Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution

Alexandra Jesse[1,2], Elizabeth K. Johnson[3]

[1]University of Massachusetts, Amherst, U.S.A.

[2]Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

[3]University of Toronto, Toronto, Canada

IN PRESS at Journal of Experimental Psychology: Human Perception and Performance

AUTHOR NOTE

This work was supported by a grant from the German Science Foundation to the first author and from the Natural Sciences and Engineering Research Council of Canada to the second author. The second author was also supported by a Spinoza grant awarded to Anne Cutler. The authors thank Eelke Spaak, Michael Rogers, and all speakers, as well as the student assistants of the Language Comprehension Group for their assistance. We also thank Rochelle Newman for helpful comments on an earlier draft of this manuscript. Correspondence should be addressed to Alexandra Jesse, Department of Psychology, University of Massachusetts, Tobin Hall, Amherst, MA 01003, or to ajesse@psych.umass.edu.

Abstract

Using a referent detection paradigm, we examined whether listeners can determine the object speakers are referring to by using the temporal alignment between the motion speakers impose upon objects and their labeling utterances. Stimuli were created by videotaping speakers labeling a novel creature. Without being explicitly instructed to do so, speakers moved the creature during labeling. Trajectories of these motions were used to animate photographs of the creature. Participants in subsequent perception studies heard these labeling utterances while seeing side-by-side animations of two identical creatures, where only the target creature moved as originally intended by the speaker. Using the cross-modal temporal relationship between speech and referent motion, participants identified which creature the speaker was labeling, even when the labeling utterances were low-pass filtered to remove their semantic content or replaced by tone analogues. However, when the prosodic structure was eliminated by reversing the speech signal, participants no longer detected the referent as readily. These results provide strong support for a prosodic cross-modal alignment hypothesis. Speakers produce a perceptible link between the motion they impose upon a referent and the prosodic structure of their speech, and listeners readily use this prosodic cross-modal relationship to resolve referential ambiguity in word-learning situations.

Word count: 199/200

Keywords: Audiovisual perception, referent resolution, prosody, synchrony, speech

Prosodic temporal alignment of co-speech gestures to speech facilitates referent resolution

The arbitrary mapping between phonological word forms and their meaning has traditionally been identified as one of the hallmarks of human language (De Saussure, 1915; 1966; Hockett, 1960; but see e.g., Bloomfield, 1935; 1976; Parault & Schwanenflugel, 2006). Finding the intended referent of a label is a challenge that children and adults often face when encountering novel labels or familiar labels in situations with several possible referents. Imagine, for example, a sailing novice is ordered to give way to 'that boat', but there is both a small taxi boat and a large ferry boat on the horizon. Which boat is the skipper referring to? Worse yet, in communicative settings we routinely encounter novel labels. Imagine that the same sailing novice is told to 'watch the boom as it moves to port side' or 'watch the ticklers on the genoa'. Clearly, as illustrated by these examples, the need to work out mappings between word form and intended referent can be a challenge throughout life for even the most seasoned language users. Given the lack of a transparent mapping between sound and meaning in spoken language, how do listeners work out what speakers are referring to?

Listeners appear to be very resourceful when it comes to solving the referent-mapping problem. Much of the work done in this area has focused on children, because with their small vocabularies and limited world knowledge, children undoubtedly face referentially ambiguous situations more often than the average adult listener. The strategies children use to infer a speaker's intended referent change over the course of development (e.g., Hollich, Hirsch-Pasek & Golinkoff, 2000). Early on, infants may rely on cross-situational statistics to work out word meanings (Smith & Yu, 2008). They also show a tendency to attach a spoken label to whatever object they happen to be attending to when the label is spoken (Pruden, Hirsh-Pasek, Golinkoff, & Hennon, 2006). Later on, toddlers rely on knowledge-based strategies, such as deduction or social pragmatics (Clark, 1990; Diesendruck & Markson, 2001; Golinkoff, Hirsh-Pasek, Bailey, & Wenger, 1992; Kidd, White, & Aslin, 2011; Markman, 1990), and eventually children begin using grammatical cues (e.g., Bernal, Lidz, Millotte & Christophe, 2007; Jolly & Plunkett, 2008). By the time language users reach adulthood, they have a substantial arsenal of referential mapping tools at their disposal.

One important tool that both speaker and listener may have at their disposal to facilitate communication is the establishment of a cross-modal relation that unifies the label and the referent events to a multisensory event (Bahrick & Lickliter, 2002). Work in the developmental literature has shown that the simple temporal synchronization of the onset of word labeling and the onset of referent motion, for example, can facilitate the associative learning of a word-referent relation in a situation, where referential ambiguity is largely resolved for the listener as only one likely possible referent is present (Gogate & Bahrick, 1998; 2001). In the present study, we tested whether intermodal temporal alignment between speech and referent motion is used by adult listeners as a cue to resolve referential ambiguity. To test whether adult listeners use this cue to determine the identity of referents of novel words, two possible referents were presented in the visual scene. Importantly, we also investigated the nature of the alignment of speech and the motion imposed on the referent by the speaker. More specifically, we examined the possibility that this perceived dynamical cross-modal alignment is prosodic in nature.

Intermodal relations as a cue to referent resolution

Adult listeners routinely face a cross-modal mapping problem when determining the referent of a novel word or when determining the intended referent of a known word in the presence of multiple possible referents (e.g., when hearing “give way to that boat” in an environment with multiple boats). The arbitrariness of a label-referent relation can be reduced by establishing a cross-modal link between the auditory label and the visual referent. One way to accomplish this is by expressing properties of the referent in the acoustic realization of the label (Nygaard, Herold, & Namy, 2009). The tone of voice a word is spoken in can be systematically varied to encode referential properties, and can hence also express the word’s meaning, without changing the word’s phonological form. Variation in tone of voice consists of changes in suprasegmental acoustic properties, such as speaking rate, pitch, vocal effort, or loudness, which are realized independently of the linguistic prosodic structure. Speakers use tone of voice to express an adjective’s semantic dimension and the valence of its meaning (e.g., big-small, yummy-yucky; Nygaard et al., 2009). Speakers tend to say, for example, the nonsense adjective “blicket” in “Can you find the blicket one?” in a lower, louder, and slower voice when referring to a big as opposed to a small referent (Nygaard et al., 2009). Adult listeners (and five-year old children) can use tone of voice to determine the intended referent out of two possible referents (e.g., a small tree and a big tree) that primarily differ along the adjectival semantic dimension expressed by the tone of voice (Nygaard et al., 2009; Herold, Nygaard, Chicos, & Namy, 2011). Similarly, tone of voice can be applied to known words to resolve referential ambiguity. The tone of voice of saying “Give way to that boat!” could help identifying the intended referent in a visual scene with multiple possible referents (e.g., a harbor) by indicating the size or proximity of the referent, or perhaps the urgency of the situation. Adult listeners use tone of voice to resolve lexical ambiguity when encountering emotional homophone pairs, such as “mourning” and “morning” (Nygaard & Lunders, 2002; see also Nygaard & Queen, 2008). Tone of voice expressing the emotional state of a speaker can also resolve referential ambiguity. Four-year-old children can use a speaker’s sad tone of voice to infer that the intended referent is a broken toy (Berman, Chambers, & Graham, 2010). Tone of voice can therefore help adult and young listeners in establishing the cross-modal relationship between known or novel words and their intended referents.

Tone of voice can also express referents’ transitory properties, such as their motion. Adult speakers change their rate of speaking in relation to the referent’s speed of motion, even when there is no referential ambiguity (Shintel, Nusbaum, & Okrent, 2006). Sentences expressing the horizontal movement of a dot (“It’s going left/right”) were spoken more slowly when seeing the dot moving at a slower than at a faster rate. Similarly, adult speakers change their pitch to express the direction of vertical movement (Shintel et al., 2006). Speakers said “up” with a higher pitch to describe the motion of an upwards-moving dot than they said “down” to describe the motion of a downwards-moving dot. Critically, these pitch differences were not due to phonetic differences between the two labels as pitch did not differ in a control condition with phonetically-matched nonsense labels (“bup”, ”bown”). This supports the idea that these pitch differences reflect the encoding of the referent’s motion. It is unclear, however, whether speakers temporally align variation in their speech to the dynamics of the referent’s motion. Also, it is currently unclear whether listeners use these acoustical expressions of the referent’s motion to determine the intended referent in cases of referential ambiguity.

Temporal intermodal alignment

Tone of voice reduces the arbitrariness between labels and their referents by expressing properties of referents in the acoustic realization of labels. Cross-modal relationships between an auditory and a visual event can also be established by temporally synchronizing labeling with the listener's visual attention to an object. For example, to establish temporal cross-modal contiguity, caregivers monitor their young infants' eye gaze and pointing gestures to label the object the infant is currently attending to (e.g., Collis & Schafer, 1975; Harris, Jones, & Grant, 1983; Masur, 1982; Messer, 1978). Temporal contiguity between labels and their corresponding referents presumably helps establish joint attention between adults and children. Establishing joint attention, in turn, helps children learn new words (Baldwin, 1991; Baldwin, Markman, Bill, Desjardins, Irwin, & Tidball, 1996; Brooks & Meltzoff, 2005; Hirotani, Stets, Striano, & Friederici, 2009).

