Cognition 118 (2011) 78–86

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Perception of intersensory synchrony in audiovisual speech: Not that special

Jean Vroomen *, Jeroen J. Stekelenburg

Tilburg University, Department of Medical Psychology and Neuropsychology, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

ARTICLE INFO

Article history:
Received 22 February 2010
Revised 17 September 2010
Accepted 4 October 2010

Keywords:
Audiovisual speech perception
Sine-wave speech
Simultaneity judgment
Temporal order judgment
Multi-sensory integration

ABSTRACT

Perception of intersensory temporal order is particularly difficult for (continuous) audiovisual speech, as perceivers may find it difficult to notice substantial timing differences between speech sounds and lip movements. Here we tested whether this occurs because audiovisual speech is strongly paired (“unity assumption”). Participants made temporal order judgments (TOJ) and simultaneity judgments (SJ) about sine-wave speech (SWS) replicas of pseudowords and the corresponding video of the face. Listeners in speech and non-speech mode were equally sensitive judging audiovisual temporal order. Yet, using the McGurk effect, we could demonstrate that the sound was more likely integrated with lipread speech if heard as speech than non-speech. Judging temporal order in audiovisual speech is thus unaffected by whether the auditory and visual streams are paired. Conceivably, previously found differences between speech and non-speech stimuli are not due to the putative “special” nature of speech, but rather reflect low-level stimulus differences.

1. Introduction

Most of our real-world perceptual experiences are specified by multisensory perceptual attributes, as in the case of a talker who can be heard and seen at the same time. The multisensory nature of perception though, raises the question about how the different sense organs cooperate and are integrated so as to form a coherent representation of the world. The most commonly held view among researchers is what has been referred to as the “assumption of unity”. It states that as information from different modalities share more (amodal) properties, the more likely the brain will treat them as originating from a common object or source (see, e.g., Bedford, 1989; Bertelson, 1999; Radeau, 1994; Stein & Meredith, 1993; Welch, 1999; Welch & Warren, 1980) Without doubt, the most important amodal property is commonality in time (e.g. Radeau, 1994), though commonality in space, association based upon co-occurrence, or semantic congruency may also be of importance.

Research in support of the unity assumption has mainly used the “ventriloquist illusion” where the apparent location of a sound is shifted toward a concurrently presented visual event. Here, it has been found that a sound is shifted more if the sound and the visual event conceivably come from the same source. However, most of this research has been criticized because response biases have confounded the interpretation of the research in this area (see de Gelder and Bertelson (2003), for a review). A less conflated prediction that follows from the unity assumption, though, is that for strongly paired intersensory stimuli, the relative temporal order of the components is lost as they are “ventriloquized in time” so as to form an integrated and synchronized percept (Morein-Zamir, Soto-Faraco, & Kingstone, 2003; Scheier, Nijhawan, & Shimojo, 1999; Vroomen & de Gelder, 2004).

One piece of evidence supporting this notion is that sensitivity for intersensory timing differences is relatively poor in naturally produced audiovisual speech if compared to simple artificial stimuli like flashes and beeps.

* Corresponding author. Tel.: +31 13 466 2394; fax: +31 13 466 2067.
E-mail address: J.Vroomen@uvt.nl (J. Vroomen).

0010-0277/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2010.10.002

Presumably, heard and lipread speech are more strongly paired than beeps/flashes, and the pairing merges the information streams in time. As an example, in a study by van Wassenhove, Grant, and Poeppel (2007) observers judged whether congruent audiovisual speech stimuli or incongruent McGurk-like speech stimuli (McGurk & MacDonald, 1976) were synchronous or not. The lags at which observers started to notice that the auditory and visual information were out-of-sync – the "temporal window of integration" – was estimated at 203 ms for phonetically congruent pairs and 159 ms for incongruent pairs. Other studies using a temporal order judgment (TOJ) task rather than simultaneity judgments found that the just noticeable difference (JND) – indexing the sensitivity for intersensory timing differences – for AV speech is in the range of 70–150 ms (Stekelenburg & Vroomen, 2007; Vatakis & Spence, 2006a,b, 2007), which is high if compared to the much lower JNDs typically found for simple flashes and beeps that are mostly below 50 ms (Hirsh and Sherrick (1961), Vroomen and Keetels (2010) for a review on intersensory synchrony).

Others have also compared intersensory timing of audiovisual speech with audiovisual events like music instruments (guitar and piano) and object actions (e.g. smashing a television set with a hammer, or hitting a soda can with a block of wood) (Vatakis & Spence, 2006a,b) and have found audiovisual speech to be particularly difficult. Vatakis and Spence (2006a,b) concluded that the sensitivity for audiovisual timing differences improves for stimuli of "lower complexity" in comparison with stimuli having continuously varying properties like syllables, words and/or sentences. Stekelenburg and Vroomen (2007) also compared sensitivity for audiovisual timing of audiovisual speech (the pronunciation of the syllable /bi/) with that of natural non-speech events (a video of a handclap). Again, sensitivity for audiovisual timing differences was much better for the non-speech events (64 ms) than for speech (105 ms), possibly because intersensory pairing of audiovisual speech is particularly strong.

