (2005) suggest that it is possible to account for both cross-modal effects and multisensory integration within a modality-specific framework. This is consistent with recent theories of multisensory integration that propose integration occurs via association between modality specific representations (Bernstein, 2005). In the literature on multisensory integration, there is considerable debate regarding whether audiovisual integration, and the effects that arise from it (e.g. the McGurk effect), is the product of convergence or association. In theories of convergence (e.g. Summerfield, 1987; Rosenblum, 2005; Fowler,
 multisensory integration is achieved by transforming each unimodal signal into a common format. For example, audio and visual components of speech share a number of linguistic features that are comparable across modalities (Summerfield, 1987), as well as amodal features such as rate, duration, rhythm, and intensity (Gibson, 1966); thus, it is possible to translate audio and visual signals into a single audiovisual signal on the basis of these common features (Schwartz, Robert-Ribes, & Escudier, 1998). Theories of association (e.g. Bernstein et al., 2004; Bernstein, 2005; Bernstein et al., 2008) alternatively propose that audiovisual speech perception results in separate, modality-specific representations that then become associated upstream in processing—stimulus information is never transformed into a common format, consistent with the findings of Besle et al. (2005). Support for association over convergence comes from evidence of top-down effects on the McGurk illusion. Perceivers that are familiar with the speaker do not experience a McGurk illusion (Walker, Bruce, & O’Malley, 1995). Such top-down effects are incompatible with theories of convergence, which make the prediction that early bottom-up processing of speech features (i.e. the McGurk effect) should not be sensitive to high-level, top-down information such as familiarity with the speaker.

Associative theories of multisensory integration suggest that while perception is integrated across modalities, perception remains modality-specific. Thus, a mixture-model of both sensory-specific and cross-modal influences on statistical learning is consistent with the predictions of an associative account. In this account, multisensory statistical learning would result in the encoding of sensory-specific representations, yet learning in each modality would not be independent due to cross-modal association. Thus, it is possible that, when segmenting multimodal stimuli, statistical learning is both modality-specific and prone to cross-modal effects on learning. A mixture model would provide a means to reconcile the modality-specific effects
in Conway and Christiansen (2005) with the cross-modal effects on multisensory statistical learning observed in the present studies. Differences in the mechanisms supporting statistical learning in each modality would give rise to qualitative and quantitative differences across modalities (as in Conway & Christiansen, 2005), while simultaneous segmentation of multiple, multimodal streams would be sensitive to the cross-modal relationship between streams. Planned ERP research will test the predictions of this model by adapting the MMN paradigm used by Besle and colleagues (2005) to examine the time-course of integration during multisensory statistical learning. If multisensory statistical learning is guided associations across modality-specific mechanisms, then I would expect to find similar results to Besle et al. (2005), observing a difference in MMNs between unimodal and multimodal conditions, while also finding evidence that MMNs for each modality are generated in different cortical regions.

**Does Visual Speech Facilitate Speech Segmentation?**

The previous sections reviewed evidence that statistical learning mechanisms are not modality independent, and accordingly these mechanisms integrate facial information with a speech stream during segmentation. The current section considers whether the integration of facial information may impact speech segmentation. A number of studies have demonstrated that talking faces contain a wealth of linguistic information (e.g. Kuhl & Meltzoff, 1982, 1984; Yehia et al., 2002) that can be used to enhance speech perception (Sumby & Pollack, 1954). In addition, there is a growing body of evidence that language processes may be facilitated by presence of synchronous talking faces (e.g. Weikum et al., 2007; Soto-Faraco, et al. 2007; Patterson & Werker, 2003; Hollich et al., 2005). I therefore examined whether information in the speaker’s face facilitates speech segmentation.
In Chapter IV, I tested adult learners’ capacity to use visual cues in a talking face display to segment a speech stream. Since stress and pitch cues to word boundary, which learners can use to segment speech (Jusczyk et al., 1999), are available in talking face displays (Yehia et al., 2002), I predicted that learners would be able to use these visual prosodic cues to segment a synchronous speech stream. Consistent with these predictions, participants were more accurate at segmenting an artificial speech stream when the visual display cued the location of word boundaries than when the visual cues did not coincide with the word boundary or when the speech stream was presented in isolation. This indicates that participants were integrating visual prosodic cues, such as head nodding and lip aperture, with the speech signal and then using this prosodic structure to segment the speech stream.

The results from Chapter IV provide the first evidence that adults can utilize facial cues to resolve the segmentation problem. These findings contribute to the growing body of evidence that faces play an important role in language development. Infants as young as 2-months-old are able to match facial cues with phonetic features (Patterson & Werker, 2003), and 4 and 6-month old infants can use this information to discriminate between languages (Weikum et al., 2007), a particularly challenging obstacle for bilingual language learners. Furthermore, the temporal synchrony between talking faces and speech facilitated 7.5-month-old infants’ ability to separate speech streams in the presence of multiple speakers (Hollich et al., 2005). Although the present studies tested adults rather than infants, the findings indicate that information available in the speaker’s face may provide an additional segmentation strategy to language learners that might complement the array of segmentation cues available in the auditory input. While there is evidence that facial cues may not be a primary segmentation strategy (e.g. faces do not appear to facilitate segmentation performance when statistical cues are also available; Mitchel & Weiss,
2010; Sell & Kaschak, 2009), facial cues may play a more prominent role in noisy environments, where auditory-based strategies may be hindered.

