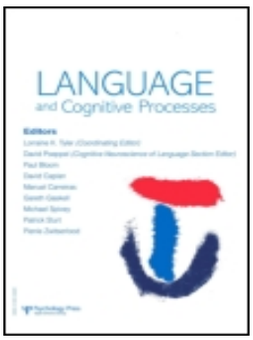

Routledge
Taylor & Francis Group

Language and Cognitive Processes

ISSN: 0169-0965 (Print) 1464-0732 (Online) Journal homepage: http://www.tandfonline.com/loi/plcp20

What's in a face? Visual contributions to speech segmentation

Aaron D. Mitchel & Daniel J. Weiss

To cite this article: Aaron D. Mitchel & Daniel J. Weiss (2010) What's in a face? Visual contributions to speech segmentation, Language and Cognitive Processes, 25:4, 456-482, DOI: 10.1080/01690960903209888

To link to this article: http://dx.doi.org/10.1080/01690960903209888

Published online: 26 Oct 2009.

Submit your article to this journal

Article views: 195

View related articles

Citing articles: 8 View citing articles

Full Terms & Conditions of access and use can be found at
http://www.tandfonline.com/action/journalInformation?journalCode=plcp20

Download by: [Penn State University] **Date:** 12 January 2016, At: 08:08

LANGUAGE AND COGNITIVE PROCESSES
2010, 25 (4), 456–482

Psychology Press
Taylor & Francis Group

What's in a face? Visual contributions to speech segmentation

Aaron D. Mitchel and Daniel J. Weiss
Department of Psychology, Pennsylvania State University, University Park, PA, USA

Downloaded by [Penn State University] at 08:08 12 January 2016

Recent research has demonstrated that adults successfully segment two interleaved artificial speech streams with incongruent statistics (i.e., streams whose combined statistics are noisier than the encapsulated statistics) only when provided with an indexical cue of speaker voice. In a series of five experiments, our study explores whether learners can utilise visual information to encapsulate statistics for each speech stream. We initially presented learners with incongruent artificial speech streams produced by the same female voice along with an accompanying visual display. Learners successfully segmented both streams when the audio stream was presented with an indexical cue of talking faces (Experiment 1). This learning cannot be attributed to the presence of the talking face display alone, as a single face paired with a single input stream did not improve segmentation (Experiment 2). Additionally, participants failed to successfully segment two streams when they were paired with a synchronised single talking face display (Experiment 3). Likewise, learners failed to successfully segment both streams when the visual indexical cue lacked audio-visual synchrony, such as changes in background screen colour (Experiment 4) or a static face display (Experiment 5). We end by discussing the possible relevance of the speaker's face in speech segmentation and bilingual language acquisition.

Keywords: Face processing; Indexical cues; Multiple representations; Speech segmentation; Statistical learning.

Correspondence should be addressed to Aaron D. Mitchel, 608 Moore Building, University Park, PA 16802, USA. E-mail: adm241@psu.edu *and* Daniel J. Weiss, E-mail: djw21@psu.edu

We thank Beth Buerger, Molly Jamison, and Troy Gury for conducting experiments. We also thank Marissa Weyer and Chip Gerfen for help in assembling the visual stimuli. We are grateful to Rich Carlson and Chip for helpful comments and to NIH R03 grant HD048996-01 for support of this research.

© 2009 Psychology Press, an imprint of the Taylor & Francis Group, an Informa business
http://www.psypress.com/lcp DOI: 10.1080/01690960903209888

Downloaded by [Penn State University] at 08:08 12 January 2016

INTRODUCTION

One of the earliest challenges of language learning is to segment words from a continuous speech stream. Until recently, laboratory studies of speech segmentation have focused on the underlying processes involved in segmenting a single input stream. Given that more than half of the world's population learns to speak more than one language (Crystal, 1997), it is important to consider how learners contend with input from multiple languages. A recent study of statistical learning in speech segmentation began to address this gap by sequentially presenting learners with multiple artificial input speech streams, simulating multilingual input (Weiss, Gerfen, & Mitchel, 2009). The findings from this research suggest that adult learners are capable of successfully segmenting multiple interleaved input streams, provided the streams are delineated with an adequate indexical cue, which we define as any consistent cue to language (such as speaker voice in Weiss et al., 2009). Indexical cues facilitate the formation of multiple representations by allowing learners to encapsulate information contained within each input stream and thereby perform separate computations. In the present study, we extend this line of investigation to explore whether learners are able to use indexical cues in the visual modality (synchronised video displays of talking faces) to trigger the formation of multiple representations in an auditory speech segmentation task.

Statistical learning in speech segmentation

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). While there are a number of acoustic cues to word boundaries in any given language, such as stress and other prosodic cues (Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000; Jusczyk, Houston, & Newsome, 1999; Mattys & Juscyzk, 2001), there are no invariant acoustic cues across all languages (Cole & Jakimik, 1980; Klatt, 1979). Given that learners (during development) do not have *a priori* knowledge about which language system will be acquired, they must determine which subset of cues will be effective. This issue is further complicated by the fact that many available cues (such as phonotactic or prosodic patterns, e.g., Friederici & Wessels, 1993; Jusczyk et al., 1999) cannot be used without having a foothold into the segmentation pattern. For example, some languages tend to favour stressing word onsets (e.g., English), while others favour penultimate stress (e.g., Spanish); therefore, in order to

use stress as a cue to segmentation, the learner must identify where stress falls within a word. Since such statistical patterns are available in all languages, statistical learning may provide an initial foothold into speech segmentation, subsequently facilitating the integration of additional language-specific cues (Thiessen & Saffran, 2003).

In order to test this theory, Saffran et al. (1996b) familiarised adult participants to an artificial language that had been stripped of any markers of word boundaries other than the transitional probabilities between sounds, which were consistently high within words and lowest at word boundaries. Following brief exposure to this input, participants reliably identified statistically defined words suggesting that adults are sensitive to statistical regularities and can use them to segment a fluid speech stream. These findings were replicated with 8-month-old infants whose performance was similar to the adult learners (Saffran et al., 1996a). Follow-up studies have demonstrated that at 6 months of age, infants prefer using transitional probabilities to segment speech streams even when stress cues are available (Thiessen & Saffran, 2003). Together, this body of work provides support for the theory that statistical cues may provide traction for early speech segmentation which subsequently allows the learner to incorporate language-specific cues.

Downloaded by [Penn State University] at 08:08 12 January 2016

Simulating bilingual segmentation

When the linguistic input consists of speech streams from different languages, the challenge for learners is to realise that sounds in one language may pattern differently than sounds in a second language (see Mehler, Jusczyk, Lambertz, Halsted, Bertoncini, & Amiel-Tison, 1988). If learners combine statistical information across languages, the noise in the statistical patterns of each language increases, thereby decreasing the likelihood of segmenting either language correctly (or minimally, delaying success). An array of cues (such as stress and phonotactic patterns, allophonic and microphonetic detail, different speakers, etc.) can signal exposure to multiple languages. However, the challenge for the learner is to converge on the appropriate set of cues, as distinct language pairs will make use of different cues to varying degrees. Therefore, the broader goals of this research effort are to first determine the conditions under which learners are capable of forming multiple representations and then to establish the developmental timeline for these abilities.

