社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
SP2005-171 (2006-03)

Study on Noise Reduction of Ventilator Noise in Recorded Speech Signals

Shimpei KAJIMA†, Ori TAKESHITA†, Keiichi YASU†, Takayuki ARAI† and Akemi IIDA‡

†Sophia University, Tokyo, Japan

‡Tokyo University of Technology, Tokyo, Japan

E-mail: †s-kajima@sophia.ac.jp

Abstract This study aimed to reduce ventilator noise in the speech signals of patients who use a ventilator. Two applications, spectral subtraction and adaptive filtering, were examined. For the first experiment, we employed a new technique based on spectral subtraction for ventilator noise reduction and then evaluated the result using mean opinion score. The results showed that the new technique was superior to the common approach of spectral subtraction in reducing ventilator noise. For the second experiment, an adaptive filter was used in a simulated environment to reduce ventilator noise and three other signals (white noise, male speech, and female speech) from each speech signal. Adaptive filtering was able to reduce ventilator noise and, based on these results, we propose a new approach of recording speech signals in a real environment using an adaptive filter.

Keywords speech enhancement, noise reduction, spectral subtraction, ventilator

1. Introduction

Speech is one of the important ways in which we communicate with each other in daily life. However, individuals with respiratory organ disease can experience difficulty communicating verbally; among them are patients with muscle paralysis [1, 2]. Amyotrophic lateral sclerosis (ALS) is a progressive disease causing muscle paralysis [1, 2, 3], which requires tracheotomy when the respiratory muscles weaken [2]. Following tracheotomy, patients require speech rehabilitation. Any subsequent weakness of the tongue muscle can result in the loss of speech [5].

A speech synthesis system is very useful method of communication for these patients. A corpus-based speech synthesis system that synthesizes speech from pre-recorded speech signals offers the advantage of using the user's own speech, recorded before the loss of voice [3]. Examples are "Festival" [11] and "CHATR" [12] produced by ATR.

The quality of speech synthesized by the corpus-based speech synthesis system is highly affected by the quality of pre-recorded speech signals in a corpus. Speech sounds spoken by a patient who is using a ventilator are often recorded together with the noise produced by the ventilator (ventilator noise). Thus, the quality of recorded speech signals is degraded by this ventilator noise.

In the present study, we attempted to reduce ventilator noise in recorded speech signals using spectral subtraction (SS) [13] and adaptive filtering [14].

2. Ventilator noise reduction using spectral subtraction

We recorded speech sounds uttered by a Japanese male ALS patient who uses a ventilator. Following tracheotomy three years ago, the patient uses a ventilator. The speech sound recording was conducted in the patient's home and contained ventilator noise.

We analyzed the following 12 English tokens from all recorded speech sounds: "ta ba ba", "ta pa pa", "ta da da", "ta ta ta", "ta ga ga", "ta ka ka", "ta cha cha", "ta jha jha", "ta hha hha", "ta tha tha", "ta dha dha", and "ta fa fa" (the patient was not able to pronounce all of the tokens correctly). In this study, we mainly used the speech sample "ta tha tha" (actual pronunciation was [ta ga ga]).

2.1. Property of ventilator noise

Figure 1 shows the waveform and spectrogram of a ventilator noise sample. Amplitude increases in the first half of the time frame and decreases thereafter. Ventilator noise occurred periodically with an almost constant duration; however, the endpoints of ventilator noise were often difficult to detect. The spectrogram shows two properties of the noise: 1) strong energy between approximately 2 and 3 kHz; and 2) fluctuating amplitude at approximately 70 to 100 Hz.

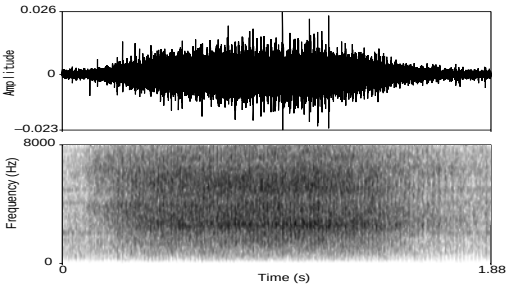

Figure 1 Waveform of ventilator noise (upper panel) and the corresponding spectrogram (lower panel).

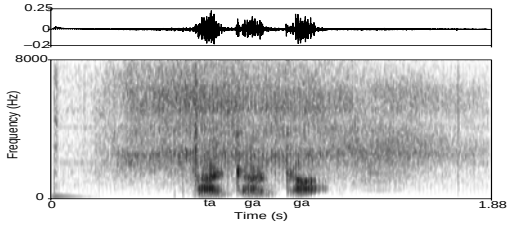

Figure 2 Waveform of recorded utterance "ta tha tha" (upper panel) and the corresponding spectrogram (lower panel).

