J Multimodal User Interfaces (2008) 2: 199–203
DOI 10.1007/s12193-009-0018-4

ORIGINAL PAPER

Effect of speed difference between time-expanded speech and moving image of talker's face on word intelligibility

Shuichi Sakamoto · Akihiro Tanaka · Komi Tsumura · Yôiti Suzuki

Received: 31 August 2008 / Accepted: 13 July 2009 / Published online: 1 August 2009

Abstract This study investigated effects of asynchrony between speech signal and moving image of talker's face induced by time-expansion of the speech signal on speech intelligibility. Word intelligibility test was performed to younger listeners. Japanese 4-mora words were uttered by a female speaker. Each word was processed with STRAIGHT software to expand the speech signal by from 0 to 400 ms. These signals were combined with moving image of talker's face which was kept at original speed. This test was performed under three conditions: visual-only, auditory-only, and auditory-visual (AV) condition. Results showed that intelligibility scores under AV condition were statistically higher than those under auditory-only condition even when the speech signal was expanded by 400 ms. These results suggest that moving image of talker's face is effective to enhance speech intelligibility if the lag between the speech signal and moving image of talker's face does not exceed 400 ms.

Keywords Time-expanded speech · Moving image of talker's face · Word intelligibillity · Audio-visual interaction · Lip-reading

S. Sakamoto (✉) · K. Tsumura · Y. Suzuki
Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University,
2-1-1 Katahira, Aoba-ku, Sendai, Miyagi 980-8999, Japan
e-mail: saka@ais.riec.tohoku.ac.jp

A. Tanaka
Department of Psychology, Tilburg University, Warandelaan 2,
PO Box 90153, 5000 LE Tilburg, The Netherlands

1 Introduction

"Speaking slowly" and "moving mouth cleary" are very good way to talk to older adults, especially under noisy condition. The effects of them have been addressed by many studies. Tanaka *et al.* showed that intelligibility increased in younger and older adults when the length of speech signal was expanded [1]. The cue of moving image of talker's face is usually called "lip-reading" information, and this cue improves speech intelligibility for normal hearing listeners under a low signal-to-noise ratio (S/N) condition [2]. Based on these knowledge, a speech rate conversion technique has been proposed and applied to broadcasting systems [3, 4].

In this system [3, 4], speech rate conversion is realized by slowing down "only" the speech signal without accumulating a time delay. In particular, the total duration is kept constant by deleting the pauses between phrases to an amount equal to the duration of the speech expanded. For example, when the speech signal is expanded by 200 ms longer than the original speech duration, pause duration is deleted to 200 ms shorter than the original pause length. The results of subjective experiments showed the effectiveness of this system.

However, there exists an asynchrony between auditory and visual speech signals in this system [3, 4]. 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 [5–13]. In these studies, the amount of asynchrony between auditory and visual speech signals was constant from the onset to the offset of the signals. However, in the multimodal speech-rate conversion system [3, 4], the audio lag grows progressively toward the end, whereas the

auditory and visual signals are synchronized at the beginning because only auditory signal was time-expanded. In this situation, a visual advantage may no longer occur in spoken language comprehension. No criterion exists for the maximum length of expansion to use visual information effectively. For that reason, it is important to investigate how people integrate speech sounds and moving image of talker's face at different presentation rates.

In this study, we examined the effect of speed difference between time-expanded speech and moving image of speaker's face on word intelligibility. If human can integrate these signals, the performances would be higher in audio-visual conditions than in auditory-only conditions. Also if integration depends on the amount of speed difference, there would be an interaction between the amount of speech expansion and presentation modality (i.e., auditory-only or audiovisual). In contrast, the speed difference might cause even a negative effect. In that case, the performance would be lower in audio-visual conditions than in auditory-only conditions. The results of this study have practical importance for the design of a multimodal speech-rate conversion system and are expected to give certain criteria about how long the length of speech signal can be expanded without degradation of the effect of visual information in such a speech-rate conversion system.

2 Word intelligibility test

2.1 Participants

Participants were 12 undergraduate and graduate students (22.7 ± 1.0 y). All had normal or corrected-to-normal vision and had normal hearing. All were native Japanese speakers.