In speech perception, perceptual benefits of talking face displays are greater in noisy environments than in quiet environments (Sumby & Pollack, 1954; Grant & Seitz, 2000), suggesting that learners may weight facial cues above other segmentation cues in noisy environments. Because natural learning environments are typically noisy (Picard & Bradley, 2001), facial cues may play a significant role in speech segmentation; thus, future work should investigate the role of visual speech on segmentation in noisy environments. One way to investigate visual speech segmentation in noise would be to replicate the experiments reported in Chapter IV in a noisy environment (e.g. Hollich et al., 2005 used a 10db signal-to-noise ratio). If learners attend to facial cues to a greater extent in the presence of noise, then performance in the aware audiovisual condition should be enhanced in the noisy environment. Additionally, one could test how facial cues are weighted against other segmentation cues. One possibility would be to use a competing cue paradigm (Weiss, Gerfen, & Mitchel, 2010), in which two segmentation cues are available to learners, but these cues are incompatible, such that they signal the learner to segment the stream in separate locations. In this type of task, it would be possible to pit facial cues against other segmentation cues and then gradually decrease the signal-to-noise ratio (i.e. increase the level of noise) to determine the point at which segmentation preferences might switch.

**Broader Applications of Visual Speech.** The results from Chapter IV provide evidence that visual speech facilitates speech segmentation in typical language development. If this is the case, what are the consequences of visual speech on atypical language development? For example, if visual cues play a role in language acquisition, then it might be expected to observe
language delays in children that are congenitally blind. Though there has been relatively little empirical research on this topic, there seem to be few overall language delays associated with congenital blindness (Landau & Gleitman, 1985; McConachie and Moore, 1994). Blindness is associated with slight delays in word comprehension ability (Mills, 1988), yet these deficits are more likely a consequence of reduced and altered linguistic input from caretakers than to visual deficits (Anderson, Dunlea, & Kekelis, 1993). However, there is some evidence that blind children experience delays in the acquisition of phonemes that are visually marked (e.g. /p/, /b/, and /m/), suggesting that there are some specific linguistic deficits associated with the absence of visual speech information (Mills, 1983). Moreover, lack of access to visual information appears to be associated with delays in development of pragmatic aspects of language (Tadic, Pring, & Dale, 2010). This provides indirect evidence that cues in the speakers’ face play an important role in language acquisition.

In addition, visual speech information can be used in training scenarios to facilitate speech perception and production in individuals with hearing loss as well as late learners of a second language (Massaro & Light, 2004, 2003). Massaro and Light (2004) used a computer-animated talking head (Baldi) as a language tutor for a group of individuals with hearing loss between the ages of 8 and 13 who had severe deficits in speech perception and production. Baldi was used to illustrate articulatory gestures associated with different minimal pairs (e.g. /f/ vs. /v/). After lengthy training (21 weeks), all participants showed significant gains in their ability to both produce and perceive these contrasts. Similarly, the same authors (Massaro & Light, 2003) used Baldi to train Japanese late-learners of English on /r/ and /l/ contrasts, one of the more well-known difficulties with a second language. After just 6 days of training, perception and production accuracy for this contrast was significantly improved, and this facilitation
generalized to novel words. The results of these training studies indicate that visual speech not only plays a role in language acquisition, but it can also be used to help overcome deficits in speech production and perception.

Finally, visual speech may have implications for the treatment of broader deficits in language acquisition. Given the prominence of facial cues in noisy environments, one might imagine that visual cues provided by the speaker’s face may play an increased role in the language development of children that have deficits in auditory comprehension. Children diagnosed with Central Auditory Processing Disorder (CAPD) typically have global deficits in speech comprehension, including speech segmentation, that result in delays in linguistic development (Keith, 1999). In addition, suggested treatment and management strategies often include emphasizing visual cues (Chermak & Musiek, 1992). If, through future research, it is possible to identify the specific facial cues that facilitate speech segmentation, it may be possible to use this information to augment treatment approaches for children diagnosed with CAPD both by training children to attend to these cues as well as training caretakers to emphasize the relevant facial cues.

Conclusions

Two fundamental concerns for research in language acquisition are the types of information available to learners in language input and the extent to which language acquisition relies on this input. In this thesis, I have presented evidence that processes supporting language acquisition are not limited to the auditory modality, but instead encompass information across the senses, especially information provided by the face of a speaker. This research therefore adds to a growing body of work illustrating the foundational role of visual input in language
acquisition (e.g. Weikum et al., 2007; Hollich et al., 2005), requiring theories of language
acquisition to reconsider the nature of input to language learners.
References


(Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 3-25). Hove, U.K.: Psychology Press.


VITA

Aaron D. Mitchel

Education

2011 (Expected)  Ph.D. in Psychology, The Pennsylvania State University
2008  M.S. in Psychology, The Pennsylvania State University
2005  B.A. with honors in Psychology, Macalester College

Selected Honors and Awards

2009-2010  Graduate Teaching Fellowships, PSU
2009  Language Science Scholar, Center for Language Science, PSU

Selected Publications


Selected Presentations