2.2. Proposed techniques

2.2.1. Technique 1

In a preliminary study, we attempted to reduce ventilator noise (Figure 1) in speech signals (Figure 2)

using the conventional technique of spectral subtraction (SS). However, the processed signal contained a large amount of "musical noise" [15].

In the present study, we therefore proposed a new technique based on SS to improve noise estimation. First, syllable endpoints were detected and then, in each detected syllable, the root mean square (RMS) of noise was estimated from the surrounding non-speech frames. Finally, the RMS of the speech frames in each syllable was reduced to the estimated RMS of noise by compressing the amplitude of the speech frames. Figure 3 shows the block diagram of this method (Technique 1).

Only an input signal at high frequency band above 2 kHz was processed because the energy of ventilator noise is relatively strong compared to the energy of speech in that frequency region. Therefore, the input signal was fed to a high-pass filter with a cut-off frequency of 1500 Hz. The upper panel of Figure 4 shows the waveform of the high-passed signal.

The lower panel of Figure 4 shows the estimated noise by Technique 1 and the upper panel of Figure 5 shows the resultant signal (in this case, the subtraction coefficient α [15] was set to 1.0). Compared with the original signal (Figure 2), each syllable was enhanced. Reduced musical noise was also apparent when listening to the processed signals.

We then attempted to increase the value of α to reduce ventilator noise completely. The spectrograms of Figure 5 show the resultant signals with α=1.0 (upper panel) and α=1.2 (lower panel). There was greater reduction in ventilator noise for α=1.2 than for α=1.0. However, some speech components were also lost at high frequency band (A in Figure 5). Further, some of the components resulted from ventilator noise remaining at low frequency band and this degraded the quality of processed signal. We, therefore, proposed another method (Technique 2).

2.2.2. Technique 2

Figure 6 shows the block diagram of Technique 2. In this method, the input signal was split into 1/3 octave subbands. In each band, the amplitude of speech frames was compressed as in the case of Technique 1. The processed signal from each subband was then added to obtain the estimated noise. Figure 7 of the spectrogram of the resulting signal shows that Technique 2 reduced ventilator noise almost completely in non-speech frames without loosing speech components.

2.3. Subjective evaluation experiment

Subjective impressions of the processed signals were evaluated using mean opinion score (MOS). Six subjects were presented with 36 randomly presented stimuli (12 tokens x 3 types of processing [SS, Technique 1, Technique 2])for evaluation by MOS (5 = excellent, 1 = very poor). The listening experiment was conducted five times for each subject. Table 1 and Figure 8 shows the experimental result.

Table 1 shows that Technique 2 produced the best quality of processed signals among the three methods.

The differences were statistically significant (p < 0.01) between SS and Technique 2. As shown in Figure 8, the stimuli processed by Technique 2 had the best MOS for all types of stimuli.

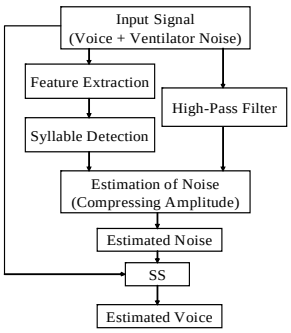

Figure 3 Block diagram of Technique 1.

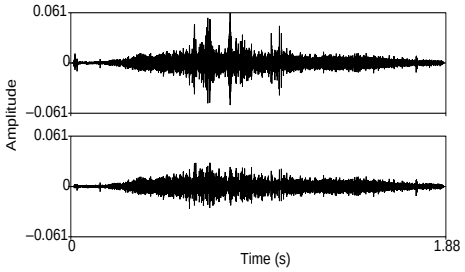

Figure 4 Waveforms of high-passed speech "ta tha tha" (upper panel) and the estimated noise (lower panel)

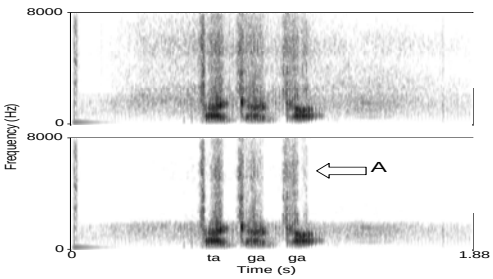

Figure 5 Spectrograms of the processed signal "ta tha tha" by Technique 1 (upper: α=1.0; lower: α=1.2).

