Aging effect on audio-visual speech asynchrony perception: comparison of time-expanded speech and a moving image of a talker's face

Shuichi Sakamoto[1], *Akihiro Tanaka*[2], *Shun Numahata*[1], *Atsushi Imai*[3], *Tohru Takagi*[3], *Yôiti Suzuki*[1]

[1]Research Institute of Electrical Communication and
Graduate School of Information Sciences, Tohoku University, Sendai, Japan
[2]Graduate School of Humanities and Sociology, The University of Tokyo, Tokyo, Japan
[3]NHK Science and Technical Research Laboratories, Tokyo, Japan
[1]{saka,numa,yoh}@ais.riec.tohoku.ac.jp, [2]a.tanaka@uvt.nl [3]{imai.a-dy,takagi.t-fo}@nhk.or.jp

Abstract

In this study, we measured detection and tolerance thresholds of auditory-visual asynchrony between time-expanded speech and a moving image of the talker's face for older adults. During experiments, words were presented under two conditions: asynchrony by time-expanded speech (expansion condition, EXP) and simple timing shift (asynchronous condition, ASYN). We used 16 Japanese shorter words (four morae) and 20 Japanese longer words (seven or eight morae). For EXP, auditory speech signals were expanded and combined with the visual signals so that the onset of the utterance was synchronous. For ASYN, the auditory speech signal was simply lagged behind the visual speech signal. Detection and tolerance thresholds for auditory-visual asynchrony obtained for older adults was higher than these obtained for younger adults, which suggests that older adults are tolerant of audio-visual asynchrony.

Index Terms: lip-reading, auditory-visual asynchrony, time-expanded speech, detection and tolerance thresholds, aging

1 Introduction

"Speaking slowly" is a good way to speak to older adults, especially under noisy conditions. Based on this benefit, a speech rate conversion technique has been proposed and applied to broadcasting systems [1, 2]. In this system, however, only the speech sound was expanded. Therefore, the synchrony between the speech sound and moving image of talker's face is broken. Consequently, improved comprehension attributable to the effect of lip-reading might be decreased. For that reason, it is important to investigate how people integrate speech sounds and the moving image of the talker's face.

Our previous studies revealed the effect of speed difference between time-expanded speech and a moving image of a talker's face on spoken word recognition [3, 4]. Moreover, the experimental results suggest that detection and tolerance thresholds of auditory-visual asynchrony between time-expanded speech and a moving image of the talker's face might depend on the ratio of the expansion rate to word duration [5]. However, the results of earlier studies show that the effect of time-expanded speech on word intelligibility was different between younger and older adults. Moreover, when speech rate conversion system was evaluated subjectively by older adults, many did not mind the asynchrony between auditory and visual signals, which implies that older adults are tolerant of this asynchrony.

For this study, we measured the detection and tolerance thresholds of older adults' auditory-visual asynchrony between time-expanded speech and a moving image of the talker's face and compared the results with the results obtained by younger adults.

2 Measurement of detection and tolerance thresholds for shorter words

2.1 Methods

Almost all experimental methods were indentical to those used for measurement of detection and torelance thresholds of younger adults in previous study [5]. In this section, we will describe the experimental methods in more detail.

2.1.1 Participants

Participants were 10 older adults (70.3 ± 2.7 years old). All had normal or corrected-to-normal vision and normal hearing (mean hearing level: 18.4 ± 3.8 dB). All were native Japanese speakers.

