Brief article

Audio–visual speech perception is special

Jyrki Tuomainen\textsuperscript{a,b,*}, Tobias S. Andersen\textsuperscript{a},
Kaisa Tiippana\textsuperscript{a}, Mikko Sams\textsuperscript{a}

\textsuperscript{a}Laboratory of Computational Engineering, University of Technology,
P.O. Box 3000, FIN-02015 Helsinki, Finland
\textsuperscript{b}Phonetics Lab (Juslenia), University of Turku, FIN-20014, Finland

Received 9 June 2004; accepted 18 October 2004

Abstract

In face-to-face conversation speech is perceived by ear and eye. We studied the prerequisites of audio–visual speech perception by using perceptually ambiguous sine wave replicas of natural speech as auditory stimuli. When the subjects were not aware that the auditory stimuli were speech, they showed only negligible integration of auditory and visual stimuli. When the same subjects learned to perceive the same auditory stimuli as speech, they integrated the auditory and visual stimuli in a similar manner as natural speech. These results demonstrate the existence of a multisensory speech-specific mode of perception.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Audio–visual speech perception; Sine wave speech; Selective attention; Multisensory integration

A crucial question about speech perception is whether speech is perceived as all other sounds (Fowler, 1996; Kuhl, Williams, & Meltzoff, 1991; Massaro, 1998) or whether a specialized mechanism is responsible for coding the acoustic signal into phonetic segments (Repp, 1982). “Speech mode” refers either to a structurally and functionally encapsulated speech module operating selectively on articulatory gestures (Liberman & Mattingly, 1985), or to a perceptual mode focusing on the phonetic cues in the speech signal (Remez, Rubin, Berns, Pardo, & Lang, 1994).

* Corresponding author. Phonetics Lab (Juslenia), University of Turku, FIN-20014, Finland. Fax: +358 2 333 6560.
E-mail address: jyrtuoma@utu.fi (J. Tuomainen).

0022-2860/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2004.01.004
A compelling demonstration of a speech mode was provided by Remez, Rubin, Pisoni, and Carrell (1981) who used time-varying sine wave speech (SWS) replicas of natural speech. SWS stimuli consist of sine waves positioned at the centres of the lowest three or four formant frequencies (i.e. vocal tract resonances) of natural speech. The resulting sine wave replicas lack all other cues typical to natural speech such as regular pulsing of the vocal cords, aperiodicities, and broadband formant structure. Naïve subjects perceived SWS stimuli mainly as non-speech whistles, bleeps or “computer sounds”. When another group of subjects was instructed about the speech-like nature of the SWS stimuli, they could easily assign a linguistic content to the same stimuli.

In face-to-face conversation, speech is perceived by ear and eye. Watching congruent articulatory gestures improves the perception of acoustic speech stimuli degraded by presenting them in noise (Sumby & Pollack, 1954) or by reducing them to sine wave replicas (Remez, Fellowes, Pisoni, Goh, & Rubin, 1998). In some instances, observing talker’s articulatory gestures that are incongruent with the acoustic speech can change the auditory percept, even when the acoustic signal is clear (McGurk & MacDonald, 1976). For example, when subjects see a face articulating /ga/ and are simultaneously presented with an acoustic /ba/, they typically hear /da/. This “McGurk effect” provides an example of multisensory integration where subjects combine the visual articulatory information with the acoustic information in an unexpected manner at a high level of complexity. A non-speech example is the audio–visual integration of the “plucks” and “bows” of cello playing reported by Saldan˜a and Rosenblum (1993). This suggests that not only speech, but also other ecologically valid combinations of auditory and visual stimuli can integrate in a complex manner. Even though audio–visual speech perception has been suggested to provide evidence for a special mode of speech perception (Liberman & Mattingly, 1985), to date there is no convincing empirical evidence showing that this type of integration would be specific to speech.

In this paper we investigate whether subjects’ expectations about the nature of the auditory stimuli has an effect on audio–visual integration. Sine wave replicas of Finnish nonwords /omso/ and /onsol/ were presented to the subjects either alone or dubbed onto a visual display of a congruent or incongruent articulating face. In Experiment 1, in non-speech mode, the subjects were trained to classify the SWS stimuli in two arbitrary categories and were not told about their speech-like nature. In speech mode, the same subjects were trained to perceive the same SWS stimuli as speech. We studied whether subjects integrated the acoustic and visual signals in a similar way in these two modes of perception. Our hypothesis was that if audio–visual speech perception is special, then integration would only occur when the subjects perceived the SWS stimuli as speech. For comparison, natural speech stimuli were also employed. The subjects were required to always report how they heard the auditory-only and audio-visual stimuli. Audio–visual integration was defined here as the amount of visual influence on auditory perception (Calvert, 2001; Stein & Meredith, 1993; Welch & Warren, 1980) although we are aware that this definition may not hold if the mechanism of integration is highly non-linear (Massaro, 1998). Performance was quantified by calculating the percentage of correctly identified auditory part of the stimuli (henceforth “correct identification”). For incongruent audio–visual stimuli, a low percentage of correct identifications would indicate strong integration.
as integration would cause illusory percepts (the McGurk effect). Experiment 2 was
designed to ensure that learning effects could not account for the results of Experiment 1.