In an initial attempt to simulate multilingual input, Weiss and colleagues (2009) used the logic from the statistical learning studies to determine whether learners are capable of forming multiple representations when confronted with two interleaved speech streams. Adult participants were presented with two artificial languages that had incongruent statistical

structures (i.e., languages whose combined statistical regularities are noisier relative to the statistics contained within each language in isolation, resulting in increased difficulty for segmentation). Consequently, in order to successfully segment both artificial languages, participants had to encapsulate the statistics for each language rather than combine statistical information across languages. The findings from this study indicated that learners could succeed at segmenting both languages at above chance levels when provided with the indexical cue of speaker voice (one stream was presented in a male voice while the second stream was presented in a female voice), but failed to correctly segment both languages in the absence of such indexical cueing. Instead, when no indexical cue was given, participants learned only one of the two languages, and this asymmetry in learning is somewhat consistent with other studies familiarising participants to multiple input sources (see Gebhart, Aslin, & Newport, 2009). An outstanding issue from this study is determining the types of cues that learners can use indexically. The present study begins to address this issue by providing learners with visual indexical cues such as talking face displays and background colour displays.

Downloaded by [Penn State University] at 08:08 12 January 2016

Visual speech

There is a large body of research suggesting that talking faces are extremely salient for infants and adults, and that visual information can impact speech perception in both populations (Kuhl & Meltzoff, 1982, 1984; Vatikiotis-Bateson, Munhall, Kasahara, Garcia, & Yehia, 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). Recent research on language discrimination with infants has demonstrated that visual information alone is sufficient to facilitate discrimination between native and unfamiliar languages in infants as young as 4 months of age (Weikum, Vouloumanos, Navarra, Soto-Faraco, Sebastián-Gallés, & Werker 2007). Moreover, there is evidence that this ability to discriminate languages on the basis of visual speech extends into adulthood. Adult learners were presented with visual displays of faces producing two distinct languages (Spanish and Catalan) with no accompanying audio input. The participants were able to discriminate the languages on the basis of the visual display alone provided that they were familiar with at least one of the languages (Soto-Faraco, Navarra, Weikum, Vouloumanos, Sebastián-Gallés, & Werker, 2007). Together, these studies suggest that visual information from talking faces alone provides learners with a rich source of information that may be used to discriminate between language pairs.

The ability to use such facial information may also be particularly useful in noisy environments. In environments with a low signal-to-noise ratio, a synchronous visual display of the talker's face can enhance the discriminability of spoken messages (Sumby & Pollack, 1954) and may increase speech

perception abilities in bilinguals in their second language environment (Navarra & Soto-Faraco, 2007; Reisberg, McLean, & Goldfield, 1987). The benefits of visual information under noisy conditions accrue even in infancy. Hollich, Newman, and Jusczyk (2005) found that presenting 7.5-month-old infants with a synchronous audio-visual display facilitated their ability to recognise a known word embedded in a continuous speech stream.

In sum, talking faces appear to be particularly useful to language learners. Evidence to date suggests that information from talking faces suffices for language discrimination, a prerequisite step for bilingual language segmentation. In order to segment multiple languages, learners must first discriminate between the languages and then form distinct representations that allow for the encapsulation of the statistical information relevant to each language. It is unlikely that language discrimination alone is sufficient to successfully segment multiple languages (see Weiss et al., 2009). Therefore, in this study we seek to establish whether visual speech may facilitate the process of forming multiple representations for computing statistics across two input streams with alternating presentation. Specifically, we explore whether learners compute separate sets of statistics for speech streams that correspond to the appearance of particular faces.

The present study

As mentioned, previous studies have reported that changes in speaker voice allow learners to encapsulate the statistics of multiple input streams. Here we extend these studies by investigating whether the individual identity cues in the visual domain can facilitate the formation of multiple representations in the absence of auditory indexical cues. In Experiment 1, we present learners with two sequential incongruent artificial languages produced in the same female voice. Each speech stream is paired with a synchronous, dynamic video display of a talking female face, one producing each stream (alternating in two-minute blocks). These faces represent a possible indexical cue for separating the languages. In Experiment 2 we explore the contribution of pairing a single dynamic face video with a single input stream to determine whether this contribution might account for the effects observed in Experiment 1. In Experiment 3, we pair a single face display with the two input streams used in Experiment 1 in order to assess whether the indexical nature of the talking face cues in Experiment 1 facilitated learning. In Experiment 4, we explore whether any visual indexical information can suffice to facilitate successful segmentation of both streams by providing learners with a simple static visual cue of background colour on a video monitor. Finally, in Experiment 5 we test the effectiveness of an indexical cue consisting of a static image of the speakers' face in effort to explore the role of audio-visual synchrony in indexical cues.

Downloaded by [Penn State University] at 08:08 12 January 2016

Downloaded by [Penn State University] at 08:08 12 January 2016

EXPERIMENT 1

Experiment 1 adapted the methods used in previous research (Weiss et al., 2009), which interleaved artificial languages. As mentioned above, when the input streams contained incongruent statistical relationships, learners were only successful in segmenting both streams in the presence of an indexical cue (speaker voice). In Experiment 1, we presented participants with two interleaved, incongruent artificial languages, paired with a synchronous visual display of two different female speakers lip-synching to two artificial languages (one face per language). Both languages were presented in the same voice, and thus only the talking face provided an indexical cue to language. If learners can form multiple representations based on a synchronous display that encodes the visual identity of the speaker, then we predicted that learners should be able to segment both artificial languages at above-chance levels. However, if learners are not sensitive to this type of visual information, or cannot combine it with the auditory input, then we predicted that one or both of the languages would not be learned at above-chance levels.

Method

Participants. Forty undergraduate Introductory Psychology students from The Pennsylvania State University participated for class credit and were included in the analysis (19 male, 21 female). All participants were monolingual English speakers. Four additional participants were excluded due to technical failure (1) or failure to follow instructions (3).

Stimuli. The familiarisation stimuli were two artificial audio speech streams identical to those used in Weiss et al. (2009). Each language consisted of four trisyllabic words, with a CV.CV.CV. structure (see Figure 1). The words were formed from eight consonants and eight vowels. These were combined to form 20 CV syllables. The CV syllables were created by digitally

	Words	Part-words
Language 1	bə tɪ gu	vi bə tɪ
	sɪ tʃə vi	tʃa sɪ tʃ ə
	vʊ bo sæ	gu vʊ bo
	to gʊ tʃ a	sæ to gʊ
Language 2	gu pæ tə	sɪ gu pæ
	dʒi ga pʊ	bo dʒi g a
	sæ dʒu bo	tə sæ dʒu
	ta bi sɪ	pʊ ta bi

Figure 1. Test words and part-words for both artificial languages.

recording a female speaker producing CVC syllables, with the coda being one of four possible places of articulation (bilabial, alveolar, palatoalveolar, and velar). The CVC syllable was then hand-edited in Praat©, removing the coda consonants and controlling for duration to create the CV syllable. Creating the CV syllables in this manner preserved the vowel-to-consonant transitions when the CV syllables were combined to create the words. The syllables were also recorded without a coda consonant in order to create the test items.

All words were normalised using SoundForge©, controlling for any salient loudness differences, and then resynthesised using Praat©. The same f0 contour (pitch contour), shaped from the second syllable in a CV.CV.CV. token to maintain a natural sound, was overlaid onto the CV tokens, removing any pitch cues to segmentation. The resynthesised words were concatenated in random order into continuous speech streams, with each word presented the same number of times. The speech streams had a duration of 1 minute 56 seconds, and consisted of 96 words (288 syllables). One stipulation that governed the ordering of the speech stream was that the same word never occurred twice in a row. The only cues to word boundaries were the transitional probabilities. Transitional probabilities are a conditional probability statistic between successive syllables (Aslin, Saffran, & Newport, 1998). Given two sounds, X and Y, the probability of Y occurring after the occurrence of X, is equal to the frequency of co-occurrence of X and Y divided by the overall frequency of X. In other words, transitional probability can be seen as a relative frequency of co-occurrence; the primary difference is that transitional probabilities account for the overall frequency of the word, while a simple co-occurrence probability does not (Aslin, Saffran, & Newport, 1998).