2.2 Stimuli

We used Familiarity-controlled word lists 2003 (FW03) as stimuli [14]. The FW03 consists of 20 lists of 50 words in four word-familiarity ranks (i.e., 4000 words in all). All words had four morae and the same pitch-accent type (low-high-high-high pitch for respective morae). Mora is a unit of sound used in phonology that determines a syllable weight. Japanese is a language famous for its moraic qualities; it uses morae as the basis of the sound system rather than syllables. In this experiment, we used word lists within the range of middle-high word familiarity, i.e., between 5.5 to 4.0. For example, gu-n-ba-i (sumo referee's fan), ko-wa-i-ro (impersonation), and so on.

A female speaker repronounced the words in an anechoic room. The utterance was recorded using a DV camera (AG-DVX100A; Panasonic Inc.). Auditory speech was collected using a 1/2 inch condenser microphone (Type 4165; B&K) and digitally recorded on the DV. The mean speech rate was 7.1 mora/s (563 ms in average duration). Auditory speech was digitized at 48 kHz, with 16-bit amplitude resolution. Visual signals were recorded with a digitization rate of 29.97 frames/s (1 frame = 33.33 ms). Pink noise was used as the noise signal. Pink noise is a signal with a frequency spectrum such that the power spectral density is proportional to the reciprocal of the frequency and each octave carries an equal amount of power. All the auditory speech was presented in noise. The S/N was 0 dB.

For expansion conditions, auditory speech signals were analyzed and resynthesized to change the duration of the words using STRAIGHT [15]. As we wrote in Introduction, the size of a temporal window of auditory-visual speech integration is around 200 ms when an auditory signal lags a visual signal. Moreover, Grant and Greenberg [10] indicated that intelligibility declined to an almost identical level as the auditory-only condition when the auditory signal lagged the visual signal by 400 ms. Therefore, the maximum length of expansion was decided as 400 ms and the auditory signals were time-expanded 0, 50, 100, 150, 200, 250, 300, 350, or 400 ms longer than the original. The synthesized speech signal was combined with visual signal so that the onset of the utterance be 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.

2.3 Experimental conditions

In all, 21 experimental conditions were used. Auditory speech signal was time-expanded (expansion: 0, 50, 100, 150, 200, 250, 300, 350, or 400 ms). Therefore, the rate of auditory speech signal was slower than that of moving image of talker's face. Speech signals were presented either through auditory modality ("auditory-only condition") or through auditory and visual modalities ("auditory-visual condition"). In addition to these conditions, three control conditions were applied. Original auditory and visual speech was presented in the "ORG(= original)(auditory-visual)" condition. Original auditory speech was presented in the "ORG (auditory-only)" condition. Original visual speech was presented in the visual-only condition. Which word list was presented in which experimental condition was counterbalanced among participants.

2.4 Procedure

In each of the 21 experimental conditions, 50 words were presented. The conditions included nine auditory-only conditions, nine auditory-visual conditions, and three control conditions. Therefore, the total of presented words in all conditions was 1,050. Each of five sessions consisted of 210

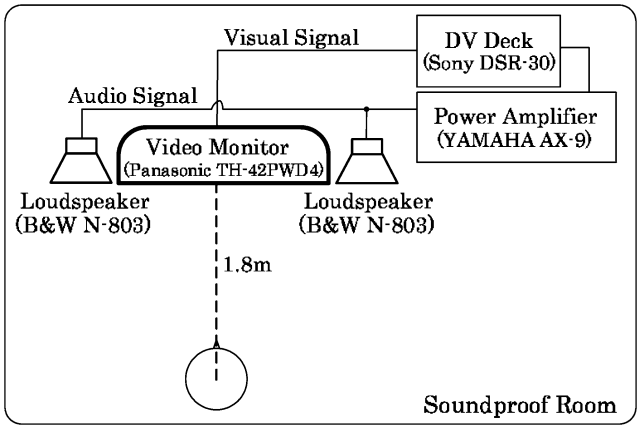

Fig. 1 Experimental setup of word intelligibility test

words. The inter-trial-interval was 6 s. The order of the sessions and that of words within sessions were randomized.

Figure 1 shows the experimental setup. Participants were seated facing a display in a sound-proof room. Sounds were presented through a pair of loudspeakers (N-803; B&W) at 60 dB (A-weighted equivalent continuous sound pressure level) using a DV deck through an amplifier. Visual signals were presented on a 42-inch display (TH-42PWD4; Panasonic Inc.). The horizontal width of the mouth was about 4.5 degree of visual angle.