1. Experiment 1

1.1. Methods

1.1.1. Subjects

Ten students of the Helsinki University of Technology were studied. All reported
normal hearing and normal or corrected-to-normal vision. None of the subjects had earlier
experience with SWS stimuli. Two subjects were excluded from the subject pool because
they reported perceiving the SWS stimuli as speech before being instructed about their
speech-like nature.

1.1.2. Stimuli

Four auditory stimuli (natural /omso/ and /onso/ and their sine wave replicas) and
digitized video clips of a male face articulating /omso/ and /onso/ were used. These stimuli
were chosen because, for natural speech, incongruent audio–visual combinations of /m/
and /n/ have been shown to produce a strong McGurk effect so that the visual component
modifies the auditory speech percept (MacDonald & McGurk, 1978). In addition, based on
an informal pilot study, inclusion of the fricative /s/ increased the distinctiveness of the
sine wave speech stimuli. The natural speech tokens produced by one of the authors (JT)
were videotaped in a sound-attenuating booth using a condenser microphone and a digital
video camera. The audio channel was transferred to a microcomputer (digitized at
22,050 Hz, 16 bit resolution) and sine wave replicas of both /omso/ and /onso/ were
created by Praat software (Boersma & Weenink, 1992–2002) with a script provided by
Chris Darwin (http://www.biols.susx.ac.uk/home/Chris_Darwin/Praatscripts/SWS). 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.

Four audio–visual stimuli were created for both natural speech and SWS conditions by
dubbing the auditory stimulus to the articulating face using the FAST Studio Purple video-
editing software by replacing the original acoustic utterance with either natural or SWS
audio track: two unedited congruent /omso/ and /onso/ stimuli in which both the face and
the auditory signal were the same, and two incongruent stimuli, in which auditory /onso/
was dubbed to visual /omso/ and auditory /omso/ was dubbed to visual /onso/. In addition,
for a visual-only control task, two visual stimuli of the face articulating /omso/ and /onso/
without accompanying sound were created.

1.1.3. Procedure

The experiment consisted of six tasks, which were always performed in the following
order:

1. Training in non-speech mode. Subjects were taught to categorize the two sine-wave
speech tokens into two non-speech categories without knowledge of the speech-like
nature of the sounds. The subjects were told that they would be hearing two different (perhaps strange sounding) auditory stimuli. They were asked to press a button labelled “1” if they heard stimulus number one (sine wave replica of /omso/), and “2” if they heard stimulus number two (sine wave replica of /onso/). The two sounds were played back several times and on each presentation a correct response code was demonstrated. When the subjects felt that they had learned the correspondence, classification performance was tested by presenting both stimuli 10 times in random order. All subjects learned to classify the stimuli accurately.

2. **SWS in non-speech mode.** SWS tokens were presented alone or audio-visually with a congruent or incongruent visual articulation. Each stimulus was repeated 20 times. Subjects’ task was to focus on the moving mouth of the face displayed on a computer screen and to listen to what was played back in the loudspeakers. Subjects were never told that the mouth movements were actually articulatory gestures, but were only informed that they would see a face with a moving mouth. They were instructed to indicate by a button press whether they heard stimulus “1” or “2”. After the test, subjects were asked questions about the nature of the SWS stimuli to find out if they had spontaneously perceived any phonetic elements in the SWS stimuli. Two subjects reported hearing speech sounds /omso/, /onso/ or /oiso/, and they were excluded from the subject pool.

3. **Natural speech.** The same test as in the second task was administered but now the auditory stimuli consisted of natural tokens of /onso/ and /omso/. Subjects were told to indicate by using the keyboard whether the consonant they heard was /n/, /m/ or something else.

4. **Training in speech mode.** A similar training session as in the first phase in non-speech mode was administered but now the subjects (of which eight were still under the impression that the SWS stimuli were non-speech sounds) were taught to categorize the SWS stimuli as /omso/ and /onso/. Learning was tested by presenting both stimuli 10 times in random order. All subjects learned to categorize the SWS stimuli as /omso/ and /onso/. They were also asked how they heard the stimuli, and all reported that now they perceived them as speech sounds.

5. **SWS in speech mode.** The same test as in the second task was administered but the subjects responded as in the third task.

