HHS Public Access

Author manuscript

Mem Cognit. Author manuscript; available in PMC 2015 October 15.

Published in final edited form as:

Mem Cognit. 2009 September ; 37(6): 889–894. doi:10.3758/MC.37.6.889.

Does Visual Speech Information Affect Word Segmentation?

Andrea J. Sell and **Michael P. Kaschak**

Florida State University

Abstract

We present an experiment in which we explored the extent to which visual speech information affects learners' ability to segment words from a fluent speech stream. Learners were presented with a set of sentences consisting of novel words, in which the only cues to the location of word boundaries were the transitional probabilities between syllables. They were exposed to this language through the auditory modality only, through the visual modality only (where the learners saw the speaker producing the sentences but did not hear anything), or through both the auditory and visual modalities. The learners were successful at segmenting words from the speech stream under all three training conditions. These data suggest that visual speech information has a positive effect on word segmentation performance, at least under some circumstances.

A key step in children's language acquisition is to find the word units of their language. Succeeding at this task can be quite difficult, because, most often, the words must be extracted from a fluent stream of speech that does not provide obvious cues (such as pauses) to the boundaries between words (e.g., Saffran, Aslin, & Newport, 1996). The statistical learning approach to language acquisition proposes that children are able to successfully segment words from a fluent stream of speech by exploiting statistical regularities in their linguistic input (e.g., Aslin, Saffran, & Newport, 1998; Saffran et al., 1996). In support of this approach, it has been shown that infants are able to use the transitional probabilities between syllables (i.e., the likelihood of one syllable following another) in order to find the boundaries between words in a fluent speech stream (Aslin et al., 1998; Saffran et al., 1996). The use of transitional probabilities between linguistic units has also been shown to be sufficient to begin solving other problems in language acquisition. For instance, children and adults can use the transitional probabilities between word classes to acquire the rudiments of syntax (e.g., Kaschak & Saffran, 2006; Saffran & Wilson, 2003; Thompson & Newport, 2007), and statistical word segmentation processes have been shown to facilitate the development of links between objects and their labels (e.g., Estes, Evans, Alibali, & Saffran, 2007; Mirman, Magnuson, Estes, & Dixon, 2008).

Although numerous researchers have demonstrated that the statistical properties of the linguistic input can be exploited in the service of language acquisition (e.g., Gomez & Gerken, 1999; Saffran et al., 1996), fewer have explored how other factors interact with

Correspondence concerning this article should be addressed to A. J. Sell, Department of Psychology, Florida State University, Tallahassee, FL 32306 (asell@psy.fsu.edu).

We analyzed the data with all of the participants (except the ones that were discarded because of an experimenter's errors), and the results of these analyses were virtually identical to the results reported here based on the reduced sample.

statistical information in language learning. As one example, Thiessen and Saffran (2003, 2007) traced the developmental course of infants' use of statistical and prosodic cues to word boundaries in tasks in which both types of cues are present. Their results show that infants weight the importance of statistical and prosodic cues differently across time, initially giving more weight to statistical cues but later giving more weight to prosodic cues. Results such as these suggest that there are many layers of information that must be considered in a statistical approach to language learning. Transitional probabilities between syllables provide one layer of statistical information, and the interaction between these probabilities and the presence of prosodic cues (such as stress) in the linguistic input provides another layer of statistical information. Given the multiplicity of cues (and layers of statistical information) that are potentially available to the language learner, it is important to understand how and when (if at all) language learners exploit the information around them in acquiring their language (see Mattys, White, & Melhorn, 2005).

The goal of the present work is to examine the extent to which one type of linguistic information—the availability of visual speech information (i.e., the ability to lip read or speech read the speaker)—affects learners' ability to segment novel words from a fluent speech stream. Visual speech information refers to the range of information that can be gleaned from a speaker by watching him or her produce language, including facial expression and the movement of both the jaw and (to some extent) the tongue. It has long been known that the presence of visual speech information has beneficial effects for the perception and comprehension of speech. Sumby and Pollack (1954) were among the first to demonstrate this. They collected speech intelligibility scores from participants who listened to speech that was masked with noise. Whereas intelligibility scores tended to decline as the level of the noise increased, the decline in intelligibility was curbed by the presence of visual speech information. The finding that speech reading can improve the perception of speech in noise has been replicated numerous times (e.g., Dodd, 1977).