2.1.2 Stimuli

As shorter words, we used 16 Japanese four-mora words selected from a database of lexical properties of Japanese [6]. Mean familiarity (rated between 1 (low) and 7 (high)) of the words was 4.88. All words were of the same pitch-accent type (type 0 (flat) accent). A trained female speaker pronounced the words in an anechoic room. The utterances were recorded using a digital video camera (AG-DVX100A; Panasonic Ltd.). Auditory speech was recorded digitally using a 1/2-inch condenser microphone (Type 4165; Bruel & Kjaer) and a DV camera. The mean speech rate was 6.83 morae/s. Auditory speech was digitized at 48 kHz, with 16-bit quantization resolution. Visual signals were digitally recorded at 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. The SNR was determined for each participant (-4.7 ± 1.1 dB). For expansion conditions (EXP), auditory speech signals were analyzed and resynthesized to change the duration of the words using STRAIGHT [7]. The auditory signals were time-expanded so that they were between 120 ms and 840 ms longer than the original signals. Synthesized speech signals were combined with the visual signals using a nonlinear editing system (Avid Xpress Pro/Mojo; Avid Technology Inc.) so that the onset of the utterance was synchronous. Consequently, 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. For asynchronous conditions (ASYN), the auditory speech signal was simply lagged behind the visual speech signal between 120 ms and 840 ms.

2.1.3 *Experimental condition*

In expansion conditions (EXP), the auditory speech was time-expanded. Therefore, the rate of presentation was slower in auditory modality than in the visual modality. In asynchronous conditions (ASYN), the auditory speech signal lagged the visual speech signal. The rate of presentation was maintained as constant between modalities, leading to a fixed audiovisual time lag. The auditory and visual signals themselves were unchanged, except for the timing relation of these signals.

Table 1 shows experimental conditions for the determination of detection and tolerance thresholds. In the detection threshold measurement, 11 experimental conditions were used: 6 EXP, 6 ASYN, and 1 control condition. Regarding tolerance threshold measurements, 15 experimental conditions were used: 7 EXP, 7 ASYN, and 1 control condition. The control condition in all experimental conditions was 0-ms expansion.

2.1.4 *Procedure*

Figure 1 portrays the experimental setup. Participants were seated facing a display in a soundproof room designed according to the ITU-R BS.1116-1 criteria. Auditory signals were presented through a pair of loudspeakers (N-803; Bowers & Wilkins) located on right and left sides of the display. The sound pressure level of the speech signal was determined for each participant individually according to their own comfortable level of listening (60–62 dB, A-weighted equivalent continuous sound pressure level) and presented from a DV tape deck (DSR-30; Sony Corp.) through an amplifier (AX-9; Yamaha Corp.). Visual signals were presented on a 42-inch flat plasma display (TH-42PWD4; Panasonic Inc.). The horizontal width of the talker's mouth was about 4.5° of the visual angle.

Every word was presented five times at each condition. Consequently, the total amount of the presentation was 624 (16 words $\times$ 3 times $\times$ 13 conditions) for measurement of the detection threshold and 720 (16 words$\times$ 3 times $\times$ 15 conditions) for measurement of tolerance threshold. Each measurement consisted of six sessions (208 trials or 240 trials each). In each trial, the inter-trial interval was 5 s. The order of the sessions and that of words within each session were randomized. No feedback was provided.

Participants were asked whether or not they were able to detect the lag between auditory and visual signals in the detection threshold measurement. They were asked whether or not they were able to tolerate the lag in the tolerance threshold measurement. A cumulative distribution function was fitted to the data and detection and tolerance thresholds were defined as 50% of the proportion of the "detection" response and that of the "tolerance" response.

Table 1: Experimental conditions for detection and tolerance threshold measurements (shorter words)

	Cond.	Expansion/Auditory delay [ms]
Detection Threshold	EXP	120 240 360 480 600 720
	ASYN	120 240 360 480 600 720
Tolerance Threshold	EXP	120 240 360 480 600 720 840
	ASYN	120 240 360 480 600 720 840