Another form of establishing a temporal cross-modal relationship is for a speaker to align the onset of labeling to the onset of motion the speaker imposes on the referent. Most of the research examining the role of this type of cross-modal synchrony has been conducted with children acquiring their first language. When asked to teach label-referent relations to their children, mothers appeared to temporally synchronize the onset and offset of their labeling with the onset and offset of motion they imposed on the object (Gogate, Bahrick, & Watson, 2000). This simple form of temporal synchronization was found more often for the label to be taught than for other labels in the mothers' speech. Young infants can use this simple form of intermodal temporal synchronization to learn word-referent relationships in situations with little referential ambiguity, when only one object is presented. 7-month-old children, for example, only learned word-referent associations when the labeling and referent motion were temporally synchronized during training (Gogate & Bahrick, 1998; 2001). When the onset of the object's motion and the onset of labeling were asynchronous or when the object did not move, these children failed at learning the relations. Similar results were found already for 2-month-old infants when learning a one syllable-object pair (Gogate, Prince, & Matatyaho, 2009). The degree to which mothers produce this type of temporal synchrony is correlated with their six- to eight-month-old infants' success in learning word-referent associations in a word-learning setting (Gogate, Bolzani, & Betancourt, 2006). Temporal synchrony thus establishes an important link between otherwise arbitrarily related labels and referents. This link helps young infants with associative label-referent learning in situations where only the intended referent is presented, that is, when the problem of determining the intended referent has been minimized for the child. But it is unclear whether temporal alignment also helps with referent resolution by establishing the novel label-referent relationship in situations with referential ambiguity; i.e. when more than one likely possible referent is present in the visual scene.

In the present study, we investigated whether temporal cross-modal alignment can be used by adults as a cue to infer the intended referent of a novel label. As a first step in examining the role of temporal cross-modal alignment in referent resolution, we tested whether adult listeners can determine the intended referent in a situation with referential ambiguity by relying solely on the temporal cross-modal alignment speakers naturally establish between the imposed referent motion and the produced speech.

Prosodic temporal intermodal alignment

A second aim of this study was to examine the nature of the cross-modal temporal alignment of motion imposed on the referent and the accompanying speech. Previous studies showed that the caregivers synchronized the onset of labeling and the referent's motion (Gogate et al., 2000) and that infants were sensitive to this simple form of intermodal synchronization (Gogate & Bahrick, 1998, 2001; Gogate et al., 2009). We hypothesize, however, that the nature of the temporal alignment is more complex. More specifically, we suggest that the cross-modal temporal alignment of referent motion and labeling is prosodic in nature.

This prosodic cross-modal temporal alignment hypothesis is supported by evidence from studies examining the temporal relationship of speakers' body movement and the prosody of their speech, since a referent object held by a speaker can be seen as an extension of a speaker's body (Hirose, 2002). Body movement and some types of manual co-speech gestures have often been postulated to be linked to the prosodic structure of accompanying speech (Condon, 1976; Dittmann, 1972; Kendon, 1972). The empirical evidence, although often based only on the detailed analyses of a few speakers, suggests that indeed such a relationship exists between the speakers' motion and the prosodic structure of their accompanying speech. The movement of speakers' bodies coincides, for example, with prosodic boundaries, and the extent to which body parts are involved indicates the prosodic hierarchy (Kendon, 1972). Body movement is also linked to the assignment of sentence-level stress (Bull & Connelly, 1985; Hadar, Steiner, & Clifford Rose, 1984; Hadar, Steiner, Grant, & Rose, 1983, 1984; Kendon, 1980; Levelt, Richardson, & La Heij, 1985; but see McClave, 1994), and possibly to the rhythmic hierarchy of the accompanying speech (Condon, 1976; Condon & Ogston, 1966;

Dittmann, 1972; Kendon, 1972). Speakers' manual beat gestures, that is, speakers' simple repetitive gestures that do not convey meaning (Feyereisen, Van de Wiele, & Dubois, 1988; McNeill, 2000), are also linked to the rhythm of speech (Efron, 1941; Ekman & Friesen, 1969; Freedman & Hoffman, 1967). The movement of speakers' heads tends to co-occur with lexical stress placement (Scarborough, Keating, Mattys, Cho, & Alwan, 2009) and seems to convey intonation (Cave, et al., 1996; Yehia, Kuratate, & Vatikiotis-Bateson, 2002). The movement of speakers' bodies and of their body parts seems therefore be related to the prosodic structure of their produced adult-directed speech. It could thus be the case that the motion imposed by the speaker on the referent object in a word learning situation is temporally linked to the prosodic structure of the accompanying speech. That is, the motion imposed on the referent should be linked to stress, rhythm, and intonation of the accompanying speech.

Although body and body part movement seems to be aligned to the prosodic structure of the accompanying speech, only a few studies have shown that listeners are indeed sensitive to this prosodic alignment. Explicit prominence judgments are affected by perceiving visual beat gestures conveyed by eyebrow, head, or hand movement (Bernstein, Eberhardt, & Demorest, 1989; Dohen, Loevenbruck, Cathiard, & Schwartz, 2004; Granstrom & House, 2005; Krahmer & Swerts, 2007; Risberg & Lubker, 1978; Thompson, 1934). Adult listeners thus use these visual prosodic cues when explicitly asked to judge the prosodic structure of speech. Evidence that seeing body movements related to prosody may play a role in speech perception is scarce, however. Seeing head movements seems to improve word recognition in sentences in Japanese, for example (Munhall, Jones, Callan, Kuratate, & Vatikiotis-Bateson, 2004). It is unclear, however, whether head movements may have helped segmentation by signaling the moraic rhythm that helps with the segmentation of Japanese speech (e.g., Cutler & Otake, 1994; Otake, Hatano, Cutler, & Mehler, 1993) or may have simply provided a direct timing signal for segmentation. In summary, these perceptual studies suggest that some perceptible prosodic link may exist in the production of speech and body movement. Whether listeners implicitly use this link in language processing has not been fully established. In the present study, we tested whether a temporal intermodal prosodic link exists between the motion speakers impose on a referent and their speech and whether adult listeners use this cross-modal prosodic temporal link in referent resolution.

In a series of five experiments using a referent detection task, we investigated whether temporal cross-modal alignment of speech and motion can help in referent resolution. On each trial of the referent detection task, adult listeners were asked to indicate which out of two moving objects a speaker was referring to with a novel name. The motion of the referent object was the motion the speaker had naturally imposed on the object during the production of the speech in a prior recording session. It was hence the motion that was naturally aligned to the accompanying speech. The motion of the competing object was not aligned to the accompanying speech as it followed the referent object's motion path reversed in time. The two objects were otherwise visually identical. Listeners had to use the intermodal temporal relationship between the referent and the label in order to identify the intended referent reliably (Experiment 1; see Table 1 for an overview of all experimental manipulations). We ruled out an alternative explanation, namely, that listeners chose the referent object based on motion information alone (Experiment 4). Listeners may possibly perceive a difference in the naturalness of the two objects' motion and select the object following the more natural-looking motion. This was tested by animating the competing object along motion paths recorded originally with other labeling utterances.

To examine the perceived nature of the alignment of speech and motion imposed on the referent we systematically manipulated the linking information available to listeners by modifying the accompanying speech. To test for a link of motion to the prosodic structure of the speech, the speech track was low-pass filtered in Experiment 1 (and in Experiment 3). Low-pass filtering retains mostly prosodic information. The phonetic information retained in low-pass filtered speech is typically not sufficient to recognize words. If recovering the phonetic content is not necessary to use cross-modal alignment, then listeners should still reliably detect the intended referent in this condition. This would furthermore support our prosodic cross-modal temporal alignment hypothesis. If suprasegmental cross-modal alignment is sufficient, then listeners should also perform better than chance in detecting the intended referent when both speech track and referent motion are reversed in time (Experiment 2 and 3). Speech reversed over large intervals is generally unintelligible and has no prosodic structure. Reversing both auditory and video signals together retains any form of temporal simultaneity between pitch, amplitude, rate changes, speech onset/offset and referent motion. If this simple simultaneity between changes in suprasegmental features and referent motion is sufficient to determine the intended referent, then listeners should still perform better than chance in the time-reversed speech condition (Experiment 2 and 3). If the perceived intermodal temporal alignment is more complex, namely reflecting an alignment of referent motion to the prosodic structure, then listeners should perform better in the low-pass filtered than in the time-reversed condition (Experiment 3). Last, we replaced the speech track with a sine-wave tone following the pitch and amplitude of the original speech track (Experiment 5). This tone version thus only contained the original suprasegmental variation of pitch, amplitude, and rate. If listeners perform better than chance in this condition, then this would strongly suggest that the prosodically-mediated variation of suprasegmental features in the speech signal is temporally aligned to the motion imposed on the referent object by the speaker. In summary, this pattern of results would strongly support the prosodic cross-modal temporal alignment hypothesis.

Experiment 1

In Experiment 1 we examined whether there is a perceptible temporal intermodal relationship between the motion speakers impose on an object while labeling it and speech patterns they produce in their referential utterances. In a referent detection task, adult participants listened to speakers teaching the name of a novel object, while watching two moving objects on a screen. One of the objects followed the motion that the speaker had originally imposed on the object during recording. The other object followed the same motion path, but reversed in time. Participants had to indicate which object a speaker was referring to. The linguistic content of the presented speech was not informative about the intended referent. Participants should be able to detect the intended referent, only if there was a perceptible temporal alignment of the motion imposed on the object and the accompanying speech.

The second purpose of Experiment 1 was to test whether this temporal cross-modal relationship is prosodic in nature. Motion imposed on the object may be temporally linked to suprasegmental changes in the speech that indicate prosodic structure, such as changes in amplitude, pitch, and duration. On half of the trials, participants were hence presented with low-pass filtered versions of the speech tracks. Low-pass filtering removes the higher frequency bands in speech, so that the remaining phonetic information is generally not sufficient to recognize words. Prosodic structure is, however, largely retained. The remaining lower frequency bands provide listeners with prosodic information about intonation and phrasal boundaries but less so with lexical prosodic information, such as the lexical stress pattern or word length (Grant & Walden, 1996). If the temporal audiovisual relationship between the object's motion and the variation in the speech signal can be perceived without recovering the linguistic content and be sufficiently driven by information contained in the lower-frequency bands, that is, by prosodic information, then perceivers should still be able to recover the intended referent in the low-pass filtered condition.

Method

Participants.

Twenty-nine native Dutch participants (three men and 26 women) from the Max Planck Institute's participant pool were paid for their participation. Their average age was 21.6 years. All participants were right-handed and reported no hearing or language deficits. All had normal or corrected-to-normal vision.

Materials.