Visual speech information has been shown to have beneficial effects on the performance of linguistic tasks beyond the increased ability to perceive speech in noisy contexts. As one example, Soto-Faraco et al. (2007) showed that visual information alone is sufficient for Catalan– Spanish bilinguals to differentiate sentences produced in each of their languages. Although most researchers of speech reading have focused on the benefits provided by visual speech information when the linguistic signal is degraded (e.g., it is masked by noise, or the auditory portion of the speech act is taken away), others have shown that speech reading can have effects on language performance when the linguistic input is perfectly intelligible. Arnold and Hill (2001) demonstrated that the ability to speech read helps comprehenders understand the message that is being conveyed, particularly in cases in which the message is complex (see also Reisberg, McLean, & Goldfield, 1987). In addition, the McGurk effect shows that visual speech information can affect speech perception in cases in which the auditory signal is perfectly intelligible (McGurk & MacDonald, 1976).

The question of interest in this article is whether (and under what circumstances) visual speech information affects word segmentation performance. In general terms, there are good reasons to suspect that visual speech information should aid learners in the word segmentation process. Teinonen, Aslin, Alku, and Csibra (2008) reported a study in which

6-month-old infants were presented with a set of syllables along the continuum between /*ba*/ and /*da*/. When infants were presented with a visual /*ba*/ or /*da*/ with each syllable (depending on where each syllable fell on the /*ba*/–/*da*/ continuum), they learned to distinguish the two syllable types; when they were presented with one type of visual information (either a token of /*ba*/ or a token of /*da*/), they did not learn to distinguish the two syllable types. Teinonen et al. (2008) suggested that visual speech information plays a role in learning phonetic boundaries. Visual speech information may act in a similar manner during word segmentation. For example, when learners are presented with a fluent stream of speech containing unfamiliar words, they may occasionally mishear syllables (such as confusing particular tokens of /*ba*/ and /*da*/) because of suboptimal listening conditions (e.g., the learner is in a noisy room) or a lack of attention. These mistakes will lead them to incorrectly recover the statistical structure of the linguistic input and will thereby hurt word segmentation performance. Visual speech information can serve to protect against such problems and allow learners to more successfully recover the structure of the linguistic input.

Arnold and Hill's (2001; Reisberg et al., 1987) demonstration that visual speech information improves the comprehension of complex linguistic input points to a second reason that visual speech information may improve word segmentation. Given that the speech stream presented to the participants in a word segmentation experiment is largely unfamiliar to the learners, visual speech information may lighten the load associated with processing the linguistic input and may thereby increase their attention to the aspects of the training set relevant for learning (see Toro, Sinnett, & Soto-Faraco, 2005, for evidence that attention is an important prerequisite for word segmentation).

The experiment reported below represents an initial exploration of the effects of visual speech information on word segmentation. Because this is an initial exploration, we chose to test adult participants in our study. We presented the participants with a word segmentation task similar to those used in earlier studies (e.g., Saffran et al., 1996; Thiessen, Hill, & Saffran, 2005). The participants were exposed to a series of sentences that consisted of four training words, plus beginning and ending syllables that were irrelevant to the words (see below). The sentences were generated such that the only cues to word boundaries in the speech stream were the transitional probabilities between syllables. The participants were trained in one of four conditions (created by crossing the presence or absence of auditory information at training with the presence or absence of visual information at training): no exposure to either the auditory or visual components of the training sentences (control condition); exposure to only the auditory component of the training sentences (auditory-only condition); exposure to only the visual component of the training sentences (visual-only condition); or exposure to both the auditory and the visual components of the training sentences (auditory-and-visual condition).

On the basis of previous work using similar tasks and the same input language (e.g., Thiessen et al., 2005), it is expected that the auditory input will be enough for the participants to at least begin to segment the words from the sentences that are presented. The question of interest is whether giving the participants the ability to speech read the speaker producing the training sentences will provide any benefits in performance above and beyond

what is produced by the auditory training. If this is the case, it will provide some of the first evidence that visual speech information can produce beneficial effects in language learning (see Teinonen et al., 2008, for evidence that visual speech information can benefit phonetic learning in infants).