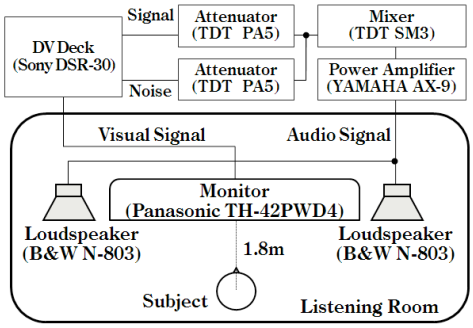


Figure 1: Experimental setup

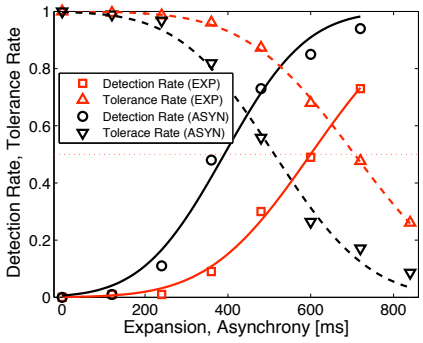


Figure 2: Detection and tolerance rates for shorter words

2.1.5 *Results*

Figure 2 portrays the detection and tolerance rates in all conditions. Calculated detection thresholds were 599 ms for EXP and 391 ms for ASYN. Calculated tolerance thresholds were 673 ms for EXP and 488 ms for ASYN. These thresholds were slightly higher in the expansion condition than in the asynchronous condition. Two-way repeated measures ANOVA with stimulus type (EXP or ASYN) and measurement type (detection or tolerance threshold) revealed a significant main effect of stimulus type ($F(1,9) = 64.88, p < .01$) and measurement type ($F(1,9) = 18.88, p < .01$).

3 Measurement of detection and tolerance thresholds for longer words

3.1 Methods

Almost all methods including participants were identical to those used for measurement of detection and tolerance thresholds for shorter words, except as described below.

3.1.1 *Stimuli*

As longer words, we used 20 Japanese words. Ten consisted of seven morae; the remaining ten consisted of eight morae. These words were selected from the same database. Mean familiarity of

the words used was 4.80. All words were of the same pitch-accent type (type 0 (flat) accent). The speaker and equipment of recording were the same as those used in the previous experiment. The mean speech rate was 7.16 morae/s. All auditory speech was presented in pink noise. The SNR was decided for each participant (-1.9 ± 1.1 dB). For expansion conditions (EXP), the auditory signals were time-expanded so that they were between 160 ms and 1,120 ms longer than the original signals. For asynchronous conditions (ASYN), the auditory speech signal was simply lagged behind the visual speech signal between 160 ms and 960 ms.

3.1.2 Experimental condition

Table 2 shows experimental conditions used for determination of the detection and tolerance thresholds. For detection threshold measurement, 12 experimental conditions were used: 6 EXP, 5 ASYN, and 1 control condition. In the tolerance threshold measurement, 13 experimental conditions were used: 6 EXP, 6 ASYN, and 1 control condition. The control condition in all the experimental conditions was 0-ms expansion.

3.1.3 Procedure

Every word was presented three times at each condition. Consequently, the total amount of presentation was, respectively, 720 (20 words $\times$ 3 times $\times$ 12 conditions) for measurement of the detection threshold and 780 (20 words $\times$ 3 times $\times$ 13 conditions) for measurement of the tolerance threshold. Each measurement consisted of six sessions (240 trials or 280 trials each). In each trial, the inter-trial interval was 5 s. The order of the sessions and that of words within each session were randomized. No feedback was provided.

3.2 Results

Figure 3 shows the detection and the tolerance rates in all conditions. Calculated detection thresholds were 763 ms for EXP and 379 ms for ASYN. Calculated tolerance thresholds were 914 ms for EXP and 513 ms for ASYN. These thresholds were higher in the expansion condition than in the asynchronous condition. A two-way repeated measures ANOVA with stimulus type (EXP or ASYN) and measurement type (detection or tolerance threshold) revealed a significant main effect of stimulus type ($F(1,9) = 70.50, p < .01$) and measurement type ($F(1,9) = 12.64, p < .01$). However, the interaction between stimulus type and measurement type was not statistically significant ($F(1,9) = 0.27, n.s.$).

