Visual speech improves the intelligibility of time-expanded auditory speech

Akihiro Tanaka, Shuichi Sakamoto, Komi Tsumura and Yôiti Suzuki

This study investigated the effects of intermodal timing differences and speed differences on word intelligibility of auditory–visual speech. Words were presented under visual-only, auditory-only, and auditory–visual conditions. Two types of auditory–visual conditions were used: asynchronous and expansion conditions. In the asynchronous conditions, the audio lag was 0–400 ms. In the expansion conditions, the auditory signal was time expanded (0–400 ms), whereas the visual signal was kept at the original speed. Results showed that word intelligibility was higher in the auditory–visual conditions than in the auditory-only condition. The results of auditory–visual benefit revealed that the benefit at the end of words declined as the amount of time expansion increased, although it did not decline in the asynchronous conditions. *NeuroReport* 20:473–477 © 2009 Wolters Kluwer Health | Lippincott Williams & Wilkins.

NeuroReport 2009, **20**:473–477

Keywords: asynchrony, audio-visual, intelligibility, multimodal, speech perception, speech-rate conversion

Research Institute of Electrical Communication, Tohoku University, Sendai, Japan

Correspondence to Dr Akihiro Tanaka, Cognitive and Affective Neuroscience Laboratory, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands Tel: +31 13 466 3644; fax: +31 13 466 2067; e-mail: a.tanaka@uvt.nl

Received 14 November 2008 **accepted** 15 December 2008

Introduction

In naturalistic settings, listeners use the visual information from the speaker's mouth for speech perception as well as auditory speech. Lipreading is especially useful in noisy environments [1–3] and is one example of multisensory integration. Multisensory integration requires temporal coordination of signals from multiple sensory modalities [4]. These signals, however, do not need to be precisely synchronous to be integrated at least in the speech domain. Earlier studies have investigated a temporal window of auditory–visual speech integration during which auditory and visual speech signals are integrated, and have suggested that the size of the temporal window is around 200 ms when an auditory signal lags a visual signal [5–10].

In these studies, the amount of asynchrony between auditory and visual speech signals was constant from the onset to the offset of the signals. This is ecologically valid because the lag between multiple modalities is constant in a naturalistic setting. However, recent advances in signal processing technology have brought about a situation in which such constancy is broken down. In a speech-rate conversion system for broadcasting [11], the rate of an auditory speech signal is slowed down even though that of the visual signal remains unchanged. In this case, speech expansion brings a new type of asynchronization between auditory and visual speech signals; the lag increases toward the end of the stimulus.

Another study from our group showed that visual information is effective in enhancing the speech intelligibility of a time-expanded auditory signal [12]. However, the temporal property of integration between the visual signal and the time-expanded auditory signal has not been investigated in a comparable manner. What is the difference between the auditory–visual integration in the case of intermodal timing difference and speed difference? In this study, we examined the effect of intermodal speed difference on auditory–visual spoken word recognition and compared it with that of intermodal timing difference.

Methods

Participants

Participants were 10 undergraduate and graduate students (19.8 ± 1.0 years). All had normal or corrected-to-normal vision and had normal hearing (mean hearing level 3.6 ± 3.2 dB). All were native Japanese speakers. This experiment was conducted with written informed consent of each participant. The ethics committee of the Research Institute of Electrical Communication of Tohoku University formally approved the experiments.

Stimuli

We used 20 minimal pairs (40 words) of Japanese words. All words had four morae and the same pitch-accent type (low–high–high–high pitch for respective morae). Paired words differed by only one consonant and required different mouth movements. Five minimal pairs (10 words) were used for each of the four mora positions. Manipulating the mora position in which mouth movement differs, we can examine the changes in visual contributions from the first to the last mora of the words.

0959-4965 © 2009 Wolters Kluwer Health | Lippincott Williams & Wilkins DOI: 10.1097/WNR.0b013e3283279ae8

Words were selected from a database of lexical properties of Japanese [13]. Mean familiarity, rated between 1 (low) and 7 (high), of the words used was 5.18. Average familiarity of the ten words consisting of five minimal pairs for the list of each mora position was matched within the range between 5.06 and 5.28. The difference in the familiarity of the paired words was less than 1.0.