In addition to manipulating the information that was presented during training, we also manipulated whether the speaker who produced the test items was the same as or different from the speaker who produced the test stimuli. We did this largely because of reports from the literature on speech perception suggesting that speech reading effects may be speaker specific (i.e., training on one speaker does not transfer to other speakers; Rosenblum, Miller, & Sanchez, 2007). Therefore, we wanted to ascertain whether any speech-reading effects in our task would transfer across speakers. Our participants heard test items either from the same (female) speaker who produced the test stimuli, from a different female speaker, or from a male speaker. The question of interest is whether the participants would do a better job recognizing the words that they segmented from the speech stream when the same speaker is used at training and test than when the speakers change between training and test.

METHOD

Participants

Two hundred and forty-three undergraduate students from Florida State University participated in this study for class research credit. All of the participants were native English speakers with normal hearing. After completion of the study, we had to discard the data from 4 participants because of errors on the part of our research assistants. The loss of these participants unbalanced the number of participants per condition, and so we discarded the last 1 or 2 participants (depending on the original sample size of the condition) that were run in each condition. This resulted in a total of 19 participants in each of the 12 cells of the design (created by crossing the presence or absence of auditory information with the presence or absence of visual information with three test speaker conditions), for a grand total of 228 participants.1

Materials

The artificial language used for this experiment was adapted from Thiessen et al. (2005). The language consisted of four words: *nifopa*, *dibo*, *kuda*, and *lagoti*. These words were arranged into 12 training sentences, such that each word appeared after each other word an equal number of times across the training set. The transitional probability between syllables within words was 1.0, and the transitional probability between syllables at word boundaries was .25. Each of the 12 sentences was presented once during training, which resulted in a training set approximately 1 min in length. The sentences were produced by a female speaker with extensive musical training. She was instructed to produce the words as a constant monotone stream of syllables, without spaces in between syllables and without intonation or inflection of voice. Each sentence began with *mo* and ended with *fa* to eliminate the use of beginnings and endings of sentences as word boundary cues. For the test of the participants' acquisition of the words, four word pairs were generated. Each pair had one word and one nonword. The nonwords consisted of two or three syllables that appeared

in the training words but did not appear together in the same word. The transitional probabilities between syllables in the nonwords varied, with the highest transitional probability between syllables being .25 and the lowest transitional probability between syllables being 0. The word– nonword test pairs were *nifopa–nilaku*, *dibo–padi*, *kuda–paku*, and *lagoti–labogo*. The test materials were recorded by the same female speaker who produced the training stimuli, by a different female speaker, and by a male speaker. The word member of each test pair was recorded separately for the test (as opposed to being spliced from the speech stream used for training).

Virtually all studies of word segmentation have used synthesized speech in order to ensure that transitional probabilities are the only cues available to find word boundaries. The nature of our experiment necessitated the use of a live human speaker, and we therefore wanted to verify that our speaker did not unwittingly introduce any word boundary cues into the training set. We did this in two ways. First, we performed a norming study in which each sentence was presented individually to participants not included in any of the experiments reported here. The participants listened to one sentence and were asked to pick out what they thought could be the potential words of the artificial language. They did so by verbally producing any set of syllables that they thought made a word in the sentence. Out of 69 total responses recorded from all of the participants, only 2 were actual words in the language. Thus, it appears that the individual training sentences do not provide learners with cues to the identities of the words in the language.

To further ensure that our speaker did not produce cues to word boundaries in the training set, we analyzed the pitch, amplitude, and duration of each of the syllables in the linguistic input. We looked at these data in several ways. First, we assessed whether word-initial, word-medial, or word-final syllables differed systematically in pitch, amplitude, or duration. There were no significant differences across syllable types on any of these dimensions [duration, $F < 1$; amplitude, $F(2,7) = 2.02$, $p = .21$; pitch, $F(2,7) = 2.35$, $p = .17$]. Second, we assessed whether any of the syllables produced within a given sentence stood out from the rest of the syllables in that sentence with regard to pitch, amplitude, or duration (i.e., in any way that would signal the learners to pay attention to that syllable in the speech stream and that would thereby point them to a word boundary). We defined standing out as being any value that was more than 2 standard deviations from the mean value for pitch, amplitude, or duration for that sentence. Only one syllable from one of the four words in the language met this criterion—the syllable *fo* (from *nifopa*), produced in Sentence 2; its pitch was 2 Hz below the 2 standard deviation value for that sentence. In addition, the syllable *fa* often fell below the 2-standard-deviation value on both amplitude and pitch. This is likely because *fa* is the last syllable of each sentence and was produced with a slightly lower pitch and amplitude than the rest of the sentence as the speaker finished her production. When productions of *fa* fell below this range, they were an average of 1.03 dBs below the amplitude criterion and 13 Hz below the pitch criterion. Because *fa* was a filler syllable at the end of the sentence, we do not feel that these relatively small deviations from the production of the rest of the sentence affected word segmentation performance. Finally, we found no pauses between any of the syllables. On the basis of this analysis of the acoustic

features of our training set, we were confident that our speaker did not inadvertently introduce any cues to word boundaries into the linguistic input.