The artificial languages were identical to the incongruent condition of Weiss et al. (2009). This set of languages was created such that the combined statistics across languages were noisier than the statistics within each of the languages (i.e., the encapsulated statistics). The encapsulated syllable statistics for each individual language consisted of 1.0 word-internal transitional probabilities and a 0.33 transitional probability at word boundaries. Additionally, the transitional probabilities between individual consonant and vowel segments (Newport et al., 2004) were consistently higher within words (0.50) and lower at word boundaries (0.33). In contrast, the combined statistics of these two languages were noisier due to variation in the within-word syllable transitional probabilities. Within-word transitional probabilities fluctuated, ranging from 0.50 to 1.00 at the syllabic level, and from 0.25–1.0 at the segmental level (see Figure 2). These statistical properties preclude learners from segmenting both languages successfully if they do not encapsulate the statistics for each language (Weiss et al., 2009). These statistical properties were derived by inserting two word-final syllables from Language 1 into word-initial positions in Language 2 and by inserting

Downloaded by [Penn State University] at 08:08 12 January 2016

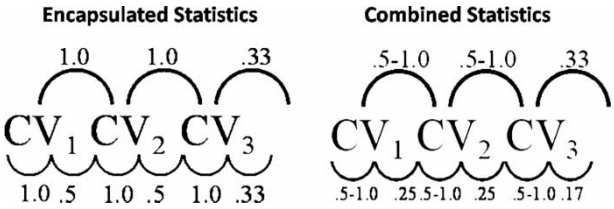

Figure 2. This figure shows the values of the transitional probabilities between adjacent syllables and adjacent segments when the statistics are encapsulated or combined.

two word final syllables from Language 2 into a word-initial position and a word-medial position in Language 1 (see Figure 1). Thus, the same sound marked a word boundary in one language, but did not indicate a word boundary in the other, thereby providing a conflicting cue to segmentation if learners attempted to combine statistics across languages.

In Experiment 1, we presented participants with an indexical cue to language, a visual display of dynamic faces. Two female assistants (the faces were female to be consistent with the female voice of the audio stream) were videotaped lip-synching each of the artificial languages. During videotaping, the assistants sat in a chair approximately 150 cm from the camera (a Sony Handicam), which was mounted on a tripod 122 cm from the ground. The familiarisation audio stream was played on a nearby computer as assistants read in synchrony from a list of the artificial words. Note that the words were separated on the list the assistants read from, thus the assistants were not naïve to word boundary. A 29-second clip of 24 words (72 syllables) was created that was subsequently looped to create a longer stream lasting 1 minute and 56 seconds (96 words, 288 syllables). Between each 29-second clip both the audio and visual streams were faded out and then faded back in over a period lasting one second. Fading the streams was necessary to remove any possible head jerks resulting from the splicing of the videos. After the videos were created, they were carefully hand-edited using Adobe Premiere© software, synchronising the onset of the video with the onset of the auditory familiarisation stream. Each assistant was also videotaped during a rest period, during which they sat in the chair and were instructed not to move their mouth. This video was designed to be displayed when faces were not active as the slight movements in the face (e.g., blinking) were intended to impart a more realistic animate representation even during periods when the individual was not lip-synching. The videos were also cropped and adjusted to approximately equate the size of the faces.

In total, four lip-synching videos (two faces each producing two languages) and two still videos (one for each face) were created. The familiarisation streams consisted of a side-by-side presentation of both faces. During presentation of the speech streams, one face was active (i.e., lip

Downloaded by [Penn State University] at 08:08 12 January 2016

Figure 3. A still frame from the visual display in Experiment 1. Note that the face on the left is talking while the face on the right is resting, providing an indexical cue to language.

Downloaded by [Penn State University] at 08:08 12 January 2016

synching) and one face was silent (though animate as described above). This was achieved by combining the two video tracks with an audio track in Adobe Premiere©. The two tracks were positioned adjacently, such that both videos were equal in size and in the middle of the screen (see Figure 3). The faces always appeared in the same location with Face 1 to the left of centre and Face 2 to the right of centre.

The videos were concatenated such that during the presentation only one face would be active (i.e., lip-synching to the audio stream) during the presentation of one language, and the second face would be active only during the second language. During presentation of the language associated with the first face, the second face remained animate (see above) but did not lip synch (see Figure 3). Given the link between a particular face and a particular language, the faces represented a consistent indexical cue to language. For example, in one familiarisation block, Face 1 was active for 1 minute 56 seconds while L1 was presented, and then Face 2 was active for 1 minute 56 seconds while L2 was presented. This 3 minute 52 second sequence was looped to create a movie lasting 7 minutes 44 seconds.

Procedure. For each experimental session, the movie was repeated three times for a total of 23 minutes and 12 seconds of familiarisation. Participants received a 1 minute break between each block (during which the screen was white and there was no sound played through the headphones). We counterbalanced the order of presentation for both the faces (Face 1 and Face 2) and the languages (L1 and L2), such that each face and language pairing was presented first an equal number of times.

Learning of the statistically defined words was tested by asking learners to discriminate words from part-words. Test items consisted of one of the eight words presented in the speech streams paired with one of eight part-words (see Figure 1). The part-words were formed by concatenating the third syllable of each of the four test words with the first two syllables from a different word. Thus, all part-words occurred within the familiarisation

stream. Each of the test words and part-words was generated using the same methods as described for creating the streams.

During test, every word was paired with two part-words from the same language, and each pairing was presented twice, counterbalancing the order of presentation. For each pair of test items, the first item was presented, followed by a one second pause, and then the second test item was presented. Between each test trial there was a four second inter-trial interval. In total, there were 32 test trials, 16 trials from each language. There was no accompanying visual display of faces present during testing.

The familiarisation stream was played using iTunes version 7.0, and the test was presented using E-Prime software (Psychology Software Tools, 2002) on a Dell PC computer (Optiplex GX280). Participants were informed that they would watch a movie and then be tested on information garnered from the audio stream. Instructions were presented both on the screen and verbally for the test phase. All verbal instructions were read from a script in order to standardise them across experimenters. Instructions given to participants were minimal. They were only instructed to watch a movie that would last roughly 30 minutes – there was no mention of the face stimuli or details of the accompanying audio stream. Research assistants stayed in the room with the participants to monitor them and ensure that the instructions were followed and that participants watched the entire movie.

During the test phase, participants were asked to identify which of the two items in a test trial represented a word from the audio stream by pressing a key on the keyboard corresponding to the first or second test item. Participants' responses, response time, and response accuracy were recorded in ePrime. After testing, participants were given a questionnaire about language use and language background.

Results

Overall, participants successfully segmented both languages at above chance performance. The mean number of correct responses was 19.85 out of a possible 32 (62%), with a standard deviation of 2.94 (see Figure 4). The mean number of correct responses for L1 was 10.10 (63%; $SD = 2.02$) and for L2 was 9.75 (61%; $SD = 2.11$). Performance was above chance overall, $t(39) = 8.28$, $p < .001$, $d = 2.65$, as well as for each of the individual languages: L1, $t(39) = 6.57$, $p < .001$, $d = 2.10$, and L2, $t(39) = 5.25$, $p < .001$, $d = 1.68$.

