Journal of Experimental Psychology:
Human Perception and Performance

0096-1523/10/$12.00 DOI: 10.1037/a0019952

Audiovisual Asynchrony Detection in Human Speech

Joost X. Maier, Massimiliano Di Luca, and Uta Noppeney
Max Planck Institute for Biological Cybernetics

Combining information from the visual and auditory senses can greatly enhance intelligibility of natural speech. Integration of audiovisual speech signals is robust even when temporal offsets are present between the component signals. In the present study, we characterized the temporal integration window for speech and nonspeech stimuli with similar spectrotemporal structure to investigate to what extent humans have adapted to the specific characteristics of natural audiovisual speech. We manipulated spectrotemporal structure of the auditory signal, stimulus length, and task context. Results indicate that the temporal integration window is narrower and more asymmetric for speech than for nonspeech signals. When perceiving audiovisual speech, subjects tolerate visual leading asynchronies, but are nevertheless very sensitive to auditory leading asynchronies that are less likely to occur in natural speech. Thus, speech perception may be fine-tuned to the natural statistics of audiovisual speech, where facial movements always occur before acoustic speech articulation.

Keywords: speech, multisensory integration, synchrony judgment, temporal order judgment, spectral rotation

Human speech conveys meaning through both auditory (voice) and visual (facial movement) information. Combining these sources of information can enhance and even restore the intelligibility of speech in many, especially noisy, listening situations (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954; Summerfield, 1992). One important factor determining the integration of signals is their relative timing. Even though multisensory signals do not have to be precisely physically synchronous, they have to co-occur within a certain temporal 'window of integration' to be integrated (Stein & Meredith, 1993). For simple, transient audiovisual signals, this temporal window of integration is in the order of tens of milliseconds (Hirsh & Sherrick, 1961; Stone et al., 2001; Zampini, Guest, Shore, & Spence, 2005; Zampini, Shore, & Spence, 2003). In contrast, for speech signals this window has been shown to be much wider (in the order of several hundred milliseconds). For instance, Dixon and Spitz (Dixon & Spitz, 1980) showed that audiovisual asynchrony was only detected when the visual speech signal leads the auditory speech signal by at least 250 ms or follows the auditory signal by 130 ms. Similarly, intelligibility of audiovisual speech has been shown to be tolerant to large temporal offsets between the component signals (Campbell & Dodd, 1980; Jones & Jarick, 2006; Massaro & Cohen, 1993; Massaro, Cohen, & Smeele, 1996; Munhall, Gribble, Sacco, & Ward, 1996; Pandey, Kunov, & Abel, 1986; van Wassenhove, Grant, & Poeppel, 2007).

This raises the question whether audiovisual integration of speech is 'special' or relies on generic mechanisms (Tuomainen, Andersen, Tiippana, & Sams, 2005; Vatakis, Ghazanfar, & Spence, 2008). Previous studies have tried to investigate why and how audiovisual integration may be different for speech and nonspeech signals (Conrey & Pisoni, 2006; Dixon & Spitz, 1980; Vatakis & Spence, 2006a, 2006b) by comparing the temporal integration window of speech signals with multiple classes of nonspeech stimuli such as simple light/tone pairings (Conrey & Pisoni, 2006), musical instruments (Vatakis & Spence 2006a, 2006b), object actions (Dixon & Spitz, 1980; Vatakis & Spence, 2006b) and reversed speech (Vatakis & Spence, 2006b). Given the considerable variability in spectrotemporal structure and familiarity of the stimulus materials used, it is not surprising that these studies have provided quite inconsistent results: while some studies have demonstrated clear differences between speech and nonspeech signals (Dixon & Spitz, 1980; Vatakis & Spence, 2006a, 2006b), others have observed no differences (Conrey & Pisoni, 2006; Vatakis & Spence, 2006b). This variability is further increased as speech materials of different lengths were used, including syllables/bi-syllables (Jones & Jarick, 2006; Vatakis & Spence, 2006a, 2006b [Experiment 2]), single words (Conrey & Pisoni, 2006), and complete sentences (Vatakis & Spence, 2006b [Experiment 1]). This difference in stimulus types can alter the amount of audiovisual correlation, segmental (e.g. manner of articulation), and suprasegmental (e.g. syllabification) cues for synchrony perception.

This study more systematically investigates the factors that render temporal perception of audiovisual speech different from nonspeech signals using three approaches.

First, speech may be different in the way that the incoming complex time-varying acoustic and visual signals can be mapped

Joost X. Maier, Massimiliano Di Luca, and Uta Noppeney, Cognitive Neuroimaging Group, Max Planck Institute for Biological Cybernetics.

Joost X. Maier is now with the Department of Psychology at Brandeis University.

Special thanks go to Karin Pilz for help with stimulus generation, and Mario Kleiner for help with stimulus editing. We thank Sebastian Werner for help with testing the audiovisual misalignment of the camera, and Asif Ghazanfar for useful comments on an earlier version of this manuscript. This work was supported by the Max Planck Society (J. X. M., M. D. L., & U. N.), the DFG (U. N.), and EU Grant 27141 ImmerSence (M. D. L.).

Corresponding concerning this article should be addressed to Joost X. Maier, Department of Psychology, Brandeis University, 415 South Street, MS 062, Waltham, MA 02453. E-mail: joost.maier@gmail.com

onto linguistic (e.g. phonological, semantic) representations. In this study, we investigated whether the availability of these linguistic representations affects the temporal integration window. To this end, we compared speech to spectrally inverted (rotated) and temporally reversed speech. Spectral (Scott, Blank, Rosen, & Wise, 2000; Warren et al., 2006) and temporal (Bedard & Belin, 2004) inversion are two complementary methods that are classically used in auditory speech perception to render the original speech signal unintelligible, while retaining the spectrotemporal complexity of the auditory signal. *Rotated speech* is rendered unintelligible by inverting the spectrum of the original auditory speech signal (see Figure 1). Spectral rotation preserves the spectral complexity and to some degree the temporal envelope of the original speech stimuli, but changes the frequency distribution. This changes the fine-tuned natural mapping between different frequency bands and the visual stimulus and hence slightly decreases the temporal audiovisual correlations. Therefore, we used *reversed speech* as an additional control stimulus. Temporal reversion reverts the time-courses of both the original auditory and visual speech signals and hence completely preserves temporal audiovisual correlations. Both spectral and temporal inversion may either widen or narrow the temporal integration window. Within the framework of the *unity assumption* (Welch & Warren, 1980; Vatakis & Spence, 2007), the availability of linguistic representations in natural connected speech should emphasize the common source of the auditory speech signal and the associated facial movements. This should facilitate binding of the auditory and visual signals and lead to a wider integration window for intelligible relative to nonintelligible speech. In line with this hypothesis, incongruent audiovisual speech stimuli have previously been associated with a more narrow temporal integration window compared to natural congruent audiovisual pairings (van Wassenhove, Grant, & Poeppel, 2007). Conversely, one may argue that humans have been fine-tuned to natural speech statistics through sustained exposure during their lifespan. Under this hypothesis, even small changes to the speech statistics that are induced by spectral and temporal inversion will render fine-tuned synchrony detection less effective, leading to a widened temporal integration window for nonintelligible relative to intelligible speech.

Second, audiovisual integration of speech may be different in the way that auditory speech signals are spectrally and temporally complex signals that evolve over time. For example, the temporal relationship between auditory and visual speech signals can vary considerably across different utterances and even realizations of the same utterance (e.g. in terms of voice onset time, but also segmental and supresegmental features within a sentence). To evaluate the importance of the continuous nature of speech for synchrony detection, we systematically manipulated the amount of temporal information provided by the speech signal. More specifically, we compared complete sentences with fragments that were reduced to the sentence on- and offsets. Again, this manipulation may either widen or narrow the temporal integration window. Having access to the entire time course of the sentence may lead to increased binding of the auditory and visual signals, and a wider temporal integration window compared to on/offset sentences. Alternatively, availability of temporal audiovisual correlations throughout the entire sentence may enable humans to make more precise and reliable synchrony judgments, leading to a more narrow integration window relative to on/offset sentences.

