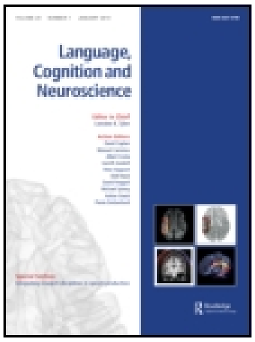


Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: http://www.tandfonline.com/loi/plcp21

Visual speech segmentation: using facial cues to locate word boundaries in continuous speech

Aaron D. Mitchel & Daniel J. Weiss

To cite this article: Aaron D. Mitchel & Daniel J. Weiss (2014) Visual speech segmentation: using facial cues to locate word boundaries in continuous speech, Language, Cognition and Neuroscience, 29:7, 771-780, DOI: 10.1080/01690965.2013.791703

To link to this article: http://dx.doi.org/10.1080/01690965.2013.791703

Language, Cognition and Neuroscience, 2014
Vol. 29, No. 7, 771–780, http://dx.doi.org/10.1080/01690965.2013.791703

Routledge
Taylor & Francis Group

Visual speech segmentation: using facial cues to locate word boundaries in continuous speech

Aaron D. Mitchel[a]* and Daniel J. Weiss[b]

[a]*Department of Psychology, Bucknell University, Lewisburg, PA 17837, USA;* [b]*Department of Psychology and Program in Linguistics, The Pennsylvania State University, 643 Moore Building, University Park, PA 16802, USA*

(*Received 24 June 2012; final version received 15 March 2013*)

Speech is typically a multimodal phenomenon, yet few studies have focused on the exclusive contributions of visual cues to language acquisition. To address this gap, we investigated whether visual prosodic information can facilitate speech segmentation. Previous research has demonstrated that language learners can use lexical stress and pitch cues to segment speech and that learners can extract this information from talking faces. Thus, we created an artificial speech stream that contained minimal segmentation cues and paired it with two synchronous facial displays in which visual prosody was either informative or uninformative for identifying word boundaries. Across three familiarisation conditions (audio stream alone, facial streams alone, and paired audiovisual), learning occurred only when the facial displays were informative to word boundaries, suggesting that facial cues can help learners solve the early challenges of language acquisition.

Keywords: speech segmentation; visual prosody; audiovisual speech; language acquisition; multisensory integration

Input to language learners is typically not restricted to the auditory modality (Massaro, 1998). In particular, the speaker's face is both highly salient to language learners (e.g. Morton & Johnson, 1991) and linguistically informative (e.g. Patterson & Werker, 2003; see below). While previous studies have demonstrated that learners may be sensitive to patterns occurring in talking faces (Weikum et al., 2007), to the best of our knowledge no study has established that learners are capable of utilising regularities associated with visual cues in talking faces to help solve the early challenges of language acquisition. Thus, the goal of the present study is to examine whether learners can segment an auditory stream based primarily on cues occurring in the visual domain.

Speech segmentation

Speech segmentation is one of the earliest obstacles confronting infants, as they must determine which combinations of sounds constitute words. Mastering this complex task may involve numerous strategies incorporating acoustic (e.g. Christophe, Gout, Peperkamp, & Morgan, 2003; Jusczyk, Houston, & Newsome, 1999), phonotactic (e.g. Friederici & Wessels, 1993), distributional (e.g. Saffran, Newport, & Aslin, 1996) and pragmatic cues (e.g. Brent & Cartwright, 1996). Prosodic cues, such as stress and intonation, appear to be particularly prominent by 8 months of age (Johnson & Jusczyk, 2001). For instance, many languages follow a syllabic stress pattern (e.g. strong-weak), and infants can use these patterns to identify the location of word boundaries in familiar and novel languages (Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000; Johnson & Jusczyk, 2001). Further, both infants (Gout, Christophe, & Morgan, 2004) and adults (Christophe, Peperkamp, Pallier, Block, & Mehler, 2004; Endress & Hauser, 2010) are capable of using phrasal intonation patterns (e.g. falling pitch at phrase boundaries) to segment novel languages.

The aforementioned research suggests that learners are acutely sensitive to prosodic information in the speech stream. However, cues to prosody are not restricted to the auditory modality as it is also possible to extract prosodic cues from watching the speaker's face (Graf, Cosatto, Strom, & Huang, 2002; Yehia, Kuratate, & Vaitikiotis-Bateson, 2002; see below). Given that typical learning environments have low signal-to-noise ratios (see Hollich, Newman, & Jusczyk, 2005), learners may frequently rely on facial cues to augment auditory cues for speech segmentation. The present study asks whether it is possible to use visual cues to acoustic prosody (hereafter *visual prosody*) to segment continuous speech when auditory cues to

*Corresponding author. E-mail: adm018@bucknell.edu segmentation are insufficient. Before describing the experiments, we briefly review the types of information conveyed in the face of the speaker that learners might use to segment a speech stream.