Two phonotactically legal monosyllabic nonwords of Dutch were created to serve as proper names of the novel creatures used in the experiments. These novel names were *Kag* ([kɑx]) and *Zeut* ([zøt]). Eight female native speakers of Dutch were video recorded teaching these two names to two-year-old children shown in a video. The video was presented on a computer screen 50 cm in front of them. All speakers were given the toy creature depicted in Figure 1. The video shown to the speakers consisted of a 20-second-long silent video clips of individual toddlers watching TV inattentively. Speakers were instructed to imagine that they were situated in a distracting environment (e.g., in a noisy daycare center), where they had to attempt to keep the children's attention. This naturally encouraged speakers to use a lively attention-getting voice as well as to move the novel creature they had been given to hold (see Figure 1 for a picture of a typical recording session). Using a silent video of two-year olds rather than live children was essential as it allowed us to obtain recordings in a controlled way and where only the speaker was audible.

Each speaker was recorded three times for 20 seconds teaching the name *Kag*, and three times teaching the name *Zeut*. Speakers were naive in regard to the purpose of the study investigating the link between motion and speech. They were simply informed that their recordings would be used as materials in a word learning study with infants. For this purpose, speakers were asked to attempt to name the object in every sentence. Recordings were not scripted, but speakers were told not to refer to any defining feature of the creature (e.g., their color) or to an action imposed on it (e.g., jumping).

In order to be able to track the motion imposed by speakers on the object, all recordings were conducted with the speakers sitting in front of a black screen wearing a dark sweater and black gloves. Videos recorded a frontal view of the speakers. The collected PAL video recordings were digitized as uncompressed AVI files. Motion paths of the creature were extracted using Adobe After Effect Professional 6.5's Parallel Corner Pin tracking method. The program tracked the motion of one of the creature's eyes and of two white stickers placed above the eyes. The Parallel Corner Pin tracking method regards these tracking points as three of the corners of a parallelogram, for which it estimates a fourth corner point. This method captures when an object is skewed, rotated, and scaled with depth, and preserves the relative distances of the tracking points. That is, any rotation or back-and-forth motion of the object is also reflected in the animations. All obtained motion paths were verified by hand. Figure 1 shows an example of the four tracking points and their motion paths over time.

The obtained motion paths were used to animate a photo of the toy creature against a black background. In these animations, the creature still followed over time the motion originally imposed by the speaker, but the speaker was no longer visible. The display size of these target animations was halved to create videos consisting of a target animation side-by-side with a competitor animation (see Figure 2 for an example frame of a final video). A competitor animation was created for each target animation by animating the photo of the same creature along the target motion trajectories reversed in time. Target and competitor animations were therefore equated in terms of their overall amount of motion during a given trial, but only the target animation followed the original motion path over time. Each target and its assigned competitor animation were arranged side-by-side and exported along with the original soundtrack as one video. Two versions of each video were created where the target animation was shown either on the left or the right side in order to control target position in the experiment. The name of a creature was hence not informative about the identity of the referent or its presentation side. These 96 videos (8 speakers x 6 tokens x 2 sides) served as stimuli in the normal speech condition and were the basis for the construction of the materials in all other conditions in this series of perceptual experiments. For Experiment 1, versions of these videos with speech tracks low-pass filtered at 600 Hz were created in Adobe Audition for the low-pass filtered condition.

Procedure.

Participants were tested individually in a sound-attenuated room. The experiment was run by the NESU software on a PC. Audio was presented diotically at a comfortable listening level over headphones. Participants were alerted that sometimes the speech might sound "altered" ("vervormd"). Participants were instructed to watch the video presented on each trial and to specify by button press at the end of the trial which object the speaker was referring to. The response was indicated by pressing the button on the side that corresponded to the half of the screen the participant thought the target object was shown. Participants had to provide a response in order to continue with the next trial. After a response was given on a trial, participants were also asked to rate their confidence in their response on a scale ranging from one to seven. One end of the scale was labeled "very sure" ("heel zeker"), the other end "not very sure" ("heel niet zeker"). Assignment of labels to endpoints was counterbalanced across participants. No feedback was given.

The experiment consisted of one block containing six videos from each of the recorded eight speakers. Overall, half of the trials were presented to a participant under each speech-type condition (normal or low-pass filtered), such that all of the videos from a recorded speaker were presented under the same speech-type condition. Listeners could therefore not benefit from hearing a speaker in the normal condition for the low-pass filtered condition. Assignment of videos to a speech-type condition was counterbalanced across participants. Across participants, each video was presented equally often with the target at each side. For each participant, within all videos taken from a given speaker, the intended referent was located equally often at each side of the screen. Presentation order was randomized for each participant.

Results

Five participants were excluded from the analyses due to equipment failure during the experiment. Figure 3 shows the mean percentage of correct responses as a function of speech-type condition. One-sample t-tests over subjects (t_1) and items (t_2) compared performance in each speech-type condition to chance (50%). Recognition of the intended referent was significantly better than chance in the normal-speech condition (*M*=75.7%, *SD*=11.88%; t_1(23)=10.19, *p*<.001, *d*=2.08; t_2(47)=8.48, *p*<.001, *d*=1.22) and in the low-pass filtered speech condition (*M*=76.5%, *SD*=11.46%; t_1(23)=11.31, *d*=2.31, *p*<.001; t_2(47)=12.30, *p*<.001, *d*=1.78). That is, in both speech-type conditions, participants were able to correctly detect the intended referent. Two-sample dependent means t-tests over subjects (t_1) and items (t_2) comparing performance across speech- type conditions showed that performance did not vary as a function of speech type (t_1(23)=.54, *p*=.60, *d*=0.11; t_2(47)=.93, *p*=.36, *d*=0.13). Participants were also equally confident in their decisions in these two speech-type conditions (average confidence ratings based on all responses in the normal-speech condition: *M*= 4.12, *SD*=.86; low-pass filtered speech: *M*=4.30, *SD*=.84; (t_1(23)=.94, *p*=.36, *d*=0.2). Whether provided with normal speech or with low-pass filtered speech, participants were equally good at detecting the object the speaker was referring to.

Discussion

Results from Experiment 1 showed that adults were able to correctly detect the intended referent from perceiving a temporal relationship between the motion imposed on the object and acoustic variation in the speech. Thus, while teaching the name of an object, speakers move the object in a way that is temporally linked to their speech. Furthermore, the results from Experiment 1 showed that listeners are still sensitive to this cross-modal temporal relationship when presented with low-pass filtered speech. Temporally-varying acoustic information in the lower frequency bands, that mainly contain prosodic information about intonation and phrasal boundaries (Grant & Walden, 1996), was sufficiently temporally linked to the motion imposed on the target object in order for listeners to detect speaker intent. This suggests that speakers temporally align the motion imposed on a referent object and the dynamics of suprasegmental variation in the accompanying speech. Listeners are sensitive to this cross-modal temporal alignment in referent detection.

Experiment 2

The results obtained in Experiment 1 demonstrated that participants can resolve referent ambiguity from the alignment of motion imposed on the referent object and the accompanying speech. This cross-modal temporal relationship is sufficiently retained when the speech track is low-pass filtered, suggesting that the cross-modal temporal alignment between speech variations and motion is prosodic in nature. One alternative explanation, however, is that listeners are just sensitive to the temporal synchronization of speech onset/offset to motion onset/offset. Speakers may simply move the object while they talk and rest the object while they are not talking. Onset/offset synchronization of speech and motion could hence facilitate referent resolution. In addition, the motion imposed on the object could also be temporally linked to suprasegmental changes in the speech, such as changes in amplitude, pitch, and duration, but not require an analysis of prosodic structure. If that were the case, then listeners should still be sensitive to these cross-modal temporal correlations, even when they are reversed in time.

To test whether listeners are simply sensitive to cross-modal synchrony of onset and offset of speech and motion and/or of suprasegmental changes and motion, a time-reversed speech condition was added to Experiment 2. In this time-reversed condition, the speech tracks of the videos were reversed overall in time. Speech that is reversed in time over longer windows, as is the case here, is no longer comprehensible (Saberi & Perrott, 1999). The animations were not altered. Videos still consisted of one animation following the original motion path over time and one animation following the motion path reversed in time. The time-reversed animation thus became the temporally-aligned target. Any synchrony between onset/offset of motion of the toy creature and the acoustic onset/offset of speech is retained for the reversed speech track and the reversed animation. Reversing speech and motion in time also retains any temporal synchrony between pitch, intensity, and speaking rate changes and the (reversed) motion of the object. Critically, this rendered the cross-modal temporal synchrony prosodically nonsensical. For example, prosodic cues to upcoming phrasal boundaries would now follow such boundaries. If, for example, the referent object was raised along with a rise in pitch towards the end of a phrase to indicate a question, then in the time-reversed condition, the object would be lowered along with a drop in pitch at the beginning of a phrase. That is, the cross-modal link would be preserved but not reflect the natural prosodic structure of the language.

We therefore tested in Experiment 2 whether non-prosodic cross-modal temporal synchrony was sufficient for listeners to infer the referent. If this was the case, then the toy creature following the motion path reversed in time should be perceived as the aligned target. Critically, if this type of synchrony is the only cue participants use, then performance should be the same in the time-reversed condition as in the normal-speech condition.

Method

Participants.

Twenty-four new participants (eight men and 16 women) from the same population as in Experiment 1 were tested (average age: 21.5 years).

Materials.

The same video materials as in the normal speech condition in Experiment 1 were used here. Time-reversed versions were created by reversing the complete speech track of each video in time. Videos still consisted of a time-reversed animation and the original animation. Procedure and design were the same as in Experiment 1. Participants received half of the trials under each speech-type condition (normal or time-reversed), such that all of the videos from a recorded speaker were presented under the same speech-type condition. This assignment was counterbalanced across participants.

Procedure.

The procedure was the same as in Experiment 1.