The most direct evidence in support of the unity assumption comes from a study by Vatakis and Spence (2007). They had participants judge the temporal order of audiovisual speech that was matched or mismatched in gender (e.g., the sound of a male/female /bi/ dubbed onto a male/female face saying /bi/) or phonemic content (the syllable /ba/ or /da/ dubbed onto a face saying /ba/ or /da/). The authors found that sensitivity for temporal order was worse if the auditory and visual streams were matched rather than mismatched, presumably because the matched stimuli were more strongly paired. More recently, though, Vatakis, Ghazanfar, and Spence (2008) qualified these findings and reported that the congruency effect may be specific for human audiovisual speech, because there was no difference between matching or mismatching call-types of monkeys ("cooing" versus "grunt" or "threat" calls), and no difference between matching and mismatching human vocalizations of the monkey calls. Vatakis and Spence (2008a) also found no difference between matching and mismatching recordings of music and object events (e.g., the visual signal of a hammer smashing a block of ice combined with the sound of a ball bouncing), and based on these findings, it has been suggested that intersensory pairing in audiovisual speech may be "special" (van Wassenhove et al., 2007; Vatakis & Spence, 2008a).

At this stage, though, it seems that this conclusion is premature because all the previously mentioned studies suffer from the same confound, namely that comparisons are made across different and sometimes quite arbitrary chosen stimulus classes that differ on a number of low-level acoustic and visual dimensions. For example, a low-level factor that needs to be taken into account if one wants to make sensible comparisons across stimulus classes is the extent to which the auditory and visual signal are transient. van der Burg, Cass, Olivers, Theeuwes, and Alais (2010) reported that visual search became more efficient if a modulating visual target was paired with a synchronous auditory signal. Crucially, slow sinusoidal audiovisual modulations did not support efficient search, and benefits were only obtained if the changes in the component signals were both synchronized and transient. Judgments of temporal order are also affected by whether the component signals are transient or slowly changing, and judging temporal order in audiovisual speech may be particularly difficult if it lacks abrupt changes like stop consonants that provide clear temporal markers (Conrey & Pisoni, 2003; Vroomen & Keetels, 2010).

As another example, it has been demonstrated that judging the temporal order of audiovisual stimuli becomes difficult if the stimulus pairs are presented above ~4 Hz (Benjamins, van der Smagt, & Verstraten, 2005; Fujisaki & Nishida, 2005). Above this rate, observers are no longer able to discriminate whether the auditory and visual stimulus elements are synchronous, and the two modality streams are perceived as being segregated with no order between them. This limit at ~4 Hz is rather low if compared with the unimodal perception of synchrony (e.g., deciding whether two flickering visual signals are in- or out-of-phase breaks down above ~25 Hz) (Fujisaki & Nishida, 2005). The ~4 Hz is also approximately the average rate at which syllables are produced in fluent speech, and judging audiovisual temporal order in fluent speech may be difficult because the presentation rate is fast, rather than that it is speech-like nature of the stimulus per se.

Yet another factor that makes a direct comparison between even matched stimuli difficult is that in normally-matched audiovisual continuous speech, there is a continuous temporal correlation between the time-varying characteristics of the auditory and visual streams, especially in the 3–4 Hz range (van Wassenhove et al., 2007). In case there is a lag between the two streams, there is still the (time-shifted) correlation between sound and vision (Munhall, Gribble, Sacco, & Ward, 1996). This correlation may induce a so-called "temporal ventriloquist" effect (Morein-Zamir et al., 2003). The basic phenomenon in temporal ventriloquism is that a sound presented shortly before or after a light (~100 ms) can attract the perceived temporal occurrence of that light. The purpose of temporal ventriloquism may be to reduce differences in transmission and processing times of the different senses so that naturally occurring lags are perceived as being simultaneous (Vroomen & Keetels, 2010). Temporal ventriloquism may explain why sensitivity for audiovisual temporal order is

better for incongruent than congruent audiovisual speech because it is likely that the fine temporal correlation in incongruent speech is disrupted, thus preventing temporal ventriloquism to occur. Small lags in continuous audiovisual speech may thus be unnoticeable because there is more temporal ventriloquism for congruent than incongruent audiovisual speech.

For discrete rather than continuous events – like a hammer hitting an ice cube – there is no such inherent time-varying correlation between the auditory and visual streams, and perceivers will have to rely primarily on the temporal coincidence of auditory and visual onsets. If in this situation a discrete sound is replaced by another incongruent but discrete sound, congruency is unlikely to affect perception of audiovisual synchrony because there is still an auditory onset that can be judged relative to the visual onset. So, congruency between the auditory and visual streams may affect continues speech but not discrete events because only continues speech is affected by the time-varying correlation (Vroomen & Keetels, 2010). It seems therefore reasonable that each of these stimulus characteristics – and likely many others – needs to be controlled if one wants to compare intersensory timing across stimuli in a non-arbitrary way. Finally, if "unity between the senses" is indeed crucial, it seems important to have an independent measure of whether information streams were indeed paired or not, because otherwise it becomes a circular argument.