Third, it is important to note that the audiovisual temporal integration window will be estimated through subjects' decisions in the context of a specific task, rendering the results dependent on attentional and strategic effects (van Eijk, Kohlrasch, Juola, & van der Par, 2008). To gain some insight into the role of contextual effects, we compared temporal perception in two commonly used tasks: synchrony judgment (SJ) (Conrey & Pisoni, 2006; Jones & Jarick, 2006) and temporal order judgment (TOJ) (Vatakis & Spence, 2006a, 2006b). These tasks have previously been shown to

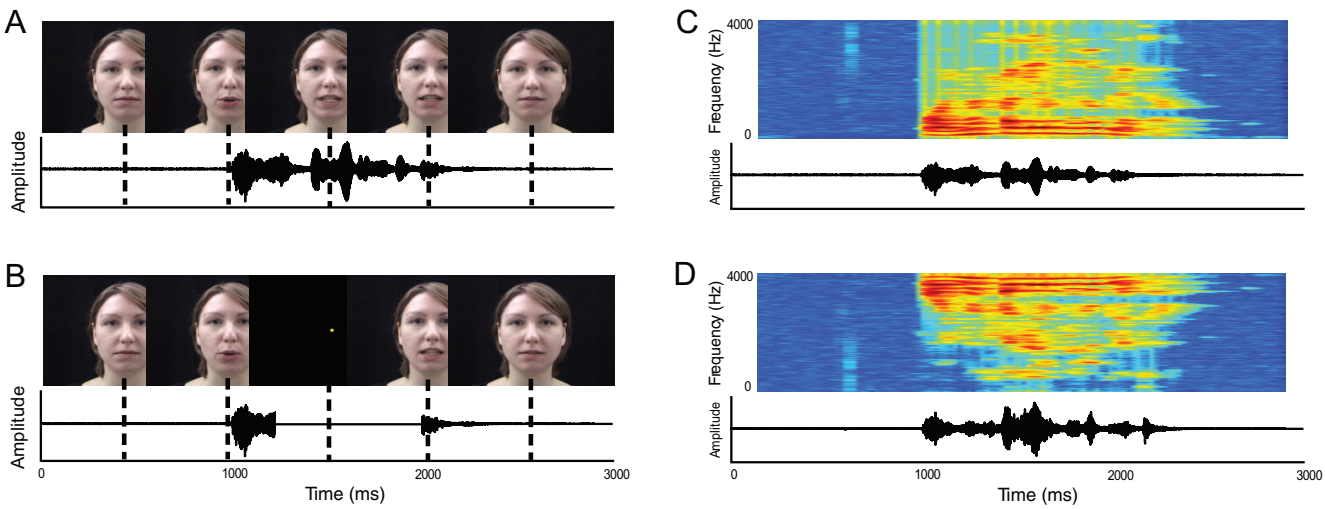

Figure 1. Visual and auditory components of one example stimulus used in this study. (A, B) Bottom panels show time-amplitude waveforms of the auditory component in the complete sentence (A) and on/offset only (B) conditions. Dotted lines connect to representative, temporally corresponding video frames in the top panels. (C, D) Time-amplitude (bottom panels) and time-frequency (top panels) representations of the auditory component in the normal (C) and rotated (D) speech conditions.

influence subjects' performance in different ways (Soto-Faraco & Alsius, 2007; Vatakis, Navarra, Soto-Faraco, & Spence, 2008).

In summary, subjects took part in two experiments in which they judged either the synchrony or the temporal order of spectrally complex audiovisual stimuli. In both experiments we factorially manipulated the audiovisual asynchrony, the spectrotemporal structure of the stimuli (normal speech vs. spectrally inverted speech vs. temporally reversed speech), and the stimulus length (complete sentence vs. onset/offset sentence).

Method

Subjects

Ten subjects (5 female, 5 male; between 18 and 35 years old, mean: 22.5) gave informed consent to participate in the experiments. All subjects reported normal or corrected-to-normal vision, normal hearing, and were German native speakers. Each subject was completely naïve to the hypotheses and goals of the experiment and was paid for participation. The study was approved of by the research review committee of the Max Planck Society.

Stimuli

Stimulus material was taken from close up video recordings of a female face looking straight into the camera, uttering short sentences (see Appendix A). Sentences were 5-word long neutral statements in German. Audio and video was recorded with a Sony Handycam Vision DCR-TRV900E digital video camera (www.sony.net). Video was acquired at 30 frames per second (720 × 480 pixels); audio was acquired at 44 kHz, 16 bit resolution in stereo. Clips were cropped (Adobe Premiere Pro, www.adobe.com) to one single complete sentence, preceded and followed by 15 frames of neutral facial expression during which no sound was presented (total duration, between 2,767 and 3,433 ms, mean ± *SD* = 3,057 ± 185 ms). To remove background noise, audio tracks were then separated from the video tracks and edited (Adobe Audition). The subsequent processing steps were performed in Matlab (www.mathworks.com). Asynchronous stimuli were created by removing neutral frames/silent audio from the on- and offset of the video/audio tracks. Audiovisual asynchronies included 0 (synchronous), 67, 133, 200, 267, and 333 ms in both directions (auditory and visual leading). However, additional analyses have revealed that the camera and the presentation software did not precisely align the visual and auditory tracks of the movie.

Artifactual asynchronies can be created by different latencies in the recording and/or presentation of stimuli (e.g., see Vatakis & Spence, 2006c). If the latency is equal for audio and visual streams, physically synchronous events are reproduced as synchronous stimuli in the experiment. Conversely, if the latency is unequal for audio and visual streams, physically synchronous events are reproduced as asynchronous stimuli in the experiment. To estimate any artifactual differences between the auditory and visual signals, we employed the following procedure. First, we constructed an electronic circuit that produced simultaneous auditory and visual signals. Using an NE556 chip we created two Schmitt-trigger a-stable oscillators: one oscillator provided a 1 Hz square-wave signal that was used to gate the second 1 kHz oscillator. The resulting signal was an interrupted 1 kHz carrier that was connected to an LED and a piezoelectric speaker, and should thus produce simultaneous flashes and beeps with a duration of 500 ms and an interstimulus interval of 500 ms. To ensure that flashes and beeps were indeed synchronized, they were recorded via a photodiode (OSRAM BPW21, rise- and fall-time = 1.5 μs) and a microphone that were directly connected to two identical preamplifiers and two synchronous A/D converters (the stereo line-in channels of a computer sound card). Indeed, by placing the photodiode in front of the LED and the microphone in front of the speaker the recordings confirmed that there was no detectable asynchrony between flashes and beeps produced by the circuit. Second, these simultaneous flash and beep signals were then used to measure the total latency differences that may be induced by the video camera, the acquisition and manipulation software, and/or the reproduction in our experimental setup. To this end, we filmed the flashes and beeps using the same video camera and acquisition software, and displayed the video using the same experimental setup as in our main experiment. The flashes and beeps were recorded by placing the photodiode in front of the screen and the microphone in front of the headphones. Thus, the asynchrony measured between the video of the flashes and beeps captured the total sum of latency differences between the auditory and visual channels that may have arisen during stimulus recording, manipulation, and reproduction.