A between-subjects 2 (language order) × 2 (face order) factorial ANOVA revealed no significant order effects, neither for the order of face presentation, $F(1, 36) < 1$, nor for the order of language presentation, $F(1, 36) < 1$. Furthermore, there was no significant interaction between face order and language order, $F(1, 36) = 2.96$, $p = .094$, $\eta^2 = .08$.

Downloaded by [Penn State University] at 08:08 12 January 2016

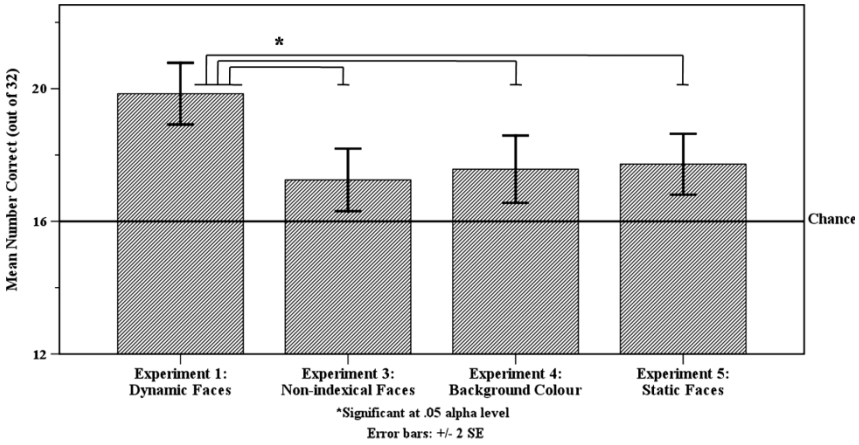

Figure 4. Results from Experiments 1, 3, 4, and 5. Performance is plotted as the mean number of items correct (out of 32).

Performance on the same languages without a face cue or any indexical cue in Experiment 3 of Weiss et al. (2009)[1] was 17.22 (54%; $SD = 4.29$; N = 18) out of 32 overall, and 9.22 (58%; $SD = 2.23$) and 7.78 (49%; $SD = 3.07$) out of 16 for L1 and L2, respectively. Performance in Experiment 1 was significantly greater overall, $t(56) = 2.72$, $p = .009$, $d = 0.73$. While there was no significant difference in Language 1 performance, $t(56) = 1.37$, $p = .177$, $d = 0.37$), performance in Language 2 was significantly better in Experiment 1, $t(56) = 2.72$, $p = .009$, $d = 0.73$.

Discussion

The results of Experiment 1 revealed that faces, as indexical cues to language, facilitate the simultaneous segmentation of multiple input streams in a statistical learning task with artificial languages. Performance was significantly above chance for both languages, indicating that learners could use the visual information in order to encapsulate the statistical information within each language Previous studies demonstrated that in the absence of an indexical cue, learners will successfully segment only one of the two artificial languages (Weiss et al., 2009). However, when an indexical cue of speaker voice was incorporated into the speech stream, participants learned both languages. Experiment 1 extends the previous findings by demonstrating that the visual information contained in indexically presented talking faces can

[1] Note that both experiments were conducted in the same laboratory using an identical auditory familiarisation stream and test files. Given this similarity (excepting the visual cues of the present experiment), we analysed the differences in performance.

Downloaded by [Penn State University] at 08:08 12 January 2016

trigger the encapsulation of statistics within languages, thereby facilitating segmentation.

It should be noted that the design of this experiment did not rule out possible 'ventriloquism effects' (see de Gelder & Bertelson, 2003, for a review) in which auditory localisation is displaced by a synchronous, yet spatially disparate visual signal. Since individual faces in this experiment always occurred on the same side of the screen, it is possible, though in our view less likely, that such ventriloquism effects might perceptually induce a spatial indexical cue (i.e., L1 is presented on the left, L2 is presented on the right), enabling participants to encapsulate statistics.

There are a number of other possible causes for the findings observed in Experiment 1. As mentioned above, to the best of our knowledge, this is the first study to pair a dynamic video display of faces with an auditory speech stream in an artificial speech segmentation task. Thus, it is possible that the presence of talking faces paired with the auditory stream facilitates learning in statistical learning paradigms (for example, by increasing attention; see below), and the facilitation did not hinge on the indexical presentation of talking faces. In Experiment 2 we tested this idea by presenting learners with a visual display in the context of a more standard statistical learning task using a single speech stream.

Downloaded by [Penn State University] at 08:08 12 January 2016

EXPERIMENT 2

Experiment 1 found that an indexical, synchronous talking face display facilitated segmentation of multiple input streams. Since dynamic faces provide a rich source of information for language processing, their presence alone might account for the observed facilitation of learning in Experiment 1. That is, the prosodic and phonemic information in the faces (Kuhl & Meltzoff, 1982) may have, independent of their role as an indexical cue, facilitated the segmentation of two input streams by making the segmentation task less difficult. Alternatively, faces may have increased attention to the task, a factor that has been shown to affect the outcome of statistical learning (Toro, Sinnet, & Soto-Faraco, 2005).

The goal of Experiment 2 was to explore the contribution of dynamic face displays to statistical learning of speech segmentation by pairing a single talking face display with a single auditory stream. Given our interest in determining how faces may facilitate auditory speech segmentation, this experiment was designed to parallel the seminal statistical learning experiments mentioned above (e.g., Saffran et al., 1996a, 1996b), thereby eliminating potential confounding effects of interleaving two auditory streams. Further, we chose to use an auditory stream with a level of learning

(when presented alone or paired with other languages) that has been documented in our other recent work (see Weiss et al., 2009).

Method

Participants. Thirteen undergraduate Introductory Psychology students from The Pennsylvania State University participated for course credit.[2] There were 5 males and 8 females, and all were monolingual English speakers. Two additional subjects were excluded from the analysis due to technical failure (1) or failure to follow instructions (1).

Stimuli. The stimuli in Experiment 2 were analogous to those used in Experiment 1. However, the audio stream was the male-voice L1 from Weiss et al. (2009). The structure and syllables, as well as the speech synthesis and syllable creation were identical to the female-voice L1 from Experiment 1 (see Figures 1 and 2). The only differences between the audio stream in Experiment 1 and the audio stream here are the voice of the speaker and the inclusion of only one artificial language. Since there was only one input stream, the total duration of the familiarisation phase was reduced by 50%. Thus, the 1 minute 56 second audio stream was looped into 3 minute 52 second blocks. Each participant received three blocks of exposure with one minute of silence between each block, for a total familiarisation of 11 minutes 36 seconds. Although the duration of familiarisation overall was shorter in this condition, the duration of the familiarisation to any given language was held constant. Participants in Experiment 2 listened to L1 for the same amount of time as participants in Experiment 1.

The talking face display was created in an identical manner to the displays used in Experiment 1. A male assistant[3] was video-recorded reading along with the audio stream from a list of words. This video was then hand-edited in Adobe Premiere and synchronised with the audio stream. Unlike in Experiment 1, in which both talking face videos were presented on the screen, the video here was displayed alone in the centre of the screen.

Procedure. After the familiarisation phase, performance was tested with a 16 item two-alternative forced-choice test that was identical in structure to

[2] The other experiments in this study require greater numbers of participants in order to counterbalance face and language presentation order. Because the critical comparison for Experiment 2 is with Weiss et al., and not with other experiments here, we chose to keep the sample size comparable to Weiss et al., which included 13 participants.

[3] Unlike the previous experiments, the same person whose voice was used to create the audio stream was used to create the face video. Thus, if learners are able to pick up speaker-specific details (e.g., fundamental frequency) from the speech stream, then these cues should be, to a first approximation, compatible with the face display.