Here we thought to alleviate all these problems in the most rigorous way, namely by using *identical* audiovisual stimuli that were either paired or not paired depending on whether the sounds themselves were perceived as speech or non-speech. For that purpose, we used sine-wave speech (SWS) and combined it with lipread information. In SWS, the natural richness of the auditory signal is reduced to a few sinusoids that follow the centre frequency and the amplitude of the first three formants. These stimuli sound highly artificial, and most naïve subjects perceive them as "non-speech" sounds like whistles or sounds from a science fiction movie. Typically, though, once subjects are told that these sounds are actually derived from speech, listeners cannot switch back to a non-speech mode again and continue to hear the sounds as speech (Remez, Rubin, Pisoni, & Carrell, 1981). Critical for our purpose is that if SWS sounds are combined with lipread information, then naïve subjects in non-speech mode show no or only negligible intersensory integration when asked to identify the sound, while subjects who learned to perceive the same auditory stimuli as speech do integrate the auditory and visual stimuli in a manner similar to natural speech (Tuomainen, Andersen, Tiippana, & Sams, 2005; Vroomen & Baart, 2009). This situation provides the ideal platform to put the unity assumption to test because it predicts that listeners in non-speech mode should be more sensitive to audiovisual timing differences than listeners in speech mode (Experiments 1 and 2), while all low-level stimulus factors are equated. To anticipate, under these conditions we did not observe any difference between listeners in speech and non-speech mode. In Experiment 3, we then checked whether listeners in speech mode did in fact more likely integrate the auditory and visual information than listeners in non-speech mode by looking at the effect of incongruent lipread information on sound identification. Here, we expected – and indeed observed – that listeners in speech mode were more affected by lipread information than listeners in non-speech mode. The core assumption of the unity hypothesis – intersensory pairing impairs judgements of temporal order – could thus not be corroborated in the most stringent experimental situation.

2. Experiment 1

2.1. Participants

Ninety healthy participants (undergraduate students) with normal hearing and normal or corrected-to-normal vision participated after giving written informed consent. Their age ranged from 18 to 28 years with mean age of 19.5 years. They were equally divided into three between-subjects conditions (i.e., natural speech, SWS speech mode, and SWS non-speech mode). Note that a between-subject design was required because once participants perceive an SWS sound as speech, they cannot switch back to a non-speech mode again.

2.2. Stimuli

The experiment took place in a dimly-lit and sound-attenuated room. Visual stimuli were presented on a 17-in. monitor positioned at eye-level, 70 cm from the participant's head. The sounds came from a loudspeaker directly below the monitor. The stimulus was the Dutch pseudoword /tabi/ pronounced by a Dutch male speaker whose entire face (from top of the shoulders to top of the head) was visible on the screen. Peak intensity of the auditory stimulus was 70 dB(A), duration was 627 ms. The videos were presented at a rate of 25 frames/s with an auditory sample rate of 44.1 kHz. The size of the video frames subtended 14° horizontal and 12° vertical visual angle.

2.3. Procedure

There were three between-subject conditions: Either the original (natural) audio recording was used (serving as a control condition), or the sound was transformed into SWS by running a script provided by Chris Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS) in the Praat software (Boersma & Weenink, 2005). The script creates a three-tone stimulus by positioning time-varying sine waves at the centre frequencies of the three lowest formants of the natural speech tokens. These SWS stimuli were presented either in a "speech mode" or a "non-speech mode". Participants in speech mode were trained to perceive the SWS stimuli as speech. This was done by alternating the original audio recording and the SWS token ten times before the start of the experiment. Participants in the non-speech mode condition heard the SWS sound equally often, but they were told that it was an artificial computer sound. The stimulus onset asynchronies (SOA) between the auditory (A) and visual (V)

information varied from −320 ms (A-first) to +320 ms (V-first) in 40-ms steps (17 SOA's). For each SOA, a total of 24 randomized trials were administered across four separate blocks. Participants judged whether A or V was presented first using two designated buttons. The next trial started 1 s after a response was detected. The duration of the first video frame containing the still face with mouth closed varied randomly from trial-to-trial (100–500 ms) so that participants could not judge audiovisual temporal order on the basis of the silence at the start of the video alone.

Following the training session, participants were asked about the nature of the SWS stimuli to examine whether they had spontaneously perceived any phonetic element in the SWS stimuli. If so (eight participants in total), they were assigned to either the SWS speech-mode group or to the natural speech group. This procedure allowed us to exclude participants who heard the SWS sounds as speech in a spontaneous way from the non-speech group. At the end of the experiment, participants of the SWS non-speech group were asked about the nature of the sound. Although most participants guessed that the sound has something to do with the actor on the screen, none of them reported to have heard the SWS sound as /tabi/.

2.4. Results and discussion

The proportion of "V-first" responses was calculated for each participant, and these data were submitted to a MANOVA for repeated measures with as within-subjects variable SOA (17 levels) and as between-subjects variable Condition (natural speech, SWS speech mode, and SWS non-speech mode). As shown in Fig. 1, a typical S-shaped psychometric curve was obtained in each condition, but importantly, there was *no* difference between these conditions. There was a main effect of SOA, $F(16, 72) = 215.84$, $p < 0.001$ because – unsurprisingly – the more "V-first" responses were given the more V was presented before A. More importantly, though, judgments of temporal order did not depend on whether listeners heard the sound as speech or non-speech, as there was no main effect of Condition and no interaction between SOA × Condition (both F's < 1).

To estimate sensitivity in a more direct way, a logistic function was fitted on the individual data so as to determine the just noticeable difference (JND) for each condition. The JND was derived by taking the difference between the SOAs at which 25% and 75% of the responses were "V-first", and dividing it by two. The average JNDs were virtually identical and were 101 ms for the natural speech group, 101 ms for the SWS speech group, and 100 ms for the SWS non-speech group ($F < 1$). The JNDs were comparable to previous studies (i.e., 106 ms) (Stekelenburg & Vroomen, 2007) when using the same experimental set-up but different speech stimuli.

