TECHNICAL REPORT

# Time-reversed reverberation yields lower speech recognition rate by human and machine

Takayuki Arai*

*Department of Information and Communication Sciences, Sophia University,*
*7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan*

(*Received 28 May 2012, Accepted for publication 27 September 2012*)

**Abstract:** We first compared a speech signal with two reverberations, normal reverberation and its time-reversed version, that have the same modulation transfer function. Results showed that intelligibility of speech with the time-reversed reverberation was significantly less than that with the normal reverberation. We then compared the results of human speech recognition (HSR) with those of automatic speech recognition (ASR) to see whether a similar tendency could be observed in both cases. Results showed the similar asymmetry in ASR, but we found the HSR was more tolerant even though reverberation becomes longer. Finally, we discussed factors of asymmetric temporal properties in speech production and perception that current speech recognizers do not have.

**Keywords:** Time-reversed reverberation, Human speech recognition, Automatic speech recognition, Modulation spectrum

**PACS number:** 43.71.Es, 43.55.Hy, 43.70.Mn [doi:10.1250/ast.34.142]

## 1. INTRODUCTION

Houtgast and Steeneken showed that the modulation transfer function (MTF) can predict the degradation of intelligibility of speech due to noise and reverberation [1]. The speech transmission index (STI) is a common predictor of intelligibility of speech. Many studies have been done in several fields on the MTF and STI, such as architectural acoustics and speech perception. Drullman *et al.* reported the effect of temporal filtering of the time trajectories of the spectral envelope on the intelligibility of reconstructed speech [2,3]. Their results showed that the low- and high-modulation frequencies of the magnitude spectrum are not essential for the intelligibility of speech. Arai *et al.* reported that the modulation frequencies of 1 to 16 Hz in the modulation spectrum are important for speech perception [4,5]. Thus, the MTF and the modulation spectrum itself have been used as predictors of intelligibility of speech. This concept is further applied in automatic speech recognition as a part of the frontend processing for robust recognition, such as RASTA processing [6], which passes modulation components between 1 and 12 Hz unattenuated and suppresses the components that change more slowly or quickly than the modulation frequency range, and modulation filtering [7].

Griesinger proposed an objective measurement for estimating intelligibility of reverberant speech with syllable counting by focusing on each syllable onset [8]. For estimating intelligibility of speech in a room, the STI is widely used. However, the STI is based on the magnitude modulation spectrum of temporal envelopes of speech, so the STI yields the same result as “time-reversed reverberation” even though a speech signal with time-reversed reverberation is less intelligible than a speech signal with non-time-reversed reverberation. Thus, Griesinger pointed out that the STI is not a perfect predictor due to this discrepancy. Unlike with the STI, the method of counting syllable onsets yields different predictions for normal and time-reversed reverberations. In other words, syllable onsets are weighted more heavily with normal reverberation but less heavily with time-reversed reverberation. Griesinger also pointed out that this counting method well matches subjective human evaluation [8].

Although some have discussed asymmetric temporal properties in speech in previous studies, such as Griesinger [8], speech recognition with time-reversed reverberation has not been fully discussed besides Arai [9] and Longworth-Reed *et al.* [10]. In this paper, we review Arai [9] and further discuss the asymmetric temporal properties in speech production and perception. Furthermore, we also compare the results of human speech recognition (HSR) with those of automatic speech

*e-mail: arai@sophia.ac.jp

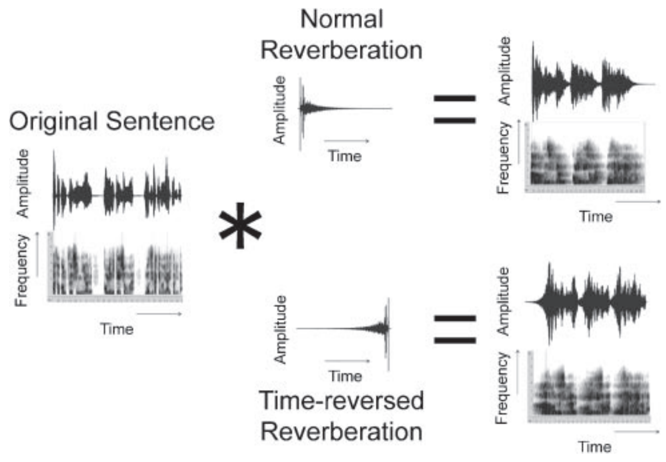

**Fig. 1** Two reverberant sentences with normal reverberation and time-reversed reverberation (adapted from [9]).

recognition (ASR) to see whether there is a mechanism for human speech perception to compensate for reverberation that current speech recognizers do not have.