A trained female speaker pronounced the words in an anechoic room. The utterance was recorded using a DV camera (AG-DVX100A; Panasonic Inc., Osaka, Japan). Auditory speech was collected using a half-inch condenser microphone (Type 4165; Brüel and Kjaer, Naerum, Denmark) and digitally recorded on the DV. The mean speech rate was 6.9 morae/s (583 ms average duration). Auditory speech was digitized at 48 kHz, with a 16-bit quantization resolution. Visual signals were digitally recorded with a frame rate of 29.97 frames/s (1 frame=33.33 ms). All auditory speech was presented in pink noise to avoid the ceiling effect and floor effect in the word intelligibility. The signal-to-noise ratio was 10 dB.

For the expansion conditions, auditory speech signals were analysed and resynthesized to change the duration of the words using the STRAIGHT algorithm [14]. The auditory signals were time expanded 0, 100, 200, 300, or 400 ms longer than the original. Synthesized speech signals were combined with the visual signal so that the onset of the utterance was synchronous.

Experimental conditions

Seventeen experimental conditions were used in total. They can be classified as five asynchronous conditions, 10 expansion conditions, and two control conditions. In the asynchronous conditions, the auditory speech signal lagged the visual speech signal (audio delay: 0, 100, 200, 300, or 400 ms). The amount of audio delay corresponds to the exact timing in the original recording. The rate of presentation between modalities was kept constant. The auditory and visual signals themselves were unchanged, except for the timing relation of these signals. In the expansion conditions, the auditory speech was time expanded (expansion: 0, 100, 200, 300, or 400 ms). Therefore, the rate of presentation was slower in the auditory modality than it was in the visual modality. Speech signals were presented either through the auditory modality (expansion-A conditions) or through the auditory and visual modalities [expansion-audio-visual (AV) conditions]. In the expansion-AV conditions, auditory and visual speech signals were synchronous at the onset of the stimuli and asynchronous at the offset of the stimuli, according to the amount of the expansion. In addition to the asynchronous and the expansion conditions, two control conditions were used. The original auditory speech and the original visual speech were presented in the auditory-only and the visual-only control conditions, respectively.

Procedure

Each of the 17 sessions was assigned to one of the 17 experimental conditions. In each session, all 40 words were presented six times (i.e. 240 trials per condition). The intertrial interval was 6 s. The order of the sessions and that of the words within each session were randomized.

Participants were seated facing a display in a soundproof room designed according to the International Telecommunications Union Recommendation BS.1116-1 criteria. Sounds were presented through a pair of loudspeakers (N-803; Bowers & Wilkins, Worthing, West Sussex, UK) at 60 dB (A-weighted equivalent continuous sound pressure level) by a DV tape-deck through an amplifier. Visual signals were presented on a 42-inch flat plasma display (TH-42PWD4; Panasonic Inc.). The horizontal width of the talker's mouth was approximately 4.5° of the visual angle.

For each trial, an incomplete word was written on the paper (e.g. mi-zu-a-__). A participant listened to and/or looked at the stimulus. Participants were instructed to fill in the blank after the word presentation. No feedback was provided, and no training was given before testing.

Results

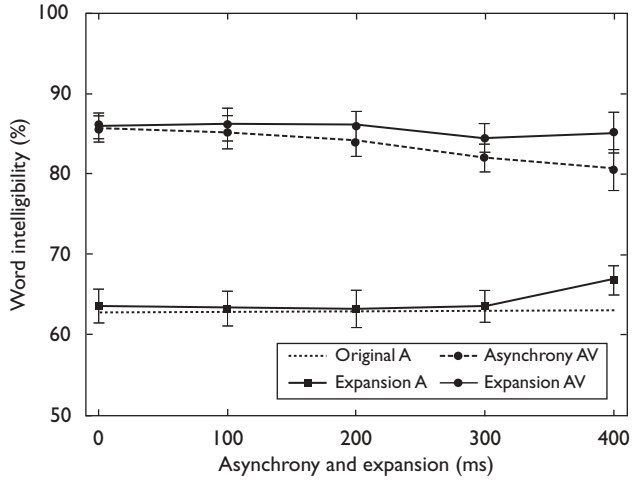
Word intelligibility

The percentages of correct responses were averaged for each participant and each condition. An overall average in each condition (i.e. word intelligibility) was calculated across participants (see Fig. 1).

