

ACOUSTICAL LETTER

Masking speech with its time-reversed signal

Takayuki Arai*

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

(*Received 27 May 2009, Accepted for publication 17 August 2009*)

Keywords: Speech privacy, Sound masking, Time-reversed signal, Speech masker
PACS number: 43.55.Hy, 43.71.Es, 43.66.Dc [doi:10.1250/ast.31.188]

1. Introduction

Many approaches to speech privacy and sound masking have been discussed [1,2], and it has been pointed out that a target signal can be effectively masked by using the target speaker's speech as a masking signal [3,4]. Ito *et al.* (2007) [4] obtained a masking signal by dividing the target signal into time frames, reversing each of the frames along the time axis, and concatenating them in random order. Although a certain amount of the target signal should be stored in memory, this method effectively masks the target signal as long as the same speaker is speaking. However, the masking performance can drop when the speaker changes. Therefore, in the present study, a masking signal is obtained by reversing the time frame preceding the target frame so that the masking signal is always obtained from the same speaker's speech signal; thus, performance reduction due to speaker change can be avoided. In addition, the required memory size decreases, leading to easier implementation of the system on a digital-signal-processor (DSP) chip for real-time processing. We describe a pilot experiment to test this method of obtaining a masking signal.

2. Pilot experiment

To test the proposed algorithm for obtaining a masking signal, we conducted a pilot experiment.

2.1. Target signals

In this experiment, we used three target signals. Each target speech signal consisted of a nonsense consonant-vowel (CV) syllable embedded in the Japanese carrier phrase, "Daimoku to shite wa ____ to iimasu" (The title is ____). The vowel was /a/ and the consonants were /p/, /t/, or /k/. The speech samples were originally obtained from the Japanese ATR Speech Database. The CV syllables were obtained from the monosyllable data set in the database, whereas the carrier phrase was a combination of two partial sentences obtained from the sentence data set in the same database. The starting point of the vowel in the embedded CV syllable was adjusted to 50 ms from the conclusion of the first half of the carrier phrase, and the second half of the carrier phrase started 35 ms after the CV syllable. The root mean square (RMS) of the CVs was adjusted to the RMS in the carrier phrase.

2.2. Masking signals

The main masking signals used in the pilot experiment were derived from the target signals. Each of the masking signals was obtained by dividing the corresponding target signal into short time frames, reversing each of the preceding time frames along the time axis, delaying them by one time frame, and concatenating them in the original order. In this way, the target time frame is always masked by the time-reversed version of the time frame preceding the target frame. The frame lengths used in this pilot experiment were 40 to 240 in 40 ms steps. Babble noise from the NOISEX database [5] was also used as a masking signal.

2.3. Stimuli

A target signal and the corresponding masking signal were summed to achieve different values of the target-to-masking signal ratio (TMR). The TMRs used in this pilot experiment were from −15 to 10 dB in 5 dB steps. We obtained 126 stimulus sentences (3 target signals × 7 masking signals × 6 TMRs).

2.4. Procedure

Three naive listeners participated in the pilot experiment. After a brief training session, the main experiment was conducted. During the training session, the sound level was adjusted to the listener's comfort level. The main experiment was divided into 21 subsessions, each of which corresponded to one combination of the 3 target signals and the 7 masking signals. In each subsession, six icons of a loudspeaker corresponding to the six TMRs in ascending order were displayed from the top to the bottom on a PC screen. When an icon was double-clicked using a mouse, a stimulus sound was played from the PC via the digital-to-analog (D/A) converter of Onkyo SE-U33GX through Sennheiser HD595 headphones. The listener was asked to click and play each stimulus three times and answer whether the stimulus was /pa/, /ta/, or /ka/. He/she was instructed to go from the top to the bottom in each subsession and was not permitted to relisten to any previous stimuli.

3. Results and discussion

The identification rates (in percentages) in the pilot experiment are shown in Table 1. The proposed masking signals with the time frame of 80–200 ms were more effective than babble noise. The masking signal proposed in this study is a result of the same modification discussed by Saberi and Perrott [6]. Their study shows that time-reversed signals with a frame length of 50 ms yield nearly perfect speech intelligibility (the intelligibility score for the 50 ms frame was higher than 95% in Fig. 1 in Saberi and Perrott's study [6]). In other words, such signals are not appropriate as masking