## 2. EXPERIMENTS

### 2.1. Human Speech Recognition

We conducted a perceptual experiment to test whether speech with time-reversed reverberation is less intelligible than speech with normal reverberation.

2.1.1. Speech samples

The original speech samples used in the experiment were selected from 1,000 phoneme-balanced Japanese sentences (NTT-AT). First, we selected sentences 4–5 seconds long out of 1,000 uttered by speaker MYK; then we selected 36 sentences that contained no proper names or difficult words with low familiarity. Finally, 12 and 24 sentences were grouped as the training and test sets, respectively. The number of morae within a sentence ranged from 25 to 33 (the average was approximately 28).

We conducted a perceptual experiment under artificial reverberant environments achieved by convolving speech samples with impulse responses. The reverberation times (T60) of the three impulse responses we used were 1.0, 1.5, and 2.0 s. These impulse responses were created from a single impulse response (measured at the Higashi-Yamato City Hamming Hall) by multiplying an exponential decay as in a previous study [11]. The time-reversed version of an impulse response was obtained by simply reversing the time axis of the impulse response. This created six impulse responses in total. The reverberant speech stimuli were obtained by convolving the original speech samples with the six impulse responses as shown in Fig. 1. Finally, the training set had 72 sentences and the test set 144. The sampling frequency of all speech samples was 16 kHz.

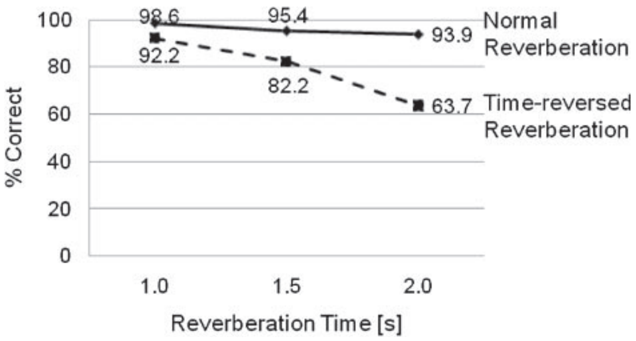

**Fig. 2** Correct rate for the HSR experiment.

2.1.2. Participants

Twenty-four young participants with normal hearing (16 males and 8 females, 19 to 24 years old, 21.8 years old on average) participated in the experiment. All were native speakers of Japanese.

2.1.3. Procedure

The experiment was conducted in a soundproof room. Stimuli were presented diotically through headphones (STAX SR-303) connected to a computer. The sound level was adjusted to each listener's comfort level during a training session prior to the experiment. A stimulus was presented in each trial and the listeners were instructed to write down what they heard on an answer sheet. The experiment was carried out at each listener's pace.

For each listener, 12 stimuli originally from 12 different sentences were presented in the training session, and 24 stimuli originally from 24 different sentences were presented in the test session. The combinations of sentences and impulse responses were counter-balanced among participants in the test session.

For each trial, the participant listened to the same stimulus three times. The participant was asked to write down an answer in kana orthography and was allowed to make corrections after the stimulus presentations.

2.1.4. Experimental results

Figure 2 shows the mean percent of sentence intelligibility of the perceptual experiment for the six reverberant conditions. The solid line shows the results for the normal reverberation, and the dashed line the time-reversed reverberation. The sentence intelligibility was measured as the percentage of correctly identified morae within each sentence.

As expected, the intelligibility of speech decreased as T60 became longer, and it was lower with the time-reversed reverberation than with the normal reverberation. An ANOVA was carried out with the three reverberant conditions and two time directions of the impulse response (normal and time-reversed) as repeated variables and the mean scores as the dependent variable. Results showed that the mean scores significantly differed across reverberation

$[F(1,23) = 149.074,\ MSe = 74.815,\ p < 0.001]$, and the scores were 1.0 s (95.4%), 1.5 s (88.8%), and 2.0 s (78.8%) in descending order. The mean scores were also significantly higher for the normal condition (96.0%) than for the time-reversed condition (79.4%) $[F(1,23) = 149.074,\ MSe = 66.378,\ p < 0.01]$. The interaction between the reverberant conditions and the time directions of the impulse response was also significant $[F(1,23) = 60.876,\ MSe = 56.363,\ p < 0.001]$. For this interaction, we tested the simple main effect among the reverberant conditions for each time direction of the impulse response. In the normal reverberation condition, the simple main effect was significant $[F(1,23) = 18.852, MSe = 13.665, p < 0.001]$; the mean scores were 1.0 s (98.6%), 1.5 s (95.4%), and 2.0 s (93.9%) in descending order. In the time-reversed reverberation condition, the simple main effect was significant $[F(1,23) = 74.626, MSe = 130.917, p < 0.001]$; the mean scores were 1.0 s (92.2%), 1.5 s (82.2%), and 2.0 s (63.7%) in descending order.