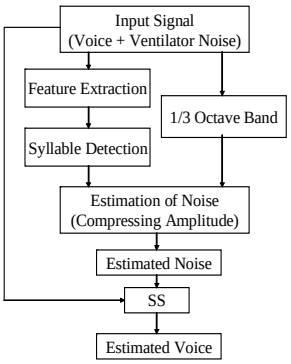

Figure 6 Block diagram of Technique 2.

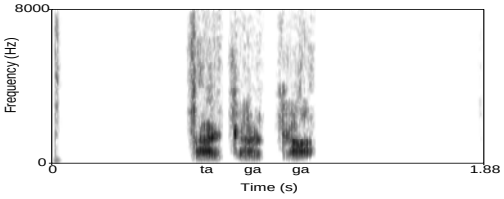

Figure 7 Spectrogram of the processed signal "ta tha tha" by Technique 2 (α=1.0).

The MOS values of the stimuli “ta ga ga” and “ta ka ka” were low because the fricative noise resembles ventilator noise remaining in these stimuli. Figure 9 shows the spectrogram of the processed signal “ta ga ga” by Technique 2. Fricative noise remains around Arrow B. In the process of RMS estimation in Techniques 1 and 2, we assumed that the estimated RMS of ventilator noise was constant within each syllable. However, in fact, because the amplitude of ventilator noise fluctuates, it might yield this kind of estimation error. The same is true for consonants. Consonants “g” and “k” have strong energy between about 1.5 and 4.0 kHz. In this frequency band, the energy of ventilator noise is also strong. Therefore, ventilator noise remained for these consonants.

Another estimation error should be considered for the method of detecting syllable endpoints. As Arrow C indicates in Figure 10, “t” of the initial syllable was lost. As Arrow D indicates, ventilator noise remained just before “k” of the final syllable. In this study, syllable endpoints were detected based on the intensity of the input signal. Although the optimal threshold of intensity for each speech signal was set, among the three syllables within the signal, the intensity of each syllable is different. Thus, the detected syllable endpoints have some errors. Manual detection of each syllable endpoint may resolve this problem.

In conclusion, Technique 2 markedly reduces ventilator noise in recorded speech signals. If the afore-mentioned problems can be resolved, the quality of processed signal should be improved.

3. Proposed method for recording by adaptive filtering

In this section, we propose a recording method based on ventilator noise reduction in the recorded speech signal by adaptive filtering [14]. We recorded ventilator noise (Figure 1) and three other signals (white noise, male speech, and female speech) in a simulated environment (Figure 11) and attempted to reduce the noise signal by adaptive filtering.

3.1. Adaptive filter

Figure 12 shows the block diagram of adaptive filtering used in this study. In Figure 12, s(k) is the speech signal, n(k) and n’(k) are noise signals, d(k) is the desired signal, e(k) is the error signal, and y(k) is the output signal. The adaptive filter adjusts the filter coefficients to minimize the energy of the error signal (e(k) = d(k) – y(k)).

3.2. Recording signals

Three speakers (BOSE, MM-1), two directional microphones (SONY, EMC-23F5) and a digital recorder (Marantz, PMD670/F1B) were used to record signals in a soundproof room. Figure 11 shows the recording environment. Microphone B was enclosed by a box to isolate the noise signal. Speech and noise signals were recorded by microphone A. Male speech and female speech were used as speech signal s(k), and ventilator noise (shown in Figure 1) and white noise as noise signal n(k). n’(k) in Figure 12 was the signal recorded by microphone B and was used as the referred input signal in adaptive filtering. All signals were recorded digitally at 16 kHz sampling and 16 bit quantization.

Table 1 MOS values of the listening experiment.

	SS	Technique 1	Technique 2
MOS	1.90	1.80	3.06

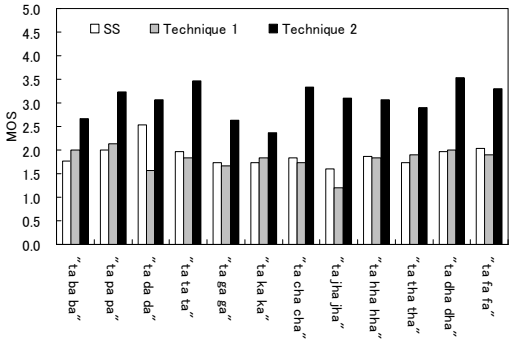

Figure 8 MOS values for each stimulus in the subjective evaluation

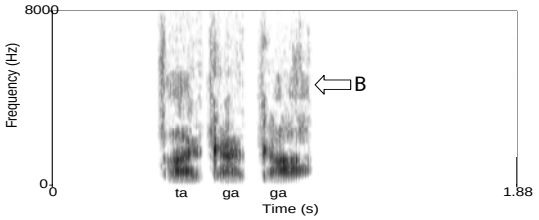

Figure 9 Spectrogram of the processed signal “ta ga ga” by Technique 2.