For completeness, we also examined whether the point of subjective simultaneity (PSS, i.e. the SOA where sound and light are judged as appearing simultaneously) would differ between conditions. The PSS was estimated from the psychometric functions by calculating the SOA at which 50% "V-first" responses were given. The average PSS across conditions was +2 ms and did not differ between conditions ($F < 1$). At this stage, there is thus no sign that sensitivity for audiovisual temporal order is worse if SWS sounds are heard as speech rather than non-speech.

3. Experiment 2

The results of Experiment 1 are important because to date they provide, in our view, the most rigorous test of the unity assumption. Yet, it could be argued that, despite that a large number of subjects were tested, it is in essence a null-result. We therefore considered it important to replicate these findings using a different methodology. In Experiment 2, we used a simultaneity judgement (SJ) task as it has been suggested that judgments about simultaneity versus temporal order are based on different information sources reflecting possibly different underlying

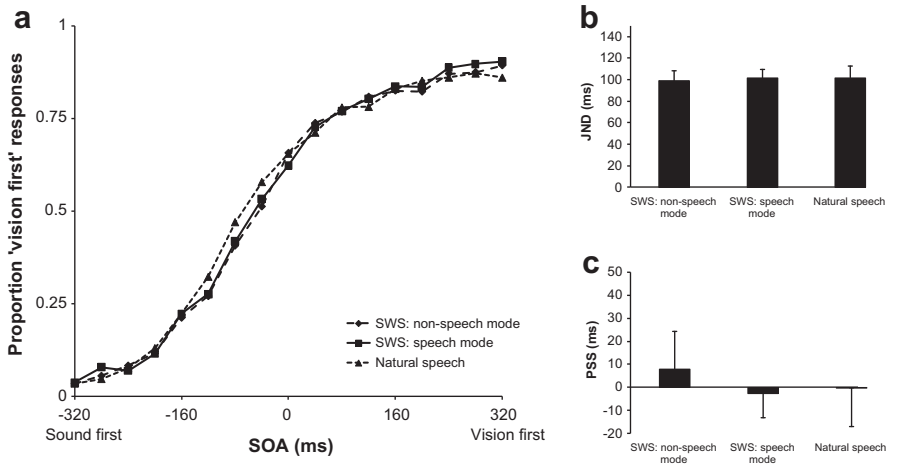

Fig. 1. (a) Mean proportion "vision first" responses as a function of stimulus onset asynchrony (SOA) for sine-wave speech (SWS) in speech mode, SWS in non-speech mode, and natural speech. (b) The just noticeable difference (JND). (c) The point of subjective simultaneity (PSS). Error bars represent 1 Standard Error of the Mean (SEM).

mechanisms (van de Par, Kohlrausch, & Juola, 2002; van Eijk, Kohlrausch, Juola, & van de Par, 2008; Vatakis, Navarra, Soto-Faraco, & Spence, 2008). Participants in Experiment 2 thus decided whether A and V were "synchronous" or "asynchronous", rather than which came first. With this task, there were actually two possible outcomes that might lend support the unity assumption: (1) There might be more overall "synchronous" responses if the SWS sound is heard as speech than non-speech, because there is more pairing for speech; or (2) the effect of pairing might be manifest in the mid-range of SOAs where the heard and lipread information conceivable belong together. This then should show up as lower sensitivity for SWS heard as speech than non-speech.

3.1. Participants

Sixty new students (18–26 years, mean 19.6 years) participated.

3.2. Stimuli and procedure

Stimuli and procedures were identical to Experiment 1, except that participants (20 per condition) now judged whether A and V were synchronous or asynchronous by pressing a left or right button, respectively.

3.3. Results and discussion

Fig. 2 shows the proportion of synchronous responses as a function of the SOA. As has been reported before, curves were asymmetrical since participants more often judged the stimuli to be synchronous when the sound came after the visual stimulus rather than before (e.g., Conrey & Pisoni, 2006; Stevenson, Altieri, Kim, Pisoni, & James, 2010; van Wassenhove et al., 2007). More importantly, though, is that the proportion of synchronous judgments in the three conditions was almost identical. This was corroborated in a MANOVA for repeated measures with as within-subjects variable SOA (17 levels) and as between-subjects variable Condition (natural speech, SWS speech mode, and SWS non-speech mode). As before, there was a main effect of SOA, $F(16, 42) = 557.80$, $p < 0.001$, but – crucially – no main effect of Condition and no interaction between Condition and SOA (both Fs < 1).

The data were also fitted with a Gaussian function of which the standard deviation (SD) was taken as a measure of sensitivity. The average SD was 170 ms and did not differ between natural speech, SWS in speech mode, and SWS in non-speech mode ($F < 1$). There was thus again no sign that sensitivity was worse (a higher SD) for participants in speech mode rather than non-speech mode. For completeness, we also analyzed the peak of the Gaussian which corresponds to the PSS. The average PSS was 132 ms, which is shifted from zero as participants were likely to judge the lagging sound as being synchronous, but there was again no difference between conditions in the PSS, ($F < 1$). The difference in PSS between Experiment 1 (2 ms) and Experiment 2 (132 ms) is remarkable but not unusual. PSS estimates derived from a TOJ task often differ from those derived from a SJ task, with auditory-leading PSS values reported mainly for the TOJ task (van Eijk et al., 2008). Our findings corroborate the notion that judgments of temporal order and judgments of simultaneity are fundamentally different.