Thus, we showed that the intelligibility of speech decreased as T60 became longer, and it also decreased more with the time-reversed reverberation than the normal reverberation. Furthermore, the scores of stimuli with the time-reversed reverberation decreased more rapidly than those of the normal reverberation.

### 2.2. Automatic Speech Recognition

We conducted an ASR experiment to compare their results with the ones of the HSR experiment and to see whether we could obtain similar tendencies from both experiments.

2.2.1. Speech samples

The original speech samples used in the ASR experiment were the same 24 sentences used in the test set of the HSR. We conducted a pilot experiment under the same reverberant environments as used in the HSR. The recognition rates turned out to range between 3 and 12%, and it showed the floor effect and no difference was observed between the two reverberant conditions: normal and time-reversed reverberations. Therefore, we used much shorter reverberation times: 0.1, 0.3, and 0.5 s. The impulse responses were created in the same way as described in Sect. 2.1.1. The reverberant speech samples were obtained by convolving the original speech samples with the new six impulse responses; the ASR experiment had 144 new sentences.

2.2.2. Procedure

The ASR engine that we used was the Julius system (rev. 4.1.2), two-pass large vocabulary continuous speech recognition decoder software [12]. We used the acoustic and language models in the Julius Dictation Toolkit. For the language model, 60,000 vocabulary size and 3-gram model were used. For the acoustic model, Gaussian mixture HMMs and triphone model were used. The speech features were 25 dimensional MFCCs (12th order cepstrum + delta cepstrum + energy).

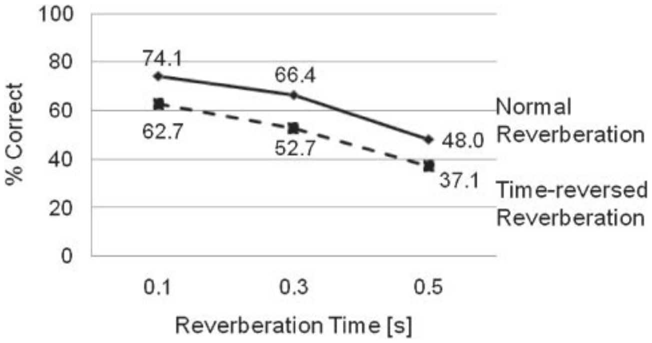

**Fig. 3** Correct rate for the ASR experiment.

2.2.3. Experimental results

Figure 3 shows the mean recognition rates of the ASR experiment for the six reverberant conditions. The dark line shows the results for the normal reverberation, and the gray line the time-reversed reverberation. The recognition rate was measured as the percentage of correctly recognized words within each sentence.

The algorithm for the ASR experiment that we used was not specially treated for reverberant speech. Therefore, the recognition rates of ASR were much lower than those in HSR. However, this is not an issue here because we were not interested in the decrease from clean to reverberant speech, but the decrease due to the temporal directions of reverberation and the declining tendency as a function of the reverberation time. The results confirmed that the ASR yields similar tendencies to the HSR. In other words, the recognition rates decreased as T60 became longer, and it was lower with the time-reversed reverberation than with the normal reverberation. However, unlike in the HSR experiment, the difference between the scores with the normal and time-reversed reverberations did not increase as reverberation became longer.

## 3. DISCUSSION

Because the STI is based on the magnitude modulation spectrum of temporal envelopes of speech, the STI predicts the same result with the normal and the time-reversed reverberation conditions. However, the two experiments show that the speech recognition rates with time-reversed reverberation are lower than with normal reverberation. In this section, we first discuss the possible reasons for this.