For each trial, the participant listened to and/or looked at the stimulus. All participants were instructed to write down the words as they heard them on the answer sheet after presentation of a word. No feedback was provided and no training was given prior to testing.

3 Results

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.

Figure 2 shows word intelligibility as a function of the amount of time-expansion. Errorbar expresses standard error. In all expansion conditions, the word intelligibility scores in the auditory-visual condition are higher than that in the auditory-only condition in spite of the asynchrony of auditory and visual stimuli.

A two-way repeated measures Analysis of variance (ANOVA) [16], in which presentation modality and time-expansion were set as factors, revealed a significant main effect of the presentation modality ($F(1, 11) = 128.6$, $p < 0.01$) and time expansion ($F(8, 88) = 2.53$, $p < 0.05$). An interaction between these two factors was not statistically significant ($F(8, 88) = 1.86$, n.s.). In the auditory-only condition, no significant difference existed between 0 ms and other time-expansion conditions. In contrast to the auditory-only condition, words were more intelligible in the auditory-visual condition when the amount of time-expansion was 150 and 250 ms than when it was 0 ms ($p < 0.05$, Dunnett's t-test [16]).

To evaluate the visual benefit at time expansion, we used another index of visual benefit on speech intelligibility: the AV benefit [17]. The AV benefit is a measure of the visual contribution to speech perception; it is calculated using the following formula:

$$AV\ benefit = (AV - A)/(100 - A), \tag{1}$$

in which A represents the intelligibility score in an auditory-only condition and AV represents that an auditory-visual condition. The AV benefit ranges between 0 (no visual benefit) and 1 (maximum visual benefit).

Figure 3 shows AV benefits. In all expansion conditions, the AV benefit is higher than 0. Therefore, the effectiveness of visual information is observed under all conditions. However, no systematic tendency prevails among the conditions. A one-way repeated measures ANOVA with a factor of the amount of time-expansion revealed no significant

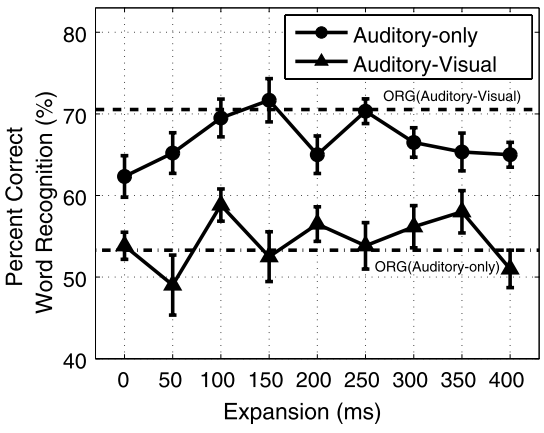

Fig. 2 Word intelligibility as a function of the amount of time expansion (S/N = 0 dB)

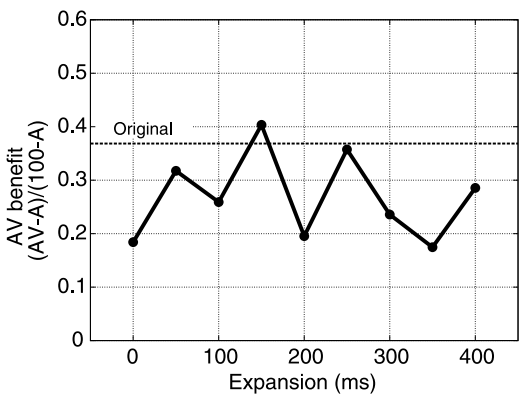

Fig. 3 AV benefit of word intelligibility

main effect of the amount of time-expansion ($F(8, 88) = 1.75$, n.s.).

4 Discussion

In word intelligibility test, the rate of auditory speech signal was much slower than that of moving image of talker's face, especially for the 400 ms time-expansion condition. Despite differences of the presentation rates between auditory and visual speech signals, the word intelligibility score in the auditory-visual condition was significantly higher than that in the auditory-only condition. Moreover, AV benefits are higher than 0 in all conditions. These facts suggest that visual information might be effective for word intelligibility even if the auditory speech signal is delayed 400 ms from the visual speech signal, although significant difference compared with 0 ms expansion was observed only in 150 ms and 250 ms expansion in auditory-visual condition. This lag is greater than that obtained by Grant and his colleagues [7]. In their experiments, auditory speech signal lagged visual speech signal and the rate of presentation was kept constant between modalities, leading to a fixed audiovisual time lag. They reported that sentence intelligibility decreases when the time lag between auditory and visual speech signals is greater than 200 ms [7]. One difference between Grant's results and our results is the synchrony at the beginning of stimuli. In our experiment, auditory and visual speech signals were synchronous at the onset of the stimuli, although the lag between auditory and visual speech signals increased gradually according to the degree of the expansion. This synchronicity at the onset and gradual change of the lag might be related to the improvement of word intelligibility.