Procedure

The participants were assigned to one test condition (same speaker, different female speaker, or male speaker) and then randomly assigned to one of four training conditions: control (no auditory or visual input), auditory training only, visual training only, and both auditory and visual training (i.e., a factorial design crossing the presence or absence of auditory information with the presence or absence of visual information). The participants in the auditory training, visual-training, and auditory-and-visual training conditions were presented with the training sentences as specified by their assigned condition. The participants in the visual-training condition saw a video of the speaker producing the sentences with the sound muted. The participants in the auditory-training condition heard a sound file of the training presentation but did not see the speaker. The participants in the auditory-and-visual condition saw the training presentation video, which contained the sound file as well as the video. The audio and video files used in the experiment were culled from the same original video of the speaker. The participants in the control condition received no training input.

After the training presentation, the participants were given a forced-choice discrimination test to assess their knowledge of the words presented in the training set. The discrimination test consisted of four trials, each a pairing of a word in the artificial language with a nonword generated from the syllables used in the training set (see above). The test items were presented in the auditory modality only, and the same test was presented after all training conditions. Although the removal of visual information at test produces a training–test mismatch in the conditions in which visual information was present at training, it is worth noting that this mismatch actually works against the hypothesis that visual speech information affects word segmentation. In the auditory-training condition and the auditory-and-visual-training condition, the participants were instructed to pick the member of each pair that sounded most like the speech that they had just heard. In the visual-training condition, the participants were told to pick the member of each pair that sounded most like it could have been from the language that they had seen the speaker produce during training. For the test condition in which the participants had no training presentation, they were told to pick the member of each pair that sounded best.

Design and Analysis

The proportion of correct responses on the test of word knowledge was analyzed with a 3 (test speaker: same speaker, different female speaker, male speaker) × 2 (auditory information: present, absent) × 2 (visual information: present, absent) ANOVA. All factors were between participants.

RESULTS

The proportion of correct responses on the forced-choice test of word knowledge (collapsed across test-speaker conditions) is presented in Table 1. The participants performed above chance on the task in the visual-training condition [$t(56) = 3.37$, $p = .001$], the auditory-

training condition [$t(56) = 9.72$, $p < .001$], and the auditory-and-visual condition [$t(56) = 10.06$, $p < .001$]. Performances in the control condition did not differ from chance [$t < 1$].

Table 2 presents the means from each training condition, separated by test speaker. The three-factor ANOVA revealed a main effect of auditory information [$F(1,216) = 59.80$, $p < .001$], with the participants who received auditory training ($M = 78\%$) outperforming those who did not ($M = 55\%$). There was a main effect of visual information [$F(1,216) = 6.90$, $p = .009$], with the participants who received visual training ($M = 70\%$) outperforming those who did not ($M = 62\%$). The main effect of test speaker was not significant ($F < 1$), suggesting that word segmentation performance was equivalent across the different speakers used to produce test items. None of the interactions were significant (Fs < 2.12, p > .12), further suggesting that the pattern of effects for the auditory-and-visual-information conditions did not differ across test-speaker conditions. Follow-up analyses revealed that the presence of auditory information at training benefited learners in both the absence [$F(1,216) = 39.23$, $p < .001$] and the presence [$F(1,216) = 21.35$, $p < .001$] of visual information. The presence of visual information during training benefited the learners in the absence of auditory input [$F(1,216) = 7.13$, $p = .008$] but not in the presence of auditory input [$F(1,216) = 1.05$, $p = .31$].

Within each training condition, there were fluctuations in test performance across speakers (see Table 2). The performance in the auditory-training condition was lower in the same-speaker condition than in the two different speaker conditions, and performance in the auditory-and- visual condition was higher in the same-speaker and different-female- speaker conditions than in the male speaker condition. We wanted to assess whether these fluctuations represent significant changes in performance across speaker conditions. To do so, we analyzed the data from each training condition with a single-factor ANOVA to look for a main effect of test speaker. The results were not significant for any of the training conditions [control and visual only, $F < 1$; auditory only, $F(2,216) = 1.92$, $p = .15$; auditory-and-visual, $F(2,216) = 1.54$, $p = .22$]. This suggests that the differences across test speakers in each training condition were random fluctuations in performance and not meaningful differences being driven by the test-speaker factor.