Speech sounds themselves have asymmetric temporal properties. A speech sound can be viewed as a sequence of syllables, and a syllable consists of an onset, a nucleus, and a coda. Typically, a syllable has the structure of a vowel (nucleus) with higher loudness surrounded by consonants with lower loudness (onset and coda). Consonants in the onset position tend to be more clearly articulated and the pronunciation tends to be less variant and more stable, while consonants in the coda position tend to be less clearly articulated and deletion and/or reduction is often observed [13]. For example, a stop consonant in the coda position is not always accompanied with the release of the burst and often has several allophones [14]. In fact, a study of the modified rhyme test for English monosyllables showed that the correct rate for the consonants in the coda position was only 68%, whereas the correct rate for consonants in the onset position was 80% [15]. Therefore, the syllable onset plays a more important role, and this view of the discussion is based on the asymmetric temporal properties in speech production.

On the other hand, it has been reported that self-masking and overlap-masking are the two major factors that decrease the intelligibility of speech in reverberant environments [16–18]. For self-masking, each phone is internally smeared, and the transients, such as onset and offset, are smeared. For overlap-masking, on the other hand, reverberant components of prior speech segments mask successive segments. As a result, speech segments following reverberating segments are more difficult to understand. As the energy of the prior segments increases, the effect of overlap-masking also increases. This fact is particularly crucial when the reverberating segment is a vowel (which has more power) and the subsequent segments are consonants (which have less power) [19,20].

Therefore, because of the combination of the asymmetry in speech production and the masking effects of reverberation, recognition rates are different between normal and time-reversed reverberations. With normal reverberation, the recognition rate of the coda can be predicted to decrease more than that of the onset within a syllable. This is because of a reverberation tail from the previous nucleus producing overlap masking affects on the following coda. However, with the time-reversed counterpart, the recognition rate of the *onset* can be predicted to decrease more than that of the *coda*. This is because, unlike normal reverberation, a time-reversed reverberation tail from the following nucleus affects the previous onset. As a result, the total speech recognition rate might be more affected with time-reversed reverberation than with normal reverberation, because codas are less reliable than onsets, in general, as described earlier, and the more reliable onsets are more affected with time-reversed reverberation.

Furthermore, from Figs. 2 and 3, we found HSR was more tolerant than ASR even though reverberation becomes longer. In the rest of this section, we discuss the possible reasons for this. According to the Haas effect (or the precedence effect), initial reflection sounds up to 50 ms are not perceived as independent from the direct sound; rather, they are fused with the direct sound [21–23]. This fusion occurs when the reflection sounds are weaker than the direct sound, and this occurs in natural reverberation. For time-reversed reverberation, on the other hand, the weaker sounds are perceived first, and, as a result, less fusion seems to occur. Thus, this reduced fusion is considered to cause higher perceptual masking and less intelligibility of speech. The Haas effect has mainly been reported in dichotic listening conditions; however, the fusion phenomenon of the direct and initial reflection sounds is reported in monaural and diotic listening conditions [23–25] (the experiment in this study was conducted diotically). In any case, this view of the discussion is based on the asymmetric temporal properties of the human auditory system, and this mechanism of human speech perception to compensate for normal reverberation is not incorporated in current speech recognizers.

## 4. SUMMARY

In this study, we first showed that the intelligibility of speech with time-reversed reverberation is lower than that of speech with normal reverberation. We then showed similar asymmetry in ASR as we had observed in HSR; but furthermore, we found the HSR was more tolerant even though reverberation becomes longer. This discrepancy was discussed from the viewpoint of asymmetric temporal properties in speech production and perception. In the future, we will be able to design a speech recognizer with a mechanism of human speech perception to compensate for reverberation that current speech recognizers do not have.

## ACKNOWLEDGMENTS

I would like to thank Hiroshi Riquimaroux of Doshisha University who made suggestions when I started this study, and Hideki Tachibana then of the University of Tokyo, and the members of his lab, especially Kanako Ueno and Sakae Yokoyama (at that time) for providing the impulse response data. I also acknowledge the assistance of the members in the Arai Lab. and the Cognitive Psychology Lab. at Sophia University, especially Emi Nishidate, Takuma Okuda, Nao Hodoshima and Hirofumi Kamata. This work was partially supported by a Grant-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for the Promotion of Science and Sophia University Open Research Center from MEXT.

## REFERENCES

[1] T. Houtgast and H. J. M. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, **77**, 1069–1077 (1985).

[2] R. Drullman, J. M. Festen and R. Plomp, "Effect of temporal envelope smearing on speech reception," *J. Acoust. Soc. Am.*, **95**, 1053–1064 (1994).