In the auditory-only control condition, word intelligibility was 62.9%. In the visual-only control condition, word intelligibility was 47.0%. Overall, intelligibility was higher in asynchronous conditions (83.5% in average across five conditions) than in the auditory-only control condition. Intelligibility in asynchronous conditions declined as the audio delay increased. A one-way analysis of variance (ANOVA) with repeated measure revealed that the main effect of asynchrony was significant [$F(4,36)=8.83$, $P<0.01$]. Post-hoc analysis (Dunnett's t-test) showed that words were less intelligible when the audio delay was 300 and 400 ms than when it was 0 ms ($P<0.01$).

In expansion-A conditions, the intelligibility was slightly higher when speech was expanded 400 ms longer. As in the asynchronous conditions, the intelligibility was higher in the expansion-AV conditions than in the expansion-A conditions. A two-way repeated-measures ANOVA with presentation modality and time expansion revealed a significant main effect of the presentation modality

Fig. 1



Word intelligibility in all of the conditions. Dotted line with filled circles shows word intelligibility as a function of audio delay in the asynchronous audio-visual (AV) conditions. Solid lines show word intelligibility as a function of the amount of time expansion in the expansion-A (filled squares) and expansion-AV (filled circles) conditions. The horizontal dotted line (no markers) shows word intelligibility in the auditory-only control condition.

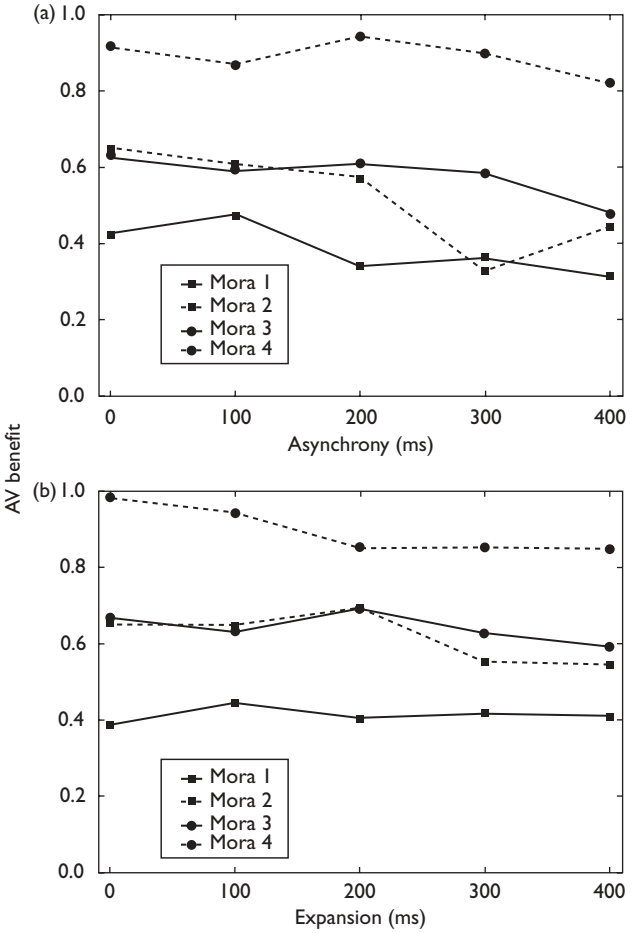
[F(1,9)=254.84, $P < 0.01$] and an interaction between the two factors [F(4,36)=3.06, $P < 0.05$]. In the expansion-A conditions, words were more intelligible when the amount of time expansion was 400 ms than when it was 0 ms ($P < 0.05$, Dunnett's *t*-test). In contrast to the asynchronous conditions, no significant difference existed among the expansion-AV conditions.

Audio-visual benefit

To evaluate the visual benefit at each mora position, we used another index of visual benefit on speech intelligibility: the AV benefit [2]. The AV benefit is a measure of the visual contribution to speech perception; it is calculated using the following formula: $AV\ benefit = (AV - A)/(100 - A)$, where A represents the intelligibility score in an auditory-only condition, and AV represents that of an auditory–visual condition. In asynchronous conditions, A corresponds to the intelligibility score in the auditory-only control condition, and AV corresponds to each asynchronous AV condition; in expansion conditions, A corresponds to the intelligibility score in each expansion-A condition, and AV corresponds to the corresponding expansion-AV conditions. The AV benefit ranges between 0 (no visual benefit) and 1 (maximum visual benefit). The AV benefit of the 0 ms condition of each mora was compared with those of other time conditions for each mora using Dunnett's multiple comparisons to evaluate the effects of timing differences and speed differences for each mora.