Downloaded by [Penn State University] at 08:08 12 January 2016

the test used in Experiment 1, excluding the items corresponding to a second input stream. All other aspects of the procedure were identical to Experiment 1.

Results

The mean number of items correctly identified in Experiment 2 was 10.77 (out of 16 – 67%) with a standard deviation of 1.79. This level of performance was significantly above chance (50%, or 8 out of 16): $t(12) = 5.91$, $p < .001$, $d = 3.41$. When the same stream was presented alone, without a display of talking faces, mean performance was 10.62 (66%) out of 16 (N = 13), with a standard deviation of 2.69 (taken from Weiss et al., 2009), and this level of performance was not significantly different from performance in Experiment 2, $t(24) = 0.17$, $p = .863$, $d = 0.07$. Additionally, performance in Experiment 2 did not differ from performance on L1 in Experiment 1, $t(51) = 1.08$, $p = .287$, $d = 0.30$.

Discussion

While learners in Experiment 2 were successful in segmenting the input stream at above chance levels, the results from this experiment suggest that the addition of a talking face display does not increase the level of performance relative to conditions restricted to auditory input alone. It is possible that the lack of facilitation was due to learners being at ceiling performance. Consequently, the benefits to speech segmentation provided by visual speech may emerge only in conditions in which learners' performance is well below ceiling. We are currently testing this hypothesis. Nevertheless, in Experiment 2, participants do not appear to benefit from visual cues to word boundary (e.g., prosodic markers) in the face display, nor from any additional attention that such displays may invoke. Given this result, we conclude that the *indexical information* from the face displays in Experiment 1 likely facilitated the formation of multiple representations. In Experiment 3, we further test this hypothesis by presenting learners with the same two speech streams used in Experiment 1 paired with only a single talking face display, thereby removing the visual indexical information.

EXPERIMENT 3

In Experiment 1, we found that the addition of an indexical talking face display facilitated the learner's ability to successfully segment both speech streams. The results of Experiment 2 suggested that the addition of talking faces to an auditory speech stream was not sufficient to facilitate segmentation performance. Experiment 3 explored whether the indexical nature of the

Downloaded by [Penn State University] at 08:08 12 January 2016

visual presentation facilitated performance. In this experiment participants listened to the streams from Experiment 1 paired with a synchronous dynamic face during familiarisation *without* indexical information. If indexical information was critical to the success observed in Experiment 1, then we predicted that the learners would not be successful in segmenting both streams in Experiment 3.

Method

Participants. Forty undergraduate Introductory Psychology students from The Pennsylvania State University participated for class credit and were included in the analysis (8 male, 32 female). All participants were monolingual English speakers. Five additional participants were excluded for failure to follow instructions.

Stimuli. The familiarisation stream consisted of one of the talking face displays used in Experiment 1. Rather than pairing one face with one input stream, in Experiment 3 a single face was active for both audio streams. When either L1 or L2 was presented, Face 1 was active in synchrony with the audio stream. Thus, the only differences between Experiment 1 and Experiment 3 were the removal of the second face display and that the single face display was centred on the screen. The same face display was used for all participants. All other aspects of the stimuli were identical to Experiment 1, including test items. Language presentation order was counterbalanced across subjects.

Procedure. The procedure in Experiment 3 was identical to the procedure from Experiment 1.

Results

The mean number of correct responses overall was 17.25 out of a possible 32 (54%; $SD = 2.99$, see Figure 4). The mean number of correct responses for L1 was 9.23 (58%; $SD = 2.66$) out of 16 and for L2 was 8.03 (50%; $SD = 2.62$) out of 16. Performance was significantly above chance overall, $t(39) = 2.64$, $p = .012$, $d = 0.85$, but only above chance for L1, $t(39) = 2.92$, $p = .006$, $d = 0.94$, while performance on L2 was not significantly above chance, $t(39) = 0.06$, $p = .952$, $d = 0.02$. A paired-sample t-test revealed that there was a significant difference in performance between languages, $t(41) = 2.04$, $p = .048$, $d = 0.64$. There was no significant language order effect, $F(1, 38) < 1$.

An independent samples t-test revealed that performance overall in Experiment 3 was significantly lower than performance in Experiment 1, $t(78) = 3.92$, $p < .001$, $d = 0.89$. Performance on L1 was not significantly different, $t(78) = 1.66$, $p = .101$, $d = 0.38$, while performance on L2 was

Downloaded by [Penn State University] at 08:08 12 January 2016

significantly lower in Experiment 3, $t(78) = 3.25$, $p = .002$, $d = 0.74$. There was no significant difference between the results of Experiment 3 and performance on the same languages without an indexical cue in Experiment 3 of Weiss et al. (2009): overall: $t(56) = -0.03$, $p = .977$, $d = -0.01$; L1: $t(56) = -0.00$, $p = .997$, $d = 0$; L2: $t(56) = -0.32$, $p = .754$, $d = -0.09$.

Discussion

The results of Experiment 3 indicated that there was a significant decrease in performance when the talking faces did not provide an indexical cue to language relative to their performance in Experiment 1 in which faces were used indexically. This suggests that the successful learning of both languages in Experiment 1 cannot be explained by the presence of synchronised, talking faces alone. Consequently, we argue that the indexical nature of the talking face cues in Experiment 1 allowed the participants to form multiple representations for each input language, consistent with the findings from our previous work in which speaker voice served as the indexical cue (Weiss et al., 2009). In Experiment 4, we test whether learners are selective with respect to the types of indexical cues that can be used to form multiple representations for input language streams.

EXPERIMENT 4

How selective are learners in their use of indexical cues? During the course of acquisition, there are myriad potentially available indexical cues. For example, there may be environmental cues such as where the language is heard (e.g., English at home, French at school), speaker-specific cues (i.e., cues correlating an individual with a specific language), language-specific phonetic cues such as pitch or stress patterns, and phonotactic cues. However, not all environmental cues are reliable, suggesting that learners must be selective in finding and attending to appropriate cues. The process by which these indexical cues are selected remains to be formally specified.

Accordingly, the goal of Experiment 4 was to establish whether learners are selective with respect to visual indexical cues and to identify some of the features that make cues effective. Research in related areas suggests that learners may be sensitive to two properties of indexical cues: whether the cue provides information about speaker identity (Krajlic & Samuels, 2007) and whether the cue is synchronous with the audio stream (Hollich et al., 2005). Experiment 4 explored whether these properties were important for success in our task by investigating whether learners could use background colour as an effective visual indexical cue. Background colour is a highly salient visual cue, but does not provide a cue to speaker identity nor is it temporally synchronous with the audio stream. Thus, if learners were to succeed in

Downloaded by [Penn State University] at 08:08 12 January 2016

segmenting both input streams in this experiment, it would lend support to the idea that learners are not selective with respect to the types of indexical cues they can use to form multiple representations.

Method

Participants. Forty-two undergraduate Introductory Psychology students from The Pennsylvania State University participated for class credit and were included in the analysis (20 male, 22 female). All participants were monolingual English speakers. An additional four participants were excluded. Two participants were excluded due to technical failure, and two were excluded due to failure to follow instructions.

Stimuli. The familiarisation stream consisted of the same audio stream from Experiment 1 and Weiss et al. (2009). In Experiment 4, however, the familiarisation stream was no longer synched with a dynamic visual display; rather, the indexical cue was the background colour of the screen. During the familiarisation, participants viewed a screen that switched between two background colours, purple and teal, with no other visual stimuli. The colour switches coincided with language switches (with both switches occurring after every 1 minute, 56 second block), and in this way the background colour provided a consistent, reliable indexical cue, such that each screen colour was coupled with one language. During L1 presentation, the background colour was purple, and during L2 presentation the background colour was teal. The test was identical to Experiments 1 and 3 and did not contain any visual cueing.