4 Discussion

We proceeded to a combined analysis between two experiments to investigate the difference of thresholds between shorter and

Table 2: Experimental conditions for detection and tolerance threshold measurements (longer words)

	Cond.	Expansion/Auditory delay [ms]
Detection Threshold	EXP	160 320 480 640 800 960
	ASYN	160 320 480 640 800
Tolerance Threshold	EXP	320 480 640 800 960 1120
	ASYN	160 320 480 640 800 960

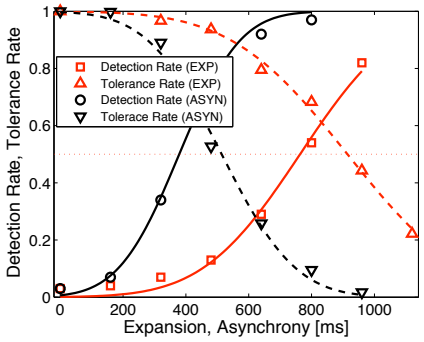
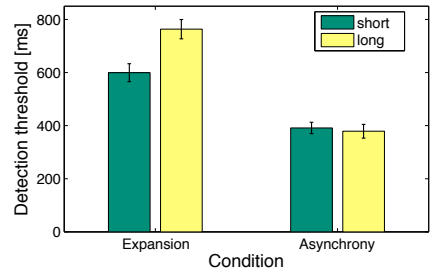


Figure 3: Detection and tolerance rates for longer words

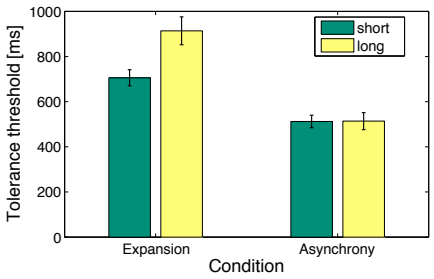
longer words. Figure 4 shows the difference of detection and tolerance thresholds between shorter and longer words. A three-way ANOVA with word length (short or long), stimulus type (EXP or ASYN), and measurement type (detection or tolerance threshold) revealed a significant interaction between word length and stimulus type ($F(1,9) = 28.64, p < .01$). The main effect of measurement type was also statistically significant ($F(1,9) = 26.36, p < .01$). The simple main effect of word length was statistically significant only in EXP condition (EXP: $F(1,18) = 51.24, p < .01$, ASYN: $F(1,18) = 0.00, n.s.$). This result is contrastive to the result in our previous study that word length affected both EXP and ASYN condition of younger adults [5].

For detailed analysis of the aging effect to detection and tolerance thresholds, the results of these experiments were compared with the results of our previous experiment, which were performed with younger adults [5]. Figure 5 shows the difference of detection and tolerance thresholds between younger and older adults. In all conditions, thresholds obtained by older adults are higher than those obtained by younger adults. A three-way ANOVA with word length (short or long), stimulus type (EXP or ASYN) and age (younger or older) was applied to the detection or tolerance threshold. The results revealed a significant interaction between the stimulus type and age (detection threshold: $F(1,18) = 36.01, p < .01$, tolerance threshold: $F(1,18) = 20.68, p < .01$). The interaction between word length and stimulus type was also statistically significant (detection threshold: $F(1,18) = 36.01, p < .01$, tolerance threshold: $F(1,18) = 54.41, p < .01$). The simple main effect of age was statistically significant in all stimulus types ($p < .01$). These results suggest that older adults have difficulty perceiving audio-visual asynchrony. In consequence, older adults are more tolerant of this asynchrony than younger adults. This result might suggest that older adults are less sensitive to auditory-visual lag than younger adults.