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

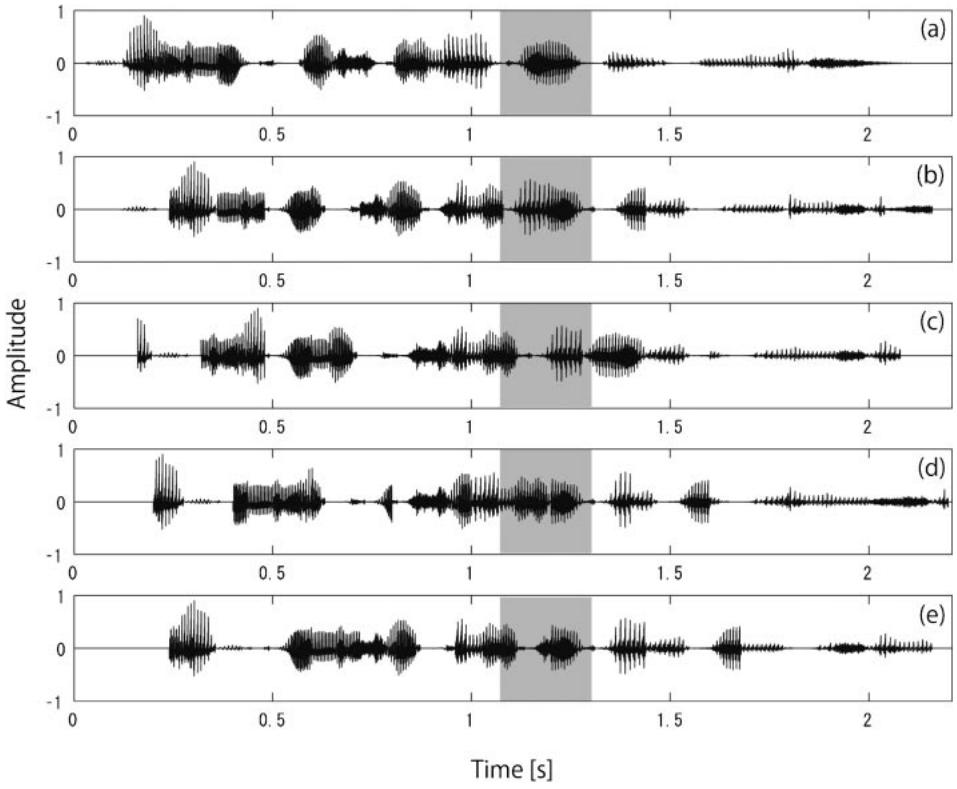

Fig. 1 The original sentence with the target CV syllable of /ka/ and masker signals of different frame lengths: (a) original signal, (b) masking signal (120 ms), (c) masking signal (160 ms), (d) masking signal (200 ms), and (e) masking signal (240 ms). The shaded section indicates where the original target CV syllable of /ka/ is located.

Table 1 Identification rates (%) in pilot experiment.

TMR	Type of masking signal						
	Proposed (frame length)						Babble noise
	40 ms	80 ms	120 ms	160 ms	200 ms	240 ms	
−15 dB	67	44	33	56	44	78	22
−10 dB	67	56	33	67	44	78	67
−5 dB	78	44	11	78	33	78	67
0 dB	100	44	56	89	33	100	78
5 dB	100	78	67	89	56	100	100
10 dB	100	89	89	100	67	100	100

signals because they are too intelligible. As the frame length increases, the identification rates decrease up to the frame length of 120 ms (Table 1). However, the identification rates fluctuate when the frame length is between 120 and 240 ms. This can be explained by the "local" TMR. Figure 1 shows the original sentence of /ka/ (a) and the proposed maskers with frame lengths from 120 ms (b) to 240 ms (e). The shaded section indicates where the original target CV syllable (/ka/, in this case) is located. From this figure, one can observe that the identification rate drops when a masker has high energy in the shaded section (the local TMR is low). The optimal frame length is between 120 and 240 ms, the average syllable duration; in this case, the nuclei (vowels) in a masking signal can effectively mask the consonants in the target signal.

It has been pointed out that annoyance increases as TMR decreases [7]. Therefore, we use a masker that has high masking efficiency, even when the TMR is high. Table 1 shows that the identification rates increase as TMR increases. The optimal TMR is approximately −5 dB, because it is in the high end of the range where the identification rates are still low.

4. Summary

We tested how a time-reversed signal effectively masks a target speech signal by reversing the time frame preceding the target frame to obtain a masking signal. We confirmed that this proposed masker, which is robust against speaker change and has advantages for implementation on a DSP chip, can effectively mask an input speech signal. In fact, a proposed masker was recently implemented with a DSP in our laboratory. We will further investigate the performance of the algorithm to produce the masking signal by increasing the number of stimuli and listeners.

Acknowledgments

This research was partly supported by a Grant-in-Aid for Scientific Research (18530762) from the Japan Society for the Promotion of Science and by Sophia University Open Research Center from MEXT.

References

[1] W. J. Cavanaugh, W. R. Farrell, P. W. Hirtle and B. G. Watters, "Speech privacy in buildings," *J. Acoust. Soc. Am.*, **34**, 475–492 (1962).

[2] A. W. Bronkhost and R. Plomp, "Effect of multiple speech-like maskers on binaural speech recognition in normal and impaired hearing," *J. Acoust. Soc. Am.*, **92**, 3132–3139 (1992).
[3] D. S. Brungart, B. D. Simpson, M. A. Ericson and K. R. Scott, "Informational and energetic masking effects in the perception of multiple simultaneous talkers," *J. Acoust. Soc. Am.*, **110**, 2527–2538 (2001).
[4] A. Ito, A. Miki, Y. Shimizu, K. Ueno, H. J. Lee and S. Sakamoto, "Oral information masking considering room environmental condition; Part 1: Synthesis of maskers and examination on their masking efficiency," *Proc. Inter-Noise*, Istanbul (2007).
[5] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, **12**, 247–251 (1993).
[6] K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," *Nature*, **398**, 760 (1999).
[7] K. Ueno, H. J. Lee, S. Sakamoto, A. Ito, A. Miki and Y. Shimizu, "Oral information masking considering room environmental condition; Part 2: Subjective assessment for 'Masking efficiency and Annoyance'," *Proc. Inter-Noise*, Istanbul (2007).