Results

Figure 3 shows the mean percentage of correct responses as a function of testing condition. One-sample t-tests compared performance under each speech type to chance (50%). Participants were able to detect speaker intent correctly when presented with normal speech (*M*=68.3%, *SD*=17.5%; t_1(23)=5.12, *p*<.001, *d*=1.05; t_2(47)=7.35, *p*<.001, *d*=1.06), but also when speech was reversed in time (*M*=58.6%, *SD*=15.6%; t_1(23)=2.70, *p*<.013, *d*=0.55; t_2(47)=3.16, *p*<.003, *d*=0.46). Two-sample dependent means t-tests showed, however, a significant difference in performance between the two speech-type conditions (t_1(23)=2.59, *p*<.016, *d*=0.53; t_2(47)=2.27, *p*<.028, *d*=0.33). Participants performed better when presented with normal speech than when presented with time-reversed speech. Participants were also more confident in their responses when presented with normal speech (*M*= 3.97, *SD*=.86%) than when presented with time-reversed speech (*M*=3.18, *SD*=1.07; t_1(23)=3.22, *p*<.004, *d*=0.65).

Cross-experiment comparisons showed that the intended referent can be significantly more accurately inferred when presented with low-pass filtered speech than when presented with time-reversed speech (t_1(46)=4.53, *p*<.0001, *d*=4.86; t_2(47)=4.71, *p*<.0001, *d*=0.68). This could suggest that listeners use additional temporal alignment cues in the low-pass filtered speech condition. Performance seems, however, also to be somewhat higher in the normal-speech condition in Experiment 1 than in Experiment 2. This difference is significant in the item and not in the subject analysis (t_1(46)=1.47, *p*=.15, *d*=1.66; t_2(47)=2.42, *p*<.019, *d*=0.36).

Discussion

The results obtained for the normal-speech condition in Experiment 2 replicate those obtained in Experiment 1 showing that adult participants can indeed resolve referential ambiguity by using an audiovisual temporal relationship between the object's motion and variation in speech. Furthermore, results suggest that the alignment between the time-reversed speech and the (time-reversed) motion was sufficient to determine the intended referent. Listeners provided with cross-modal temporal synchronization of speech and object motion that was not prosodically mediated were thus still able to detect the intended referent object. Performance was, however, lower than in conditions where listeners were exposed to prosodically-mediated acoustic variation, such as in the normal-speech condition and in the low-pass filtered speech condition in Experiment 1. This could suggest that although simple cross-modal temporal synchrony is a sufficient cue here to detect the linked referent, prosodic temporal cross-modal alignment provides additional information for the adult listener.

Experiment 3

The results of the first two experiments have shown that listeners were better at detecting the intended referent in the low-pass filtered condition in Experiment 1 than in the time-reversed condition in Experiment 2. Listeners more reliably detected the intended referent in conditions where prosodic information was available, and hence a prosodically-mediated cross-modal temporal alignment could exist. Performance in the normal-speech condition, however, was also lower in Experiment 2 than in Experiment 1. This difference in performance could be due to list effects. Namely, performance in Experiment 2 was lowered in the normal-speech condition through the presence of the more difficult time-reversed trials. It is, for example, feasible that participants used different types of cross-modal alignments in the time-reversed condition than in the normal and low-pass filtered speech conditions. Participants could be using prosodic temporal alignment in the normal and low-pass filtered conditions, but in its absence, as is the case in the time-reversed condition, participants may switch strategies and use simple temporal synchrony. Switching strategies across trials in the mixed list presentations in Experiment 2 could have therefore lowered overall performance in Experiment 2 compared to Experiment 1. To directly compare performance across these conditions, Experiment 3 tested participants in both the low-pass filtered speech condition and in the time-reversed speech condition. The normal-speech condition was also added as a control.

Method

Participants.

Twenty-seven new participants (five men and 22 women) from the same population as in the previous experiments were tested (average age: 20.9 years).

Materials.

The same stimuli materials as in Experiment 1 and 2 were used.

Procedure.

As in the previous two experiments, normal speech was presented on half of the trials and manipulated speech was presented on the other half of the trials. Half of these manipulated speech tracks contained low-pass filtered speech, the other half time-reversed speech. As in the previous experiments, participants were presented with all videos from a respective speaker under the same speech-type condition. Assignment of speaker to speech-type condition was counterbalanced across participants. The rest of the design and procedure was also the same as in Experiment 1 and 2, with the exception that the labels of the confidence scale were now always “not confident at all” at endpoint 1 (“heel niet zeker’”) and “very confident” (”heel zeker”) at endpoint 7.

Results

Data from three participants were excluded from the data analyses due to experimenter errors. Figure 4 shows the average correct detection of speaker intent for the three speech-type conditions. Comparisons to chance show that participants were able to detect the intended referent when presented with normal speech (*M*=66.57%, *SD*=18.75%; $t_1(23)=4.33$, $p<.0001$, $d=0.88$; $t_2(47)=5.65$, $p<.0001$, $d=0.82$) and when presented with low-pass filtered speech (*M*=66.26%, *SD*=19.98%; $t_1(23)=3.99$, $p<.001$, $d=0.81$; $t_2(47)=5.81$, $p<.0001$, $d=0.84$). Participants failed to resolve referential ambiguity when presented with time-reversed speech (*M*=53.19%, *SD*=16.31%; $t_1(23)=.96$, $p=.35$, $d=0.2$; $t_2(47)=1.11$, $p=.27$, $d=0.16$).

One-way ANOVAs with speech type as within-subject and within-item factor indicated a significant effect of speech type ($F_1(2,46)=6.69$, $p<.003$, $\eta_G^2=0.23$; $F_2(2,94)=8.80$, $p<.0001$, $\eta_G^2=0.16$). Planned pair-wise comparisons showed no difference between performance in the low-pass filtered and in the normal speech condition ($t_1(23)=.10$, $p=.93$, $d=0.02$; $t_2(47)=.35$, $p=.73$, $d=0.05$). Performance in the time-reversed speech condition differed, however, from performance in the normal speech condition ($t_1(23)=3.08$, $p<.005$, $d=0.63$; $t_2(47)=3.40$, $p<.001$, $d=0.49$) and in the low-pass filtered condition ($t_1(23)=2.75$, $p<.012$, $d=0.56$; $t_2(47)=3.67$, $p<.001$, $d=0.53$).

Analyses on confidence ratings supported these results. A one-way ANOVA with speech type as within-subject factor showed a significant effect of speech type ($F(2,46)=33.90$, $p<.0001$, $\eta_G^2=0.60$). Confidence ratings for responses in the time-reversed speech condition ($M=2.82$, $SD=1.06$) were lower than in the low-pass filtered speech ($M=3.95$, $SD=1.10$; $t_1(23)=7.00$, $p<.0001$, $d=1.22$) or in the normal speech condition ($M=4.02$, $SD=1.05$; $t_1(23)=5.94$, $p<.0001$, $d=1.43$). Confidence ratings in the low-pass filtered speech condition and the normal speech condition did not differ from one another ($t_1(23)=.56$, $p=.58$, $d=0.12$).

Discussion

The results of Experiment 3 demonstrated once more that there is a perceptible temporal cross-modal relationship between the motion imposed on an object by a speaker and variation in the speaker's speech. Listeners successfully used this temporal cross-modal relationship to determine the intended referent when presented with normal speech but also to the same degree when only presented with the lower-frequency bands of the speech tracks. Lower-frequency bands primarily contain intonational and phrasal prosodic information and make word recognition no longer possible. The retrieval of phonological representations was therefore not necessary to interpret acoustic variation to be linked to the referent object's motion.

Critically, the referent object was not readily detected when only simple audiovisual temporal synchrony was provided. In time-reversed conditions, where only simple and not prosodically-mediated synchrony between acoustic variation (onset/offset of speech, pitch movement, and rate and intensity changes) and the object's motion was retained, listeners failed to detect the intended referent in Experiment 3. Listeners can, however, somewhat use this non-prosodic form of audiovisual synchrony: In Experiment

2, where listeners were given twice as many time-reversed trials as in Experiment 3, a weak but significant effect was found. Listeners detected the correct referent in the time-reversed condition 58.6% of the time. Critically, however, when given the same number of trials in the time-reversed and in the low-pass filtered speech conditions in Experiment 3, participants performed worse in the time-reversed condition than in the low-pass filtered condition, where prosodic structure is retained. This suggests, that although listeners can detect temporal cross-modal relationships that are not mediated by prosody, prosodic temporal cross-modal alignment can be more readily used to resolve referential ambiguity.

Experiment 4

Results from the first three experiments showed that there is a perceptible temporal relationship between speech and the imposed motion on an object. Participants could have, however, relied on a strategy based on visual information alone: The forward-moving target object could have been perceived as moving more naturally than the time-reversed competitor object. Participants could have simply selected the target animation exhibiting the most natural motion as the intended referent. This could also explain why participants performed less well in the task in the time-reversed condition, where the time-reversed object (exhibiting what might be considered a less natural motion trajectory) had to be selected.

In Experiment 4, we tested whether listeners could infer the intended referent when presented with two natural-moving objects, where again only one of the objects is linked temporally to the presented speech. In this *natural-moving competitor condition*, target stimuli from the same speaker were combined to target-competitor pairs. That is, on a given trial, the same speaker had originally produced the motion underlying both animations, but only one animation had been produced with the presented speech track and thus served as the target. For a comparison, we also tested participants on half of the trials (*time-reversed competitor condition*) with competitors following the time-reversed target motion paths, as done in Experiments 1 through 3.

If participants used the cross-modal temporal relationship between speech and the motion imposed on the target object to resolve referential ambiguity, then performance should be above chance level in both the natural-moving competitor and the time-reversed competitor condition. In contrast, if participants in Experiments 1 through 3 were simply choosing the natural-moving object as the referent, then performance in the natural-moving competitor condition should be at chance level because in this case both the target and competitor are animated with a naturally-produced motion.

Method

Participants.

Twenty-four new participants (seven men and 17 women) from the same population as in the previous experiments were tested (average age: 20.8 years).

Materials.

For the time-reversed competitor condition, videos from the normal-speech condition in the previous experiments were used. For the natural-moving competitor condition, each original target animation of a speaker was combined with each of the other target animations from the same speaker. Each video was then once saved with each of its two original audio tracks. Target position was controlled. A total of 480 stimuli were created for the natural-moving competitor condition (15 animation pairs x 2 audio tracks x 2 sides x 8 speakers).