DISCUSSION

The purpose of our study was to explore the extent to which visual speech information affects word segmentation. The main effect of visual information observed here, coupled with the lack of interactions between this factor and the auditory-information and test-speaker factors, suggest that, overall, visual speech information aids word segmentation. Within this broader conclusion, there are several things to note. First, the presence of visual speech information alone appears to be sufficient to allow learners to segment words from a fluent speech stream. The participants in the visual-training condition performed above chance on the test of word knowledge. Furthermore, performance in the visual-training condition was identical in the same-speaker and different-female-speaker test conditions, which suggests that learning based on visual speech information can transfer across speakers. This latter point is qualified somewhat by the data from the male-speaker test condition: Although performance in the visual- training condition did not differ across test

speakers (see above), performance in the male-speaker test condition was not significantly different from chance. The diminished performance in the male-speaker test condition suggests that the generalization of learning based on visual speech information may not be uniformly strong across all speakers.

A second noteworthy aspect of our data is that, whereas there is a main effect of visual information overall, the effect of visual information in the presence of auditory training was not statistically reliable. Examining these data by individual test speakers, we found that the effect of visual information in the presence of auditory training was significant in the same-speaker condition ($p = .04$) but not in the different-female-speaker and male-speaker conditions ($Fs < 1$). Interestingly, the benefits of visual information in the same-speaker condition seem to be the result of a decline in performance in the auditory-only training condition (relative to the other test-speaker conditions) rather than an increase in performance in the auditory- and-visual training condition (see Table 2). The safest conclusion to draw at this point appears to be that whereas there may be circumstances under which visual speech information can benefit word segmentation in the presence of auditory training, the present study does not provide strong evidence for this claim. Indeed, the weakness of the visual-information effect in the presence of auditory information may be due to the nature of our task: The speech stream was clear, intelligible, and presented over headphones to obviate the intrusion of extraneous noise, and the word segmentation task was a comparatively easy one (using only four words). If the participants were able to successfully segment the words from the speech stream using auditory information alone (and the generally high levels of performance in the auditory-training condition suggest that this is the case), there may not have been much room for visual speech information to further improve performance.

In light of these findings, what can we say about the role of visual speech information in word segmentation? It is well established that both auditory and visual information play a role in speech perception (see Rosenblum, 2008, for a discussion), although the weight given to each modality may differ across speech perception contexts (e.g., Massaro & Friedman, 1990). The relative import of the auditory and visual modalities in speech perception should determine the influence of visual speech information on word segmentation performance. In cases in which the acoustic elements of the speech stream are unambiguous and easy to identify, and in which the word segmentation task is not particularly difficult, visual speech information may exert a relatively weak effect on word segmentation. In cases in which the acoustic elements of the speech stream are degraded (e.g., the speech is difficult to hear in a noisy environment) or difficult to interpret (e.g., listening to a speaker with an unfamiliar accent), visual speech information will play a larger role in speech perception and will therefore play a larger role in the word segmentation process. Our data bear these expectations out. When the word segmentation task can be done successfully with auditory information alone, visual speech information does not reliably contribute much to performance above and beyond what can be done with the auditory information itself. When the word segmentation task cannot be done with auditory information alone (in our case, because this information was not presented), visual speech information plays a larger role. Given that at least some early language learning (and word

segmentation), occurs in environments that are less than optimal for the perception of the acoustic speech signal (e.g., a noisy playroom), and that some of the acoustic signal may be ambiguous to a child who is just beginning to acquire phonetic categories, it seems likely that visual speech information is a cue that plays a supporting role in language acquisition (see Teinonen et al., 2008).

Another issue with respect to the role of visual speech information in word segmentation concerns the generalizability of the learning that occurs. Our data clearly show that learning based on auditory information generalizes across speakers (see the data from our auditory-training condition). However, the data on the generalizability of visual-speech-based learning suggest that generalization is not uniformly strong across all speakers. It has been demonstrated that infants can generalize their auditory learning across speakers under certain circumstances (e.g., Houston & Jusczyk, 2000, 2003), and our data suggest that visual-speech-based learning may generalize in a similar manner. It may be that getting a complete picture of the patterns of generalization for visual-speech-based learning across speakers will require modifications of the research design employed here. For instance, it might be possible to get a different look at generalization effects in cases in which the experiment is structured to heighten the role of visual speech information in word segmentation performance (e.g., by obscuring the auditory speech stream in noise). Additionally, Houston and Jusczyk's (2000, 2003) research suggests that stronger generalizations of visual-speech-based learning across speakers may be found if more than one speaker is used in the training set. Addressing this issue will be important for defining the role of visual speech information in language learning.