Visual speech

An abundance of linguistic content is conveyed to listeners by viewing a talking face (known as *visemic* information), including 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 lip movements using only the spectral information (i.e. formant structure) in the vowels (Kuhl & Meltzoff, 1982; Patterson & Werker, 1999). In addition, visual speech displays have been shown to facilitate speech perception (Sumby & Pollack, 1954) and phoneme discrimination (Teinonen, Aslin, Alku, & Csibra, 2008), suggesting that observers are able to utilise visemic information to assist in acquiring 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 prosodic cues include *x*-axis rotation of the head (head nodding) and lip aperture (how far apart the lips are), as both head-nods and peak lip aperture tend to coincide with stressed syllables and phrase boundaries (Graf et al., 2002; Yehia et al., 2002). Since auditory prosody is a prominent cue for speech segmentation (see Jusczyk et al., 1999), if a talking face display contains visual prosodic cues, it is reasonable to speculate that such cues might also support segmentation.

Several prior studies have investigated the role of faces in the context of speech segmentation, though to the best of our knowledge, none have addressed whether learners are able to use visemic cues to directly segment speech when boundary cues in the auditory stream are inaccessible. In one of the first studies to explore facial cues in speech segmentation, Hollich and colleagues (2005) familiarised infants to a set of familiar target words (e.g. *cup*) embedded within a fluid speech stream. The target speech stream was presented in a noisy environment (a second speech stream produced by a different voice reading the methods section of a previous paper), resulting in a low signal-to-noise ratio. The authors paired the streams with a variety of visual displays; a synchronous dynamic face, an asynchronous dynamic face, a static face and a synchronous oscilloscope display. Synchronous talking faces and oscilloscope displays facilitated infants' ability to identify and segment the target speech stream. Therefore, the authors proposed that redundancy in amodal information (e.g. rhythm, rate, duration, etc; see Bahrick & Lickliter, 2000) afforded by synchronous audiovisual displays enhanced attention to the target stream and the relevant acoustic cues to segmentation (Hollich et al., 2005). Similarly, Cunillera, Càmara, Laine, and Rodríguez-Fornells (2010) presented adults with an artificial speech stream containing transitional probability cues to word boundaries. This speech stream was paired with a corresponding visual stream composed of static pictures of unrelated objects. The authors found that segmentation of the speech stream was enhanced (relative to an auditory-only baseline) when the onset of the pictures was contiguous with the onset of words in the speech stream. Likewise, Thiessen (2010) discovered that speech segmentation was enhanced for adults (though not for 8-month-old infants) when word boundaries in the speech stream coincided with the appearance of distinct objects in the visual stream. However, in these two prior studies, performance was significantly above chance in the auditory baseline condition, as well as when the onset of words and pictures was misaligned (Cunillera et al., 2010), suggesting that learners were relying on auditory cues for segmentation. Similar to the study by Hollich and colleagues (2005), the visual cues served to enhance attention to relevant auditory information.

The preceding studies demonstrate that synchronous audiovisual information can facilitate multimodal speech segmentation by augmenting auditory information (see also Mitchel & Weiss, 2011). However, these studies did not assess whether visemic cues may directly contribute to speech segmentation. Two recent studies have investigated whether visemic cues from a talking face display can enhance statistical learning of audiovisual speech. In Mitchel and Weiss (2010), 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 and Kaschak (2009) tested the effect of visual speech on statistical learning by presenting adults with an artificial language in three conditions: audio-only (a speech stream in isolation), an audiovisual condition (a speech stream 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. Thus, like Mitchel and Weiss (2010) these findings did not provide evidence that visual speech facilitates statistical learning. However, in each of these studies, the presence of robust statistical cues to word boundaries 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 (Grant & Seitz, 2000; Hollich et al., 2005; Sumby & Pollack, 1954), visual speech may play a greater role in speech segmentation if cues in the auditory stream are insufficient to support robust levels of speech segmentation. That is, if the statistical patterns in the auditory stream are noisy, providing only minimal cues to word boundaries, then learners may utilise visually cued prosodic information to segment speech.

In the present study, we explore this issue by presenting learners with a speech stream in which the auditory cues to word boundaries (e.g. stress, distributional information, etc.) are inadequate to effectively support speech segmentation when presented in isolation. We test participants' ability to segment an artificial language in four separate (between-subjects) display conditions. In Condition 1 (audio-only), we determine the baseline ability to segment an audio speech stream with minimal cues to segmentation. In Condition 2 (visual-only), we test participants' ability to segment visual speech streams in isolation (i.e. there was no auditory stimuli during familiarisation or test), produced by speakers that were either aware of (Condition 2a) or misinformed (Condition 2b) about the location of word boundaries (see below). In Condition 3 (audiovisual), we pair the auditory speech stream with each visual display (aware and misinformed), thereby familiarising participants to an aware audiovisual display (Condition 3a) and a misinformed (Condition 3b) audiovisual display. We then test participants on the same auditory-only test from Condition 1. In both Condition 2a and 2b, we predicted that participants would successfully segment the visual display and audio stream, respectively. Conversely, in Conditions 2b and 3b, performance on the visual display produced by the misinformed assistant should not result in successful segmentation. Finally, in Condition 4 we control for potentially confounding methodological factors (described below).