Procedure.

Each participant was presented with two trials from each speaker in each competitor condition. Natural-moving animations used as targets were not presented as competitors to the same participant. This assignment was random but counterbalanced across participants. Each participant thus only received a total of 32 trials here. All other aspects of the design and the procedure were the same as in the previous experiments.

Results

Figure 5 shows the average correct detection of the referent object performance for the two competitor types. Participants were able to detect the correct referent object when the competitor object's motion was reversed in time (*M*=69.55%, *SD*=18.03%; t_1(23)=5.31, *p*<.0001, *d*=1.08; t_2(47)=6.07, *p*<.0001, *d*=0.88) and when the competitor moved naturally (*M*=66.44%, *SD*=13.57%; t_1(23)=5.93, *p*<.0001, *d*=1.21; t_2(47)=6.66, *p*<.0001, *d*=0.96). Performance did not differ between these two competitor conditions (t_1(23)=.70, *p*=.49, *d*=0.14; t_2(47)=.88, *p*=.38, *d*=0.13). Participants were also equally confident in both conditions (time-reversed competitor condition: *M*=3.93, *SD*=.87; natural-moving competitor condition: *M*= 4.01, *SD*=.92; t_1(23)=.74, *p*=.46, *d*=0.16).

Discussion

Experiment 4 showed that participants were not determining the referent by simply selecting the more natural-moving object. Participants were equally able to detect the object linked to the speech when the competing object also moved naturally. The referent object was hence detected based on the audiovisual alignment of the target object's motion and the speech.

Experiment 5

The experiments reported so far suggest that listeners use a prosodic temporal link of motion and speech to detect speaker intent. Participants' performance was unaffected when they were provided with only the lower-frequency bands of the speaker's labeling utterances, suggesting that listeners were relying on the prosodic structure of the utterances to work out the referential intention of speakers' statements. There is, however, a caveat in applying a low-pass filter to child-directed speech, as done here. The prosody of child-directed speech is exaggerated, with higher average pitch and increased pitch variation relative to adult-directed speech (Fernald & Simon, 1984; Jacobson, Boersma, Fields, & Olson, 1983). Therefore, the cut-off level chosen for the low-pass filter in Experiment 1 was somewhat higher than the cutoff normally used for adult-directed speech, as the intention was to preserve the pitch information. It is hence possible that participants may have been able to partially understand some of the speakers labeling utterances. Even though the linguistic content per se was not informative about the correct referent, more than just prosodic information may have been available for use by the listeners. In Experiment 5, we therefore provide a more stringent test of whether listeners use the prosodic cross-modal temporal alignment to resolve referential ambiguity. In a tone condition, the original audio track of each video was replaced by a sine-wave tone following the original pitch track and amplitude. If the motion imposed on the object is temporally linked to changes in pitch and amplitude, then participants should still succeed when tested in this condition. As in the previous experiments, performance was compared to that in a normal-speech condition.

Method

Participants.

Twenty-five new participants (six men and 18 women) from the same population as in the previous experiment were tested (average age: 19.42 years).

Materials.

For the normal-speech condition, the same videos as in the previous experiments were used. To create the stimuli for the tone condition, the audio tracks of the normal-speech videos were read into PRAAT (Boersma & Weenink, 2005) to extract their pitch tracks using PRAAT's autocorrelation method. For this method, a measurement interval of .005 seconds was used. The algorithm was run with a Gaussian window of a length of a sixth of the pitch floor. The pitch range was set from 100 Hz to 800 Hz for all but one speaker for whom the range was set from 100 Hz to 700 Hz. The standard parameter values of the algorithm were used as values for silence threshold (.03), voicing threshold (.45), octave costs (.01), octave-jump costs (.35), and voiced/unvoiced cost parameters (.14). Octave jumps were hand-corrected, considering the shape of the harmonics. Microprosody was not altered but smoothed with a smoothing algorithm (bandwidth 10 Hz). A script then created a sine-wave tone track that followed pitch points in frequency and intensity over time. The resulting tone track thus retained the original temporal relationship of pitch and intensity to the object's motion as in the original recorded speech. Final audio tracks were then saved with the original videos.

Procedure

The procedure was the same as in the previous experiments. Participants were presented on half of the trials with stimuli from the tone condition, on the other half with stimuli from the normal-speech condition. All stimuli from the same speaker were presented under the same condition to a participant, but this assignment was counterbalanced across participants.

Results

Data from one participant who failed to understand the task was excluded. Figure 6 shows the results for the remaining participants. Critically, participants were able to correctly infer the referent object when presented with tones (M=65.94, SD=10.78; t_1(23)=7.24, p<.0001, d=1.48; t_2(47)=5.89, p<.0001, d=0.85). Participants were also able to do this when presented with normal speech (M=72.28, SD=13.08; t_1(23)=8.35, p<.0001, d=1.70; t_2(47)=7.40, p<.0001, d=1.07). A paired-sample t-test showed a difference in performance between these two speech-type conditions (t_1(23)=2.47, p<.02, d=0.50; t_2(47)=2.14, p<.04, d=0.31). Participants were better at detecting the intended referent when presented with normal speech than with tones. Participants were also more confident in their decisions when presented with normal speech (M=3.96, SD=.65) than when presented with a tone track (M= 3.20, SD=.79; t_1(23)=6.03, p<.0001, d=1.23).

A cross-experiment comparison of performance to Experiment 1 showed that performance was better in the low-pass filtered condition in Experiment 1 than in the tone condition in Experiment 5 (t_1(46)=3.28, p<.002, d=3.16; t_2(47)=4.48, p<.0001, d=0.65). The performance in the normal-speech condition did not differ across these experiments (t_1(46)=.67, p=.51, d=0.69; t_2(47)=.92, p=.37, d=0.13).

Discussion

Results from Experiment 5 provide strong evidence that the motion imposed on the referent object by the speaker is temporally linked to the prosodic structure of the speech.

The sine-wave tone manipulation in Experiment 5 only retains pitch and amplitude information, and hence only prosodic structure. Participants were able to infer the intended referent and thus provided evidence for a prosodic link between the motion imposed on the referent and the accompanying speech. Performance in the tone condition was, however, lower than in both the normal-speech condition in Experiment 5 and in the low-pass filtered condition in Experiment 1. This suggests that although the prosodic alignment of speech and motion is sufficient to detect speaker intent and hence the correct referent object, listeners are better at resolving referential ambiguity when presented with a wider range of speech frequencies.

General Discussion

In a series of referent detection experiments, we investigated whether the motion speakers impose on a referent object is temporally aligned to their accompanying speech and whether listeners use this cross-modal temporal link to resolve referential ambiguity. Secondly, we examined the perceptual nature of the temporal alignment of speech and motion. In particular, we tested whether this alignment is prosodic in nature.

Our results demonstrate that listeners detect the intermodal temporal relationship between referent motion and acoustic variation in speech and use it to infer a speaker's intended referent. In all five experiments, listeners reliably identified the object the speaker was referring to with a novel name out of two moving objects. Only the referent object followed the motion the speaker had imposed on the object during the production of the speech track and was thus temporally aligned with the accompanying speech. The competing object's motion was not temporally aligned with the speech, as it either followed the target's recorded motion path reversed in time (Experiment 1 through 5) or came from another recording of the same speaker (as in the natural-moving competitor condition in Experiment 4). The linguistic content of the presented speech did not contain any cues to the identity of the referent object and both objects were identical novel toy creatures. When teaching the novel name of an object, speakers thus move the referent in a way that is temporally aligned with acoustic variation in their speech. Listeners are sensitive to this intermodal temporal alignment and use it to resolve referential ambiguity. The natural alignment of referent motion and accompanying speech thus establishes an intersensory link that helps with referent resolution.

To assess the perceived nature of the intermodal alignment of speech and motion imposed on the referent, we systematically manipulated the cross-modal linking information available to listeners. Listeners were able to use the intermodal alignment in referent resolution when presented with low-pass filtered speech (Experiment 1 and 3). The cross-modal temporal relationship between motion and speech was thus sufficiently retained in the lower-frequency bands of speech that mainly contain suprasegmental prosodic information (i.e., changes in pitch, amplitude, and duration) about intonation and phrasal boundaries (Grant & Walden, 1996). We showed that the temporal link between suprasegmental variation and referent motion needs to follow the prosodic structure of the language in order to be reliably beneficial for referent resolution. Listeners failed to detect the intended referent in Experiment 3, when presented with time-reversed speech aligned to a time-reversed target object but not when presented with low-pass filtered speech. The time-reversed condition provides a critical test as it maintains any temporal synchronization of referent motion and suprasegmental variation in speech, but this suprasegmental variation of time-reversed speech no longer matches the prosodic structure of the listeners' native language. Phrase-final pitch raises at the end of questions would become, for example, phrase-initial pitch drops. This does not correspond to a familiar prosodic structure in Dutch. Listeners can glean some referential information from this temporal alignment though, when provided with sufficient exposure. When listeners were provided with twice as many trials in Experiment 2 than in Experiment 3, listeners performed better than chance in the time-reversed condition. This effect was, however, numerically small (*M*=58%, with a chance level of 50%). Critically, this performance was also significantly worse than in the low-pass filtered condition in Experiment 1, with the same number of trials presented.

Temporal cross-modal synchrony can hence be a sufficient cue to referent resolution, even when not mediated by prosody. But prosodic temporal cross-modal alignment can be more readily used to resolve referential ambiguity. Our prosodic cross-modal temporal alignment hypothesis is further supported by the results of Experiment 5 that showed that the audiovisual temporal relationship between speech and motion persists, if the speech track is replaced by a tone following the pitch, amplitude, and rate of the original speech track. Listeners were still able to detect the intended referent in this tone condition. Performance, however, was lower in this tone condition than in the low-pass filtered condition in Experiment 1. Prosodic alignment of the pitch track and motion is hence sufficient to detect speakers' intended referent objects, but information from a wider range of speech frequencies provides additional help. This could suggest that the dynamics of temporal alignment of referent motion to speech are more complex and follow multiple phases across different frequency bands. Time-varying information in the higher frequency bands could be aligned on a different or similar time scale to the referent motion and thus provide additional information to the identity of the referent.