Across multiple measurements, the present setup misaligned the auditory and visual tracks with an auditory lead (mean ± *SD* = 60 ± 23 ms; maximum = 127 ms; minimum = 32 ms). Because of the variability of the measured delay, the levels of asynchrony in the present study are labeled as if no artifactual asynchronies were present. This means that movies labeled having 0 ms lag refer to stimuli that are modified by an additional asynchrony with an auditory lead of approximately 60 ms. Conversely, movies labeled having a visual lead of 67 ms refer to conditions where stimuli are much closer to veridical timing.

Experimental Design

The experiment conformed to a 2 × 2 × 2 factorial design with the within-subjects factors Speech (normal vs. rotated), Sentence (complete sentence vs. onset/offset), and Task (synchrony judgment vs. temporal order judgment).

Speech (Normal vs. Rotated Speech)

To generate auditory stimuli that are comparable to ordinary speech in terms of temporal and spectral complexity, we transformed the original speech into unintelligible speech-like stimuli using spectral rotation and temporal inversion. To this end, speech stimuli were first low-pass filtered at 4 kHz. This initial low-pass filtering is required to effectively apply spectral rotation (around 2 kHz) to speech stimuli. Spectral rotation inverts the frequency spectrum of the low-passed filtered speech signal (see Figure 1). Low-pass filtered speech is completely intelligible and does not differ qualitatively from natural speech. Therefore, we used low-pass filtered speech in our normal speech condition and will refer to it as normal speech. Importantly, spectral rotation makes the speech sound unintelligible, but largely preserves temporal structure and spectral complexity and certain phonetic features (e.g. the distinction between voiced and unvoiced sounds). Normal and

spectrally rotated audio tracks were equated with respect to RMS energy. The method of spectral rotation has been extensively described elsewhere (Blesser, 1972), and has previously been used as a control for speech sounds (Scott et al., 2000; Warren et al., 2006). As prior presentation of the original sentences may prime the subjects and render the corresponding rotated sentences intelligible, two sets of sentences were used: one for the normal, and one for the spectrally rotated condition. Each set of sentences encompassed 5 different sentences. Within the normal and spectrally rotated condition, the same sentences were used for all asynchrony levels. Across subjects, the sentence items were counterbalanced across the normal and spectrally rotated conditions (see Appendix A).

Sentence (Complete Sentence vs. Onset/Offset)

To investigate whether subjects use information available within the continuous speech stream for their synchrony and temporal order judgments, we presented complete sentences and sentence fragments (i.e., sentences that were reduced to their on- and offsets). On/offset fragments were created by removing audio and video between 12 frames after audio onset and 12 frames before audio offset, and replacing the video in that period with a yellow fixation spot (see Figure 1).

Task (Synchrony Judgment vs. Temporal Order Judgment)

All subjects performed two tasks: a synchrony judgment (SJ) task and a temporal order judgment (TOJ) task. In the synchrony detection task, subjects judged whether the audio and video track were synchronous or asynchronous. In the TOJ task, subjects judged whether the audio track or the video was leading.

Each of the 5 sentences was presented 5 or 6 times at each asynchrony level under each of the 4 conditions (normal vs. rotated; and complete sentence vs. onset/offset) for each task. The two tasks were performed in different sessions (separated by 2 to 7 days) and the order in which the tasks were performed was counterbalanced across subjects. All factors (5 sentences, 4 conditions, 11 asynchronies, and 5 or 6 repetitions) were randomly interleaved within a session to control for learning effects and response biases.

Procedure

Subjects were seated in a darkened room (background noise: ~35 dB SPL, A-weighted) at 50 cm distance from a choice response time (CRT) monitor used to present visual stimuli. Auditory stimuli were presented through headphones (binaurally). The audiovisual stimuli were presented with an intertrial interval of 2 s. Subjects silently viewed the video and listened to the audio track. They were told that the video and the sound track (even if unintelligible) emanated from the same event, but could be temporally offset by various amounts. They were instructed to respond at the end of each sentence. Subjects indicated their response as accurately as possible through a two choice key-press. Visual feedback with regard to timing of the response (the words "Too late" or "Too early") was provided if subjects responded either during the sentence or more than 2 s after offset of the video. No feedback was provided with respect to the correctness of the response. Early and late responses were excluded from the analysis.

Data Analysis

Based on visual inspection, the data from the SJ and TOJ judgments did not conform to Gaussian or cumulative Gaussian distributions. In particular, the distribution of responses was asymmetric. Hence, fitting Gaussian and cumulative Gaussian psychometric functions to the data may bias the estimate of the peak away from the mean of the distribution and towards the median (Sternberg & Knoll, 1973). To accommodate the problems induced by deviation from normality, we provide two types of analysis: First, our main approach is nonparametric and refrains from making any distributional assumptions. Instead, we derive four assumption-free indices (peak performance and location, width, and asymmetry) for each subject and enter them into a two-stage summary statistic (random effects analysis). Second, for comparison with results previously reported in the literature on perceived simultaneity of auditory-visual speech stimuli, we also fitted the data from individual subjects using Gaussian and cumulative Gaussian functions.

Nonparametric Analysis

For each subject, we computed the proportion of synchronous responses (PSR) for SJ and the proportion of correct responses (PCR) for TOJs at each asynchrony level, ranging from −333 ms (auditory leading) to 333 ms (visual leading). For TOJ, we used s PCR. Given the artifactual misalignment of the auditory and visual signals, nominally synchronous stimuli (i.e., 0 lag in the figures) were in fact auditory leading with a lag of 60 ms (see Method). Therefore, subjects' "auditory leading" responses for nominally synchronous trials were counted as correct responses. Because the mean misalignment of 60 ms was smaller than the asynchrony levels of 67 ms, assignment of correctness of the responses at the other asynchrony levels was not affected by the artifactual misalignment. In other words, at all other asynchrony levels, physically and nominally auditory (or visually) leading were in agreement.

For TOJ, we used PCR rather than the more frequently used proportion of "visual leading" responses to render the shapes of the psychometric functions comparable across the two tasks. In this way, the analysis performed on the data is comparable across the TOJ and SJ tasks.

Moreover, to refrain from making any distributional assumptions about the data, we did not fit a parametric psychometric function (e.g., cumulative Gaussian) to the responses. Instead, we characterized the psychometric function by four indices that were computed as follows:

1. *Peak performance*, the extreme point of the psychometric function

(SJ) peak performance = max(PSR)

(TOJ) peak performance = min(PCR)

2. *Peak location*, the level of audiovisual asynchrony associated with the extreme point of the psychometric function, which corresponds to the point of subjective simultaneity. Positive values indicate that simultaneity is perceived with stimuli where auditory signals were delayed.

Using the values of peak performance obtained for each subject, we then calculated the adjusted proportion of responses (AP) for the two experiments as follows:

$$\text{(SJ) AP} = 1 - \max(\text{PSR}) + \text{PSR}$$

$$\text{(TOJ) AP} = 1 + \min(\text{PCR}) - \text{PCR}$$

We then subdivided the values at each level of audiovisual asynchrony in AP + and AP − where

AP + are the values of AP at asynchrony < peak location

AP − are the values of AP at asynchrony > peak location

We trimmed AP + or AP− so that we would have the same number of elements for each subject. From these we computed:

3. *Width,* determined by adding the mean proportion correct (or synchronous) responses on the right side of the peak and the mean proportion correct (or synchronous) responses on the left side of the peak. This value increases with the width of the psychometric function of responses is around its peak.

$$\text{width} = \text{mean(AP+)} + \text{mean(AP−)}$$

4. *Asymmetry*, determined by subtracting the mean proportion correct (or synchronous) responses on the right side of the peak from the mean proportion correct (or synchronous) responses on the left side of the peak. This value indicates the direction in which the psychometric function is skewed. Positive values indicate that the proportion of correct (or synchronous) responses is greater on the right of the peak location, that is, towards visual leading asynchrony levels:

$$\text{asymmetry} = \text{mean(AP−)} - \text{mean(AP+)}$$

Peak performance, peak location, width, and asymmetry were entered into separate two-way repeated measures ANOVAs with factors Speech (normal vs. spectrally rotated) and Sentence (complete vs. on/offset). To evaluate task effects, we also performed a three-way repeated measures ANOVA with Task (SJ vs. TOJ) as an additional factor.