The comparison of age group indicated that the effect of word length was smaller in older adults. Especially, we observed no effect of word length in ASYN condition of older adults. However, the results of our previous study suggest that thresholds might depend on the ratio of the expansion rate to word duration for younger adults [5]. These two facts imply an existance of two different mechanisms for detection of the audio-visual asynchrony: one depends on the amount of absolute lag, the other depends on the whole length of speech signal. Moreover, the dominant mechanism might differ between younger and older adults.

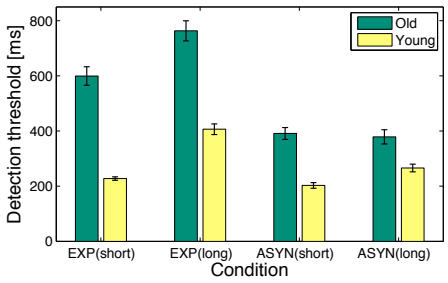


(a) Detection threshold

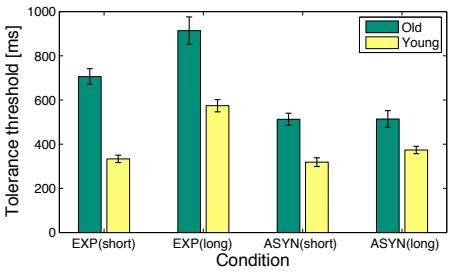


(b) Tolerance threshold

Figure 4: Detection and tolerance thresholds between shorter and longer words



(a) Detection threshold



(b) Tolerance threshold

Figure 5: Aging effect to detection and tolerance thresholds

5 Conclusions

In this study, we measured detection and tolerance thresholds of auditory-visual asynchrony between time-expanded speech and a moving image of the talker's face for older adults. The results indicate that detection and tolerance thresholds for auditory-visual asynchrony obtained by older adults were higher than that obtained for younger adults. These results suggest that older adults are tolerant of audio-visual asynchrony. Thie knowledge is important to apply the speech-rate conversion technique to the universal communication system.

6 Acknowledgements

This work was supported by a Grant-in-Aid from MEXT Japan for Specially Promoted Research No. 19001004 and for Young Scientists (B) No. 20700193. The authors wish to thank Dr. Hideki Kawahara for permission to use the STRAIGHT vocoding method.

References

[1] E. Miyasaka, A. Imai, N. Seiyama, T. Takagi, and A. Nakamura, "A new technology to compensate degeneration of hearing intelligibility for elderly individuals –development of a portable real-time speech rate conversion system–," in *Proceedings of the Third Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan*, 1996, pp. 267–272.

[2] A. Imai, R. Ikezawa, N. Seiyama, A. Nakamura, T. Takagi, E. Miyasaka, and K. Nakabayashi, "An adaptive speech rate conversion method for news programs without accumulating time delay," *The Journal of the Institute of Electronics, Information, and Communication Engineers (in Japanese with English figure captions)*, vol. 83-A, pp. 935–945, 2000.

[3] S. Sakamoto, A. Tanaka, K. Tsumura, and Y. Suzuki, "Effect of speed difference between time-expanded speech and talker's moving image on word or sentence intelligibility," in *Proceedings of International Conference on Auditory-Visual Speech Processing 2007 (AVSP2007)*, 2007, pp. 238–242.

[4] A. Tanaka, S. Sakamoto, K. Tsumura, and Y. Suzuki, "Visual speech improves the intelligibility of time-expanded auditory speech," *NeuroReport*, vol. 20, pp. 473–474, 2009.

[5] S. Sakamoto, A. Tanaka, S. Numahata, A. Imai, T. Takagi, and Y. Suzuki, "Effect of audio-visual asynchrony between time-expanded speech and a moving image of a talker's face on detection and tolerance thresholds," in *Proceedings of International Conference on Auditory-Visual Speech Processing 2008 (AVSP2008)*, 2008, pp. 79–82.

[6] S. Amano and T. Kondo, *Nihongo-no Goi-Tokusei (Lexical properties of Japanese) I.* Sanseido, 1999.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.