Method

Participants

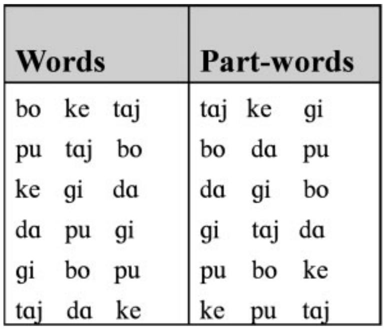
A total of 149 undergraduate students at The Pennsylvania State University participated in this study for course credit and were included in the analysis. All participants were monolingual English speakers. Participants were assigned to one of four display conditions. In Condition 1 (audio only), there were 30 participants (16 female, 14 male). In Condition 2 (visual only), there were 59 participants (48 female, 11 male); of these, 29 were in Condition 2a and saw the *aware* visual display and 30 were in condition 2b and saw the *misinformed* visual display (see below). In Condition 3 (audiovisual), 60 participants (35 female, 25 male) completed the audiovisual condition; participants were divided evenly between Condition 3a and Condition 3b. In Condition 4 (audiovisual control), there were 32 participants (25 female and 7 male). Across all conditions, 24 additional participants were excluded from analysis. Participants were excluded for a self-rated effort[1] at 5 or below on a scale of 10 (13), for failing to follow instructions (5), due to a technical malfunction (4), and we excluded one participant who reported being deaf for the first 5 years of life. In addition, we performed an outlier analysis on test scores for each experimental condition, and excluded one score in Condition 3 for being a statistical outlier.[2]

Stimuli

Condition 1: audio stimuli

The auditory stimuli consisted of an artificial language comprised of six trisyllabic words (see Figure 1). Six consonants and six vowels were combined to form a total of six CV syllables. Each syllable was created by synthesising natural speech syllables and removing any acoustic cues to word boundaries (for further details on synthesis methods, see Weiss, Gerfen, & Mitchel, 2009).

Each syllable was used three times and occurred in every possible word position (i.e. onset, medial and coda positions). In addition, there were ordering constraints ensuring that transitions within words did not also occur at word boundaries and that no syllable followed itself. Consequently, within-word transitional probabilities were .33, while the transitional probability between-words was .11.[3] While transitional probabilities did provide a cue to segmentation, the difference



Words			Part-words		
bo	ke	taj	taj	ke	gi
pu	taj	bo	bo	da	pu
ke	gi	da	da	gi	bo
da	pu	gi	gi	taj	da
gi	bo	pu	pu	bo	ke
taj	da	ke	ke	pu	taj

Figure 1. The words and part-words from the artificial language in Condition 1.

between within- and between-word transitional probabilities was less than typical for these types of studies (and the within-word transitions themselves were rather low), particularly given the number of words and length of exposure. The words were concatenated into a clip of 18 words lasting 15 seconds. This clip was then looped 16 times to create a 4-minute block, consisting of 288 words.

The test stimuli consisted of the six words and six 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 familiarisation stream, albeit less frequently than words.

Condition 2: visual stimuli

The visual stimuli were created by digitally video-recording two different male research assistants lip-syncing to the artificial language. During recording, the audio familiarisation stream was played on a nearby computer while the assistant read from a list of the items comprising the stream. In the aware video, the assistant read from a list in which the words were presented individually with word boundaries corresponding to the underlying structure of the language. In the misinformed video, a second assistant read from a list of part-words. 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 the statistical (i.e. frequency and transitional probability) cues.

In the aware video, the list of words the assistant read from only contained six unique words that were repeated (see Figure 1). Because the list in the misinformed video contained part-words, there were 18 unique items, making it more difficult to maintain accuracy while lip-syncing to a full-speed video. To resolve this issue, during video-recording of the misinformed video, the audio stream that the assistant was lip-syncing 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 permitted the assistant to lip-sync with the movie at a manageable rate. The presentation rate of the final familiarisation video was identical to the presentation rate of the audio stream in Condition 1. Since the misinformed list was comprised of 3-1-2 part-words, the first syllable and the final two syllables of the misinformed stream were removed to make the misinformed and aware videos compatible. That is, the two visual streams were identical in content, starting and ending in line with the audio stream. The only differences between the streams were the actors and the location of word boundaries signalled by visual prosodic cues.

For both videos, assistants were asked to minimise their head movements by keeping the back of their heads affixed to a point (the back of a thumbtack) on the wall behind them. This stipulation was included to avoid a wide range of movements that could produce large jerks of the head when concatenating the loops of the videos (see below). While this constraint prevented large movement artefacts, it did not interfere with subtle cues to prosody (e.g. lip aperture and small nods of the head; Blossom & Morgan, 2006; Graf et al., 2002).

The initial movies comprised 15-second clips of 18 words. The clips were imported into Adobe Premiere© and were faded in for 1 second in order to remove any jerky head movements that resulted from looping the clips to form the familiarisation stream. The locations of the fades within the streams were identical for both aware and misinformed videos. The clips were looped 16 times to create a 4-minute block, consisting of 288 words.

The visual test stimuli (only used in the video-only condition) consisted of six visual words and six visual part-words (consistent with the test items in Condition 1). Test items were created by extracting video segments from the two visual streams (aware and misinformed) using Adobe Premiere© software. The test items were paired in the same fashion as Condition 1.