Procedure. All aspects of the procedure were identical to that of Experiment 1, except that in Experiment 4, the entire experiment (both familiarisation and test sections) was conducted using E-prime software.

Results

The mean number of correct responses overall was 17.57 out of a possible 32 (55%), with a standard deviation of 3.28 (see Figure 4). The mean number of correct responses for L1 was 9.36 (59%; $SD = 2.74$) and for L2 was 8.21 (51%; $SD = 2.12$). Performance was significantly above chance overall, $t(41) = 3.10$, $p = .003$, $d = 0.97$. However, this effect appeared to be driven by performance in only one of the languages. A paired-sample t-test revealed that there was a significant difference in performance between languages, $t(41) = 2.04$, $p = .048$, $d = 0.64$. When performance is viewed at the level of the individual languages, participants learn L1, $t(41) = 3.21$, $p = .003$, $d = 1.00$, but not L2, $t(41) = 0.65$, $p = .517$, $d = 0.20$. Performance overall was not

Downloaded by [Penn State University] at 08:08 12 January 2016

significantly different from performance when there was no indexical cue in previous studies (Weiss et al., 2009; $t(58) = -0.34$, $p = .732$, $d = -0.09$).

Discussion

The background colour cue in Experiment 4 was a consistent, reliable cue to language, and contained sufficient indexical information to potentially allow learners to maintain multiple representations. However, learners did not exploit the indexical information, suggesting that not all indexical cues are valued evenly. It is also possible that the background colour cue, while salient, was not as engaging for learners as the synchronous faces, and therefore failed to engage participants' attention to the same degree as the faces (see General Discussion).

The results from Experiment 4 also replicate the findings from Weiss et al. (2009). In the absence of an effective indexical cue, participants were unable to successfully segment both languages, presumably because they combined the statistics across the languages (see Weiss et al., 2009). Indeed, there was no difference in performance when there was a background colour indexical cue and when there was no indexical cue at all. Similar to the results of Experiment 3 and our previous study, we observed an asymmetry in the pattern of results, with performance on segmentation at above chance levels for one language but not significantly different from chance on the other.

The results from Experiment 4 were consistent with the ideas mentioned above that temporal synchrony and individual identity represent important components of indexical cues (Hollich et al., 2005; Krajlic & Samuel, 2007). Experiment 5 was therefore designed to further examine the criteria guiding indexical cue selection by maintaining some individual identity information, but removing temporal synchrony.

EXPERIMENT 5

In Experiment 5, we presented participants with an indexical cue to language that consisted of static images of the speaker's face. Similar to Experiment 1, each language was coupled with a face, providing consistent indexical information as well as speaker identity information. However, in contrast to Experiment 1, the face display was a still image. Thus, if speaker-specific information is solely responsible for indexical cue efficacy, then static face images should facilitate segmentation. However, if temporal synchrony is a critical component of cue selection (either alone, or in conjunction with speaker identity), then static faces should result in chance performance, similar to that observed in Experiment 4.

Downloaded by [Penn State University] at 08:08 12 January 2016

Method

Participants. Forty undergraduate Introductory Psychology students from The Pennsylvania State University participated for class credit in Experiment 5 and were included in the analysis (22 male, 18 female). All participants were monolingual English speakers. An additional 5 participants were excluded from analysis for failure to follow directions (4) or due to technical failure (1).

Stimuli. The familiarisation stream was the same audio stream used in Experiment 1. In Experiment 5, each audio stream was paired with a visual display of a static face (one face per language). The face images were created by taking a still frame of the video from Experiment 1 (both faces on the screen) and overlaying a black shape over one of the faces (see Figure 5). Consistent with Experiment 1, the faces always appeared on the same side of the screen (i.e., Face 1 appeared left of centre and Face 2 appeared to the right). Thus, during familiarisation, participants viewed a single face, presented to the left or right of centre on the computer monitor, which switched in conjunction with the familiarisation stream (occurring after every 1 minute, 56 second block). For example, during L1 presentation, Face 1 would be displayed on the screen, and during L2 presentation, Face 2 would be presented on the screen. Since the images were taken directly from the videos in Experiment 1, the size and position of the images were identical to Experiment 1. Language order was counterbalanced across subjects. Participants were then given the same test as in Experiment 1.

Procedure. All aspects of the procedure were identical to Experiment 4.

Figure 5. An example of the static face displays used in Experiment 5.

Downloaded by [Penn State University] at 08:08 12 January 2016

Results

The mean number of correct responses overall was 17.73 out of a possible 32 (55%), with $SD = 2.90$ (see Figure 4). The mean number of correct responses for L1 was 8.85 (55%; $SD = 2.76$) and for L2 was 8.88 (56%; $SD = 2.30$). Performance was significantly above chance overall, $t(39) = 3.76$, $p = .001$, $d = 1.20$, but was not significantly different from performance when there was no indexical cue in previous studies (Weiss et al., 2009; $t(58) = -0.34$, $p = .732$, $d = -0.09$). Additionally, when performance is viewed at the level of the individual languages, participants learned L2, $t(39) = 2.41$, $p = .021$, $d = 0.77$, but not L1, $t(39) = 1.95$, $p = .059$, $d = 0.62$. This level of performance was moderated by the order of presentation, as evidenced by a marginally significant order effect, $F(1, 38) = 3.69$, $p = .062$, $\eta^2 = .09$. If separated by presentation order, L1 was learned, $t(19) = 2.73$, $p = .013$, $d = 1.25$, but not L2, $t(19) = 0.63$, $p = .538$, $d = 0.29$, when L1 was presented first. Likewise, when L2 was presented first, L2 was learned, $t(19) = 3.11$, $p = .006$, $d = 1.43$, but not L1, $t(19) = 0.24$, $p = .817$, $d = 0.11$.

A one-way ANOVA yielded a significant difference in overall performance across the four multilingual conditions presented here, $F(3, 158) = 6.10$, $p = .001$, $\eta^2 = .10$. Bonferroni post-hoc analyses revealed that performance was significantly greater in Experiment 1 than in Experiment 3, Experiment 4, and Experiment 5 (ps $= .001$, .005, and .013 respectively), while there were no differences between Experiment 3, Experiment 4, or Experiment 5.

Discussion

Experiment 5 indicated that a static face cue presented indexically does not facilitate the formation of multiple representations in a simulated bilingual statistical learning task. This finding provides support for the hypothesis that temporal synchrony is a critical property of indexical cues (see General Discussion).

It should be noted that the results of Experiment 5 do not rule out the importance of speaker-specific representations. Although speaker-specific information was available, learners may not have integrated such information with the speech stream. Evidence for this claim comes from neuroimaging work (Calvert & Campbell, 2003) demonstrating that stilled faces produce significantly less cortical activation than moving faces in key speech regions of the brain, including the auditory cortex, left superior temporal sulcus (STS), and the left inferior frontal gyrus (Broca's area). Given this, it is possible that the stilled faces in Experiment 5 did not effectively induce speaker-specific representations. It is also possible that a combination of speaker identity information and temporal synchrony is required for effective visual indexical cues.