Parametric Analysis

Despite its limitations and biases, we also performed a parametric analysis to allow for comparison with results previously reported in the literature.

Synchrony judgments were fitted using a Gaussian (least squares method) to the proportion of "synchronous" responses, and we derived peak location and width from the fitted functions. Peak location was defined as the mean, and width was defined as the standard deviation of the Gaussian. For TOJs, we did not use proportion correct responses, as in our nonparametric analysis, but proportion of "visual first" responses, for comparison with the literature. We fitted a cumulative Gaussian (maximum likelihood method) to the proportion of "visual first" responses, and derived the just-noticeable difference (JND) and point of subjective simultaneity (PSS) from the fitted functions. PSS was defined as the value of auditory lag at which the proportion of "auditory first" and "visual first" responses was equal. JND was defined as half the interval between the value of auditory lag corresponding to the proportion of 0.25 and 0.75 "auditory first" responses.

Results

One subject was excluded from the analysis because her peak performance was less than 70% correct in at least one of the conditions on both the synchrony and TOJ tasks. Hence, the reported results are based on the remaining nine subjects.

In brief, subjects were presented with the audio tracks and the video clips of a female speaker uttering short sentences in their native language. Audio and video tracks of the videos were presented with different audiovisual asynchronies ranging from −333 ms (audio leading) to +333 ms (video leading).

Nonparametric Analysis

In the following, we will report the results of repeated measures ANOVAs with factors Speech (normal vs. spectrally rotated speech signals) and Sentence (complete vs. on- and offset fragments of audiovisual sentences). Main effects and interactions are reported for each of the four nonparametric indices: (1) peak performance; (2) peak location; (3) width; and (4) asymmetry. Results are reported separately for the SJ and TOJ tasks, and are summarized in Tables 1 and 2. To test for effects of task, we also performed a three-way ANOVA with an additional factor Task (SJ vs. TOJ).

In the SJ task, subjects made a two-alternative forced choice: are the audio and video tracks synchronous or asynchronous? Figure 2a shows the mean (n = 9) proportion "synchronous" responses at the nine levels of audiovisual asynchronies in the four conditions. Negative values indicate that the audio track was leading, positive values that the video track was leading. As can be seen from Figure 2a, stimuli in all conditions were most often judged to be synchronous when the video was leading the audio track. This shift of peak location towards visual leading stimuli was observed for all four conditions (see Table 1) and was possibly accounted for by the audiovisual misalignment during stimulus recording (see Method). The two-way ANOVA showed no significant effects of Speech, Sentence, or their interaction on peak performance and peak location. There was a significant effect of Speech, but not Sentence on the width of the psychometric functions (the interaction term was not significant). The increased width for normal relative to rotated speech suggests that the threshold for detecting asynchrony was lower in normal compared to rotated speech. Visual inspection of the data (Figure 2a) shows that the psychometric functions were not symmetrical around their peaks. The ANOVA on asymmetry revealed a significant effect of Speech and Sentence in the absence of an interaction. Asymmetry

Table 1
Summary of Results: Mean (SEM)

	Peak location	Peak performance	Width	Asymmetry
Synchrony judgments				
Normal complete	111.67 (24.97)	0.94 (0.02)	1.11 (0.29)	−0.39 (0.07)
Normal on/off	89.33 (27.35)	0.92 (0.04)	1.30 (0.37)	−0.13 (0.09)
Rotated complete	74.44 (20.72)	0.87 (0.03)	1.40 (0.13)	−0.02 (0.07)
Rotated on/off	81.89 (21.70)	0.91 (0.03)	1.45 (0.21)	0.07 (0.04)
Temporal order judgments				
Normal complete	59.56 (26.06)	0.46 (0.05)	1.41 (0.11)	−0.05 (0.05)
Normal on/off	22.33 (15.79)	0.37 (0.04)	1.26 (0.06)	−0.07 (0.05)
Rotated complete	37.22 (11.77)	0.40 (0.03)	1.41 (0.06)	−0.01 (0.02)
Rotated on/off	29.78 (35.51)	0.41 (0.04)	1.41 (0.08)	0.03 (0.06)

was greater for normal speech and complete sentences, compared to rotated speech and on/offset sentences, respectively. To summarize, in the SJ task, normal relative to rotated speech is characterized by a narrower and more asymmetrical temporal integration window. The asymmetry effect is even more pronounced for complete compared to onset/offset sentences.

In the TOJ task, subjects made a two-alternative forced choice: Is the audio or the video track leading in time? Figure 3a shows the mean ($n = 9$) proportion correct responses at the nine levels of audiovisual asynchronies in the four conditions. The ANOVA did not reveal any significant effects on peak performance, peak location, width, or asymmetry, although normal speech was characterized by a slightly more asymmetric temporal integration window compared to rotated speech (Figure 3b).

Effect of Task

To compare performance in the SJ and TOJ tasks, we performed a three-way repeated measures ANOVA with factors Task (SJ vs. TOJ), Speech (normal vs. spectrally rotated) and Sentence (complete vs. on/offset). The results are shown in Table 3. As peak performance is not comparable across the two tasks (e.g. peak performance is ~100% in the SJ, and ~50% in the TOJ task), we performed the three-way ANOVA only for the remaining three indices. We found a significant main effect of Task on peak location, indicating that the shift in peak location towards visual leading was greater for SJs compared to TOJs. The three-way ANOVA for width revealed a main effect of Speech, corroborating our results from the two-way ANOVAs that were performed separately for each task, and an interaction between Task and Sentence. The three-way ANOVA for asymmetry revealed main effects of Speech and Sentence, as well as an interaction between Speech and Task.

These data indicate that the integration window was generally wider for rotated than for normal speech irrespective of task, and that reducing stimuli to their on/offsets had a larger effect on width in the SJ task compared to the TOJ task. The effect of our experimental manipulations on asymmetry was also modulated by task: the differences in asymmetry between rotated and normal speech were more pronounced in the SJ relative to the TOJ task.

The three-way ANOVA thus confirms that asymmetry and width indices obtained from the SJ task are more sensitive to alterations of the speech stimulus (i.e., spectrotemporal structure and reductions of the sentence to its onsets/offsets) than the indices obtained from the TOJ task. These Task × Sentence and Task × Speech interactions highlight differences between synchrony and temporal order judgments. They raise the question whether the width and asymmetry indices obtained from the two tasks may be related to similar underlying processing mechanisms. To address