In auditory-only condition, word intelligibility did not decrease even when the auditory signals were time-expanded 400 ms longer than the original. Similar experiments were performed by Nejime and Moore [18]. However, their results somewhat differed from ours. They reported that word intelligibility slightly decreases as the speech rate decreases. A difference between the studies is the speech rate of stimuli. While the slowest speech rate used in Nejime and Moore's experiment was about 3.8 mora/s, that used in our experiment was 4.2 mora/s. The rates used in our study slightly faster but the difference is not large, and thus this would not seem us a reason of the different results. A possible reason may be the softwares used to change the speech rate. In the present study, speech rates were converted by STRAIGHT. STRAIGHT is well known algorithm which can process speech sounds with high sound quality. For example, the Nitech HMM-Based Speech Synthesis System [19] is using STRAIGHT for processing speech and evaluated as the good speech processing system for its high quality of processed sounds. Moreover, STRAIGHT won the first place among four synthetic vocal systems in the blind listening test conducted by RENCON'04 [20]. On the other hand, the sound quality of the software used by Nejime and Moore is not clear. Moreover, they applied their waveform expansion algorithm only to segments of signals whose power exceeded a certain threshold. Segments whose power was below the threshold were passed without any manipulation, resulting in nonuniform time expansion. This synthesized speech sound would differ from that produced when a person actually speaks more slowly. Therefore, the synthesized signal generated by Nejime's method would degrade sound quality worse than that generated by STRAIGHT when the expansion rate becomes large, resulting in the decrease of word intelligibility as the time-expansion became long.

As mentioned before, speech rate conversion technique has been applied to broadcasting systems and the effectiveness of these systems was reported [3, 4]. However, there is no research which indicates the effectiveness of them in terms of intelligibility. Our research shows the effectiveness of these systems where only the auditory speech signal is expanded. The results suggest that presentation of the moving image of talker's face is effective to enhance speech intelligibility if the lag between the speech signal and moving image of talker's face does not exceed 400 ms. However, when we apply these findings to such systems, we have to consider various factors in addition to the effect of the lag between moving images of talker's face and speech signal.

Firstly, it is important to consider the effect of pause duration and which point we should decide as the beginning point of speech signal in continuous speech signals. We are conducting a study to answer these questions. Tanaka *et al.* [1] investigated the relationship between speech expansion and pause deletion using sentence materials. The results revealed that speech intelligibility was not decreased when pause duration was decreased by 200 ms as a consequence of 200 ms-expansion of speech signal. By combining the findings in Tanaka *et al.* and the current research, we can predict the usefulness of visual information even when we use audiovisual sentence materials. However, it is very important and interesting to investigate in a single study the relationship among the effects of moving images of talker's face, speech expansion, and pause duration on perception of time-expanded speech signal.

Secondly, we should consider the effect of non-linear expansion. In this study, linear time-expanded speech signal was combined with moving image of talker's face. However, the results of previous research [21] addressed the efficacy of a nonuniform time-expansion algorithm for improvement of speech recognition in older adults. In this research, each speech frame was classified into silence, unvoiced consonant, voiced consonant, or vowel and only unvoiced consonant was modified. It should be investigated whether the results of our paper can be applied when moving image

of talker's face was combined with such nonuniform time-expansion speech signal.

Thirdly, we should decide the limit of expansion rate to get visual information effectively. As described in this paper, the speech signal was time expanded within the range of 400 ms. It is unclear how long the length of speech signal can be expanded without degradation of the effect of visual information over 400 ms or whether the effect of visual information differs in mora position. We presume that if the length of expansion becomes too long, the effect of visual information would decrease.

Finally, the effect of cognitive load [22] should be estimated to apply this result to older adults, because age-related cognitive changes may affect the ability to efficiently process speech. We consider these topics as a subject for future work.