Condition 3: audiovisual stimuli

The familiarisation streams used in the audiovisual condition were created by combining the audio and visual streams described above. The procedure for creating the aware (Condition 3a) and misinformed (Condition 3b) audiovisual videos was identical. Each 15-second video clip was overdubbed with the 15-second audio clip, creating an aware audiovisual clip and a misinformed audiovisual clip. Each video clip was then hand edited in Adobe Premiere© to sync the onset of the lip movements with the onset of the syllables in the audio stream. The audio and visual streams were faded in over 1 second at the beginning of the clip and then faded out at the end of the clip to remove movement artefacts between clips. The duration of the fade (1 second) ensured that the fade itself did not provide a cue to word boundary. In addition, the relative position and duration of the fade were identical for the aware and misinformed streams. The clip was then looped 16 times to create a 4-minute audiovisual block. There were no differences in the content of the misinformed or aware audiovisual stimuli – the only differences were the actor and the presence of visual cues consistent with auditory word boundaries.

Condition 4: audiovisual control

The audiovisual stimuli in Condition 4 were designed to control for two potential methodological confounds (discussed below) present in Conditions 2 and 3. The audio stimuli for Condition 4 consisted of the same speech stream from Condition 1. The visual stimuli consisted of a facial display that was created in an identical manner as the misinformed facial display in Condition 2b. The original actor from the misinformed condition was brought back to the lab approximately 1 year after the initial video was created (in order to ensure that there was no memory for the previous recording). The actor was told that he was there to record lip movements to a new language, and his subjective reports after the video was recorded confirmed that he did not recognise the syllables or language from the previous recording. For the new aware facial display, the actor lip-synced to the speech stream while reading from the same list of words used to create the original aware video mounted behind the video camera. The speech stream was slowed to 50% of the original rate while the actor lip-synced, using the same method from the earlier misinformed conditions. During the recording process, the actor was aware of the word boundaries, similar to Condition 2b. The video was then processed as described above for the misinformed conditions. The facial display was synchronised with the speech stream from Condition 1 in Adobe Premiere (as described above).

Procedure

Across all conditions, familiarisation consisted of three 4-minute presentations of the audio, visual or audiovisual familiarisation streams, with a 1-minute break between blocks, for a total of 12 minutes of familiarisation (864 words). The experimenter was present throughout the experiment to ensure that participants followed instructions, and after testing participants were given a questionnaire that assessed their self-reported level of effort.

In Condition 1, the audio stream and test were presented using E-Prime software. Participants wore noise-cancelling headphones and were instructed to listen to an audio stream about which they would later be tested. During the test phase, participants were asked to discriminate words from part-words in a two-alternative, forced-choice (2afc) test. 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 1-second pause. Participants responded 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 for position.

In Condition 2, the familiarisation stream was presented using iTunes software. Approximately half the participants watched the aware video (Condition 2a) and the other half watched the misinformed video (Condition 2b). Participants were instructed to watch a 15-minute movie and were informed that they would be tested immediately following familiarisation. They were also instructed several times that the movie did not have any sound but were asked to keep their headphones on to reduce ambient noise. During the test phase, participants completed 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 1-second pause. Participants used the keyboard to indicate which item was more likely to be a word from the movie. Unlike the auditory test, the visual test did not exhaust every possible pairing of words and part-words. 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. Test items were extracted from different points in the video stream, so no two test items would have co-occurred during familiarisation.

In Conditions 3 and 4, the familiarisation stream was presented using iTunes software. Participants in each condition were instructed to watch a 15-minute movie and informed that they would be tested immediately afterward. They were instructed to keep their headphones on throughout the experiment. All participants in the audiovisual condition completed the audio test from Condition 1, using an identical test procedure.

Results

Condition 1: audio-only

The mean percent of words chosen in Condition 1 was 52.96% (SD = 8.44). This level of performance was not significantly above chance (50%), though it approached significance: $t(29) = 1.92$, $p = .064$, $d = 0.35$, all tests two-tailed. This provides a baseline level of performance for segmentation of the auditory speech stream in the absence of additional cues.

Condition 2: visual-only

The mean percent of words chosen in Condition 2a for the aware display was 58.48% (SD = 8.87%; see Figure 2). This level of performance was significantly above chance (50%), $t(28) = 5.15$, $p < .001$, $d = 0.96$. The mean percent of words chosen for the misinformed