Table 2
Summary of Results From the Two-Way ANOVA

Factor (*df*: 1,8)	Peak location	Peak performance	Width	Asymmetry
Synchrony judgments				
Speech	$F = 0.76$, $p = .408$, partial $\eta^2 = 0.087$	$F = 1.89$, $p = .206$, partial $\eta^2 = 0.186$	$F = 11.0$, $p = .011^{*}$, partial $\eta^2 = 0.579$	$F = 33.2$, $p < .001^{***}$, partial $\eta^2 = 0.806$
Sentence	$F = 0.37$, $p = .559$, partial $\eta^2 = 0.044$	$F = 0.55$, $p = .481$, partial $\eta^2 = 0.055$	$F = 1.94$, $p = .201$, partial $\eta^2 = 0.196$	$F = 5.50$, $p = .047^{*}$, partial $\eta^2 = 0.407$
Speech × Sentence	$F = 0.44$, $p = .525$, partial $\eta^2 = 0.052$	$F = 2.55$, $p = .149$, partial $\eta^2 = 0.250$	$F = 1.03$, $p = .339$, partial $\eta^2 = 0.116$	$F = 1.99$, $p = .196$, partial $\eta^2 = 0.198$
Temporal order judgments				
Speech	$F = 0.14$, $p = .719$, partial $\eta^2 = 0.017$	$F = 0.15$, $p = .709$, partial $\eta^2 = 0.013$	$F = 1.92$, $p = .203$, partial $\eta^2 = 0.193$	$F = 3.86$, $p = .085$, partial $\eta^2 = 0.322$
Sentence	$F = 1.07$, $p = .332$, partial $\eta^2 = 0.118$	$F = 3.32$, $p = .106$, partial $\eta^2 = 0.296$	$F = 1.65$, $p = .235$, partial $\eta^2 = 0.171$	$F = 0.03$, $p = .879$, partial $\eta^2 = 0.000$
Speech × Sentence	$F = 0.44$, $p = .525$, partial $\eta^2 = 0.052$	$F = 3.40$, $p = .103$, partial $\eta^2 = 0.299$	$F = 2.62$, $p = .145$, partial $\eta^2 = 0.248$	$F = 0.39$, $p = .550$, partial $\eta^2 = 0.046$

$^{*}p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$.

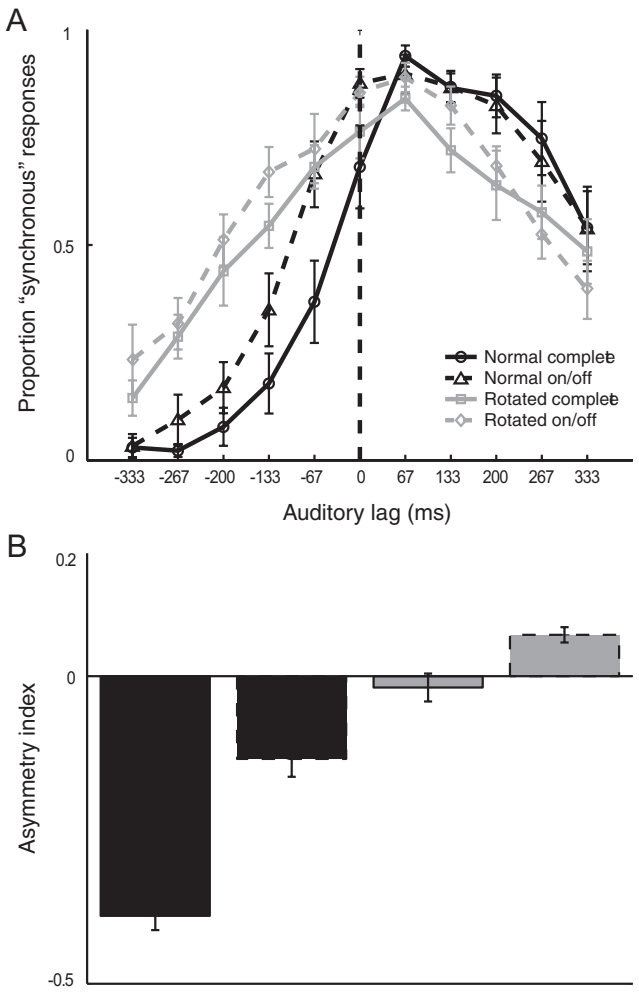

Figure 2. Performance in the synchrony judgment task. (A) Mean (n = 9) proportion "synchronous" responses as a function of stimulus onset asynchrony in the four different conditions. Negative (positive) asynchrony levels indicate audiovisual stimuli in which the auditory (visual) signal was leading the visual (auditory) signal. Synchronous stimuli (asynchrony level = 0) are indicated by the dotted vertical line. Error bars indicate ± 1 *SEM*. (B) Mean (n = 9) asymmetry index in the four different conditions. Negative values indicate better performance for auditory leading asynchrony levels compared to visual leading asynchrony levels. Error bars indicate ± 1 *SEM*.

this question we further analyzed the interrelationship of peak location, width, and asymmetry obtained from each subject for the two tasks using a canonical correlation analysis. A multivariate canonical correlation analysis was used because for each index we had to consider four independent variables (normal vs. rotated, and complete vs. on/offset for synchrony judgments) and four dependent variables (normal vs. rotated, and complete vs. on/offset for temporal order judgments). A canonical correlation analysis determines a linear combination of the independent and dependent set of variables, such that the correlation between the two sets of variables is maximized. Using Bartlett's chi-square statistic, we then tested whether the canonical correlation was significantly different from zero. A significant canonical correlation means that inter-subject variability for each index in the two tasks likely results at least in part from common underlying mechanisms. We found significant positive canonical correlation for our measures of peak location ($\chi^2(16,3.69) = 29.77$, $p = .019$) and width ($\chi^2(16,3.69) = 46.24$, $p < .001$) between the SJ and TOJ tasks, but there was no significant correlation for asymmetry ($\chi^2(16,3.69) = 21.78$, $p = .151$).

In summary, when considering mean estimates over subjects, the significant Task × Sentence and Task × Speech interactions highlight differences between synchrony and temporal order judgments and suggest that the synchrony judgments are more sensitive to changes in spectrotemporal structure of speech. However, focusing on inter-subject variability, the significant canonical correlations between the two tasks for peak location and width demonstrate that these characteristic indices reflect at least in part common processing mechanisms for synchrony and temporal order judgments.

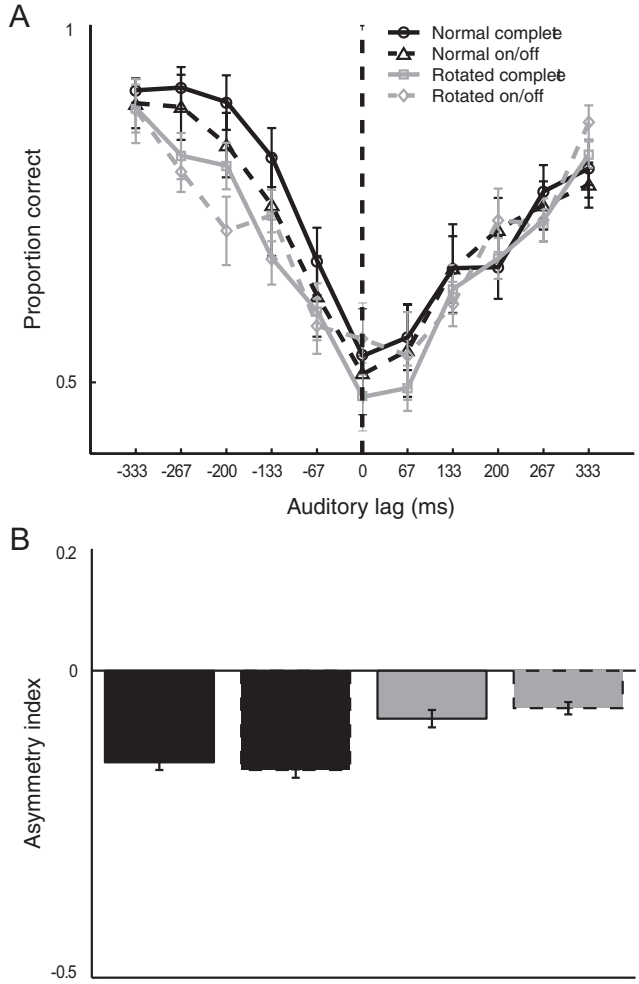

Figure 3. Performance in the temporal order judgment task. (A) Mean (n = 9) proportion correct responses as a function of stimulus onset asynchrony in the four different conditions. Same conventions as in Figure 1. Error bars indicate ± 1 *SEM*. (B) Mean (n = 9) asymmetry index in the four different conditions. Same conventions as in Figure 2. Error bars indicate ± 1 *SEM*.