As in Experiment 1, we observed that sensitivity for audiovisual timing did not differ between listeners hearing SWS in speech versus non-speech mode. These data therefore lead one to conclude that whether SWS is heard as speech or non-speech, and thus whether it is paired with lipread information, is of no consequence for perception of temporal order. This conclusion is in stark contradiction with the unity assumption. However, before accepting this, it is of importance to check that A and V were actually paired if listeners were in speech mode, but not so if in non-speech mode. In Experiment 3, we examined this by

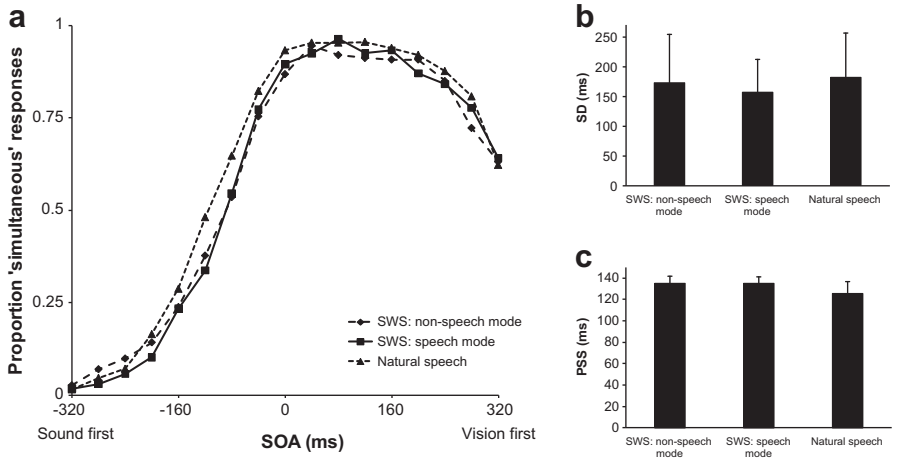

Fig. 2. (a) Mean proportion of "simultaneous" responses as a function of stimulus onset asynchrony (SOA) for sine-wave speech (SWS) in speech mode, SWS in non-speech mode, and natural speech. (b) The standard deviation (SD) of the fitted Gaussian distributions. (c). The point of subjective simultaneity (PSS). Error bars represent 1 Standard Error of the Mean (SEM).

measuring whether our instruction to hear the SWS sound as speech or non-speech did modify the strength of the McGurk effect. We expected that incongruent lipread information would strongly bias the proper identification of an SWS sound if that sound was heard as speech, but the visual influence should be greatly diminished if the sound was heard as non-speech (Tuomainen et al., 2005).

4. Experiment 3

In Experiment 3 we examined whether participants were actually more likely to integrate sound and vision if the sound was heard as speech than non-speech. For that purpose, another stimulus (/tagi/) was recorded of which the audio and video were dubbed in either a congruent or incongruent way on the original sound /tabi/. An incongruent video (e.g., the sound /tabi/ dubbed onto the video of /tagi/) normally hampers proper identification of the sound because lipread information strongly biases sound identification (McGurk & MacDonald, 1976). Here we examined whether the bias by incongruent lipread information was bigger – thus signalling more audiovisual integration – if the SWS sound was heard as speech than non-speech (Tuomainen et al., 2005).

It should be noted that we interpret the bias by incongruent lipread information as an 'integration effect' in the sense that the auditory and visual information streams are merged at a perceptual level. However, there are also alternative interpretations that do not evoke the concept of 'perceptual integration'. For example, a difference between congruent and incongruent lipread information may also reflect interference via response competition (e.g., a perceiver may hear /ba/ and see /da/, and - despite instructions - respond to the visual signal). Our stimuli gave the strong impression of an integrated percept, and we therefore assume that it reflects integration rather than competition. We acknowledge, though, that different measures of perceptual integration would be welcome.

4.1. Participants

Sixty new students (18–45 years, mean 20.6 years) participated.

4.2. Stimuli

The Dutch pseudoword /tagi/ was recorded from the same actor in the same recording session as for the original /tabi/. Both /tabi/, /tagi/, and the possible audiovisual fusions like /tadi/, /tabdi/, or /tabgi/ are all pseudowords not closely related to any real word in Dutch. The two stimuli were presented auditory-only, audiovisual congruent, and audiovisual incongruent (visual /tabi/ paired with auditory /tagi/, and visual /tagi/ paired with auditory /tabi/). The total duration and the onset of the critical consonant of /b/ and /g/ were matched in the two recordings. There were 30 repetitions per stimulus category, all presented in random order and divided across two blocks (ITI = 1.5 s).

4.3. Procedure

In the natural speech and the SWS speech-mode conditions, listeners had to label the stimuli on the basis of whether they had heard /tabi/ or /tagi/, while in the SWS non-speech condition the tokens were labeled as "1" or "2" (see Tuomainen et al., 2005). A fourth control condition was added to check whether there was interference from response labeling, as listeners in non-speech mode might have heard the SWS sounds as speech, but might have difficulties assigning the labels "1" and "2" to the tokens "tabi" and "tagi", respectively. To check whether there was indeed interference from response labeling, listeners in the fourth condition were informed that the SWS sound were actually derived from /tabi/ and /tagi/ (so they were in speech mode), but they had to label them as "1" or "2" (for /tabi/ and /tagi/, respectively). Each participant was allowed to press a third button if the sound did not match any of the two given categories. In this way, we could measure the amount of fusions (e.g., /tadi/) and blends (e.g., /tabgi/) that were expected to occur in the incongruent speech conditions.

