ICSLP, 3:586-589, 2000

DESIGNING MODULATION FILTERS FOR IMPROVING SPEECH INTELLIGIBILITY IN REVERBERANT ENVIRONMENTS

Tomoko Kitamura, Keisuke Kinoshita, Takayuki Arai, Akiko Kusumoto and Yuji Murahara

Dept. of Electrical and Electronics Eng., Sophia University
7–1 Kioi-cho, Chiyoda-ku, Tokyo, JAPAN

ABSTRACT

In this paper, we propose a new technique to design modulation filters to reduce degradation of speech intelligibility in reverberant environments. Using the inverse modulation transfer function, we design data-derived modulation filters for each speech frequency band. These filters preprocess speech signals between a microphone and a loudspeaker that radiates speech into a performance hall. Using our modulation filters, we conducted perceptual experiments with one hearing-impaired subject and two subjects with normal hearing. Test results indicate that our proposed method improves the intelligibility of reverberant speech.

1. INTRODUCTION

Many public halls are designed for multiple purposes such as musical concerts and lectures. Reverberation is usually preferred for musical performances, however, it degrades speech intelligibility.

Previous studies show that there is a relation between the modulation transfer function (MTF) and the intelligibility of speech [1]–[2]. In addition, the rapid speech transfer index (RASTI) based on the MTF is widely used for measuring speech intelligibility in auditoriums [2].

Several studies show that the important modulation frequency range for speech intelligibility is between 1 and 16 Hz, centered around 4 Hz [3]–[5]. When speech is reverberant, the peak of the modulation spectrum shifts to a lower modulation frequency and its modulation index declines [2].

For enhancement of speech degraded by room reverberation, several studies have been done using a single microphone [6]–[7]. Avendano and Hermansky [7] filtered the time trajectories of spectral bands of reverberant speech to recover original modulation. They applied a technique using the inverse MTF (IMTF) proposed by Langhans and Strube [6]. Langhans and Strube applied the theoretically derived IMTF to reduce reverberation [6], while Avendano and Hermansky applied the data-derived IMTF [7]. In other studies, microphone-array approach has been proposed. The microphone-array is composed of multiple microphones which are spatially arranged to take advantage of spatial information about sound sources to suppress noise and reverberation [8]–[11].

Kusumoto et al. [12] have recently proposed a different technique to improve speech intelligibility. They preprocess the speech signal between the microphone and loudspeaker, before the speech signal is transmitted into the hall. They empirically designed a *single* modulation filter to enhance specific frequencies of the speech modulation spectrum. Their preference tests indicated that the hearing-impaired preferred preprocessed speech signals.

In this paper, we design data-derived modulation filters for each frequency band instead of applying the same modulation filter to each channel. These filters are designed using IMTF, following the technique described in [7]. In section 2, we describe the method we design the modulation filters. In section 3, we describe our perceptual experiments with a hearing-impaired subject and subjects with normal hearing.

2. MODULATION FILTERING

We use the method of modulation filtering proposed in [12]. Fig. 1 shows the block diagram for the signal processing between a microphone and a loudspeaker. The input signal is split into 16 frequency bands corresponding to the critical bands by applying constant-Q band-pass filters (BPFs) with 1/3-octave bandwidths. For each band-passed signal, the envelope is extracted using Hilbert transform. After down-sampling (by the factor of $M = 160$), we apply the modulation filters to the temporal trajectories of the envelope. Up-sampling (by the same factor of M) converts the filtered envelope back to the original sampling rate. After up-sampling, we apply half-wave rectification to remove any negative value artifacts produced by the modulation filtering. By applying BPFs to the resultant signal, we remove frequency components outside the range of the band. Finally, the processed speech signal is obtained by summing the output signals from each band.

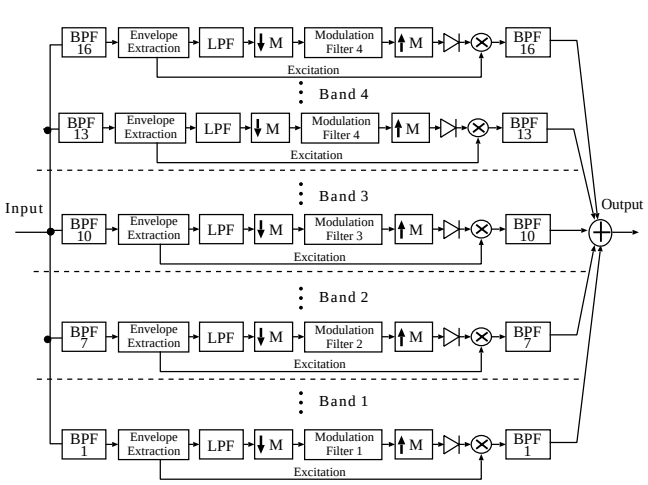

Figure 1: Block diagram of the speech-processing system.