Table 3
Summary of Results From the Three-Way ANOVA

Factor (*df*: 1,8)	Peak location	Width	Asymmetry
Task	$F = 10.8$, $p = .011^{*}$, partial $\eta^2 = 0.573$	$F = 0.36$, $p = .564$, partial $\eta^2 = 0.043$	$F = 3.24$, $p = .109$, partial $\eta^2 = 0.289$
Speech	$F = 1.16$, $p = .312$, partial $\eta^2 = 0.127$	$F = 12.6$, $p = .008^{**}$, partial $\eta^2 = 0.612$	$F = 46.0$, $p < .001^{***}$, partial $\eta^2 = 0.852$
Sentence	$F = 1.16$, $p = .312$, partial $\eta^2 = 0.127$	$F = 0.09$, $p = .774$, partial $\eta^2 = 0.010$	$F = 5.39$, $p = .049^{*}$, partial $\eta^2 = 0.402$
Task × Speech	$F = 0.17$, $p = .312$, partial $\eta^2 = 0.020$	$F = 3.10$, $p = .116$, partial $\eta^2 = 0.280$	$F = 10.8$, $p = .011^{*}$, partial $\eta^2 = 0.573$
Task × Sentence	$F = 0.47$, $p = .695$, partial $\eta^2 = 0.056$	$F = 6.28$, $p = .037^{*}$, partial $\eta^2 = 0.440$	$F = 3.46$, $p = .100$, partial $\eta^2 = 0.301$
Speech × Sentence	$F = 0.59$, $p = .512$, partial $\eta^2 = 0.068$	$F = 0.01$, $p = .929$, partial $\eta^2 = 0.000$	$F = 0.65$, $p = .444$, partial $\eta^2 = 0.073$
Three-way	$F = 0.00$, $p = 1.000$, partial $\eta^2 = 0.000$	$F = 2.35$, $p = .164$, partial $\eta^2 = 0.227$	$F = 2.26$, $p = .171$, partial $\eta^2 = 0.220$

Reversed Speech

Although spectral rotation largely preserves spectral complexity and temporal structure of the original speech signal, absolute amplitudes and frequencies can differ from the original waveform. As a consequence, correlations between visible mouth movements and spectrotemporal structure in the sound may be reduced in rotated compared to normal speech. Indeed, although the correlation between the envelopes of the normal and spectrally rotated speech signals is high, it was not perfect (mean $r^2 \pm SEM = 0.78 \pm 0.01$). To investigate the influence of audiovisual correlations on synchrony perception, five of our subjects performed the SJ task again in a third session, now on time-reversed versions of the complete normal speech sentences used in the original experiment. Temporally reversing the stimuli (both audio and video track) completely preserves zero lag cross-correlations between acoustic and visual features (i.e., correlations between two signals that are maximal when the signals are not shifted in time by delaying one of them). However, the temporal structure of the original speech signal is reversed and its intelligibility is removed. If SJs of normal speech sentences were based purely on zero lag cross-correlations between mouth movements and sound amplitude and/or frequency, no difference in performance would be expected for normal speech and reversed sentences.

Figure 4 shows the mean ($n = 5$) proportion synchronous judgments in the three conditions. As can be seen from this figure, the peak of the curve is strongly shifted towards visual leading (mean $\pm$ *SEM* = 240.20 $\pm$ 34.16 ms), making it impossible to calculate width and asymmetry indices of the psychometric function. To compare performance in the three conditions, we therefore calculated the slope of the ascending part of the psychometric function for each subject by fitting a linear function (normal speech: mean $r^2 \pm SEM = 0.92 \pm 0.02$; rotated speech: $r^2 = 0.86 \pm 0.05$; reversed speech: $r^2 = 0.94 \pm 0.02$), a steeper slope indicating better performance. We found a significant effect of Speech (rotated vs. normal vs. reversed) on slope

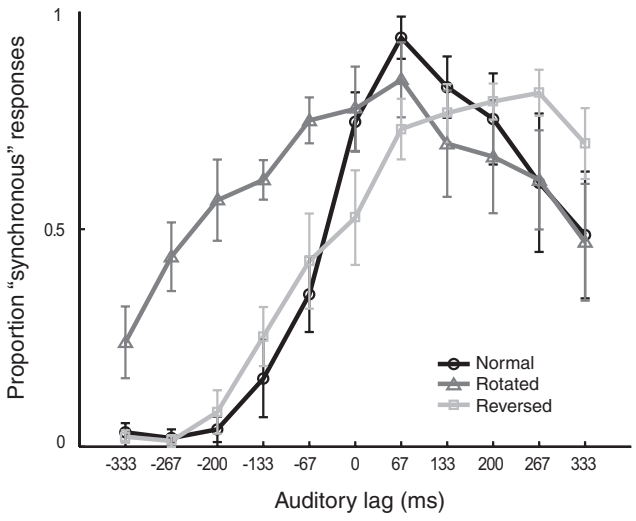

Figure 4. Performance in the synchrony judgment task. Mean ($n = 5$) proportion "synchronous" responses as a function of stimulus onset asynchrony for the three different speech signals. Same conventions as in Figures 2 and 3. Error bars indicate $\pm$ 1 *SEM*.

($F(2, 8) = 13.05$, $p = .003$). Slopes of the curves in both the rotated (mean ± *SEM* = 0.18 ± 0.01) and reversed (0.20 ± 0.03) conditions were significantly less steep compared to the normal speech (0.36 ± 0.05) condition (paired-samples *t* test: $t_{(4)} = 3.74$, $p = .02$; and $t_{(4)} = 3.94$, $p = .01$, respectively). These results demonstrate that both spectral and temporal inversion widen the temporal integration window. Both temporal and spectral changes render humans less sensitive in detecting audiovisual asynchronies. As reversed speech preserves the zero lag audiovisual cross-correlation, these data suggest that SJs of audiovisual speech signals are not based purely on zero lag cross-correlations between the auditory and visual signals.

Parametric Analysis

The values obtained from parametric fitting of the data are presented in Table 4. The values for peak location (SJ task) and PSS (TOJ task) agree well with the values for peak location obtained from the nonparametric analysis, as well as with the literature on temporal judgments of both auditory-visual speech (van Eijk et al., 2008; Vatakis et al., 2008; Vatakis & Spence, 2007; Vroomen, Keetels, de Gelder, & Bertelson, 2004) and nonspeech stimuli (Zampini, Shore, & Spence, 2003). The estimates of JNDs in the TOJ task are larger than previously reported with short stimuli, but they are comparable with previous studies that used continuous speech material (Dixon & Spitz, 1980; Grant & Seitz, 2000). We found no significant canonical correlation for our measures of peak location/PSS ($\chi^2(16,3.69) = 21.46$, $p = .162$) or JND ($\chi^2(16,3.69) = 12.82$, $p = .686$) between the two tasks.

Discussion

The present study investigated the effect of spectral rotation and temporal inversion on perceived synchrony of speech stimuli as indexed by simultaneity and TOJs. Our results demonstrate that the temporal integration window is narrower and more asymmetric for normal speech compared to rotated or reversed speech. This audiovisual asymmetry is more pronounced for complete compared to on/offset sentence fragments, indicating that subjects rely on cues during the entire sentence for their SJs. Task can modify subjects' strategies and the cues they use for audiovisual temporal judgments. While they rely on information during the entire sentence for SJs, they focus primarily on the sentence on- and offsets for TOJs.