To ensure that participants were actually watching the monitor during presentation of the video, they had to detect the occurrence of catch trials (16% of total number of trials). A catch trial contained a small superimposed white spot of 120 ms duration, either between the lips and the nose during the maximal opening of the mouth for AV presentations, or at the same position and at the same time in the A-only presentation. Participants had to refrain from responding upon detecting a catch trial.

To ensure that the SWS sounds were heard as speech in the speech mode condition, participants were presented each natural utterance followed by the corresponding SWS replica for ten times. A training session then followed in which participants learned to discriminate the two sounds. Auditory-only /tabi/ and /tagi/ were presented 10 times per stimulus, and the letters "tabi" or "tagi", or the numbers "1" or "2" for the SWS non-speech condition appeared simultaneously on the monitor. Once listeners were acquainted with the two sounds, they were trained to discriminate the two auditory stimuli using two designated buttons. Feedback was given after each trial. If the accuracy in a block of 32 trails was below criterion (80% correct), a second block was run. A short practice session (containing A-only, AV congruent and AV incongruent trials) preceded the actual experiment to familiarize the participants with the experimental task.

4.4. Results

Participants were virtually flawless in detecting the catch trials (99.7% correct) indicating that they indeed watched the monitor. The proportion of correctly identified auditory tokens was computed for each condition. As shown in Fig. 3, incongruent visual information strongly interfered with proper sound identification for the three groups that heard the sound as speech (SWS: SM; SWS: SM-num. label, and natural speech), but the interference from lipread speech was much smaller for listeners in non-speech mode (SWS: NSM). This was tested by a

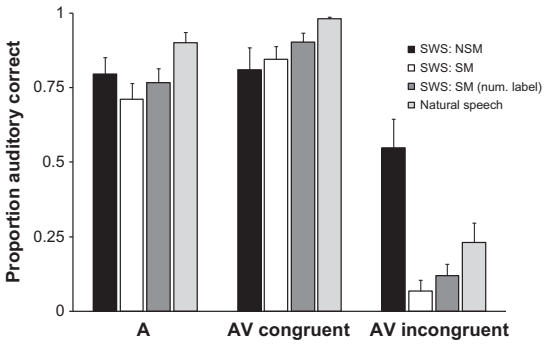

Fig. 3. Mean proportion of correctly identified auditory stimuli for auditory-only (A), audiovisual congruent and audiovisual incongruent presentations. Black and white bars denote auditory identification of sine-wave speech in non-speech mode (SWS: NSM) and speech mode (SWS: SM), respectively. Dark grey bars represent the scores for the SWS speech mode condition in which subjects labeled the sounds as "1" and "2" (SWS: SM-num. label). Light grey bars represent the scores for the natural speech condition. Error bars represent 1 Standard Error of the Mean (SEM).

MANOVA for repeated measures with as within-subject factor Modality (A-only, AV congruent, and AV incongruent) and as between-subject variable Condition.[1] There were main effects of Modality, $F(2, 55) = 166.74$, $p < 0.001$, Condition, $F(3, 56) = 4.15$, $p < 0.05$, and a significant Modality × Condition interaction, $F(6, 112) = 5.26$, $p < 0.001$. Post-hoc test revealed that for the incongruent AV stimuli, the proportion of correctly identified sounds was substantially better for SWS in non-speech mode (mean = .55) than for SWS in speech mode (mean = .07), SWS in speech mode using numerical labels (mean = .12), and natural speech (mean = .23; all p-values <0.05). There was no difference between natural speech, SWS in speech mode and speech mode using numerical labels (all p-values >0.38). These results thus demonstrate that lipread information was more strongly paired with auditory information if the sounds were perceived as speech rather than non-speech.

5. General discussion

Here we demonstrated that perception of audiovisual temporal order is not affected by whether an SWS sound is heard as speech or non-speech. Yet, the higher-order interpretation of the SWS sound as speech or non-speech had a massive effect on the proper identification of its acoustic identity, thus indicating that lipread information was paired with the SWS sound if the sound was perceived as speech, but not so if the sound was perceived as non-speech. Together, these findings demonstrate that judging audiovisual temporal order in speech is *not* affected by whether the auditory and visual streams are paired. This result is quite compelling evidence against the "unity assumption" since it is the first study on intersensory synchrony in which pairing between the auditory and visual streams was manipulated while *all* contributions from low-level stimulus differences were equated.

It is of importance to note that the support for the unity assumption from TOJ tasks has come primarily from studies using audiovisual speech (Vatakis & Spence, 2007, 2008a). Vatakis and Spence (2007, 2008a) reported several experiments in which they demonstrated that the "unity assumption" modulates performance in an audiovisual TOJ task. They used video clips of speakers uttering speech sounds or words that were either gender matched (i.e., a female face presented with a female voice) or else gender mismatched (i.e., a female face presented with a male voice). Also, a video was used that contained pronunciations of matching or mismatching phonemes (e.g., the lip movements of /ba/ together with the /da/ sound). Participants always found it more difficult (higher JNDs) to determine whether the visual lip movements or the auditory speech had been presented first in the matched speech conditions than in the mismatched conditions.

Of notice, though, the disadvantage for the matched conditions could not be replicated with non-speech stimuli. Thus, the same authors did not obtain a difference between matching and mismatching videos of musical and object action events (Vatakis & Spence, 2008a). There was also no difference between matching and mismatching videos of monkey vocalizations, and there was no difference between matching and mismatching human non-speech vocalizations (Vatakis & Spence, 2008a). This led the authors to conclude that while the "unity assumption" can influence the multi-sensory integration of speech, it does not have any such effect on the integration of non-speech stimuli.