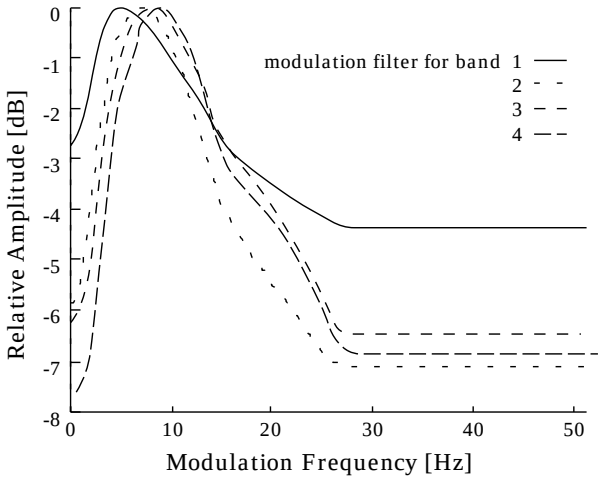

Figure 2: Frequency response of the modulation filters for bands 1 to 4.

2.1. Designing modulation filters

Houtgast and Steeneken [1]–[2] defined the MTF in reverberant environments as the reduction in the modulation index of the intensity envelope of the output signal relative to the input signal. The input signal is white noise modulated by a sine-wave. MTF measured in a reverberant environments typically shows a low-pass characteristic.

Avendano and Hermansky [7] designed a filter-bank from training data. Likewise, we design modulation filters from the sentences used in our perceptual experiments. Unlike Kusumoto et al. [12] who use the same modulation filter for each channel, we design four modulation filters for each frequency band.

To design a set of modulation filters using the IMTF, we first derive the MTF using speech signals divided into 4 frequency bands (band 1: 0–800Hz, band 2: 800–1600Hz, band 3: 1600–3200Hz, band 4: 3200–8000Hz) following Arai and Greenberg [13]. The MTF for each band is averaged over the 60 sentences used in our perceptual experiments. We obtain the frequency response of the modulation filters by inverting and taking moving averages from the derived MTF. Finally, each modulation filter is implemented as a 65-tap finite-response-impulse (FIR) filter. Fig. 2 shows the frequency characteristics of the derived modulation filters for bands 1 to 4.

3. PERCEPTUAL EXPERIMENTS

Perceptual experiments were carried out individually in the sound-proof room at Sophia University in Tokyo.

3.1. Stimuli

The 60 spoken sentences were obtained from the NTT-AT Speech Database (spoken by male and female speakers in roughly equal measure and spanning all major dialect regions of Japanese). The signals were sampled at 16kHz and quantized with 16-bit resolution. We processed these signals using the system shown in Fig. 1 by applying four modulation filters (Fig. 2) to each frequency band. To simulate the reverberant environment, the 60 processed speech signals and the 60 original speech signals were convolved with the impulse response of Osaka Symphony Hall (T=2.1s) in Osaka, Japan. We included the 60 original speech signals, giving us the 180 stimuli in total. We presented the stimuli to each subject from a pair of loudspeakers set 0.85 m from the subject's ears. The speech signal was sent to the loudspeakers from a PC.

3.2. Subjects

Our perceptual tests included one hearing-impaired subject and two subjects with normal hearing. All subjects are native-speakers of Japanese. The hearing-impaired subject has sensorineural hearing loss (hearing level without hearing aids / right: 95 dB, left: 105 dB), and wore hearing aids on both ears during experiments. The subjects with normal hearing are two females aged of 31 and 33 years old.

3.3. Experiment

3.3.1. The hearing-impaired subject

We wanted to examine speech intelligibility under the condition that the subject has some preliminary knowledge about

the topic. Consequently we gave a fill-in-the-blank test to the hearing-impaired subject. After listening to each sentence, the subject was asked to fill in a single missing word.

The hearing-impaired subject participated in three testing sessions. We divided our reverberant sentences into 2 groups, 30 sentences each of the reverberant and reverberant processed speech signals form group (a) and the remaining sentences form group (b). They were balanced in terms of the speaker (male / female), the missing word's degree of difficulty (high-predictable / low-predictable) and number of moras.

The first session of perceptual experiment had 60 stimuli including reverberant speech signals of group (a) and reverberant processed speech signals of group (b). The second session had 60 stimuli including reverberant processed speech signals of group (a) and reverberant speech signals of group (b). The last session had 60 stimuli that were all original speech signals without reverberation. We presented these 60 stimuli in random order at each session. We included each sentences only once in a session. In addition, the hearing-impaired subject took 1 day's recess between these sessions.

3.3.2. The subjects with normal hearing

The subjects with normal hearing participated in the fill-in-the-blank test and mean opinion score (MOS) test. In the MOS test the stimuli were scored on the five point scale: excellent (4 points), good (3 points), fair (2 points), poor (1 point), unsatisfactory (0 point).

For each session, 20 sentences each of the 60 reverberant and reverberant processed sentences were used. We divided these 20 sentences into 2 groups (c) and (d) of 10 sentences each. Each group was balanced in terms of speaker gender, number of moras, and degree of difficulty. We presented 20 stimuli including reverberant speech signals of group (c) and reverberant processed speech signals of group (d) for the first subject. We presented 20 stimuli including reverberant processed speech signals of group (c) and reverberant speech signals of group (d) for the second subject. We presented these stimuli in random order. The stimuli was played at the most two times to the subject.