Figure 2a shows AV benefits in the asynchronous conditions. For morae 1–3, the AV benefit decreased as the amount of asynchrony increased. In contrast, at mora 4, such a tendency was not very clear. A two-way repeated-measures ANOVA with the amount of asynchrony and mora position revealed significant main effects of the amount of asynchrony [F(4,36)=10.08, $P < 0.01$] and mora position [F(3,27)=20.87, $P < 0.01$]. An interaction was also found between the two factors [F(12,108)=5.30, $P < 0.01$]. Post-hoc analyses (Dunnett's *t*-test) showed significant decreases at morae 1–3 when the asynchrony was 400 ms. At mora 2, there was also a significant decrease in AV benefit when the asynchrony was 300 ms. In contrast, at mora 4, no significant decrease was found in any asynchronous condition.

Fig. 2



Audio-visual (AV) benefits in (a) asynchronous conditions and (b) expansion conditions. Each of the four lines shows the AV benefits at each mora position.

Figure 2b shows AV benefits in the expansion conditions. In the expansion AV conditions, the onset of the mora 1 was synchronous. The audio lag was at maximum at the offset of mora 4. Therefore, the actual amount of audio lag at the onset of each mora was smaller than the amount of time expansion. At mora 4, the AV benefit declined as

the amount of time expansion increased. A two-way repeated measures ANOVA with the amount of time expansion and mora position revealed significant main effects of the amount of time expansion [$F(4,36)=3.03$, $P<0.05$] and mora position [$F(3,27)=14.68$, $P<0.01$]. Interaction was also found between the two factors [$F(12,108)=1.92$, $P<0.05$]. Post-hoc analyses (Dunnett's *t*-test) showed no significant decline at morae 1–3 in any time-expansion condition. In contrast, at mora 4, a significant decline was found when the amount of time expansion was equal to or greater than 200 ms. These results contrast to those of asynchronous conditions.

Discussion

The purpose of this study was to examine the effect of the intermodal speed difference on auditory–visual spoken word recognition and to compare it with that of the intermodal timing difference. The results revealed that visual information contributes to the improvement of intelligibility of time-expanded words, although the rate of presentation was different between visual speech and time-expanded auditory speech. In a naturalistic setting, the auditory signal usually lags the visual signal because sound velocity is much slower than light velocity. Thus, the difference in the timing of presentation depends on the distance between the event and the observer, although the brain uses information about distance that is supplied by the visual system to calibrate simultaneity [15]. The difference in the rate of presentation between modalities is, however, negligible. Thus, the situation caused by time expansion of auditory speech signals never happens in a naturalistic setting. Nonetheless, participants could use the visual information to recognize what the speaker had said. This setting resembles that of the experiments by Recanzone [16]. He examined the auditory influences of visual rate perception and demonstrated that the perceived rate of flickering can be modulated by the rate of the sound. The results of this study offer a new finding that shows multimodal interaction between multimodal stimuli at different rates of presentation.

Intelligibility decreased as the audio delay increased in the asynchronous conditions. However, the size of the temporal window revealed in this study was larger than that of previous studies. In the study by Grant and Greenberg [7], intelligibility declined to almost the same level as the auditory-only condition when the auditory signal lagged the visual signal by 400 ms. In this study, intelligibility was still greater than in the auditory-only condition when the audio lag was 400 ms. We can point out several factors that might contribute to the different results between previous studies and this study (e.g. task). In this experiment, participants were not under a time restriction. Therefore, it was possible to rehearse the word using visual and auditory memory traces of the word presented and to guess what the speaker had said. Thus, the apparently wide temporal window obtained in this study might reflect the higher cognitive processes as well as the temporal limitation of auditory–visual integration.