5 Conclusions

This study investigated effects, on word intelligibility, of asynchronicity of a speech signal and moving image of talker's face induced by time-expansion of the speech signal.

The results of word intelligibility suggest that moving image of talker's face is effective for enhancing speech intelligibility if the lag between the speech signal and moving image of talker's face does not exceed 400 ms.

Acknowledgements This work was supported by a Grant-in-Aid for Specially Promoted Research No. 19001004 from MEXT Japan. The authors would like to thank Dr. Hideki Kawahara for permission to use the STRAIGHT vocoding method. The authors would also like to thank the members of the NHK Science and Technical Research Laboratories for their helpful comments on our research.

References

1. Tanaka A, Sakamoto S, Suzuki Y (2005) Effects of speech-rate and pause duration on sentence intelligibility in younger and older normal-hearing listeners. In: Proceedings of the 149th meeting of the acoustical society of America, 5aSC4, p 2604
2. Erber NP (1969) Interaction of audition and vision in the recognition of oral speech stimuli. J Speech Hear Res 12:423–425
3. Imai A, Ikezawa R, Seiyama N, Nakamura A, Takagi T, Miyasaka E, Nakabayashi K (2000) An adaptive speech rate conversion method for news programs without accumulating time delay. J Inst Electron Inf Commun Eng 83-A:935–945 (in Japanese with English figure captions)
4. Miyasaka E, Imai A, Seiyama N, Takagi T, Nakamura A (1996) 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 ASA and ASJ third joint meeting, 4aEA5, pp 267–272
5. McGrath M, Summerfield Q (1985) Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. J Acoust Soc Am 77:678–685
6. Pandey PC, Kunov H, Abel SM (1986) Disruptive effects of auditory signal delay on speech perception with lipreading. J Auditory Res 26:27–41
7. Grant KW, Seitz PF (1998) Measures of auditory-visual integration in nonsense syllables and sentences. J Acoust Soc Am 104:2438–2450
8. Conrey B, Pisoni DB (2006) Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. J Acoust Soc Am 119:4065–4073
9. Dixon N, Spitz L (1980) The detection of audiovisual desynchrony. Perception 9:719–721
10. Grant KW, Greenberg S (2001) Speech intelligibility derived from asynchronous processing of auditory-visual information. In: Proceedings of international conference of auditory-visual speech process, pp 132–137
11. Van Wassenhove V, Grant KW, Poeppel D (2007) Temporal window of integration in auditory-visual speech perception. Neuropsychologia 45:598–607
12. Massaro D, Cohen MM, Smeele PMT (1996) Perception of asynchronous and conflicting visual and auditory speech. J Acoust Soc Am 100:1777–1786
13. Munhall KG, Gribble P, Sacco L, Ward M (1996) Temporal constraints on the McGurk effect. Percept Psychophys 58:351–362
14. Sakamoto S, Suzuki Y, Amano S, Ozawa K, Kondo T, Sone S (1998) New lists for word intelligibility test based on word familiarity and phonetic balance. J Acoust Soc Jpn 54:842–849 (in Japanese)
15. Kawahara H, Masuda-Katsuse I, de Cheveigne A (1999) Restructuring speech representations using pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction. Speech Commun 27:187–207
16. Howell DC (2002) Statistical methods for psychology fifth edition. Duxbury, USA
17. Sumby WH, Pollack I (1954) Visual contribution to speech intelligibility in noise. J Acoust Soc Am 26:212–215
18. Nejime Y, Moore BCJ (1998) Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. J Acoust Soc Am 103:572–576
19. Zen H, Toda T, Nakamura M, Tokuda K (2007) Details of the Nitech HMM-based speech synthesis system for the blizzard challenge 2005. IEICE Trans Inf Syst 90:325–333
20. Noike K, Hiraga R, Hashida M, Hirata K, Katayose H (2005) A report of NIME04 Rencon: Listening contest to evaluate performance rendering systems. In: Proceedings of the 19th Annual conference of the Japanese society for artificial intelligence, vol 2B3-07 (in Japanese)
21. Vaughan NE, Furukawa I, Balasingam N, Mortz M, Fausti SA (2002) Time-expanded speech and speech recognition in older adults. J Rehabil Res Dev 39(5):559–566
22. Pandzic IS, Ostermann J, Millen D (1999) User evaluation: synthetic talking faces for interactive services. Vis Comput 15:330–340