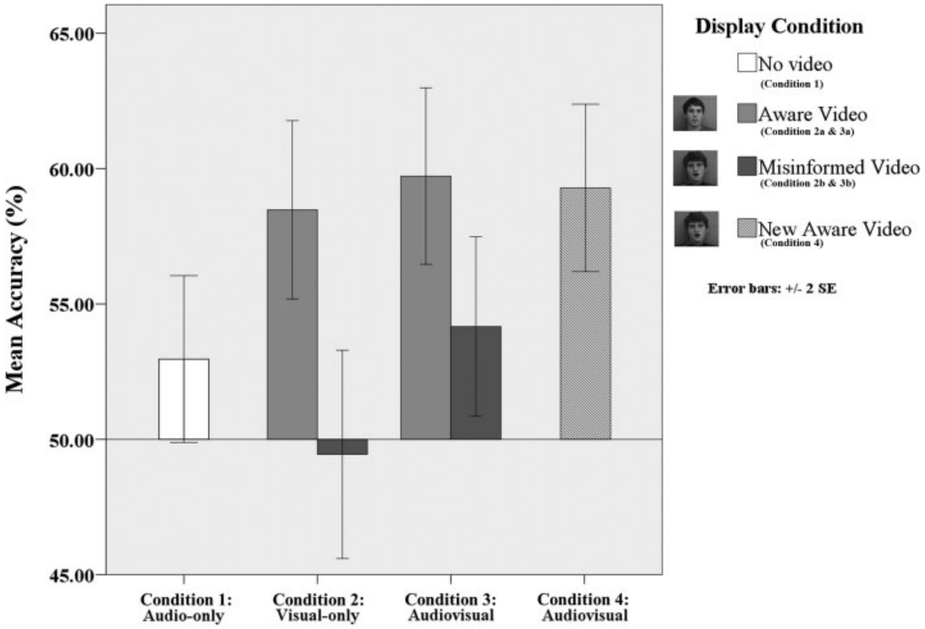


Figure 2. Percent correct identification of words across the four display conditions, plotted from chance (50%).

display in Condition 2b was 49.44% (SD = 10.54%), which was not significantly different from chance, $t(29) = 0.29$, $p = .775$, $d = 0.05$. An independent samples t-test revealed that performance for the aware display was significantly greater than performance for the misinformed display, $t(57) = 3.56$, $p = .001$, $d = 0.94$.

Condition 3: audiovisual

In Condition 3a, the mean percent of items correctly identified in the audio test following the aware audiovisual familiarisation was 59.72% (SD = 8.93), which was significantly greater than chance (50%), $t(29) = 5.97$, $p < .001$, $d = 1.09$ (see Figure 2). In Condition 3b, the mean percent of items correctly identified following the misinformed audiovisual familiarisation was 54.17% (SD = 9.07), which was significantly above chance, $t(29) = 2.52$, $p = .018$, $d = 0.46$. Independent samples' comparisons revealed that performance in Condition 3a was significantly greater than in Condition 3b ($t(58) = 2.39$, $p = .020$, $d = 0.63$) and Condition 1 ($t(58) = 3.01$, $p = .004$, $d = 0.77$). Performance for the misinformed video in Condition 3b did not differ from the baseline performance in Condition 1 ($t(58) = -0.53$, $p = .597$, $d = -0.14$). Similarly, a one-way analysis of variance revealed an overall difference in performance on the audio test across these three conditions (Conditions 1, 3a and 3b), $F(2, 87) = 5.02$, $p = .009$, $\eta^2 = 0.10$. A Tukey HSD post hoc analysis corroborated the results of the independent t-tests, revealing significant differences between the aware audiovisual condition and the audio-only and misinformed audiovisual 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 $[-.5, 1, -.5]$ for Conditions 1, 3a, and 3b, respectively, which were selected in order to compare performance in the aware audiovisual condition with performance in each of the other two conditions) confirmed that performance in the aware audiovisual condition was significantly greater than in the other two conditions, $t(87) = 3.12$, $p = .002$.

Condition 4: audiovisual control

The mean percent of words chosen in Condition 4 was 59.29% (SD = 8.74%; see Figure 2). This level of performance was significantly above chance (50%), $t(31) = 6.08$, $p < .001$, $d = 1.06$. An independent samples t-test revealed that performance in Condition 4 was significantly greater than performance in Condition 1 ($t(60) = 2.89$, $p = .005$, $d = 0.75$) and the misinformed display in Condition 3b ($t(60) = 2.26$, $p = .027$, $d = 0.58$). Further, performance in Condition 4 was not significantly different from performance in the original aware display in Condition 3a ($t(60) = 0.19$, $p = .847$, $d = 0.05$).

Discussion

In the present study, we examined whether adult participants could utilise facial cues to word boundaries to segment a continuous speech stream. In Condition 1 (audio-only), we established participants' baseline ability to segment an artificial language that contained minimal statistical cues to word boundaries. Participants had difficulty segmenting the speech stream, averaging 53% on a two-alternative forced choice task. In Condition 2 (visual-only), we tested participants' ability to segment visual speech streams produced by two assistants while lip-syncing to the auditory stream. We manipulated whether these assistants were aware (Condition 2a) or misinformed (Condition 2b) about the word boundaries during recording. We predicted that assistants would impart helpful visual prosodic contours in their facial movements if they were aware of the boundaries. Consistent with this prediction, participants were able to successfully segment the *aware* visual speech stream but failed to segment the *misinformed* visual speech stream. Notably, in Condition 2b, performance for the misinformed video was not significantly different than chance. Since the misinformed assistant likely imparted visual prosodic cues that marked the location of part-words, one might have expected participants to learn part-word boundaries, which would result in test scores significantly below chance (since the test was scored in terms of how many words were chosen relative to part-words). However, there were 18 part-words present in the familiarisation stream, each of which was observed a total of 48 times over the entire 12-minute familiarisation. By contrast, there were 6 words in the familiarisation stream that occurred a total of 144 times each during familiarisation. Given this disparity, it is not surprising that the learning of part-words in the misinformed stream did not approach the same level of learning as the words in the aware stream (see also Frank, Goldwater, Griffiths, & Tenenbaum, 2010). Regardless, the results of Condition 2 suggest that the aware facial display contained cues to the location of word boundaries that were unavailable in the misinformed stream. Further, participants were able to use this information to segment a visual speech stream, consistent with previous findings (Sell & Kaschak, 2009).