One further point requires comment. Recent studies have shown that nonlinguistic visual information (such as the movement of visual cues that are synchronized to the speech stream) can aid both learners' ability to segment auditory input into separate speech streams (Hollich, Newman, & Jusczyk, 2005) and their ability to perform word segmentation tasks (Thiessen, 2009). Although it is tempting to view visual speech information as functioning in a similar way, essentially providing a visual cue that correlates with events in the speech stream, we think that this is not an entirely apt comparison for several reasons. First, visual speech information does not perfectly correlate with the speech stream in the same way that an appearing and disappearing visual stimulus does. The movement or appearance of shapes can present a single visual signal to cue the location of word onsets, but the visual speech information that occurs at word onsets is not as consistent; essentially, it varies depending on the syllables that are used to begin and end each word. Second, although it is true that visual speech information can be used to disambiguate ambiguous auditory stimuli, it is also true that the visual speech information itself can be ambiguous. For example, the syllables /*di*/, /*ti*/, and /*ni*/ from our experiment would not be distinguishable on the basis of visual information alone. No such ambiguity exists in the nonlinguistic visual cues that have been used by Hollich et al. (2005) or Thiessen (2009). Finally, whereas the nonlinguistic cues discussed here are employed to signal the location of word boundaries, the contribution of visual speech information to word segmentation seems to be of a different sort. Rather than directly cuing the learner to the location of word boundaries, visual speech information

presumably aids word segmentation by helping the learner do a better job of recovering the information presented in the speech stream (see Teinonen et al., 2008).

Awareness of visual speech information emerges early in life (Kuhl & Meltzoff, 1982; Patterson & Werker, 2003) and plays an important role in the speech perception of children and adults (e.g., Rosenblum, 2008). The work presented in this article joins with recent studies on phonetic learning (Teinonen et al., 2008) to suggest that visual speech information may play a role in language learning. Although it is clear that we have only begun to scratch the surface with respect to defining how visual speech information affects the language learning process, it is our hope that these data will spur further interest in examining the way that this source of information functions to support language learning in both children and adults.

Acknowledgments

The work reported in this article was supported in part by NSF Grant BCS 0446637. Thanks to Juliet Leon for her contributions to an early stage of this project. Thanks also to Erik Thiessen for helpful discussions of this work. This article was improved by the comments of three anonymous reviewers.

References

Arnold P, Hill F. Bisensory augmentation: A speech reading advantage when speech is clearly audible and intact. British Journal of Psychology. 2001; 92:339–355.

Aslin RN, Saffran JR, Newport EL. Computation of conditional probability statistics by 8-month-old infants. Psychological Science. 1998; 9:321–324.

Dodd B. The role of vision in the perception of speech. Perception. 1977; 6:31–40. [PubMed: 840618]

Estes KG, Evans JL, Alibali MW, Saffran JR. Can infants map meaning to newly segmented words? Psychological Science. 2007; 18:254–260. [PubMed: 17444923]

Gomez RL, Gerken L. Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. Cognition. 1999; 70:109–135. [PubMed: 10349760]

Hollich G, Newman RS, Jusczyk PW. Infants' use of synchronized visual information to separate streams of speech. Child Development. 2005; 76:598–613. [PubMed: 15892781]

Houston DM, Jusczyk PW. The role of talker-specific information in word segmentation by infants. Journal of Experimental Psychology: Human Perception & Performance. 2000; 26:1570–1582. [PubMed: 11039485]

Houston DM, Jusczyk PW. Infants' long-term memory for the sound patterns of words and voices. Journal of Experimental Psychology: Human Perception & Performance. 2003; 29:1143–1154. [PubMed: 14640835]

Kaschak MP, Saffran JR. Idiomatic syntactic constructions and language learning. Cognitive Science. 2006; 30:43–63. [PubMed: 21702808]