This conclusion, though, naturally raises the question of what it is that drives the unity effect: why is audiovisual speech affected by congruency, but not non-speech stimuli? One possibility is that speech represents a "special" class of sensory events that is uniquely related to the articulatory gestures of speech (e.g., Liberman & Mattingly, 1985). Speech may thus be different because it is articulatory or phonetic in nature. The present study speaks to this issue as in Experiment 3 we do indeed show that intersensory pairing of a sound and lipread information is affected by whether the sound is heard as speech. Crucially, though, the pairing in the phonetic domain did not affect judgments of audiovisual temporal order. How to account for that?

One conceivable proposal is that there are two different systems at play – one related to intersensory pairing, the other related to phonetic decisions – and that there is a hierarchical relation between the two. On this view, there is first "intersensory pairing" where it is decided whether two information streams conceivably emerge from the same object/event or not. The pairing is mostly based on the low-level temporal correlation and coincidence between the two information streams (see e.g., Munhall et al., 1996). Only if there is sufficient support for "same object/event", the content of the two information streams is perceived as "synchronous" and subsequently merged at the phonetic level. This notion can, amongst others, explain why the size of the McGurk effect usually correlates well with the size of the temporal window of integration

[1] The MANOVA and subsequent test on arcsine square root transformed scores yielded similar results.

(Conrey & Pisoni, 2006; van Wassenhove et al., 2007). At the neurophysiological level, there is also evidence in support of this notion (Klucharev, Möttönen, & Sams, 2003; Stekelenburg & Vroomen, 2007). In studies using event-related potentials, it has been demonstrated that there are two qualitatively different integrative mechanisms with different underlying time courses. Early audiovisual interactions between speech and lipread information that modulate the auditory-evoked N1 are *un*affected by whether the auditory and visual information are congruent or incongruent (e.g., auditory and visual /a/ vs. auditory /a/ and visual /y/), but the effect crucially depends on the temporal relationship between visual and auditory signals (Vroomen & Stekelenburg, 2010). In contrast the mid-latency and late interactions at P2 are susceptible to informational congruency and possibly indicate multi-sensory integration at the phonetic level. Further testing of this hierarchical notion, though, is certainly needed. For example, one prediction is that the McGurk-effect only emerges if sound and vision are perceived as synchronous, but not if perceived as asynchronous. Some authors, though have reported that – at least on some occasions – an auditory /da/ and a lipread /ba/ are perceived as an integrated event 'bda', whereas the two components of this blend were frequently judged as being nonsimultaneous (Soto-Faraco & Alsius, 2009).

It remains to be explained why sensitivity for temporal order in audiovisual speech is generally low, but improves when the auditory and visual information are incongruent (Vatakis & Spence, 2007). As already mentioned, one speculative interpretation is that different stimulus factors contribute to the perception of audiovisual synchrony in speech and non-speech, and that mismatching information affects them differently. In normally-matched audiovisual continuous speech, there is the continuous temporal correlation between the time-varying characteristics of the auditory and visual streams (Munhall et al., 1996; van Wassenhove et al., 2007) that may induce a "temporal ventriloquist" effect (Martikainen, Kaneko, & Hari, 2005; Morein-Zamir et al., 2003; Vroomen & Keetels, 2010). Temporal ventriloquism may explain why sensitivity for audiovisual temporal order is better for incongruent than congruent audiovisual speech because it is likely that the fine temporal correlation in incongruent speech is disrupted so that small lags in continuous audiovisual speech become unnoticeable.

For discrete events there is no such inherent time-varying correlation between the auditory and visual streams, and perceivers will have to rely primarily on the temporal coincidence of auditory and visual transient onsets. Arguably, the monkey calls used in the study by Vatakis, Ghazanfar, et al. (2008) also contained short transient onsets with almost no visual anticipatory information. Here also, temporal judgments may thus likely be based on the temporal coincidence of the onsets rather than the time-varying co-modulation.

This idea also fits the observation that sensitivity for audiovisual temporal order is worse for multisyllabic words than for syllables (Vatakis & Spence, 2007), likely because the time-varying audiovisual temporal correlation is bigger for multisyllabic words than for monosyllables. In addition, Vatakis and Spence (2006a) showed that more complex natural non-speech events with a fine continuous temporal audiovisual correlation (such as guitar playing) resulted in a higher JND than for events with repetitive discrete actions (smashing a television set with a hammer). Fujisaki and Nishida (2009) also showed that temporal sensitivity was worse for stimuli composed of a repetitive pulse train (high time-varying correlation between the auditory and visual) than for stimuli containing a single pulse. Judging temporal order in audiovisual speech may thus differ from non-speech not because speech is "special", but because speech has a fine temporal correlation between sound and vision that induces temporal ventriloquism, and judging temporal order in audiovisual speech may for that reason thus be difficult.

Temporal ventriloquism may also explain why inversion of the video of a face, music or non-speech event has no effect on sensitivity for temporal order because the temporal structure between audition and vision remains intact if the video is turned upside-down (Vatakis & Spence, 2008b). This is remarkable because inversion of a face has been shown to be more detrimental than to the perception of other types of visual stimuli, as it results in an impairment of configural information processing that leads to slowed and reduced accuracy when performance is tested in face recognition tasks (Yin, 1969), and to a reduced visual impact in the McGurk-effect (Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994; Massaro & Cohen, 1996; Rosenblum, Yakel, & Green, 2000).