In Condition 3 (audiovisual), we asked whether participants could use these facial cues to segment a synchronous auditory stream. We paired the audio stream with the aware (Condition 3a) and misinformed (Condition 3b) visual speech streams and then presented participants with the auditory-only test (the same test used in Condition 1). In Condition 3a, the presence of visual cues to word boundaries significantly facilitated auditory speech segmentation. No facilitation was observed in Condition 3b when the talking face display did not contain visual cues to word boundaries (i.e. when the misinformed actor produced the display). Since the visual display was not present at test, the above-chance performance in Condition 3a could not have arisen as a function of visual cues to word boundaries being available during test. Rather, our results suggest that participants used visual cues during familiarisation to mark word boundaries in the auditory stream, implying that the visual input was integrated with the audio stream during learning. Furthermore, the facilitation in segmentation performance for the aware display in Condition 3a does not appear to be simply a function of increased attention to the auditory stream due to the introduction of a talking face display. If there were advantages of enhanced attention accrued from the synchronous visual display, they should have similarly benefited participants in the misinformed Condition 3b. Moreover, we observed similar levels of performance for segmentation in the aware display of Condition 2a, in which there were no auditory cues. Given the reported patterns of findings, we therefore conclude that the facilitation observed in the aware audiovisual condition (3a) did not emerge as a by-product of increased attention.

Finally, in Condition 4, we addressed two potential methodological confounds for interpreting the results of Conditions 2 and 3. The first concern was that a different actor appeared in the aware and misinformed videos. This was necessary to ensure that the actor used for the misinformed video had no prior knowledge of word 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 (see Light, Kayra-Stuart, & Hollander, 1979). Likewise, it is possible that one actor had more salient or consistent facial cues or visual prosody than the other actor. The second potential confound concerns the method of creating the videos. 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 for the aware and misinformed videos was the result of using this different recording method. However, when we controlled for these two methodological concerns in Condition 4 (using the same actor and recording methods), performance was equivalent to the two earlier aware conditions (Conditions 2a and 3a). The results of Condition 4 therefore exclude these two confounding interpretations, supporting our conclusion that learners are capable of utilising visual prosodic information to facilitate speech segmentation.

To the best of our knowledge, the pattern of results from the present study provides the first evidence that learners can rely primarily on cues from the speaker's face in order segment a novel auditory speech stream. To date, speech segmentation research has focused almost exclusively on auditory input to language learners, and we are unaware of any models for speech segmentation that incorporate visual cues from talking faces. The results of this study therefore identify an additional source of information potentially available to language learners as they attempt to overcome the challenge of speech segmentation. It is important to note that we found a benefit of facial information despite restricting the range of head movements while the actors were recorded. Prior studies investigating visual prosodic cues did not impose constraints on head movements (Blossom & Morgan, 2006; Graf et al., 2002; Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004; Sell & Kaschak, 2009; Yehia et al., 2002). Because head movements during natural discourse are likely much larger, the results of the present study may in fact underestimate the role of facial cues for speech segmentation.

Another possible contributing factor to our findings is the temporal synchrony between the visual display and audio stream. Temporal synchrony has been shown to enhance perception and learning across a variety of domains (see Lewkowicz, 2010). For example, Hollich and colleagues (2005) found that both a talking face display and an oscilloscope display, when synchronous, facilitated infants' ability to identify target words from a speech stream when presented in a noisy context. In the present study, lip-reading (see Rosenblum & Saldaña, 1996) could have provided a source of audiovisual synchrony. However, this type of synchrony alone cannot fully account for the reported pattern of results since synchrony from lip-reading was present in the misinformed audiovisual condition, yet performance was not significantly different from the audio-only baseline. This suggests that the enhanced learning in the aware audiovisual condition (Condition 3a) cannot be attributed to audiovisual synchrony. While audiovisual synchrony likely plays a critical role in early language development (see Bahrick & Lickliter, 2000), our results demonstrate that visual prosodic cues can provide a unique contribution in solving the challenge of speech segmentation.

In addition to temporal synchrony, previous work has highlighted the role of temporal contiguity between visual and auditory boundary events (Cunillera et al., 2010; Mitchel & Weiss, 2011). Cunillera and colleagues (2010) found that a contiguous visual cue (a static image) enhanced the segmentation of an auditory stream beyond the level of learning exhibited in isolation. The authors propose that the co-occurrence of onsets and offsets across modalities enhanced attention to boundary cues in the auditory stream (i.e. transitional probabilities), suggesting that temporal contiguity did not enhance learning independent of statistical mechanisms. In the present study, we extend these findings by demonstrating that adults can utilise subtle dynamic, ecologically relevant visual cues (i.e. visual prosodic cues in the speaker's face) to segment a speech stream that was difficult to parse in isolation. Moreover, given that we observed learning for the aware display of Condition 2a, we conclude that our results are consistent with the notion that visual prosodic cues enhance speech segmentation independently from auditory cues embedded in the speech stream.