Downloaded by [Penn State University] at 08:08 12 January 2016

A secondary goal of Experiment 5 was to test for potential 'ventriloquism effects' by providing spatial cues to language. In Experiment 5, we preserved the spatial properties of the face display in Experiment 1. Thus, if the effect in Experiment 1 was the product of spatial information, then learners should have had similar success in Experiment 5. However, despite the presence of spatial cues, performance in Experiment 5 was below Experiment 1, suggesting that the successful learning in Experiment 1 is not likely a function of spatial encoding. It remains possible that the stilled faces in Experiment 5 were not as effective as the moving faces in Experiment 1 at inducing a spatial indexical cue (see Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994). A future experimental condition in which a single animate face appears at two distinct locations on a screen may address this concern.

Downloaded by [Penn State University] at 08:08 12 January 2016

GENERAL DISCUSSION

In the present study, we presented participants with two artificial languages spoken by the same voice whose statistics were incongruent (i.e., combining the statistical properties of the languages resulted in noisier statistics). In Experiment 1, the artificial languages were paired with a synchronous dynamic display of two talking faces. Each face actively lip-synched during the presentation of one of the languages, and thus the faces represented an indexical cue to language. Participants in Experiment 1 scored above chance in segmenting both languages, suggesting that they formed multiple representations, encapsulating the statistics within each language. These results extend findings from previous research that reported learners are capable of successfully segmenting multiple sequential streams when provided with an adequate auditory indexical cue of speaker voice (Weiss et al., 2009). In Experiment 2, we explored the effect of talking faces on speech segmentation by presenting participants with a single auditory input stream coupled with a dynamic talking face display. While learners successfully segmented the input stream, we found that in this more traditional task, the talking face display did not facilitate performance beyond that which was observed in the absence of any visual cue. These results indicate that our findings from Experiment 1 cannot be solely attributed to performance benefits accrued from the presence of a talking face display (either increased attention or due to visual segmentation cues). Rather, the indexical nature of the visual cues appears to have facilitated performance in Experiment 1. Therefore, in Experiment 3, we tested whether the presence of a synchronous talking face display could facilitate successful segmentation of both languages even in the absence of such indexical information. We presented learners with the two incongruent artificial speech streams used in Experiment 1 paired with a single talking face display.

Participants failed to segment both languages above chance, and there was a significant decrease in performance relative to Experiment 1, further supporting the importance of indexical cues for successful segmentation. In Experiment 4, we found that learners could not segment both streams successfully when presented with an indexical cue of background colour. Performance was similar to previous studies that tested segmentation of these streams in the absence of any indexical or visual cues (Weiss et al., 2009). Finally, in Experiment 5, we extended the results of Experiment 4, testing the effectiveness of a static, indexical face cue. In this condition, as in Experiment 4, participants failed to segment both speech streams.

From this pattern of results, we draw three conclusions: first, under some conditions, visual cues can be utilised by learners to facilitate the segmentation of auditory speech streams; second, temporal synchrony and individual identity information may be critical features for effective visual indexical cues; finally, related to these points, learners appear selective with respect to the types of visual indexical information they will utilise. We discuss each of these points in turn.

Consistent with findings from recent research on infants and adults (Soto-Faraco et al., 2007; Weikum et al., 2007), language learners are capable of extracting important information from talking faces, particularly with respect to multiple-language input. The aforementioned studies indicated that learners can distinguish languages on the basis of visual displays alone. Here we demonstrated that when the talking face displays were paired with speech streams, they provided effective indexical information for multiple auditory input streams, allowing learners to perform separate statistical computations on each stream. Interestingly, the visual information provided in talking faces did not increase performance on a speech segmentation task involving only a single input stream. Though faces did not facilitate segmentation in this condition, it is possible that such visual displays could enhance performance under noisy conditions, as has been reported for other tasks (e.g., Sumby & Pollack, 1954).

The pattern of findings reported above suggests that temporal synchrony may be central to integrating visual information with the speech stream. Temporal synchrony is the co-occurrence of two phenomena in time and represents a source of amodal information (see Bahrick, 2001; Massaro, 1998). Auditory modal information was provided in the form of transitional probability cues and visual modal information was provided by the talking face display. The temporal conjunction of these two input sources provided a third source of information that may have facilitated the formation of multiple representations. In our experiments, successful segmentation of both streams occurred only when the face displays were active and temporally synched with the auditory stream. This pattern is consistent with previous findings that both dynamic faces and a synchronous

Downloaded by [Penn State University] at 08:08 12 January 2016

oscilloscope display facilitated infants' identification of words from continuous, noisy speech (Hollich et al., 2005). However, it should be noted that we did not directly manipulate the amount of synchrony between the audio stream and the visual display. The visual display in our experiments was static or synchronous, never asynchronous. Thus, it remains possible that the visual indexical cue and the audio signal only need to be perceived as originating from a common source for the indexical cue to be effective, and this may have occurred only when the face display was moving and an indexical cue was present (i.e., Experiment 1). Future experiments will test the contribution of temporal synchrony by employing an oscilloscope display similar to the Hollich et al. study.

A second important source of information in the visual display appears to be speaker identity. Work in perceptual learning and categorisation suggests that learners incorporate speaker identity into perceptual representations (Eisner & McQueen, 2005; Krajlic & Samuel, 2005, 2006, 2007; Newman & Evers, 2007; Nygaard & Pisoni, 1998). For example, in a phoneme categorisation task with multiple speakers during familiarisation, Krajlic and Samuel (2007) found that participants only exhibit perceptual learning (i.e., a token's perceptual space is shifted as a consequence of the surrounding context) when the differences between target items in the continuum are informative to speaker identity (e.g., spectral shifts). When the differences are uninformative with respect to speaker identity (e.g., temporal shifts, or voice-onset-time shifts) learners 'reset' the token's perceptual space, resulting in a lack of perceptual learning. This effect was also found with dynamic, synchronous displays of faces (Bertelson, Vroomen, & de Gelder, 2003). The findings of the current study are consistent with this claim, suggesting that learners are incorporating speaker identity to encapsulate statistics. This assertion is supported by the results of Experiment 4 in which learners did not successfully segment the speech streams when provided with an indexical cue of background colour. Alternatively, the results of Experiment 5 may not support this claim since learners did not use the static face images to form multiple representations. However, it is possible that learners did not link the static images with the languages since the situation was not naturalistic (i.e., people move their mouths when they produce speech sounds). Given prior findings in which learners could succeed at segmenting both languages when provided an indexical cue of voice (Weiss et al., 2009), we find it likely that speaker identity information is critical for indexical cues.

If speaker identity is an important cue, an open question is whether learners are representing each language individually or whether they are forming speaker-specific representations. Given the results reported here and in Weiss et al. (2009), future work must delineate more precisely how learners represent the streams when they are encapsulated. One logical extension of this problem is how learners determine when it is advantageous to combine

Downloaded by [Penn State University] at 08:08 12 January 2016

statistical information (for example, when two speakers are speaking the same language) and when it is crucial to maintain separate representations (such as in a bilingual environment; see Weiss et al., 2009 for a discussion of possible acoustic indexical cues that may help resolve this issue).

The results of the current study suggest that adult learners are selective with respect to indexical cues, successfully incorporating information from talking faces, but not from background colour or static face images. The mechanism for this process of selection is unclear. It is possible that preferences for attending to talking faces, whether through an innate subcortical mechanism (Kanwisher, 2006; Morton & Johnson, 1991; Tsao, Freiwald, Tootell, & Livingstone, 2006), or due to biases from general properties of the infant visual system (Banks & Ginsburg, 1985; Kleiner, 1987; Macchi Cassia, Turati, & Simion, 2004; Simion, Valenza, Macchi Cassia, Turati, & Umiltà, 2002; Turati, 2004), that may influence learners to pay closer attention to information provided by talking faces. Alternatively, a purely statistical process in which learners compute the effectiveness of each cue based on their co-occurrence with the input language could also account for such selectivity. Planned infant studies using the simulated bilingual statistical learning paradigm will identify the point at which infants can form multiple representations to track multiple language input, and then explore the types of indexical cues that facilitate this process, thereby lending insight into the mechanism of indexical cue selection.