6. **Visual-only.** Only the articulating face was presented with the instruction to try to speechread what the face was saying. The number of response alternatives was not restricted. As in tasks 3 and 5, /omso/, /onso/ or “something else” were given as examples of responses.

1.2. Results

The responses (percentage of correctly identified auditory part of the stimuli) were subjected to a two-way repeated measures analysis of variance (ANOVA) with two within-subjects factors, Condition with three levels (SWS in non-speech mode vs. SWS in speech mode vs. natural speech) and Stimulus Type with three levels (auditory-only vs. congruent audio–visual vs. incongruent audio–visual). The results, shown in Fig. 1,
revealed the main effects of both Condition ($F(2,14) = 12.922$, $P = 0.001$), due to higher correct identification scores for SWS stimuli in non-speech mode, and Stimulus Type ($F(2,14) = 148.959$, $P < 0.001$), due to lower identification scores for incongruent stimuli, and a significant interaction of the factors ($F(4,28) = 27.958$, $P < 0.001$). The significant interaction effect was followed up by performing one-way ANOVAs separately for the factors Condition and Stimulus Type.

The results of these analyses showed no significant differences between conditions in the auditory-only and congruent stimulus presentations (both $F$’s $< 1$) but a significant main effect in the incongruent stimuli ($F(2,14) = 26.504$, $P < 0.001$). Post hoc $t$-tests showed that this effect was due to the fact that the identification performance with the incongruent SWS stimuli in non-speech mode (84%) was significantly better than that of SWS in speech mode (29%, $t(7) = 4.271$, $P = 0.004$) and natural speech (3%, $t(7) = 24.177$, $P < 0.001$). The identification scores for SWS stimuli in speech mode and natural speech did not differ significantly from each other ($t(7) = 1.769$, $P = 0.120$, n.s.).

Separate comparisons of conditions across stimulus types revealed main effects in all conditions (SWS in non-speech mode: $F(2,14) = 8.739$, $P = 0.003$; SWS in speech mode: $F(2,14) = 26.285$, $P < 0.001$; natural speech: $F(2,14) = 522.901$, $P < 0.001$). In all conditions the pattern was similar: identification of incongruent, but not of congruent stimuli, differed from that of auditory-only baseline stimuli (all $P$’s $< 0.001$ except for SWS stimuli in non-speech mode, $P = 0.012$).

Thus, the results indicate that a strong audio–visual integration effect takes place only when the auditory stimuli are perceived as speech. An integration effect was also observed in non-speech mode, but the magnitude of it was minimal (decrease from 90 to 84%) compared with SWS stimuli in speech mode (decrease from 93 to 29%) and natural stimuli (decrease from 92 to 3%).
2. Experiment 2

In Experiment 1, the different tasks were always performed in the same order, so that
the non-speech mode always preceded speech mode for the SWS stimuli. The reason for
this was that once the subject “enters speech mode” it is impossible to hear the SWS
stimuli as non-speech. However, this procedure might have created a learning effect so that
subjects might have become more used to SWS stimuli. Then at least part of the large
integration effect observed with the incongruent stimuli could have been due to this
learning effect. To control for this, we presented to new subjects the SWS stimuli in speech
mode as a first block, and reasoned that if the subjects showed comparable performance
without lengthened prior exposure to SWS stimuli, then the large integration effects could
not be due to learning. For comparison purposes we also presented natural speech stimuli.

2.1. Methods

2.1.1. Subjects

Thirteen students of the Helsinki University of Technology who did not participate in
Experiment 1 were studied. All had normal hearing and normal or corrected-to-normal
vision. None of the subjects had earlier experience with SWS stimuli.

2.1.2. Stimuli

The same stimulus material was used as in Experiment 1.

2.1.3. Procedure

The experiment consisted of four tasks with the same instructions as in Experiment 1. The
order of the tasks, however, was different from Experiment 1. The tasks were always
performed in the following order:

1. Training in speech mode.
2. SWS in speech mode.
3. Natural speech.
4. Visual-only.

2.2. Results

Fig. 2 shows the results of Experiment 2 which replicate the finding of Experiment 1
that SWS in speech mode and natural speech give similar, low numbers of auditory
responses for incongruent audio–visual stimuli, suggesting similar, strong audiovisual
integration.

