DISCRIMINATION OF SPEECH FROM ENVIRONMENTAL NOISES USING THE MODULATION CEPSTRUM

PACS: 43.72.Dv

Doi, Takaaki; Ishida, Taeko; Goto, Takahito; Miyoshi, Tooru; Arai, Takayuki and Murahara, Yuji
Dept. of Electrical and Electronics Engineering, Sophia University
7-1 Kioi-cho, Chiyoda-ku
Tokyo, 102-8554
Japan
Tel: +81-3-3238-3417
Fax: +81-3-3238-3321
E-mail: doi-t@splab.ee.sophia.ac.jp

ABSTRACT

We introduced "the modulation cepstrum," a novel representation of an acoustic signal used to distinguish speech from environmental noises, the latter being a potential cause of performance degradation of speech related processing. The modulation cepstrum was computed by taking the inverse Fourier transform of the logarithmic modulation spectrum, a spectral representation of the temporal dynamics of the signal. An experiment showed that the modulation cepstrum enabled us to identify an input signal as speech when the position of its center of gravity was less than 125 ms on the modulation cepstral domain.

INTRODUCTION

Recognition systems typically perform poorly proportionate to interference from environmental noises. To overcome such interferences, a good deal of effort has been channeled toward designing refined recognition algorithms, robust against respective noises (e.g., [1],[2]). Our concern has been to identify a parameter that underscores the properties of noise as opposed to speech in the signal. One robust utility we have available is RASTA-PLP, which takes the properties of the human auditory system into consideration [3]. RASTA filtering is a processing for a modulation spectrum, a logarithmic spectral representation of the temporal dynamics of a bandpassed signal. It is known that the components from 1-16 Hz contain important information for speech and speaker recognition [4],[5]. In particular, for automatic speech recognition, improvements have been associated with using a component of a specific modulation frequency band [6].

In this study, we apply this modulation spectrum to both environmental noises and speech. To discuss the difference between the two, we introduce "the modulation cepstrum" derived from the inverse Fourier transform of the modulation spectrum.

MODULATION CEPSTRUM

To arrive at the modulation cepstrum, we first obtained the modulation spectrum by dividing the acoustic signal into four frequency bands. Then we computed the logarithmic spectral representation of the temporal dynamics for each frequency band. Therefore, we had the same number of modulation spectra as frequency bands.

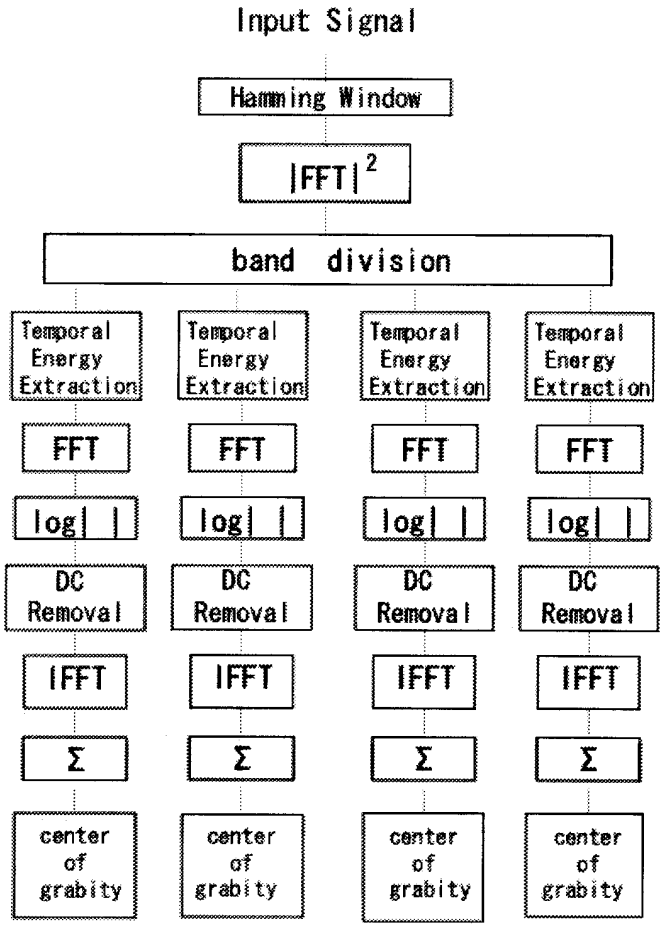

Fig. 1.- Block Diagram

We assumed there would exist a unique spectral pattern for each acoustic signal on the modulation spectrum, so we calculated the modulation cepstrum by taking the inverse Fourier transform of each modulation spectrum. In this case, we removed the direct current component. We could get the distribution of each peculiar signal by accumulating the cepstrum for each frame having a distinctive feature of total signal length, 8 sec. Next, we computed the center of gravity of each cepstrum in the qefrency domain and defined this value as the distinctive value. An overview of our signal-processing method is illustrated in Fig. 1. Details follow.

