ACOUSTICAL LETTER

Modulation cepstrum discriminating between speech and environmental noise

Tooru Miyoshi*, Takahito Goto, Takaaki Doi, Taeko Ishida,
Takayuki Arai and Yuji Murahara

Department of Electrical and Electronics Engineering, Sophia University,
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8554 Japan

(*Received 2 April 2003, Accepted for publication 26 June 2003*)

Keywords: Modulation spectrum, Modulation cepstrum, Temporal dynamics, Environmental noise, Speech/noise discrimination

PACS number: 43.72.Dv [DOI: 10.1250/ast.25.66]

1. Introduction

Environmental noises that exist around us are often the cause of degrading the performance of a system as well as doing harm to human life. Currently, many studies have been done to obtain refined speech recognition algorithms which are robust against respective kinds of noises [1] and we also need to try to find a parameter which reflects remarkable difference in properties between noise and speech signal. On the other hands, as one of the robust technique of speech recognition, there is a feature called RASTA-PLP which takes the property of human auditory system into consideration [2]. RASTA filtering is a processing for modulation spectrum which is spectral representation of the temporal dynamics of the bandpassed signal and is known to have important information on speech and speaker recognition [3,4]. In particular, for speech recognition, the improvement of the recognition rate has been achieved by using components of specific modulation frequency band [5].

In this study, we calculate the modulation spectrum from both environmental noises and speech signals in order to characterize each acoustic event. We compare and discuss the result after calculating the center of gravity of modulation cepstrum. By this method, we will see that we can obtain prominent difference between environmental noise and speech signal and confirm that this technique is efficient.

2. Method

We first divided the acoustic signal into four frequency bands. In order to obtain the modulation spectrum, we computed the logarithmic spectral representation of the temporal dynamics for each frequency band. Thus, we had four modulation spectra. In this process we removed direct current (DC) component in the modulation spectrum because the spectral pattern is not influenced with the DC component. We observed a unique spectral pattern for each acoustic signal on the modulation spectrum. To quantify this we defined the modulation cepstrum which was obtained by taking the inverse Fourier transform of each modulation spectrum. Finally, we calculated the center of gravity of accumulated cepstral pattern for each band and defined it as the distinctive index.

An overview of our signal-processing method is illustrated in Fig. 1. Details follow.

An 8,000-Hz sampled signal was first analyzed with a 32-ms Hamming window advanced in 8 ms steps. We took the fast Fourier transform (FFT) for each frame and divided it into four frequency bands (Band 1: 0–500 Hz, Band 2: 500–1,000 Hz, Band 3: 1,000–2,000 Hz, and Band 4: 2,000–4,000 Hz). We computed the sum of the energy at each frame to get the temporal dynamics for each frequency band. With this process the sampling rate becomes 125 Hz. We analyzed the bands with a 1,024-ms Hamming window advanced in 512-ms steps and performed FFT to get the modulation spectrum. Taking the inverse FFT of the modulation spectrum, we obtained the modulation cepstrum for each band. After accumulating cepstrum, we finally computed the center of gravity of the cepstrum for each band to obtain the parameter that indicates distribution pattern.

3. Experimental result

We used six types of environmental noises of "Ambient Noise Database for Telephonometry 1996" distributed by the NTT Advanced Technology and we also use "Multilingual Speech Corpus" as speech data by the Department of Eastern and Western Linguistic Culture, Tsukuba University in Japan. Six types of environmental and speech sound sources are shown in Table 1.

We performed the process described in Fig. 1. In this process we first got the temporal dynamics of the bandpassed signal. Both source signal and temporal dynamics are shown in Fig. 2. We computed the modulation spectrum that is a logarithmic spectral representation of the temporal dynamics for each frequency band. And then we calculated modulation cepstrum that was given by IFFT of modulation spectrum for each frame. We obtained the distribution that reflected the statistical feature by accumulating cepstra of each frame for signal duration. We defined the point of the center-of-gravity obtained by distribution as the feature parameter. Figure 3 shows the center of gravity point. The value on horizontal axis ranges 0–200 ms in quefrency.

4. Discussion

According to Fig. 2, we could find speech and environmental noise have own characteristics in their waveforms by themselves for each band energy. As shown in Fig. 3 the modulation spectrum has also own characteristic but it is mainly different in its shape in frequency domain. So it is difficult for us to get quantitative index from modulation

*e-mail: t_miyosh@hotmail.com

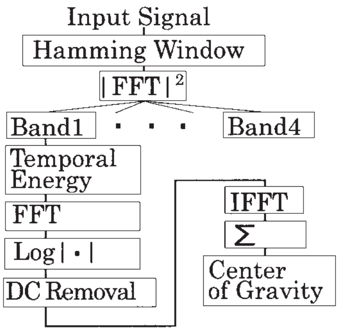

Fig. 1 Block diagram.

Table 1 Twelve sounds and their main sound sources.