Our findings also demonstrate that facial cues are integrated with the speech stream during speech segmentation, a pattern we have observed in our recent work on multisensory speech segmentation (Mitchel, Christiansen, & Weiss, in review). Mitchel et al. elicited a McGurk illusion (McGurk & MacDonald, 1976), a robust demonstration of audiovisual integration, during a statistical learning paradigm. Participants were able to integrate facial information with an artificial speech stream to produce a new statistical structure that afforded segmentation cues, thereby enhancing learning relative to the language presented in isolation. In the present study, participants used cues in the visual input to help segment an auditory speech stream, suggesting that participants integrated knowledge of word boundaries across modalities. Thus, the present study supports the notion that the mechanisms underlying speech segmentation are interactive across modalities (see Emberson, Conway, & Christiansen, 2011; Mitchel & Weiss, 2011; Mitchel, Christiansen, & Weiss, in review), and are not modality independent (cf. Seitz, Kim, van Wassenhove, & Shams, 2007).

Future work will address the developmental trajectory of this ability. It has been suggested that newborn infants have a preference for attending to faces (Goren, Sarty, & Wu, 1975; Morton & Johnson, 1991; Simion, Valenza, Macchi-Cassia, Turati, & Umilta, 2002). Furthermore, adults pattern their input to infants to enhance facial cues, exaggerating visual prosodic cues during infant-directed speech (Green, Nip, Wilson, Mefferd, & Yunusova, 2010). Facial cues therefore likely provide an early, salient cue to linguistic structure. Future infant studies will explore the extent to which infants are sensitive to visual prosodic cues to word boundaries, as well as how this sensitivity is honed with experience (see Weikum et al., 2007).

Additional research will also explore the role of visual speech cues for speech segmentation among individuals with hearing loss. For example, previous research indicates that speech segmentation is a significant obstacle for successful language development following cochlear implantation (Houston, Pisoni, Kirk, Ying, & Miyamoto, 2003). In the present study, we demonstrate that access to visual prosody in the speaker's face facilitates segmentation; thus, visual speech cues may be particularly beneficial for individuals with hearing loss. In addition, training paradigms with visual speech (e.g. Massaro & Light, 2003, 2004) could heighten access to visual speech cues, further augmenting speech segmentation abilities as well as other aspects of language acquisition.

Acknowledgements

We would like to thank Dr. Chip Gerfen for assistance in creating the auditory stimuli and Brian Giallorenzo and Kevin Weiss for assistance in creating the visual stimuli. We would also like to thank Dr. Judy Kroll and Dr. Reginald Adams for helpful comments during the preparation of this manuscript. This research was supported by NIH R03 grant HD048996-01 (DW).

Notes

1. This is consistent with previous segmentation studies (see Mitchel & Weiss, 2010; Weiss, Gerfen, & Mitchel, 2010).
2. Outliers were defined as any data point that fell outside the range specified by the following formula (see Lea & Cohen, 2004): lower bound, Quartile 1–1.5 (Quartile 3–Quartile 1); upper bound, Quartile 3 + 1.5 (Quartile 3–Quartile 1).
3. Transitional probabilities were calculated using the formula described in Aslin, Saffran, and Newport (1998): P(Y|X) = (frequency of XY)/(frequency of X).

References

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by human infants. *Psychological Science*, *9*, 321–324. doi:10.1111/1467-9280.00063

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*(2), 190–201. doi:10.1037/0012-1649.36.2.190

Blossom, M., & Morgan, J. L. (2006). Does the face say what the mouth says? A study of infants' sensitivity to visual prosody. In D. Bamman, T. Magnitskaia, & C. Zaller (Eds.), *Proceedings of the 30th annual Boston University conference on language development* (pp. 24–35). Somerville, MA: Cascadilla Press.

Brent, M., & Cartwright, T. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, *61*(1–2), 93–125. doi:10.1016/S0010-0277(96)00719-6

Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: The role of prosody. *Journal of Phonetics*, *31*, 585–598. doi:10.1016/S0095-4470(03)00040-8

Christophe, A., Peperkamp, S., Pallier, C., Block, E., & Mehler, J. (2004). Phonological phrase boundaries constrain lexical access: I. Adult data. *Journal of Memory and Language*, *51*, 523–547. doi:10.1016/j.jml.2004.07.001

Cunillera, T., Càmara, E., Laine, M., & Rodríguez-Fornells, A. (2010). Speech segmentation is facilitated by visual cues. *The Quarterly Journal of Experimental Psychology*, *63*, 260–274. doi:10.1080/17470210902888809

Emberson, L. L., Conway, C. M., & Christiansen, M. H. (2011). Timing is everything: Changes in presentation rate have opposite effects on auditory and visual implicit statistical learning. *Quarterly Journal of Experimental Psychology*, *64*, 1021–1040. doi:10.1080/17470218.2010.538972

Endress, A. D., & Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive Psychology*, *61*(2), 177–199. doi:10.1016/j.cogpsych.2010.05.001

Frank, M. C., Goldwater, S., Griffiths, T., & Tenenbaum, J. B. (2010). Modeling human performance in statistical word segmentation. *Cognition*, *117*(2), 107–125. doi:10.1016/j.cognition.2010.07.005

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

Goren, C., Sarty, M., & Wu, P. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, *56*, 544–549.