The 8000-Hz sampled signals were 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: 500-1000 Hz, Band 3: 1000-2000 Hz, and Band 4: 2000-4000 Hz). Subsequently, 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 then analyzed the bands with a 1024-ms Hamming window advanced in 512-ms steps and performed FFT to arrive at the modulation spectrum for each band. Taking the inverse FFT of the modulation spectrum, we obtained the modulation cepstrum and accumulated cepstrum for each frame. Finally, we computed the position of the center of gravity of the cepstrum for each band to arrive at the parameter that distinguishes between the modulation spectra of noise and speech.

EXPERIMENTAL RESULTS

We used six environmental noises from the "Ambient Noise Database for Telephonometry 1996" distributed by NTT Advance Technology and made available for analysis. The six environmental sounds and their main sound sources are shown in Table 1. And as speech signals, we used two sounds spoken by Japanese speakers.

The input waveforms of these six noises and two speech sounds are shown in Fig. 2. We performed the signal processing shown in Fig. 1 to the input signals in Fig. 2 to obtain the energy contours for each band as shown in Fig. 3.

To obtain the modulation spectra, we computed the logarithmic spectral representation of the energy contours for each band. Fig. 4 shows the modulation spectra of a speech signal and one of the environmental noises (the street noise). We removed the direct current component in the modulation spectrum because its spectral pattern was not influenced by our processing.

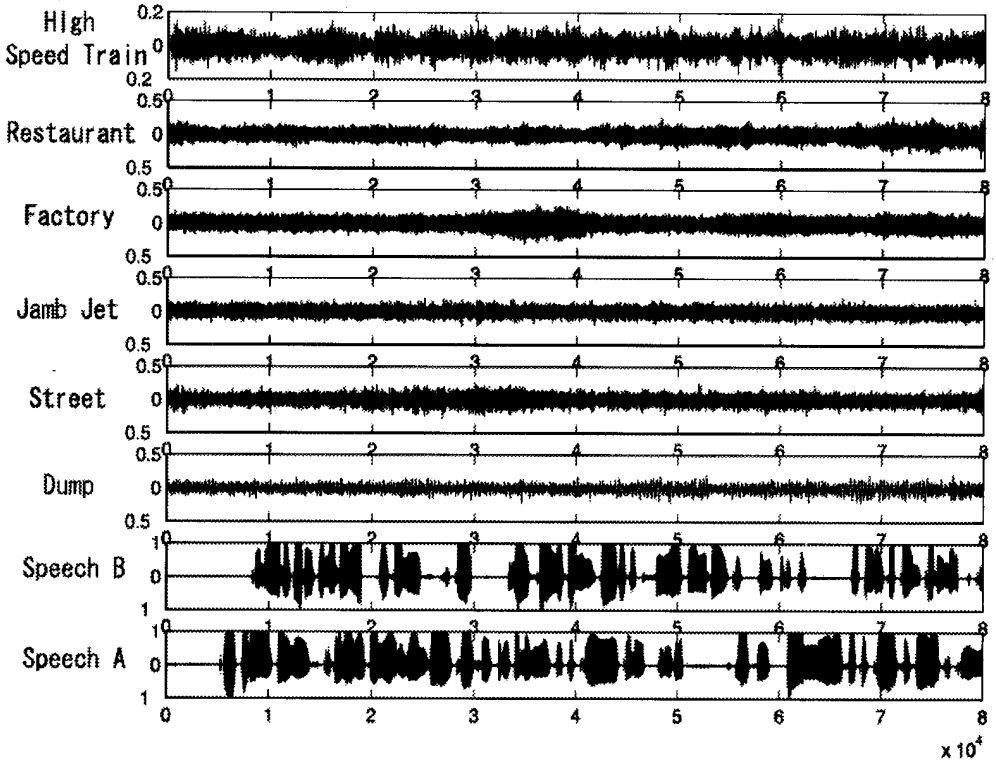

Fig. 2.- Original Signals (horizontal axes are time in sample).