Types	Main sound sources
High speed train	Noise of inside the running train
Restaurant	Noise of talking, sound of flatware and clinking table
Factory	Sound of conveyer belt and a forklift
Jumbo jet	Noise of inside the flying jet
Street	Noise of the driving automobile
Dump	Noise of the driving dump
Speech	Japanese, Chinese, English male/female

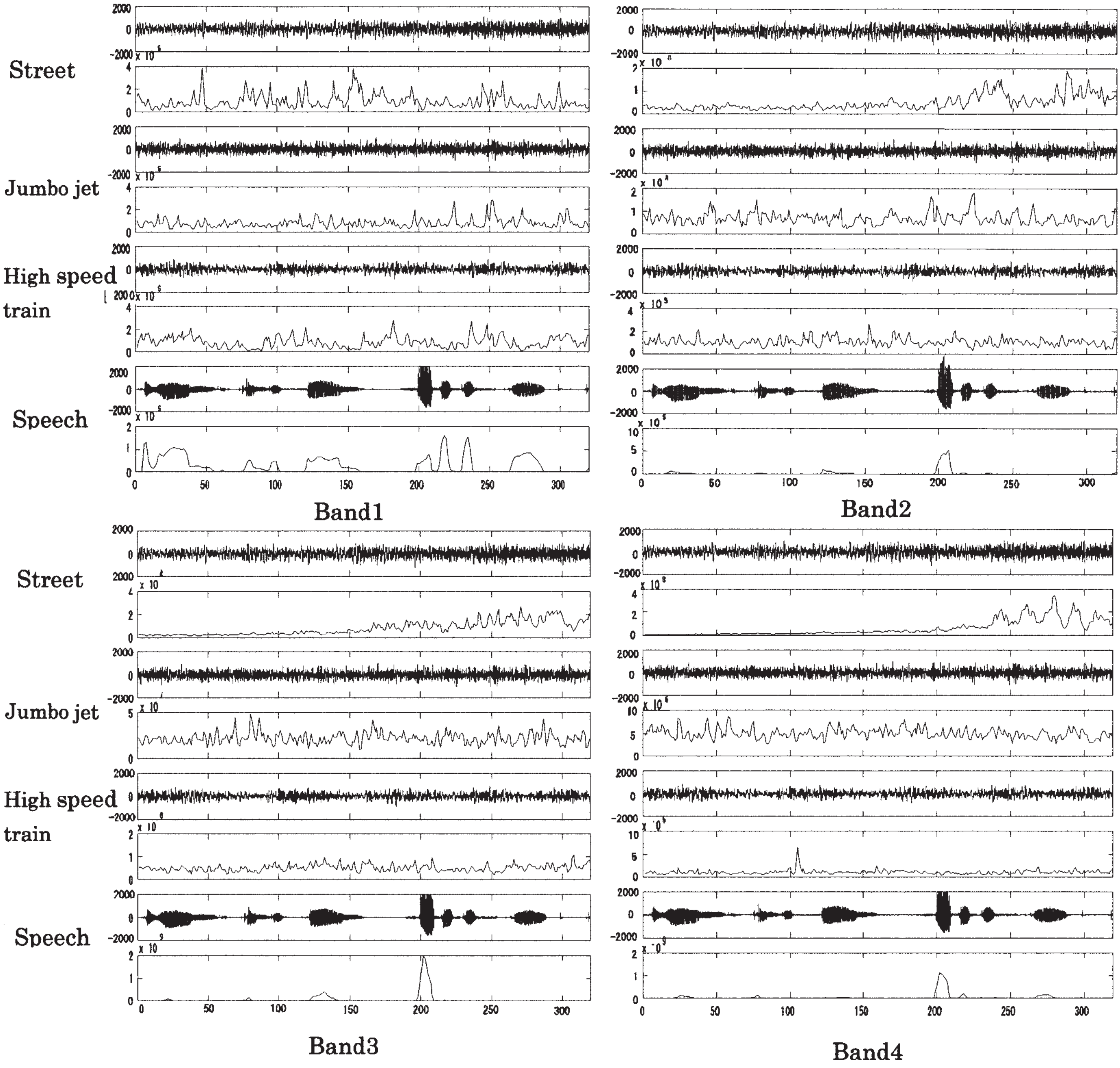

Fig. 2 Comparison between source signals and energy contours for each frequency band.

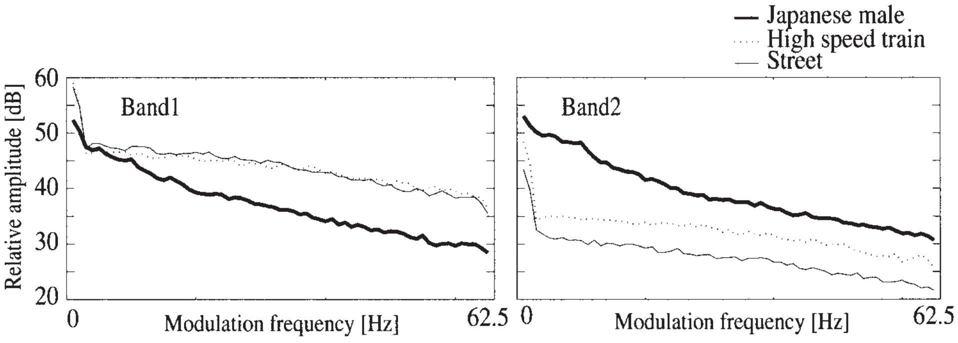

Fig. 3 Example of modulation spectra (Japanese male speech, High speed train and Street noises).