Admittedly, more research is needed to further test whether temporal ventriloquism is a cause of the difficulty in judging temporal order in audiovisual speech. The present data, though, suggest that it is the low-level sensory information that determines performance in audiovisual TOJ task rather than the higher-order interpretation of that signal.

References

Bedford, F. L. (1989). Constraints on learning new mappings between perceptual dimensions. *Journal of Experimental Psychology – Human Perception and Performance, 15*, 232–248.

Benjamins, J. S., van der Smagt, M. J., & Verstraten, F. A. J. (2005). The temporal limits of binding sound and colour. *Perception, 34*, 82.

Bertelson, P. (1999). Ventriloquism: A case of crossmodal perceptual grouping. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events* (pp. 347–362). Amsterdam: Elsevier.

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). *Exploring the relation between McGurk interference and ventriloquism*. Paper presented at the third international conference on spoken language processing (ICSLP 94).

Boersma, P., & Weenink, D. (2005). Praat a system doing phonetics by computer, v. 4.3.02 (Computer program) (Electronic Version). <http://www.praat.org/>.

Conrey, B., & Pisoni, D. B. (2003). Detection of auditory–visual asynchrony in speech and nonspeech signals. In *Research on spoken language processing progress report No. 26* (pp. 71–94). Bloomington, IN: Speech Research Laboratory, Indiana University.

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

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

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *Experimental Brain Research, 166*, 455–464.

Fujisaki, W., & Nishida, S. (2009). Audio-tactile superiority over visuo-tactile and audio-visual combinations in the temporal resolution of synchrony perception. *Experimental Brain Research, 198*, 245–259.

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

Klucharev, V., Möttönen, R., & Sams, M. (2003). Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain Research, Cognitive Brain Research, 18*, 65–75.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*, 1–36.

Martikainen, M. H., Kaneko, K., & Hari, R. (2005). Suppressed responses to self-triggered sounds in the human auditory cortex. *Cerebral Cortex, 15*, 299–302.

Massaro, D. W., & Cohen, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics, 58*, 1047–1065.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.

Morein-Zamir, S., Soto-Faraco, S., & Kingstone, A. (2003). Auditory capture of vision: Examining temporal ventriloquism. *Brain Research, Cognitive Brain Research, 17*, 154–163.

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

Radeau, M. (1994). Auditory–visual spatial interaction and modularity. *Cahiers De Psychologie Cognitive – Current Psychology of Cognition, 13*, 3–51.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*, 947–949.

Rosenblum, L. D., Yakel, D. A., & Green, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology – Human Perception and Performance, 26*, 806–819.

Scheier, C. R., Nijhawan, R., & Shimojo, S. (1999). Sound alters visual temporal resolution. *Investigative Ophthalmology & Visual Science, 40*, S792.

Soto-Faraco, S., & Alsius, A. (2009). Deconstructing the McGurk–MacDonald illusion. *Journal of Experimental Psychology – Human Perception and Performance, 35*, 580–587.

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

Stekelenburg, J. J., & Vroomen, J. (2007). Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience, 19*, 1964–1973.

Stevenson, R. A., Altieri, N. A., Kim, S., Pisoni, D. B., & James, T. W. (2010). Neural processing of asynchronous audiovisual speech perception. *Neuroimage, 49*, 3308–3318.

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

van de Par, S., Kohlrausch, A., & Juola, J. F. (2002). *Some methodological aspects for measuring asynchrony detection in audio-visual stimuli*. Paper presented at the proceedings of the forum acusticum.

van der Burg, E., Cass, J., Olivers, C. N. L., Theeuwes, J., & Alais, D. (2010). Efficient visual search from synchronized auditory signals requires transient audiovisual events. *Plos One, 5*, e10664.

van Eijk, R. L. J., Kohlrausch, A., Juola, J. F., & van de 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 auditory–visual 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 music, speech, and object actions. *Brain Research, 1111*, 134–142.

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

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

Vatakis, A., & Spence, C. (2008a). Evaluating the influence of the 'unity assumption' on the temporal perception of realistic audiovisual stimuli. *Acta Psychologica, 127*, 12–23.

Vatakis, A., & Spence, C. (2008b). Investigating the effects of inversion on configural processing with an audiovisual temporal-order judgment task. *Perception, 37*, 143–160.

Vroomen, J., & Baart, M. (2009). Phonetic recalibration only occurs in speech mode. *Cognition, 110*, 254–259.

Vroomen, J., & de Gelder, B. (2004). Temporal ventriloquism: Sound modulates the flash-lag effect. *Journal of Experimental Psychology: Human Perception and Performance, 30*, 513–518.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: A tutorial review. *Attention, Perception & Psychophysics, 72*, 871–884.

Vroomen, J., & Stekelenburg, J. J. (2010). Visual anticipatory information modulates multisensory interactions of artificial audiovisual stimuli. *Journal of Cognitive Neuroscience, 22*, 1583–1596.

Welch, R. B. (1999). Meaning, attention, and the "unity assumption" in the intersensory bias of spatial and temporal perceptions. In G. Aschersleben, T. Bachmann, & J. Müsseler (Eds.), *Cognitive contributions to the perception of spatial and temporal events*. Amsterdam: Elsevier.

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

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141–145.