In the present study, we showed that speakers align the motion imposed on a referent object to the prosody of their speech. It provides evidence for the broader hypothesis that body movement is linked to the prosodic structure of speech (Condon, 1976; Dittmann, 1972; Kendon, 1972). Previous work has suggested that caregivers align the onset of labeling with the onset of motion (e.g., Gogate et al., 2000; 2006). The use of this onset synchronization decreases, however, when the child matures (Gogate et al., 2000). This was taken to support the idea that intersensory redundancy established by onset synchronization loses its importance with age while other cues become more important (Bahrick & Lickliter, 2002). One possibility is that caregivers already produce a more complex prosodic cross-modal alignment when teaching novel label-referent relationships to young children. The previously documented onset synchronization could be a consequence of this prosodic alignment or reflect an additional alignment strategy. Another possibility is that the simple onset synchronization is replaced by the more complex, prosodic alignment when the child matures. Here, we showed videos of 2-year olds to our speakers to elicit speech. Our speakers should hence have reduced onset synchronization in teaching novel label-referent relations to children of that age (Gogate et al., 2000). As our results show, our speakers produced a prosodic alignment. The apparent decline in onset synchronization could thus reflect a qualitative change to a more complex, prosodically-driven alignment rather than reflecting a quantitative decline. This increase in complexity in the prosodically-driven alignment could be a consequence of an increase in utterance complexity of caregivers' speech with the development of the child. The alignment that speakers produce for 2-year-old children may be (or become) more like the alignment of co-speech gestures to the prosodic structure of speech in adult communication (e.g., Hadar et al., 1983). The fact that both co-speech gestures as well as speech prosody tend to be exaggerated in child-directed speech (e.g. Brand, Baldwin, & Ashburn, 2002; Brand, Shallcross, & Sabatos, & Massie, 2007) leaves room for the possibility that the intersensory relationship between gestures and prosody become less apparent and more complex, but not obsolete (e.g., Munhall et al., 2004), in adult-directed speech compared to child-directed speech.

Future research needs to determine how referent motion, that is, the movement of the manual gestures operating on the referent object, is precisely linked to the prosodic structure of speech. One possibility is, for example, that the gestural movements imposed on the referent are linked to the informational structure of speech, that is, to sentence-level stress (see e.g., Bull & Connelly, 1985; Hadar et al., 1983; 1984; Kendon, 1980; Levelt et al., 1985; but see McClave, 1994). Gestural emphasis coinciding with uttering the word label could have highlighted the label-referent link. An informal analyses of the recorded materials suggests that during utterances introducing the novel creature (e.g., “Kijk ’s, dit is Kag.”, “Look, this is Kag”), the creature is often moved with an emphasis (expressed often by a turn in motion) during the label. Another possibility seems to be that the motion imposed on the object is linked to intonational changes in the speech. For example, the object seems to be also often raised at the end of a question, seemingly following the pitch raise. For more descriptive utterances (e.g., “Hij is echt heel lief. Hij heeft mooie ogen en een mooie kleur.”; “He is very sweet. He has beautiful eyes and a nice color”), the creature was often moved repeatedly sideways, similar to swinging, seemingly aligned to the rhythm of speech (c.f. Efron, 1941; Ekman & Friesen, 1969;

Freedman & Hoffman, 1967). The motion imposed on the referent thus appears to be linked to the various aspects of prosody, that is, to stress, rhythm, and intonation of the accompanying speech. The exact nature of the prosodic alignment has yet to be formally determined.

The outcome of this study also contributes to the scarce literature showing that listeners use multisensory prosodic cues implicitly in language processing. Previous work has shown that young infants only succeed in a word-object association task when provided with label-referent motion onset synchronization (Gogate & Bahrick, 1998; 2001; Gogate et al., 2009). Here, we showed that cross-modal temporal alignment also helps, at least adults, with establishing the referent when multiple possible referents exist in the visual scene. More precisely, adults more readily used the prosodic relationship between label and referent motion, rather than simple synchronization. One direction for future research might be to see whether children are also sensitive to this type of alignment and use it to infer the speaker's intended referent.

In summary, the results of the present study have shown that the motion speakers impose on a referent object is temporally aligned with the prosodic structure of the speakers' accompanying utterances. Adult listeners are sensitive to this temporal alignment, in particular, to the prosodically-mediated aspects of the alignment. Listeners use the prosodically-mediated cross-modal alignment to establish the link between the novel label and its referent.

References

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior, 30*, 153-187.

Baldwin, D. A. (1991). Infants' contribution to the achievement of joint attention. *Child Development, 63,* 875–890.

Baldwin, D. A., Markman, E. M., Bill, B., Desjardins, R. N., Irwin, J. M., & Tidball, G. (1996). Infants' reliance on a social criterion for establishing word-object relations. *Child Development, 67*, 3135-3153.

Bernal, S., Lidz, J., Millotte, S., & Christophe, A. (2007). Syntax constrains the acquisition of verb meaning. *Language Learning & Development, 3,* 325-341.

Berman, J. M. J., Chambers, C. G., & Graham, S. A. (2010). Preschoolers' appreciation of speaker vocal affect as a cue to referential intent. *Journal of Experimental Child Psychology, 107*, 87-99.

Bernstein, L. E., Eberhardt, S. P., & Demorest, M. E. (1989). Single-channel vibrotactile supplements to visual perception of intonation and stress. *Journal of the Acoustical Society of America, 85*(1), 397-405.

Bloomfield, L. (1935/1976). *Language* (13 ed.). London: Allen & Unwin.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer [Computer program]. Version 5.1.19, retrieved 21 October 2009, from http://www.praat.org.

Brand, R. J., Baldwin, D. A., Ashburn, L. A. (2002). Evidence for "motionese": Modifications in mothers' infant-directed action. *Developmental Science, 5*, 72-83.

Brand, R. J., Shallcross, W. L., Sabatos, M. G., & Massie, K. P. (2007). Fine-grained analysis of motionese: eye gaze, object exchanges, and action units in infant-versus adult-directed action. *Infancy, 11*(2), 203-214.

Brooks, R., & Meltzoff, A. (2005). The development of gaze following and its relation to language. *Developmental Science, 8*(6), 535-543.

Bull, P., & Connelly, G. (1985). Body movement and emphasis in speech. *Journal of Nonverbal Behavior, 9*(3), 169-187.

Cave, C., Guaitella, I., Bertrand, R., Santi, S., Harley, F., & Espesser, R. (1996). *About the relationship between eyebrow movements and F0 variations.* Paper presented at the Speech and Language Processing, Philadelphia, PA, USA.

Clark, E. V. (1990). On the pragmatics of contrast. *Journal of Child Language, 17*, 417-443.

Collis, G. M., & Schafer, H. R. (1975). Synchronization of visual attention in mother-infant pairs. *Journal of Child Psychology and Psychiatry, 16*, 315-320.

Condon, W. S. (1976). An analysis of behavioral organization. *Sign Language Studies 13*, 285-318.

Condon, W. S., & Ogston, M. B. (1966). Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease, 143*(4), 338-347.

Cutler, A., & Otake, T. (1994). Mora or Phoneme? Further Evidence for Language-Specific Listening. *Journal of Memory and Language, 33*(6), 824-844.

De Saussure, F. (1915/1966). *Course in General Linguistics*. New York: McGraw-Hill.

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: a pragmatic account. *Developmental Psychology, 37*(5), 630-641.

Dittmann, A. T. (1972). The body movement-speech rhythm relationship as a cue to speech encoding. In A. W. Siegman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 131-151). Elmsford, NY: Pergamon Press.

Dohen, M., Loevenbruck, H., Cathiard, M.-A., & Schwartz, J.-L. (2004). Visual perception of contrastive focus in reiterant French speech. *Speech Communication, 44*(1-4), 155-172.

Efron, D. (1941). *Gesture and Environment*. New York: Kings Crown Press.

Ekman, P., & Friesen, W. V. (1969). The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica, 1*, 49-98.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology, 20*, 104-113.

Feyereisen, P., Van de Wiele, M., & Dubois, F. (1988). The meaning of gestures: what can be understood without speech? *Cahiers de Psychologie Cognitive, 8*(3-25).

Freedman, N., & Hoffman, S. (1967). Kinetic behavior in altered clinical states: Approach to objective analysis of motor behavior during clinical interviews. *Perceptual and Motor Skills 24*, 527-539.

Gogate, L. J., & Bahrick, L. E. (1998). Intersensory Redundancy Facilitates Learning of Arbitrary Relations between Vowel Sounds and Objects in Seven-Month-Old Infants.

Gogate, L. J., & Bahrick, L. E. (2001). Intersensory Redundancy and 7-Month-Old Infants' Memory for Arbitrary Syllable-Object Relations. *Infancy, 2*, 219-231.

Gogate, L. J., Bahrick, L. E., & Watson, J. D. (2000). A Study of Multimodal Motherese: The Role of Temporal Synchrony between Verbal Labels and Gestures. *Child Development, 71*(4), 878-894.

Gogate, L. J., Bolzani, L. H., & Betancourt, E. A. (2006). Attention to maternal multimodal naming by 6- to 8-month-old infants and learning of word-object relations. *Infancy, 9*(3), 259-288.

Gogate, L. J., Prince, C. G., & Matatyaho, D. J. (2009). Two-month-old infants' sensitivity to changes in arbitrary syllable-object pairings: the role of temporal synchrony. *Journal of Experimental Psychology: Human Perception & Performance, 35*(2), 508-519.

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Young children and adults use lexical principles to learn new nouns. *Developmental Psychology, 28*, 99-108.

Granstrom, B., & House, D. (2005). Audiovisual representation of prosody in expressive speech communication. *Speech Communication, 46*(3-4), 473-484.