Table 4
Summary of Results From Parametric Fitting: Mean (SEM)

	Peak location/PSS	Width/JND
Synchrony judgments		
Normal complete	150.23 (13.67)	164.08 (16.99)
Normal on/off	113.35 (16.53)	178.94 (12.91)
Rotated complete	70.80 (18.44)	243.38 (15.07)
Rotated on/off	57.18 (25.37)	238.07 (21.26)
Temporal order judgments		
Normal complete	59.62 (21.74)	123.91 (11.77)
Normal on/off	42.50 (27.82)	132.02 (9.95)
Rotated complete	27.50 (20.62)	144.26 (4.76)
Rotated on/off	29.10 (23.47)	157.01 (7.19)

Note. JND = just-noticeable difference; PSS = point of subjective simultaneity.

Consistent with many previous studies, our results reveal that temporal judgments of audiovisual speech are extremely tolerant to audiovisual asynchronies (Campbell & Dodd, 1980; Dixon & Spitz, 1980; Jones & Jarick, 2006; Massaro & Cohen, 1993; Massaro, Cohen, & Smeele, 1996; Munhall et al., 1996; Pandey, Kunov, & Abel, 1986; van Wassenhove, Grant, & Poeppel, 2007). In the SJ task, spectral inversion causes a widening of the temporal integration window for normal compared to rotated speech. These results may be surprising: based on the unity assumption one would expect a higher binding potential for normal speech stimuli because they enable the emergence of linguistic representations. More binding would result in a wider temporal integration window for normal compared to rotated speech because spectrally rotated speech represents semantically incongruent auditory and visual signals. In support of this hypothesis, McGurk stimuli that combine incongruent auditory and visual speech syllables (e.g., auditory "ba" and visual "ga") are associated with a narrower temporal integration window in a SJ task (van Wassenhove, Grant, & Poeppel, 2007; Soto-Faraco & Alsius, 2007). However, spectral rotation does not only preclude an intelligible speech percept, but also, to some degree alters the spectrotemporal aspects of the stimuli. These small changes in spectrotemporal structure render the spatiotemporal coincidence between visual and auditory signals less tight. Several studies have previously demonstrated the importance of audiovisual correlations for speech integration. More specifically, they have revealed correlations between facial kinematics such as mouth opening, and the acoustic envelope of the speech sound (Fairbanks, 1950; Grant & Seitz, 2000; Jiang, Auer, Alwan, Keating, & Bernstein, 2007; Munhall et al., 1996; Schwartz, Berthommier, & Savariaux, 2004; Stevens & House, 1955; Yehia, Rubin, & Vatikiotis-Bateson, 1998). However, even though reduced audiovisual correlations resulting from spectral rotation may explain the wider window for rotated compared to normal speech, they can not explain the dramatically widened integration window for reversed speech, in which audiovisual zero lag cross-correlations are completely preserved.

Taken together, our results indicate that detection of audiovisual asynchrony of speech signals is not just based on zero-lag cross-correlations between the auditory and visual speech signals. We suggest that human speech perception is fine-tuned to the specific statistics of natural speech. Therefore, even subtle changes to the time-frequency structure as induced by spectral or temporal inversion render asynchrony detection less efficient. As follows, this hypothesis is further corroborated by our analysis of the asymmetry index.

Visual inspection of the psychometric function (see Figure 2) revealed that the widening of the integration window in the rotated speech condition results primarily from increased "synchronous" responses for auditory leading stimuli. In contrast, in the normal speech condition, "synchronous" responses precipitously decline when the auditory signal leads, but remain relatively stable when the visual signal leads. In the SJ task, this asymmetry is even more pronounced for complete compared to onset/offset sentences, indicating that subjects' SJs do not rely exclusively on the temporal coincidence of on- and offset times but also on the statistical dependencies of the time varying characteristics of the auditory and visual streams. Indeed, consistent with visual inspection of the psychometric functions in Figure 2, statistical analysis of the asymmetry index showed additive effects of both normal versus rotated speech and complete versus

onset/offset sentences. Previous studies have found asymmetric performance on temporal perception of speech signals (Conrey & Pisoni, 2006; Grant, van Wassenhove, & Poeppel, 2004; Munhall et al., 1996). This asymmetry has been attributed primarily to three factors: differences in (1) reliability; (2) information content of the auditory and visual modalities; and (3) temporal relationship of the auditory and visual signals in natural speech (Grant, van Wassenhove, & Poeppel, 2004). First, for speech identification, visual information is relatively coarse and less reliable than auditory information. Hence, accumulation of visual evidence is slower compared to auditory evidence. This protracted accumulation of visual evidence may induce a coarser and delayed onset definition of the visual speech stream. Second, from a more linguistic perspective, the difference between visual and auditory asynchronies may relate to the time constants of the phonemic and syllabic cues that are carried by the auditory and visual modalities respectively. While the auditory modality provides cues about manner of articulation and voicing that are important for fine-grained phonemic analysis, vision conveys place of articulation cues that evolve over syllabic intervals of about 200 ms (de la Vaux & Massaro, 2004; Munhall & Tohkura, 1998). Third and most importantly, the asymmetry may have emerged as an adjustment to the natural timing relations between auditory and visual events in natural speech: facial movements and posturing can be observed nearly always before auditory speech articulation. Humans may have adapted to natural audiovisual speech statistics by tolerating visual leading asynchronies, but being more sensitive to the less likely event of auditory leading asynchronies. In line with this account, adaptation to temporal asynchrony of novel, artificial stimuli has been shown to shift or widen the temporal window of integration (Navarra, Soto-Faraco, & Spence, 2007; Navarra, Vatakis, Zampini, Soto-Faraco, Humphreys, & Spence, 2005), selectively in the direction of the adapted lag (Fujisaki, Shimojo, Kashino, & Nishida, 2004; Vroomen et al., 2004).

Our results suggest that because of a lifelong exposure to native speech, human audiovisual speech perception is fine-tuned to the natural mapping between facial movement and time-frequency structure in the voice. Thus, when presented with unfamiliar statistics such as rotated speech, humans cannot rely on prior experience, leading to less precise and unbiased predictions as indicated by a wider and less asymmetric integration window. This hypothesis may be further evaluated in future experiments that compare synchrony perception in native and foreign languages.

Finally, our results demonstrate that task (i.e., SJ vs. TOJ) influences subjects' decisions on audiovisual timing relations. SJ and TOJ experiments provided qualitatively different results, with the SJ task being more sensitive to our Speech and Sentence manipulations. Specifically, the effects of spectral rotation and reducing sentences to their on- and offsets were greater for SJs than for TOJs. One may attribute differences between TOJ and SJ tasks simply to biases in response strategy. In the TOJ task, visual and auditory leading stimuli are equally likely to occur, while in the SJ task, asynchronous stimuli occur more frequently than synchronous stimuli. This may bias subjects to make "asynchronous" responses, which would reduce the width estimate. Alternatively, from a frequency equalization perspective, observers may be biased to make "synchronous" responses more frequently than necessary to compensate for the unequal frequencies of "synchronous" and "asynchronous" responses, leading to a narrowing of the integration window. Importantly, our study only revealed a main effect of Task on peak location, and Task × Speech and Task × Sentence interaction on width and asymmetry. This pattern of results cannot be explained by a general response bias. Instead, the reduced sensitivity in the TOJ task to spectrotemporal alterations that preserve on/offsets indicates that subjects tend to focus more on the physical characteristics (visual motion and acoustic energy) of the stimuli at on- and offset during the TOJ task, despite being instructed to attend to the entire sentence in both tasks. This is possibly related to the fact that the TOJ task encourages subjects to focus on the component signals to determine their temporal precedence, whereas the SJ task requires judgments on the combined signals. A shift of attentional focus to the onsets of the individual component signals during the TOJ task renders subjects less sensitive to additional speech-specific temporal information in the remainder of the stimuli, as evidenced by the significant interactions between Task and Speech and Sentence manipulations. This may also explain the shift of peak position towards auditory leading in the TOJ task relative to the SJ task, as evidenced by the significant effect of Task on peak location. In natural audiovisual speech, facial movements always precede voice onset. Focusing on the physical energy at the onset of the stimuli would therefore induce a shift of peak position towards auditory leading.