Our results are consistent with the idea that a fixed-time lag is important for auditory and visual speech signals to be integrated. In the expansion-AV conditions, the AV benefit at mora 4 decreased as the amount of time expansion increased. This contrasts to the result in the asynchronous condition in which the AV benefit at mora 4 did not decline as the audio delay increased. Van Wassenhove *et al.* [8] proposed that the correlation between auditory and visual speech signals plays an important role in auditory–visual speech integration. As there is a correspondence between the area of mouth opening and the amplitude of the speech signal [7], there must be a high cross-correlation between auditory and visual speech signals when the timing between modalities is constant (i.e. asynchrony conditions). In contrast, because the constant timing lag between mouth opening and amplitude of speech signal is broken down by the time-expansion procedure, the cross-correlation should be lower in the expansion-AV conditions. The results support the possibility that the correlation between auditory and visual speech signals plays an important role in auditory–visual integration between lagged signals. However, there remain other possibilities (e.g. naturalness) that can account for the different results between asynchronous and expansion conditions. This should be tested in a future study.

The results of this study have practical importance for the design of a multimodal speech-rate conversion system. Imai *et al.* [11] showed that older observers sensed no unnaturalness when they observed a TV news program in which the phrases of the auditory speech were slowed down and pauses between the phrases were deleted in the amount equal to the amount of time expansion even though the visual signal was kept at the original speed. This study supports the usefulness of a multimodal speech rate conversion system and extends the finding in terms of the intelligibility measure.

Conclusion

Word intelligibility was higher when the visual speech signal was presented along with the time-expanded auditory speech signal. The results of AV benefits revealed that AV benefits at mora 4 decreased as the speed difference increased, although they did not decrease as the timing difference increased.

Acknowledgements

A part of this work was supported by a Grant-in-Aid for Specially Promoted Research No. 19001004 from the

Ministry of Education, Culture, Sports, Science and Technology, Japan. The authors would like to thank Hideki Kawahara for permission to use the STRAIGHT vocoding method. The authors would also like to thank Atsushi Imai and Tohru Takagi at the NHK Science and Technical Research Laboratories for their helpful comments on our research.

References

1 MacLeod A, Summerfield Q. A procedure for measuring auditory and audio-visual speech–reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br J Audiol* 1990; **24**:29–43.
2 Sumby W, Pollack I. Visual contribution to speech intelligibility in noise. *J Acoust Soc Am* 1954; **26**:212–215.
3 Ross LA, Saint-Amour D, Leavitt VM, Javitt DC, Foxe JJ. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cereb Cortex* 2007; **17**:1147–1153.
4 Stein BE, Meredith MA. *The merging of the senses*. Cambridge MA: The MIT Press; 1993.
5 Conrey B, Pisoni DB. Auditory–visual speech perception and synchrony detection for speech and nonspeech signals. *J Acoust Soc Am* 2006; **119**:4065–4073.
6 Dixon N, Spitz L, The detection of audiovisual desynchrony. *Perception* 1980; **9**:719–721.
7 Grant KW, Greenberg S. Speech intelligibility derived from asynchronous processing of auditory–visual information. *Proc Int Conf Audit–Vis Speech Process* 2001. pp. 132–137.
8 Van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory–visual speech perception. *Neuropsychologia* 2007; **45**: 598–607.
9 Massaro D, Cohen MM, Smeele PMT. Perception of asynchronous and conflicting visual and auditory speech. *J Acoust Soc Am* 1996; **100**: 1777–1786.
10 Munhall KG, Gribble P, Sacco L, Ward M. Temporal constraints on the McGurk effect. *Percept Psychophys* 1996; **58**:351–362.
11 Imai A, Ikezawa R, Seiyama N, Nakamura A, Takagi T, Miyasaka E, *et al*. An adaptive speech rate conversion method for news programs without accumulating time delay. *J IEICE* 2000; **83A**:935–945.
12 Sakamoto S, Tanaka A, Tsumura K, Suzuki Y. Effect of speed difference between time–expanded speech and talker's moving image on word or sentence intelligibility. *Proc Int Conf Audit–Vis Speech Process* 2007. pp. 238–242.
13 Amano S, Kondo T. *Nihongo–no Goi–Tokusei (Lexical properties of Japanese)*. Sanseido: Tokyo; 1999.
14 Kawahara H, Masuda–Katsuse I, de Cheveigne A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. *Speech Commun* 1999; **27**:187–207.
15 Sugita Y, Suzuki Y. Implicit estimation of sound-arrival time. *Nature* 2003; **421**:911.
16 Recanzone GH. Auditory influences on visual temporal rate perception. *J Neurophysiol* 2003; **89**:1078–1093.