Kuhl PK, Meltzoff AN. The bimodal perception of speech in infancy. Science. 1982; 218:1138–1141. [PubMed: 7146899]

Massaro DW, Friedman D. Models of integration given multiple sources of information. Psychological Review. 1990; 97:225–252. [PubMed: 2186424]

Mattys SL, White L, Melhorn JF. Integration of multiple speech segmentation cues: A hierarchical framework. Journal of Experimental Psychology: General. 2005; 134:477–500. [PubMed: 16316287]

McGurk H, MacDonald J. Hearing lips and seeing voices. Nature. 1976; 264:746–748. [PubMed: 1012311]

Mirman D, Magnuson JS, Estes KG, Dixon JA. The link between statistical segmentation and word learning in adults. Cognition. 2008; 108:271–280. [PubMed: 18355803]

Patterson ML, Werker JF. Two-month-old infants match phonetic information in lips and voice. Developmental Science. 2003; 6:191–196.

Reisberg, D.; McLean, J.; Goldfield, A. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In: Campbell, R.; Dodd, B., editors. Hearing by eye: The psychology of lip-reading. Hillsdale, NJ: Erlbaum; 1987.

Rosenblum LD. Speech perception as a multimodal phenomenon. Current Directions in Psychological Science. 2008; 17:405–409. [PubMed: 23914077]

Rosenblum LD, Miller RM, Sanchez K. Lip-read me now, hear me better later: Cross-modal transfer of talker-familiarity effects. Psychological Science. 2007; 18:392–396. [PubMed: 17576277]

Saffran JR, Aslin RN, Newport EL. Statistical learning by 8-month-old infants. Science. 1996; 274:1926–1928. [PubMed: 8943209]

Saffran JR, Wilson DP. From syllables to syntax: Multilevel statistical learning by 12-month-old infants. Infancy. 2003; 4:273–284.

Soto-Faraco S, Navarra J, Weikum WM, Vouloumanos A, Sebastián-Gallés N, Werker JF. Discriminating languages by speech-reading. Perception & Psychophysics. 2007; 69:218–231. [PubMed: 17557592]

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. Journal of the Acoustical Society of America. 1954; 26:212–215.

Teinonen T, Aslin RN, Alku P, Csibra G. Visual speech contributes to phonetic learning in 6-month-old infants. Cognition. 2008; 108:850–855. [PubMed: 18590910]

Thiessen, ED. Effects of visual information on word segmentation. 2009. Manuscript submitted for publication

Thiessen ED, Hill EA, Saffran JR. Infant-directed speech facilitates word segmentation. Infancy. 2005; 7:53–71.

Thiessen ED, Saffran JR. When cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-month-old infants. Developmental Psychology. 2003; 39:706–716. [PubMed: 12859124]

Thiessen ED, Saffran JR. Learning to learn: Infants' acquisition of stress-based strategies for word segmentation. Language Learning & Development. 2007; 3:73–100.

Thompson SP, Newport EL. Statistical learning of syntax: The role of transitional probability. Language Learning & Development. 2007; 3:1–42.

Toro JM, Sinnett S, Soto-Faraco S. Speech segmentation by statistical learning depends on attention. Cognition. 2005; 97:B25–B34. [PubMed: 16226557]

Table 1

Proportion of Correct Responses (Collapsed Across Test Speakers) and Standard Deviations on the Test of Word Knowledge

	Audio Condition			
	No Audio		Audio	
Visual Condition	***M***	***SD***	***M***	***SD***
No visual	.49	.26	.75**	.20
Visual	.60*	.23	.80**	.22

* $p = .001$.

** $p < .001$.

Table 2

Proportion of Correct Responses on the Test of Word Knowledge and Tests of Proportions Against Chance and Standard Deviations, Presented by Test Speaker

Condition	*M*	*SD*	*t*	*p*
Same Speaker at Training and Test				
Control	.47	.26	−0.44	.67
Visual only	.62	.24	2.14	.04
Auditory only	.67	.21	3.64	.002
Auditory and visual	.83	.21	6.99	<.001
Different Female Speaker				
Control	.49	.21	−0.27	.79
Visual only	.62	.24	2.14	.04
Auditory only	.80	.16	8.37	<.001
Auditory and visual	.84	.21	7.18	<.001
Male Speaker				
Control	.50	.30	0.00	1.00
Visual only	.57	.20	1.42	.17
Auditory only	.79	.21	6.05	<.001
Auditory and visual	.72	.25	3.92	<.001