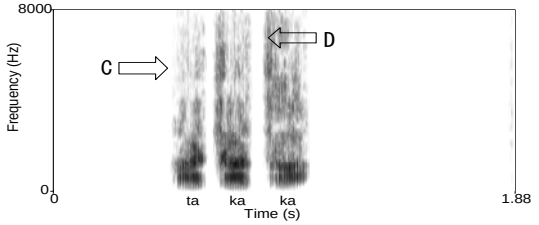

Figure 10 Spectrogram of the processed signal “ta ka ka” by Technique 2.

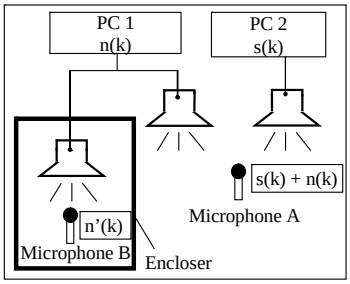

Figure 11 Recording environment.

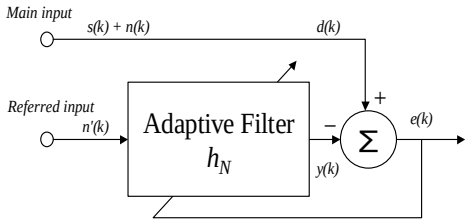

Figure 12 Block diagram of adaptive filtering.

3.3. Experiment

To reduce noise signal by adaptive filtering, the signals recorded by microphone A were used as the main input signal and the signal recorded by microphone B was used as the referred input signal. Figure 13 and 14 shows the result of this experiment using male speech s(k) and ventilator noise n(k).

Comparing the main input signal with the output signal in Figures 13 and 14, respectively, ventilator noise was especially reduced between 1.0 and 2.0 seconds. As shown in Figure 14, the energy of ventilator noise was weakened between about 2 kHz and 3 kHz. In this experiment, signal-to-noise ratio (SNR) improved by 26.4 dB.

In this study, although we could record only noise signal n'(k), we could not reduce ventilator noise completely. This appears due to the difference between n(k) and n'(k). Figure 15 shows spectrograms of n'(k) and n(k). n'(k) recorded in enclosed space lost its energy between 500 Hz and 1500 Hz, 4 kHz and 5 kHz. Therefore, the adaptive filter could not reduce ventilator noise completely. Given our results, Figure 16 shows our proposed method of recording for adaptive filtering.

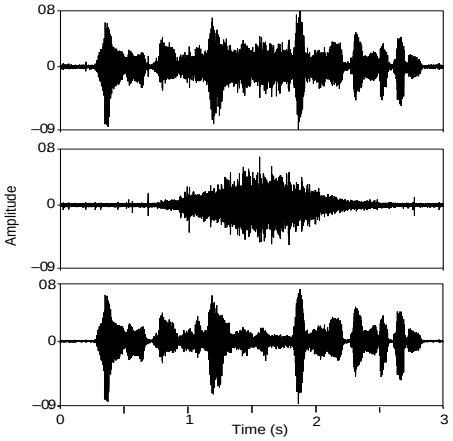

Figure 13 Main input signal (male speech + ventilator noise; upper), referred input signal(ventilator noise; middle), and output signal (lower).

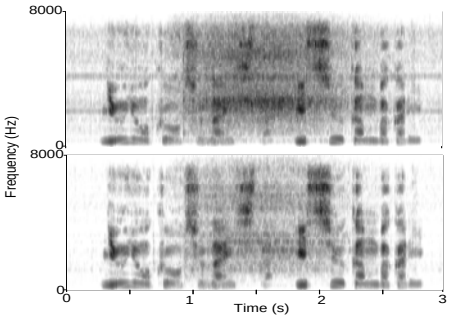

Figure 14 Spectrogram of input signal (male voice + ventilator noise; upper) and spectrogram of output signal (lower)

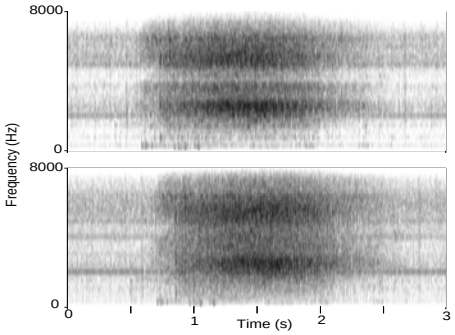

Figure 15 Spectrograms of ventilator noise: recorded in enclosed space (n'(k): upper) and recorded outside enclosed space (n(k): lower).