Gout, A., Christophe, A., & Morgan, J. L. (2004). Phonological phrase boundaries constrain lexical access: II. Infant data. *Journal of Memory and Language*, *51*, 548–567. doi:10.1016/j.jml.2004.07.002

Graf, P. H., Cosatto, E., Strom, V., & Huang, F. J. (2002). Visual prosody: Facial movements accompanying speech. In *Proceedings of the fifth IEEE international conference on automatic face and gesture recognition*, Washington DC.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*, 1197–1208. doi:10.1121/1.1288668

Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research*, *53*, 1529–1542.

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

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

Houston, D. M., Pisoni, D. B., Kirk, K. I., Ying, E. A., & Miyamoto, R. T. (2003). Speech perception skills of deaf infants following cochlear implantation: A first report. *International Journal of Pediatric Otorhinolaryngology*, *67*, 479–495. doi:10.1016/S0165-5876(03)00005-3

Johnson, E. K., & Jusczyk, P. W. (2001). Word segmentation by 8-month-olds: When speech cues count more than statistics. *Journal of Memory and Language*, *44*, 548–567.

Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159–207. doi:10.1006/jmla.2000.2755

Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*, 1138–1141. doi:10.1126/science.7146899

Lea, R. B., & Cohen, B. H. (2004). *Essentials of statistics for the social and behavioral sciences*. Hoboken, NJ: John Wiley & Sons.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66–77. doi:10.1037/a0015579

Light, L. L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory*, *5*, 212–228. doi:10.1037/0278-7393.5.3.212

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

Massaro, D. W., & Light, J. (2003). Read my tongue movements: Bimodal learning to perceive and produce non-native speech /r/ and /l/. In *Proceedings of Eurospeech* (Interspeech), *8th European conference on speech communication and technology*. Geneva, Switzerland.

Massaro, D. W., & Light, J. (2004). Using visible speech for training perception and production of speech for hard of hearing individuals. *Journal of Speech, Language, and Hearing Research*, *47*, 304–320.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748. doi:10.1038/264746a0

Mitchel, A. D., Christiansen, M. H., & Weiss, D. J. (in review). Cross-modal effects in statistical learning: Evidence from the McGurk illusion. Manuscript submitted for publication.

Mitchel, A. D., & Weiss, D. J. (2010). What's in a face? Visual contributions to speech segmentation. *Language and Cognitive Processes*, *25*, 456–482. doi:10.1080/01690960903209888

Mitchel, A. D., & Weiss, D. J. (2011). Learning across senses: Cross-modal effects in multisensory statistical learning. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *37*, 1081–1091. doi:10.1037/a0023700

Morton, J., & Johnson, M. H. (1991). CONSPEC and CONLERN: A two-process theory of infant face recognition. *Psychological Review*, *98*(2), 164–181. doi:10.1037/0033-295X.98.2.164

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility. *Psychological Science*, *15*(2), 133–137. doi:10.1111/j.0963-7214.2004.01502010.x

Patterson, M., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*, *22*, 237–247. doi:10.1016/S0163-6383(99)00003-X

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*, 191–196.

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 318–331. doi:10.1037/0096-1523.22.2.318

Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: The role of distributional cues. *Journal of Memory and Language*, *35*, 606–621. doi:10.1006/jmla.1996.0032

Seitz, A. R., Kim, R., van Wassenhove, V., & Shams, L. (2007). Simultaneous and independent acquisition of multisensory and unisensory associations. *Perception*, *36*, 1445–1453. doi:10.1068/p5843

Sell, A. J., & Kaschak, M. P. (2009). Does visual speech information affect word segmentation? *Memory and Cognition*, *37*, 889–894. doi:10.3758/MC.37.6.889

Simion, F., Valenza, E., Macchi-Cassia, V., Turati, C., & Umilta, C. (2002). Newborns' preference for up-down asymmetrical configurations. *Developmental Science*, *5*, 427–434. doi:10.1111/1467-7687.00237

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, *26*, 212–215. doi:10.1121/1.1907309

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*, 850–855. doi:10.1016/j.cognition.2008.05.009

Thiessen, E. D. (2010). Effects of visual information on adults' and infants' auditory statistical learning. *Cognitive Science*, *34*, 1093–1106. doi:10.1111/j.1551-6709.2010.01118.x

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*, 1159. doi:10.1126/science.1137686

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*(1), 30–49. doi:10.1080/15475440802340101

Weiss, D. J., Gerfen, C., & Mitchel, A. D. (2010). Colliding cues in word segmentation: The role of cue strength and general cognitive processes. *Language and Cognitive Processes*, *25*, 402–422. doi:10.1080/01690960903212254

Yehia, H. C., Kuratate, T., & Vaitikiotis-Bateson, E. (2002). Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, *30*, 555–568. doi:10.1006/jpho.2002.0165