[3] R. Drullman, J. M. Festen and R. Plomp, "Effect of reducing slow temporal modulations on speech reception," *J. Acoust. Soc. Am.*, **95**, 2670–2680 (1994).

[4] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Intelligibility of speech with filtered time trajectories of spectral envelopes," *Proc. Int. Conf. Spoken Lang. Process.*, Vol. 4, pp. 2490–2493, Philadelphia (1996).

[5] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, **105**, 2783–2791 (1999).

[6] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, **2**, 578–589 (1999).

[7] N. Kanedera, T. Arai, H. Hermansky and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Commun.*, **28**, 43–55 (1999).

[8] D. Griesinger, "Measurement of acoustic properties through syllabic analysis of binaural speech," *Proc. Int. Congr. Acoustics*, Vol. 1, pp. 29–32 (2004).

[9] T. Arai, "Degradation of speech intelligibility in time-reversed reverberation," *Trans. Tech. Comm. Psychol. Physiol. Acoust., Acoust. Soc. Jpn.*, 35(4), H-2005-41, pp. 237–242 (2005) (in Japanese).

[10] L. Longworth-Reed, E. Brandewie and P. Zahorik, "Time-forward speech intelligibility in time-reversed rooms," *J. Acoust. Soc. Am.*, **125**, EL13–EL19 (2009).

[11] N. Hodoshima, T. Arai and A. Kusumoto, "Enhancing temporal dynamics of speech to improve intelligibility in reverberant environments," *Proc. Forum Acusticum*, Sevilla (2002).

[12] Julius homepage: http://julius.sourceforge.jp/.

[13] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language?," *IEICE Trans. Inf. Syst.*, **E87-D**, 1059–1070 (2004).

[14] M. Davenport and S. J. Hannahs, *Introducing Phonetics and Phonology* (Arnold, London, 1998).

[15] A. S. House, C. E. Williams, M. H. L. Hecker and K. D. Kryter, "Articulation-testing methods: Consonantal differentiation with a closed-response set," *J. Acoust. Soc. Am.*, **37**, 158–166 (1965).

[16] R. H. Bolt and A. D. MacDonald, "Theory of speech masking by reverberation," *J. Acoust. Soc. Am.*, **21**, 577–580 (1949).

[17] A. K. Nábělek and L. Robinette, "Influence of precedence effect on word identification by normally hearing and hearing-impaired subjects," *J. Acoust. Soc. Am.*, **63**, 187–194 (1978).

[18] A. K. Nábělek, T. R. Letowski and F. M. Tucker, "Reverberant overlap- and self-masking in consonant identification," *J. Acoust. Soc. Am.*, **86**, 1259–1265 (1989).

[19] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects of suppressing steady-state portions of speech on intelligibility in reverberant environments," *Proc. Autumn Meet. Acoust. Soc. Jpn.*, Vol. 1, pp. 449–450 (2001) (in Japanese).

[20] T. Arai, K. Kinoshita, N. Hodoshima, A. Kusumoto and T. Kitamura, "Effects on suppressing steady-state portions of speech on intelligibility in reverberant environments," *Acoust. Sci. & Tech.*, **23**, 229–232 (2002).

[21] A. J. Watkins, "The influence of early reflections on the identification and lateralization of vowels," *J. Acoust. Soc. Am.*, **106**, 2933–2944 (1999).

[22] H. Haas, "Über den Einfluss eines Einfachechos an die Hörsamkeit von Sprache," *Acustica*, **1**, 49–58 (1951).

[23] R. Y. Litovsky, H. S. Colburn, W. A. Yost and S. J. Guzman, "The precedence effect," *J. Acoust. Soc. Am.*, **106**, 1633–1654 (1999).

[24] B. Rakerd, J. Hsu and W. M. Hartmann, "The Haas effect with and without binaural differences," *J. Acoust. Soc. Am.*, **101**, 3083 (1997).

[25] R. Y. Litovsky, M. L. Hawley and H. S. Colburn, "Measurement of precedence in monaural listeners," *Meeting of the American Speech and Hearing Association*, Boston, MA (1997).

**Takayuki Arai** received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Professor of the Department of Information and Communication Sciences, Sophia Univ. In 2003–2004, he is a visiting scientist at Massachusetts Institute of Technology (Cambridge, MA, USA). His research interests include signal processing, acoustics, speech and hearing sciences, spoken language processing, and acoustic phonetics.