The direct comparison of the identification performance with SWS stimuli in speech
mode and with natural stimuli between Experiment 1 and Experiment 2 was done by
performing a three-way ANOVA with Experiment with 2 levels (first vs. second) as a
between-subjects factor, and Condition with two levels (SWS in speech mode vs. natural
speech) and Stimulus Type with three levels (auditory-only vs. congruent vs. incongruent) as within-subjects factors. The results showed a main effect of Stimulus
Type \((F(2,34) = 428.273, P < 0.001)\), due to lower identification scores to incongruent stimuli, and an interaction between Stimulus Condition and Type \((F(2,34) = 8.492, P = 0.001)\) in a similar way as in Experiment 1. Most importantly, there were no main effects of Condition \((F(1,19) = 2.773, P = 0.112, \text{n.s.})\) or Experiment \((F < 1)\), and none of the interactions involving factor Experiment was statistically significant. This pattern of results suggests that the SWS stimuli in speech mode (and natural stimuli) were identified in a similar manner in Experiment 1 and Experiment 2. Accordingly, the large integration effect observed in Experiment 1 is not based on a learning effect due to the order of presentation of the stimulus conditions.

### 3. Discussion

Our results demonstrate that acoustic and visual speech were integrated strongly only when the perceiver interpreted the acoustic stimuli as speech. If the SWS stimuli had always been processed in the same way, the influence of visual speech should have been the same in both speech and non-speech modes. This result does not depend on the amount of practise with listening to SWS stimuli as confirmed by the results obtained in Experiment 2.

We suggest that when SWS stimuli were perceived as non-speech, the acoustic and visual tokens did not form a natural multisensory object, and were processed almost independently. When the SWS stimuli were perceived as speech, the acoustic and visual signals combined naturally to form a coherent phonetic percept (Remez et al., 1998, 1994). We interpret our present findings to be strong evidence for the existence of an audio–visual speech-specific mode of perception.

We have previously shown that visual speech has a greater influence on audio–visual speech perception when subjects pay attention to the talking face (Tiippana, Andersen, & Sams, 2004). Here we propose that attention may also be involved in the current case,
though in quite a different context. It has been proposed that attention may guide which stimulus features are bound to objects during the perceptual process (Treisman & Gelade, 1980). Accordingly, depending on the perceptual mode, a different set of features may be at the focus of attention. When in speech mode, attention may have enhanced processing and binding of those features in our stimuli which form a phonetic object. When the same stimuli were perceived as non-speech, attention may have been focused on some other features (such as a specific frequency band that contained prominent acoustic energy) that could be used to discriminate the stimuli. Those features in the voice or face that are less important to speech perception would not be expected to have a large influence on audio–visual speech perception (see however, Goldinger (1998) and Hietanen, Manninen, Sams, and Surakka (2001) for effects of speaker identity and face configuration on speech perception, and Kamachi, Hill, Lander, and Vatikiotis-Bateson (2003) for showing that the identity of a speaker can be extracted from vision and audition by matching faces to SWS sentences). Indeed, a difference between the spatial locations of the acoustic and visual speech influences only marginally the strength of the McGurk effect (Jones & Munhall, 1997), and the effect also occurs even when a male voice is dubbed onto female face and vice versa (Green, Kuhl, Meltzoff, & Stevens, 1991). The role of the speech mode would thus be to guide attention to speech-specific features both in auditory and visual stimuli, yielding integration only when they provide coherent information about a phonetic object (Massaro, 1998; Remez, 2003; Remez et al., 1998).

Our account can be viewed as an extension of object-based theories of selective attention in vision to the multisensory domain. Duncan (1996) suggests that when a visual object is attended, processing of all features belonging to that object is enhanced, and this enhancement influences all brain areas where relevant visual features are processed. In the present experiment, when subjects perceived the SWS stimuli as speech, attention was focused on phonetic objects. Processing of phonetic objects in the auditory domain may have enhanced processing of the corresponding phonetically relevant visual features, thus yielding strong audio–visual integration. It should be noted that we also observed a small integration effect in non-speech mode, the magnitude of which was minute compared to that in speech mode. One possible explanation is that the effect is due to weak integration of non-speech features of acoustic and visual stimuli (Rosenblum & Fowler, 1991; Saldaña & Rosenblum, 1993). The features that could be integrated in the non-speech mode could be the size of the mouth opening and loudness of the auditory stimuli (Grant & Seitz, 2000; Rosenblum & Fowler, 1991).

In conclusion, our results support the existence of a special speech processing mode, which is operational also in audio–visual speech perception. We suggest that an important component of the speech mode is selective and enhanced processing of those features in the acoustic and visual stimuli that are relevant for phonetic perception. Selectivity and enhancement may be achieved via attentional mechanisms.

Acknowledgements

The research of T.S.A. was supported by the European Union Research Training Network “Multi-modal Human–Computer Interaction”. Financial support from
the Academy of Finland to the Research Centre for Computational Science and Engineering and to MS is also acknowledged. We thank Ms Reetta Korhonen for help in data collection and Riitta Hari (Low Temperature Lab, HUT) for valuable comments on the manuscript.

References