The significant interactions Task × Speech and Task × Sentence clearly demonstrate that audiovisual integration of speech signals does not only depend on stimulus characteristics, but also at least in part on the particular context in which subjects respond to these signals. This raises the question whether we could find evidence for commonalities in the different behaviors induced by the two tasks. Previous studies have not obtained significant correlation over subjects between the estimates for the point of subjective simultaneity (PSS) between SJ and TOJ tasks, indicating that PSS values may not be independent of the experimental method (van Eijk et al., 2008; Vatakis et al., 2008). Indeed, the PSS and JND measures obtained from our parametric analysis were not correlated either between the two tasks (van Eijk et al., 2008). However, the nonparametric measures of peak location and width (but not asymmetry) were significantly correlated between SJs and TOJs, suggesting that these indices are at least in part determined by common mechanisms underlying auditory-visual speech processing. Thus, even though synchrony judgments seem more sensitive to our Speech and Sentence manipulations, both tasks also tap into a common mechanism underlying auditory-visual speech processing. Furthermore, the difference in results between the parametric and nonparametric analysis also highlights the importance of applying nonparametric approaches if response distributions do not conform to parametric assumptions.

Collectively, our results suggest that because of a lifelong exposure to natural speech, human audiovisual speech perception is fine-tuned to the natural mapping between facial movement and time-frequency structure in the voice. This fine-tuned mapping can be better characterized in synchrony detection and speech identification tasks that tap into speech-specific processing mechanisms, than in temporal order tasks where subjects focus selectively on the physical energy of the component signals, ignoring the characteristic features of natural connected speech.

References

Bedard, C., & Belin, P. (2004). A "voice inversion effect?" *Brain & Cognition, 55,* 247–249.

Blesser, B. (1972). Speech perception under conditions of spectral transformation: I. Phonetic characteristics. *Journal of Speech and Hearing Research, 15,* 5–41.

Campbell, R., & Dodd, B. (1980). Hearing by eye. *Quarterly Journal of Experimental Psychology, 32,* 85–99.

Conrey, B., & Pisoni, D. B. (2006). Audiovisual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119,* 4065–4073.

de la Vaux, S. K., & Massaro, D. W. (2004). Audiovisual speech gating: Examining information and information processing. *Cognitive Processes, 5,* 106–112.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9,* 719–721.

Fairbanks, G. (1950). A physiological correlative of vowel intensity. *Speech Monographs, 17,* 390–395.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7,* 773–778.

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

Grant, K. W., van Wassenhove, V., & Poeppel, D. (2004). Detection of auditory (cross-spectral) and audiovisual (cross-modal) synchrony. *Speech Communication, 44,* 43–53.

Hirsh, I. J., & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology, 62,* 423–432.

Jiang, J., Auer, E. T., Alwan, A., Keating, P. A., & Bernstein, L. E. (2007). Similarity structure in visual speech perception and optical phonetic signals. *Perception & Psychophysics, 69,* 1070–1083.

Jones, J. A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research, 174,* 588–594.

Massaro, D. W., & Cohen, M. M. (1993). Perceiving bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13,* 127–134.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1996). Perception of asynchronous and conflicting visual and auditory speech. *Journal of the Acoustical Society of America, 100,* 1777–1786.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351–362.

Munhall, K. G., & Tohkura, Y. (1998). Audiovisual gating and the time course of speech perception. *Journal of the Acoustical Society of America, 104,* 530–539.

Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Adaptation to audiotactile asynchrony. *Neuroscience Letters, 413,* 72–76.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25,* 499–507.

Pandey, P. C., Kunov, H., & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research, 26,* 27–41.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17,* 1147–1153.

Schwartz, J.-L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audiovisual interactions in speech identification. *Cognition, 93,* B69–B78.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123,* 2400–2406.

Soto-Faraco, S., & Alsius, A. (2007). Conscious access to the unisensory components of a cross-modal illusion. *Neuroreport, 18,* 347–350.

Stein, B. E., & Meredith, A. M. (1993). *The merging of the senses*. Cambridge, MA: MIT Press.

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. In S. Kornblum (Ed.), *Attention & performance IV* (pp. 629–685). New York: Academic Press.

Stevens, K. N., & House, A. S. (1955). Development of a quantitative description of vowel articulation. *Journal of the Acoustical Society of America, 27,* 484–493.

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. R. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society of London B, 268,* 31–38.

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

Summerfield, Q. (1992). Lipreading and audiovisual speech perception. *Philosophical Transactions of the Royal Society of London: Series B, 335,* 71–78.

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audiovisual speech perception is special. *Cognition, 96,* B13–B22.

van Eijk, R. L. J., Kohlrausch, A., Juola, J. F., & van der Par, S. (2008). Audiovisual synchrony and temporal order judgments: Effects of experimental method and stimulus type. *Perception & Psychophysics, 70,* 955–968.

van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in audiovisual speech perception. *Neuropsychologia, 45,* 598–607.

Vatakis, A., Ghazanfar, A. A., & Spence, C. (2008). Facilitation of multisensory integration by the "unity effect" reveals that speech is special. *Journal of Vision, 8,* 1–11.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments. *Experimental Brain Research, 185,* 521–529.

Vatakis, A., & Spence, C. (2006a). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience Letters, 393,* 40–44.

Vatakis, A., & Spence, C. (2006b). Audiovisual synchrony perception for music, speech and object actions. *Brain Research, 1111,* 134–142.

Vatakis, A., & Spence, C. (2006c). Evaluating the influence of frame rate on the temporal aspects of audiovisual speech perception. *Neuroscience Letters, 405,* 132–136.

Vatakis, A., & Spence, C. (2007). Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli. *Perception & Psychophysics, 69,* 744–756.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson., P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22,* 32–35.

Warren, J. E., Sauter, D. A., Eisner, F., Wiland, J., Dresner, M. A., Wise, R. J. S., Rosen, S., & Scott, S. K. (2006). Positive emotions preferentially engage a audiomotor "mirror" system. *Journal of Neuroscience, 26,* 13067–13075.

Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin, 88,* 638–667.

Yehia, H., Rubin, P., & Vatikiotis-Bateson, E. (1998). Quantitative association of vocal-tract and facial behavior. *Speech Communication, 26,* 23–43.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics, 67,* 531–544.

Zampini, M., Shore, D. I., & Spence, C. (2003). Audiovisual temporal order judgments. *Experimental Brain Research, 152,* 198–210.

(*Appendix follows*)

Appendix: Stimuli

Set 1:
Viele Leute arbeiten hart.
Eine Fliege hat Flügel.
Pferde haben vier Hufe.
Sabine schreibt ein Brief.
Bäume verlieren ihre Blätter.

Set 2:
Viele Menschen tragen Brillen.
Paul schliesst die Tür.
Bettina öffnet das Fenster.
Sebastian backt einen Kuchen.
Igel haben kurze Beinen.

Received May 6, 2009
Revision received November 11, 2009
Accepted February 8, 2010 ■