3.4. Results and Discussion

Table 1 shows the results of the hearing-impaired subject. Table 2 shows the results of the subjects with normal hearing.

We scored the degree of improvement of speech intelligibility by calculating the number of words, moras, and phonemes that the subjects identified correctly. In the correct phoneme calculation, all the phonemes were weighted equally. We did not include identification errors between long and short vowels or consonants.

Table 1: Accuracy (%) for the hearing-impaired subject.

Sentences		Reverb.	Reverb. + processed	Original
Group (a)	word	15.0	13.3	80.0
	mora	31.7	41.1	84.7
	phoneme	47.1	50.9	89.2
Group (b)	word	8.3	6.7	63.3
	mora	28.4	22.4	71.9
	phoneme	37.4	42.8	78.4
Total	word	11.7	10.0	71.6
	mora	29.6	31.9	78.3
	phoneme	42.2	47.0	83.8

Table 2: Accuracy (%) for the subjects with normal hearing and MOS on the five point scale.

	Reberb.	Reverb. + processed
Word	75.0	80.0
Mora	81.3	87.5
Phoneme	83.7	89.2
MOS	1.35	1.25

In regard to the hearing-impaired subject (Table 1), the accuracy for the word and the mora shows no striking improvement. This is partly due to the long reverberation time, which caused severe intelligibility problems for the hearing-impaired subject. On the other hand, the results for correct phoneme identification for group (a) and group (b) indicate that our method improved performance by 3.8% for group (a) and 5.4% for group (b), respectively. We divided 60 sentences into group (a) and group (b) to observe the effect of stimulus order on performance. Since each results show the same tendency of improvement regardless of the stimulus order, we can examine the results without taking stimulus order into consideration. Correct phoneme identification was raised by 4.8%. This indicates that speech intelligibility was slightly improved for the hearing-impaired.

For subjects with normal hearing (Table 2), the MOS for the reverberant processed speech signals was found to be 1.25 on the five point scale, while it was 1.35 for the reverberant speech signals. The MOS does not show any significant improvement for the processing. On the other hand, identification accuracy increased by 5% for word, 6.2% for mora and 5.5% for phoneme, respectively. This indicates that our

method improves speech intelligibility for people with normal hearing.

4. CONCLUSIONS

In this study we proposed a new technique to improve speech intelligibility in reverberant environments. We designed separate data-derived filters for each frequency band to process the speech signal between a microphone and a loudspeaker. To examine the effectiveness of our method we conducted fill-in-the-blank perceptual tests. Our results indicate that this method improves speech intelligibility for both the hearing-impaired and people with normal hearing. In future we plan to evaluate our filters in various reverberant environments.

5. ACKNOWLEDGMENTS

We express our appreciation to Prof. Hideki Tachibana, Kanako Ueno and Sakae Yokoyama of the University of Tokyo for providing impulse response data for our project. We acknowledge Haruo Watanabe of Hearing Aids Fitting Research Center for his advise on designing tests for the hearing-impaired and providing some materials for our experiment.

6. REFERENCES

[1] T. Houtgast and H. J. M. Steeneken, "The modulation transfer function in room acoustics as a predictor of speech intelligibility," *Acustica*, **28**, pp.66–73, 1973.

[2] 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(3)**, pp.1069–1077, 1985.

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

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

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

[6] T. Langhans and H. W. Strube, "Speech enhancement by nonlinear multiband envelope filtering," *Proc. IEEE ICASSP*, pp.156–159, 1982.

[7] C. Avendano and H. Hermansky, "Study on the dereverberation of speech based on temporal envelope filtering," *Proc. ICSLP*, pp.889–892, 1996.

[8] M. Omologo, P. Svaizer, and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Commun.*, **25**, pp.75–95, 1998.

[9] H. Wang and F. Itakura, "An approach of dereverberation using multimicrophone sub-band envelope estimation," *Proc. IEEE ICASSP*, pp.953–956, 1991.

[10] D.Rabinkin, R. Renomeron, and J. Flanagan, "Optimal truncation time for matched filter array processing," *Proc. IEEE ICASSP*, pp.3629–3632, 1998.

[11] T. Yamada, S. Nakamura, and K. Shikano, "Hands-free speech recognition based on 3-D viterbi search using a microphone array," *Proc. IEEE ICASSP*, pp.245–248, 1998.

[12] A. Kusumoto, T. Arai, T. Kitamura, M. Takahashi, and Y. Murahara, "Modulation filtering of speech as a preprocessing against reverberation for the hearing-impaired," *Proc. IEEE ICASSP*, pp.853–856, 2000.

[13] T. Arai and S. Greenberg, "Speech intelligibility in the presence of cross-channel spectral asynchrony," *IEEE ICASSP*, pp.933–936, 1998.