Grant, K. W., & Walden, B. E. (1996). Spectral Distribution of Prosodic Information. *Journal of Speech & Hearing Research, 39*, 228-238.

Hadar, U., Steiner, T. J., & Clifford Rose, F. (1984). The relationships between head movement and speech dysfluencies. *Language & Speech, 27*, 333-342.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1983). Head movement correlates of juncture and stress at sentence level. *Language & Speech, 26*(Pt 2), 117-129.

Hadar, U., Steiner, T. J., Grant, E. C., & Rose, F. C. (1984). The timing of shifts of head postures during conversation. *Human Movement Science, 3*, 237-245.

Harris, M., Jones, D., & Grant, J. (1983). The nonverbal context of mothers' speech to infants. *First Language 4*, 21-30.

Herold, D. S., Nygaard, L. C., Chicos, K. A., & Namy, L. L. (2011). The developing role in prosody in novel word interpretation. *Journal of Experimental Child Psychology, 108*, 229-241.

Hirose, N. (2002). An ecological approach to embodiment and cognition. *Cognitive Systems Research, 3*, 289-299.

Hirotani, M., Stets, M., Striano, T., & Friederici, A. D. (2009). Joint attention helps infants learn new words: event-related potential evidence. *Neuroreport, 20*, 600-605.

Hockett, C. F. (1960). The Origin of Speech. *Scientific American, 203*, 88-96.

Hollich, G., Hirsh-Pasek, K., & Golinkoff, R. (2000). Breaking the language barrier: An emergentist coalition model for the origins of word learning. *Monographs of the Society for Research in Child Development, 65*.

Jacobson, J. L., Boersma, D. C., Fields, R. B., Olson, K. L. (1983). Paralinguistic features of adult speech to infants and small children. *Child Development, 54*, 436-442.

Jolly, H., & Plunkett, K. (2008). Inflectional bootstrapping in 2-year-olds. *Language and Speech,* 51, 45-59.

Kendon, A. (1972). Some relationships between body motion and speech. In A. Seigman & B. Pope (Eds.), *Studies in dyadic communication* (pp. 177-216). Elmsford, New York: Pergamon Press.

Kendon, A. (1980). Gesticulation and speech: two aspects of the process of utterance. In M. R. Key (Ed.), *The relation between verbal and nonverbal communication* (pp. 207-227). The Hague: Morton.

Kidd, C., White, K. S. & Aslin, R. N. (2011). Toddlers use speech disfluencies to predict speakers' referential intentions. *Developmental Science, 14*, 925-934.

Krahmer, E., & Swerts, M. (2007). The effects of visual beats on prosodic prominence: acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language, 57*, 396-414.

Levelt, W. J. M., Richardson, G., & La Heij, W. (1985). Pointing and voicing in deictic expressions. *Journal of Memory and Language, 24*, 133-164.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*, 154-173.

Masur, E. F. (1982). Mothers' responses to infants' object-related gestures: influences on lexical development. *Journal of Child Language, 9*, 23-30.

McClave, E. (1994). Gestural Beats: The Rhythm Hypothesis. *Journal of Psycholinguistic Research, 23*(1), 45-66.

McNeill, D. (2000). *Language and gesture: window into thought and action.* Cambridge: Cambridge University Press.

Messer, D. J. (1978). The integration of mothers' referential speech with joint play. *Child Development, 49*(3), 781-787.

Munhall, K. G., Jones, J. A., Callan, D. E., Kuratate, T., & Vatikiotis-Bateson, E. (2004). Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science, 15*(2), 133-136.

Nygaard, L. C., Herold, D. S., & Namy, L. L. (2009). The semantics of prosody: acoustic and perceptual evidence of prosodic correlates to word meaning. *Cognitive Science, 33*, 127-146.

Nygaard, L. C., & Lunders, E. R. (2002). Resolution of lexical ambiguity by emotional tone of voice. *Memory & Cognition, 30*(4), 583-593.

Nygaard, L. C., & Queen, J. S. (2008). Communicating emotion: linking affective prosody and word meaning. *Journal of Experimental Psychology: Human Perception & Performance, 34*(4), 1017-1030.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable? Speech Segmentation in Japanese. *Journal of Memory and Language, 32*(2), 258-278.

Parault, S. J., & Schwanenflugel, P. J. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research, 35*, 329-351.

Pruden, S. M., Hirsh-Pasek, K., Golinkoff, R., & Hennon, E. A. (2006). The birth of words: Ten-month-olds learn words through perceptual salience. *Child Development, 77*, 266-280.

Risberg, A., & Lubker, J. (1978). Prosody and speech-reading. *Speech Transmission Laboratory Quarterly Progress Status Report, 4*, 1-16.

Saberi, K., & Perrott, D. R. (1999). Cognitive restoration of reversed speech. *Nature, 398*(6730), 760.

Scarborough, R. A., Keating, P. A., Mattys, S. L., Cho, T., & Alwan, A. (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language & Speech, 52*(2/3), 135-175.

Shintel, H., Nusbaum, H. C., & Okrent, A. (2006). Analog acoustic expression in speech communication. *Journal of Memory and Language, 55*, 167-177.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106*, 1558-1568.

Thompson, D. M. (1934). On the detection of emphasis in spoken sentences by means of visual, tactual, and visual-tactual cues. *Journal of General Psychology, 11*, 160-172.

Yehia, H. C., Kuratate, T., & Vatikiotis-Bateson, E. (2002). Linking Facial Animation, Head Motion and Speech Acoustics. *Journal of Phonetics, 30*(3), 555-568.

Table 1

Design Overview

Experiment	Manipulated Speech Conditions	Target Motion	Distractor Motion
1	Low-pass filtered	Natural	Time-reversed
2	Time-reversed	Time-reversed	Natural
3	Low-pass filtered	Natural	Time-reversed
	Time-reversed	Time-reversed	Natural
4	--- [1]	Natural	Time-reversed
		Natural	Natural, taken from other stimuli
5	Tone version	Natural	Time-reversed

[1] In Experiment 4, only normal speech was presented. A normal speech condition was also added to all other experiments.

Figure captions

Figure 1. Example of a typical recording session. Dots show the motion paths of three tracking points over time. A fourth tracking point is inferred by the software to form a parallelogram.

Figure 2. Screenshot of a typical video frame in the test trials of the experiments.

Figure 3. Average percentage correct performance in Experiment 1 and Experiment 2 as a function of speech-type condition. The dashed line indicates the 50% chance level. Error bars represent the 95% confidence intervals around the means.

Figure 4. Average percentage correct performance in Experiment 3 as a function of speech-type condition. The dashed line indicates the 50% chance level. Error bars represent the 95% confidence intervals around the means.

Figure 5. Average percentage correct performance in Experiment 4 as a function of competitor type. The dashed line indicates the 50% chance level. Error bars represent the 95% confidence intervals around the means.

Figure 6. Average percentage correct performance in Experiment 5 as a function of speech-type condition. The dashed line indicates the 50% chance level. Error bars represent the 95% confidence intervals around the means.

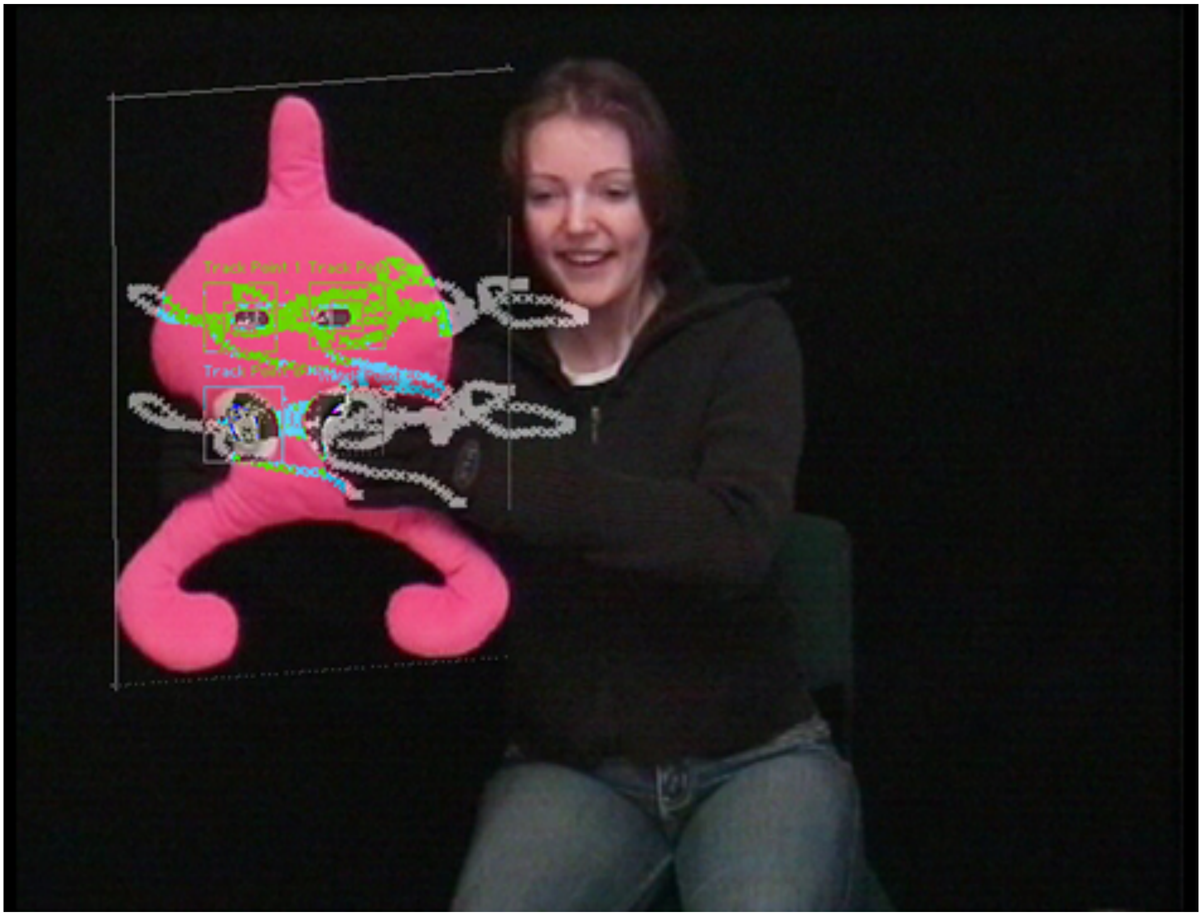

Figure 1.