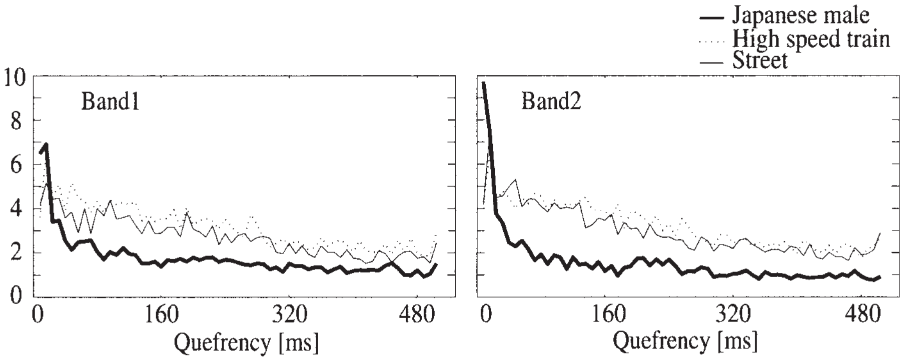

Fig. 4 Example of modulation cepstra (Japanese male speech, High speed train and Street noises).

spectrum. In order to quantify the difference in shape of modulation spectrum, we calculated the center of gravity of accumulated modulation cepstrum because we assumed that the distribution pattern of modulation cepstrum would reflect the difference in shape of modulation spectrum. As we expected, we could find manifest difference of the pattern among accumulated distributions of each modulation cepstrum, which is shown in Fig. 4. Figure 5 shows that the center of gravity of each signal reflects the difference of speech sounds and environmental noises. We can find that the point of center of gravity had critical boundary 130 ms, which does not change with regards to the difference of language and gender. If the point of center of gravity was above that critical boundary, we can recognize the acoustic signal as an environmental noise. On the other case, we can regard the signal as a speech sound.

5. Conclusion

We introduced a new concept, modulation cepstrum, and applied it to six speech sounds and six environmental noises. We discussed the center of gravity of accumulated distributions of the modulation cepstrum. By comparing the position of center of gravity, we successfully obtained the index of the discrimination between speech and environmental noise. We conclude that this feature parameter is effective for discriminating between speech and environmental noise regardless of language and gender. For future work, we would like to investigate the features more specifically not only to discriminate speech from environmental noises but also to identify different environmental noises.

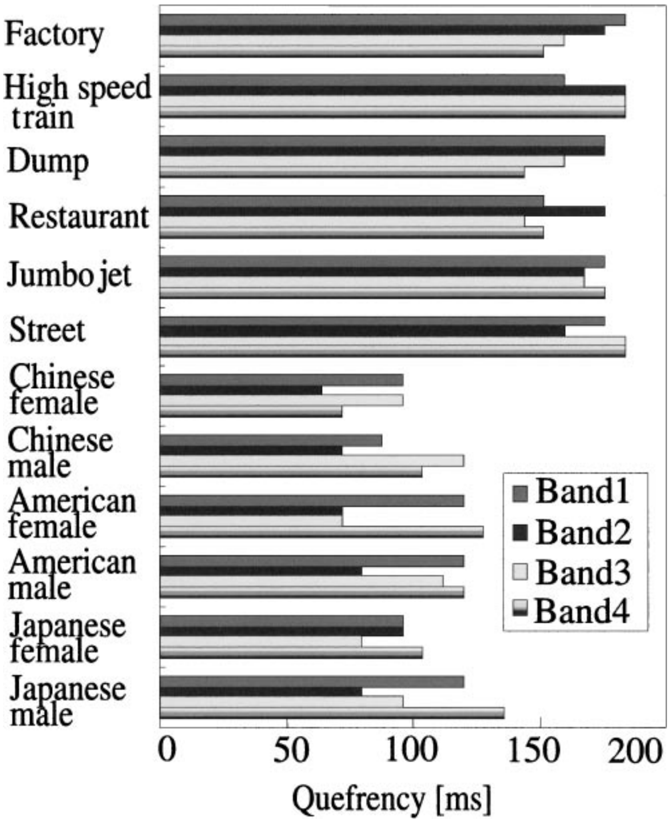

Fig. 5 Distribution of center of gravity of accumulated modulation cepstrum.

References

[1] A. Martin, D. Charlet and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC for continuous speech recognition," *Proc. Eurospeech*, Vol. 2, pp. 885–888 (2001).

[2] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, **2**, 578–589 (1994).

[3] 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).

[4] T. Arai, M. Takahashi and N. Kanedera, "On the important modulation frequency bands of speech for human speaker recognition," *Proc. ICSLP 2000*, Vol. 3, pp. 774–777 (2000).

[5] 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).