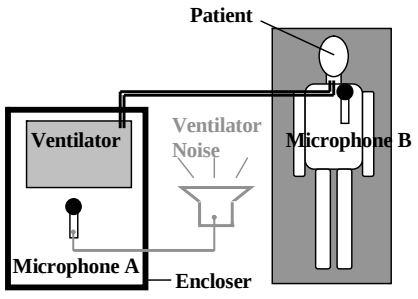

Figure 16 Proposed recording environment.

4. Conclusions

In the present study, we attempted to reduce ventilator noise in speech signals using two approaches. First, we used the conventional technique of SS, but it produced musical noise that lowered the quality of the processed signals. Therefore, we proposed a new technique that requires two steps: 1) detecting endpoints of speech in an input signal; and 2) estimating noise by compressing the amplitude of a speech portion in the input signal so that the RMS of the speech portion becomes the estimated RMS of ventilator noise. In the second technique, we processed the input signal within each 1/3 octave band. MOS evaluation revealed that Technique 2 yielded better sound quality than the conventional technique of SS.

Second, we recorded speech signals in a simulated environment and attempted to reduce ventilator noise by adaptive filtering. Using this technique, we were able to successfully reduce ventilator noise in speech signals.

5. Acknowledgement

This research was supported in part by a Grant-in-Aid for Scientific Research (A-2, 16203041) from the Japan Society for Promotion of Science. The authors would like to thank Mr. Shinichi Yamaguchi and his family for participating in this research as a speaker and also his family for their kind support.

6. Reference

[1] Y. Toyoura, Seimei no Communication, Toho shuppan, 1996.

[2] R. Tandan and W. G. Bradley, "Amyotorophic lateral sclerosis: Part 1. Clinical features, pathology, and ethical issues in management," Annals of Neurology, Vol.18, Issue 3, pp.271-280, 2004.

[3] A. Iida and N. Cambel, “Speech database design for a concatenative text-to-speech synthesis system for individuals with communication disorders,” International Journal of Speech Technology, 6, pp.379-392, 2003.

[4] Motor Neuron Disease Association, http://www.mndassociation.org/index.html.

[5] ”ima wo ikiru”, http://www.ne.jp/asahi/laconic/ikiru/index.htm.

[6] S. Furui, Onkyou-Onsei Kougaku, Kindai kagaku sha, 1992.

[7] A. W. Black and K. A. Lenzo, "FestVox Users Manual: Building sythesis voice," 2003, http://festvox.org/bsv/.

[8] W. I. Hallahan, “DECtalk software: text-to-speech technology and implementation,” Digital Technical Journal, Vol. 7, No. 4, 1995.

[9] Voiceware Co., Ltd, http://www.voiceware.co.kr/.

[10] A. Iida, N. Campbell, F. Higuchi and M. Yasumura, “A corpus-based speech synthesis system with emotion,” Speech Communication, 40, pp.161-187, 2003.

[11] P. A. Taylor, A. W. Black and R. J. Caley, "The architecture of the Festival speech synthesis system," In Third International Workshop on Speech Synthesis, Sydney, Australia, pp.147-151, 1998.

[12] A. W. Black and P. Taylor, “CHATR: A Genetic Speech Synthesis System,” Proc. COLING94, pp.983-986, 1994.

[13] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. Acoust. Speech Signal Process., ASSP-27, pp.113-120, 1979.

[14] S.Haykin, Adaptive Filter Theory, Prentice Hall, 1987.

[15] H. Tozawa, Y. Nomura, N. Yamashita, J. Lu, H. Sekiya and T. Yahagi, “Musical noise reduction using morphology process in spectral subtraction,” Workshop on Circuits and Systems, Karuizawa, 2005.

[16] Z. Goh, K.-C. Tan and B.T.G. Tan, “Postprocessing method for suppressing musical noise generated by spectral subtraction,” IEEE Trans. on Speech and Audio Processing, Vol.6, No.3, 1998.