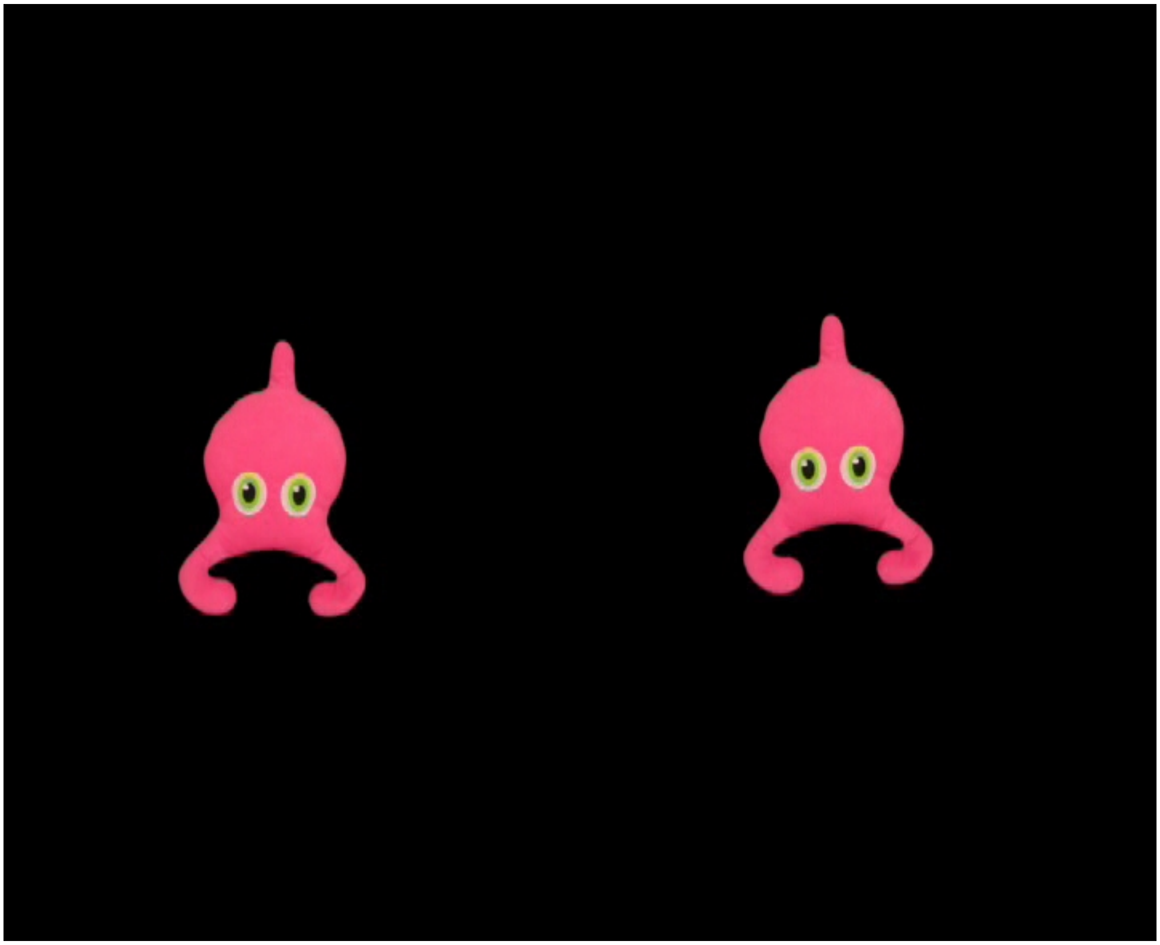

Figure 2.

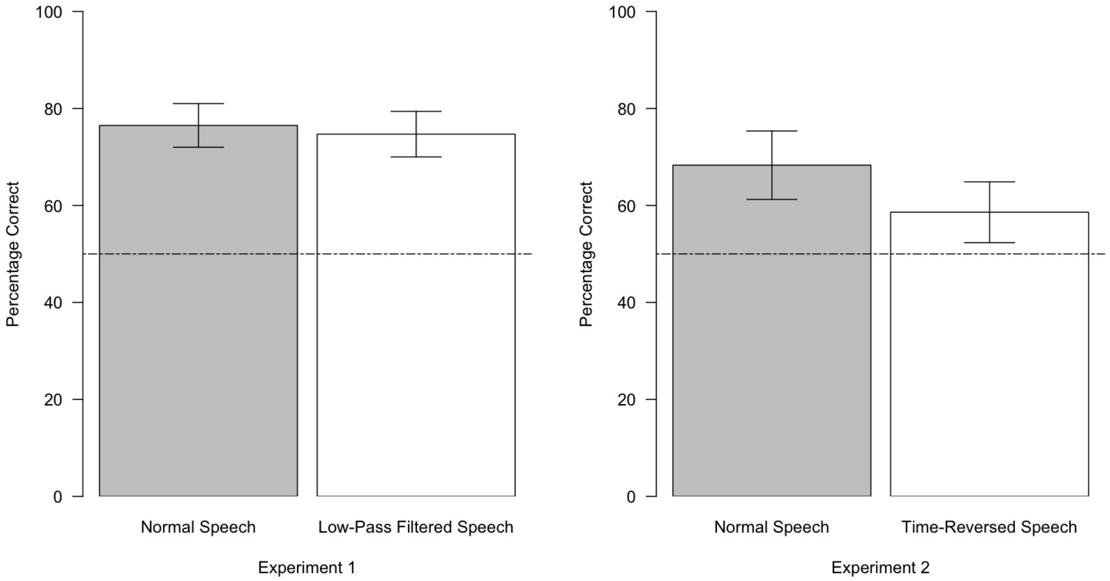
100
80
60
40
20
0
Percentage Correct
Normal Speech
Low-Pass Filtered Speech
Experiment 1
100
80
60
40
20
0
Percentage Correct
Normal Speech
Time-Reversed Speech
Experiment 2

Figure 3.

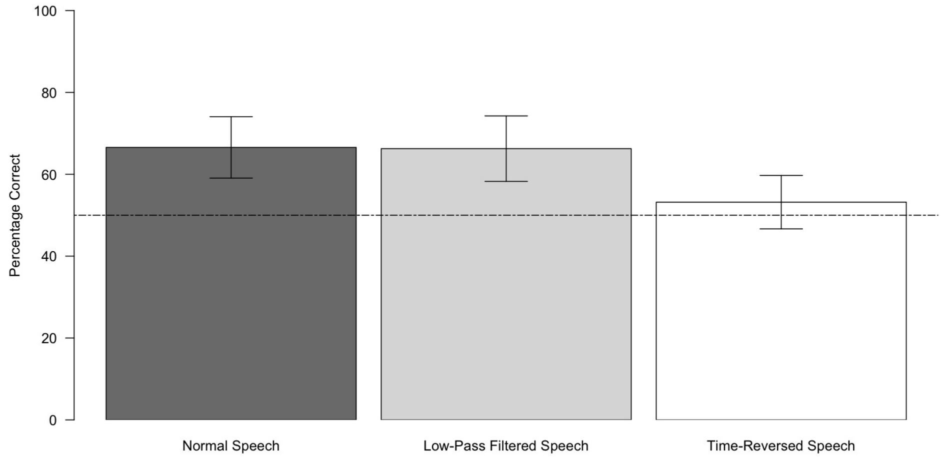
100
80
60
40
20
0
Percentage Correct
Normal Speech
Low-Pass Filtered Speech
Time-Reversed Speech

Figure 4.

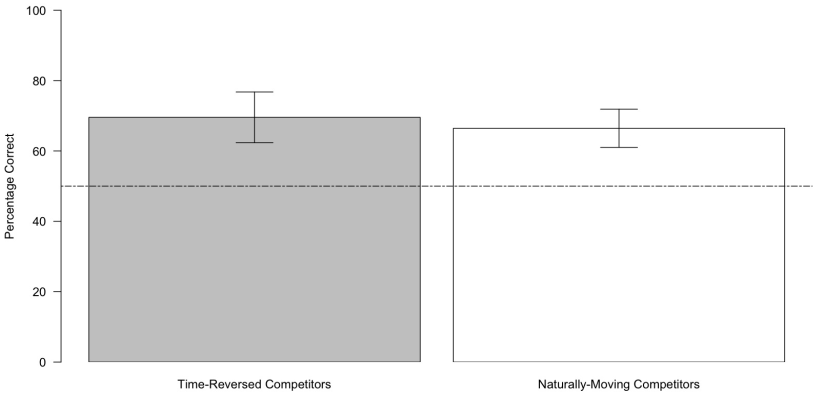
100
80
60
40
20
0
Percentage Correct
Time-Reversed Competitors
Naturally-Moving Competitors

Figure 5.

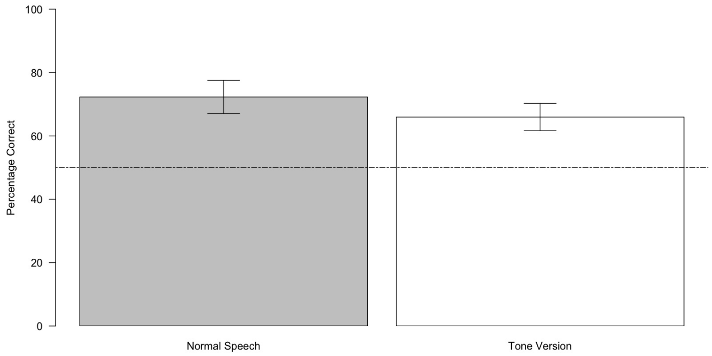
100
80
60
40
20
0
Percentage Correct
Normal Speech
Tone Version

Figure 6.