Finally, the findings presented here have broader applied implications for theories of bilingual language development. The suggestion raised in this paper that learners may benefit from forming speaker-specific representations has been a cornerstone of the One Parent – One Language (OPOL) hypothesis that has received widespread popular support despite a lack of empirical evidence (Barron-Hauwaert, 2004; Goodz, 1989). The OPOL hypothesis states that in order to reduce confusion between languages, the optimal way to raise a bilingual child is to separate the language input by parent, with each parent speaking only one language (Arnberg, 1987; Barron-Hauwaert, 2004; Döpke, 1992, 1997, 1998; Ronjat, 1913). The OPOL hypothesis, in essence, proposes that the bilingual learner benefits from having a strong indexical cue (a particular individual associated with a particular language). The findings from the present study, as well as the results from Weiss and colleagues (2009), lend support to the idea that the presence of indexical information during the early stages of acquisition may help learners to more effectively perform separate computations on different input streams. However, contrary to the OPOL hypothesis, these results do not imply an overall difference in ultimate competence; rather, they suggest that the time course of acquisition may benefit from the presence of indexical cues. Furthermore, the OPOL hypothesis extends to levels of language processing beyond initial segmentation, and thus beyond the scope of our

Downloaded by [Penn State University] at 08:08 12 January 2016

data. Since the OPOL hypothesis is essentially developmental in nature, planned studies with infants may contribute empirical data to this theory.

Manuscript received April 2008
Revised Manuscript received July 2009
First published online October 2009

REFERENCES

Arnberg, L. (1987). *Raising children bilingually: The preschool years*. Clevedon, UK: Multilingual Matters.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science*, *9*, 321–324.

Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, *79*, 253–270.

Banks, M., & Ginsburg, A. P. (1985). Infant visual preferences: A review and new theoretical treatment. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 19, pp. 207–246). New York: Academic Press.

Barron-Hauwaert, S. (2004). *Language strategies for bilingual families: The one-parent-one-language approach*. Clevedon, UK: Multilingual Matters.

Bertelson, P., Vroomen, J., & de Gelder, B. (2003). Visual recalibration of auditory speech identification: A McGurk aftereffect. *Psychological Science*, *14*, 592–597.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Exploring the relation between McGurk interference and ventriloquism. In *Third International Conference on Spoken Language Processing* (Vol. 2, pp. 559–562). Yokohama, Japan: Acoustical Society of Japan.

Calvert, G. A., & Campbell, R. (2003). Reading speech from still and moving faces: The neural substrates of visible speech. *Journal of Cognitive Neuroscience*, *15*, 57–70.

Cole, R., & Jakimik, J. (1980). A model of speech perception. In R. A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Crystal, D. (1997). *The Cambridge encyclopedia of language*. Cambridge, UK: Cambridge University Press.

de Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Science, 7*, 460–467.

Döpke, S. (1992). *One parent – one language: An interactional approach*. Amsterdam: Benjamins.

Döpke, S. (1997). Is the simultaneous acquisition of two languages in early childhood equal to acquiring each of the two languages individually? In E. Clark (Ed.), *Proceedings of the 28th Annual Child Language Research Forum* (pp. 95–113).

Dopke, S. (1998). Can the principle of 'One person – one language' be disregarded as unrealistically elitist? *Australian Review of Applied Linguistics*, *21*, 41–56.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception and Psychophysics*, *67*, 224–238.

Friederici, A. D., & Wessels, J. M. (1993). Phonotactic knowledge and its use in infant speech perception. *Perception and Psychophysics*, *54*, 287–295.

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing structures in mid-stream: Learning along the statistical garden path. *Cognitive Science*, *33*, 1087–1116.

Goodz, N.S. (1989). Parental language mixing in bilingual families. *Infant Mental Health Journal*, *10*, 25–44.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development*, *76*, 598–613.

Downloaded by [Penn State University] at 08:08 12 January 2016

Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross language word segmentation by 9-month-olds. *Psychonomic Bulletin and Review*, *7*(3), 504–509.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207.

Kanwisher, N. (2006). What's in a face? *Science, 311*, 617–618.

Klatt, D. H. (1979). Speech perception: A model of acoustic phonetic analysis and lexical access. *Journal of Phonetics*, *7*, 279–312.

Kleiner, K. A. (1987). Amplitude and phase spectra as indices of infants' pattern preferences. *Infant Behavior and Development*, *10*, 49–59.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*, 141–178.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin and Review*, *13*, 262–268.

Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language*, *56*, 1–15.

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*, 1138–1141.

Kuhl, P. K., & Meltzoff, A. N. (1984). Intermodal speech perception. *Infant Behavior and Development*, *7*, 361–381.

Macchi Cassia, V., Turati, C., & Simion, F. (2004). Can a non-specific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science*, *15*, 379–383.

Massaro, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle.* Cambridge, MA: MIT Press.

Mattys, S. L., & Jusczyk, P. W. (2001). Do infants segment words or recurring contiguous patterns? *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 644–655.

Mehler, J., Jusczyk, P. W., Lambertz, G., Halsted, G., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*, 143–178.

Morton, J., & Johnson, M. H. (1991). Conspec and Conlern: A two-process theory of infant face recognition. *Psychological Review*, *98*, 164–181.

Navarra, J., & Soto-Faraco, S. (2007). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research*, *71*, 4–12.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*, 85–103.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception and Psychophysics*, *60*, 355–376.

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

Ronjat, J. (1913). *Le développement du langage observé chez un enfant bilingue.* Paris: Champion. Reviewed in Barron-Hauwaert, S. (2004).

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996a). Statistical learning by 8-month-old infants. *Science*, *274*, 1926–1928.

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996b). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621.

Simion, F., Valenza, E., Macchi Cassia, V., Turati, C., & Umiltà, C. (2002). Newborns' preference for up-down asymmetrical configurations. *Developmental Science*, *5*, 427–434.

Soto-Faraco, S., Navarra, J., Weikum, W. M., Vouloumanos, A., Sebastián-Gallés, N., & Werker, J. F. (2007). Discriminating languages by speech-reading. *Perception and Psychophysics*, *69*, 218–231.

Downloaded by [Penn State University] at 08:08 12 January 2016

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

Thiessen, E. D., & Saffran, J. R. (2003). When cues collide: Use of stress and statistical cues to word boundaries by 7- and 9- month-old infants. *Developmental Psychology*, *39*(4), 706–716.

Toro, J. M., Sinnett, S., & Soto-Faraco, S. (2005). Speech segmentation by statistical learning depends on attention. *Cognition*, *97*(2), B25–B34.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & &. Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*, 670–674.

Turati, C. (2004). Why faces are not special to newborns: An alternative account of the face preference. *Current Directions in Psychological Science*, *13*, 5–8.

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., & Yehia, H. (1996). Characterizing audiovisual information during speech. In *Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 96)* (Vol. 3, pp. 1485–1488). New York: IEEE Press.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*, 1159.

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2009). Speech segmentation in a simulated bilingual environment: A challenge for statistical learning? *Language Learning and Development*, *5*, 30–49.

Yehia, H. C., Kuratate, T., & Vaitikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*, 555–568.

Downloaded by [Penn State University] at 08:08 12 January 2016