Table 1.- Six environmental noises and their main sound sources

Noise Type	Main sound source
High Speed Train	Train running
Restaurant	Noise of a visitor, sound of flatware clinking on table
Factory	Sound of a conveyer belt and a forklift
Jamb Jet	Jet flying
Street	Automobile driving
Dump	Dump truck dumping, and wind

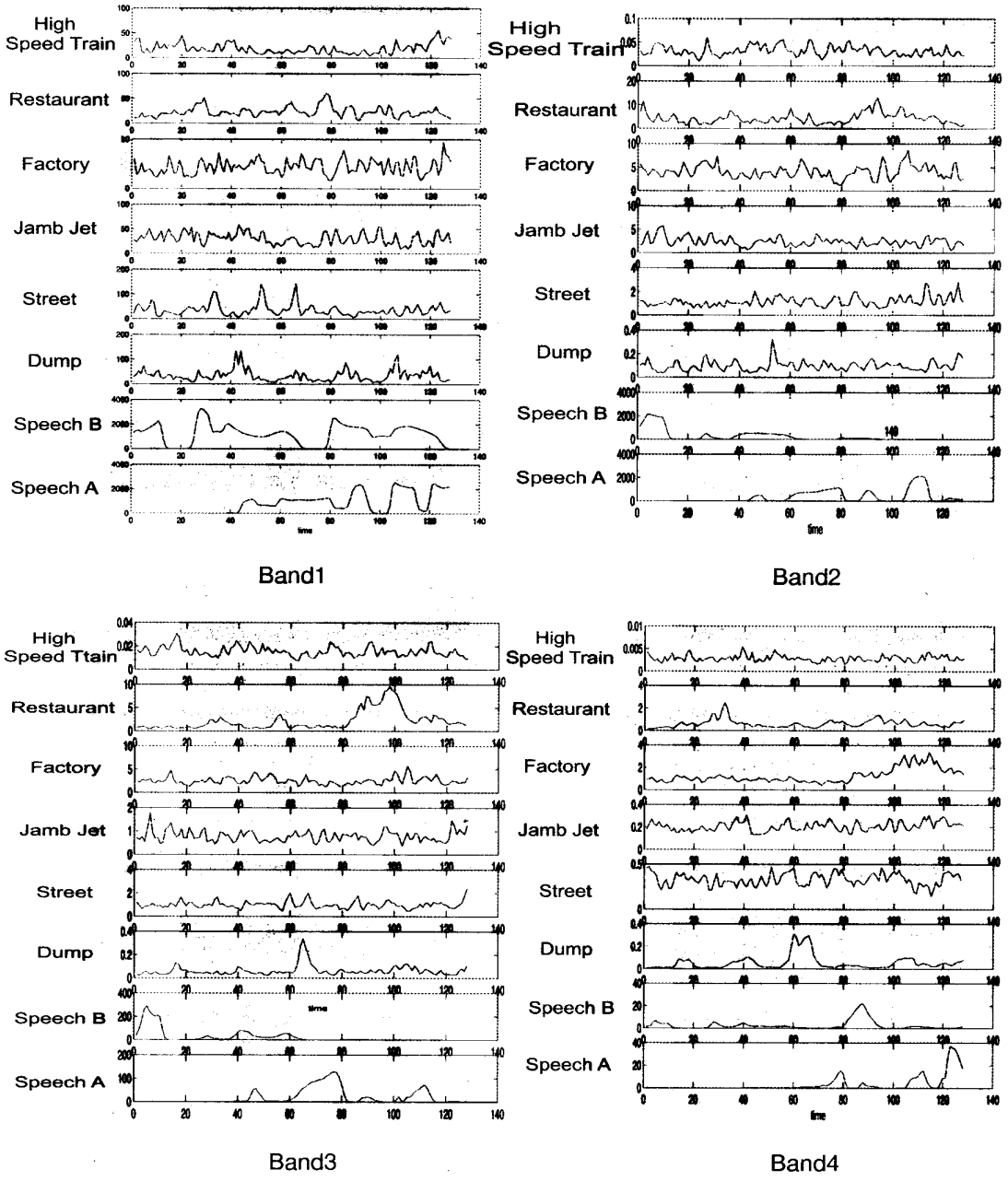

Fig. 3.- Energy contours of various signals for each band (horizontal axes are time in frame).

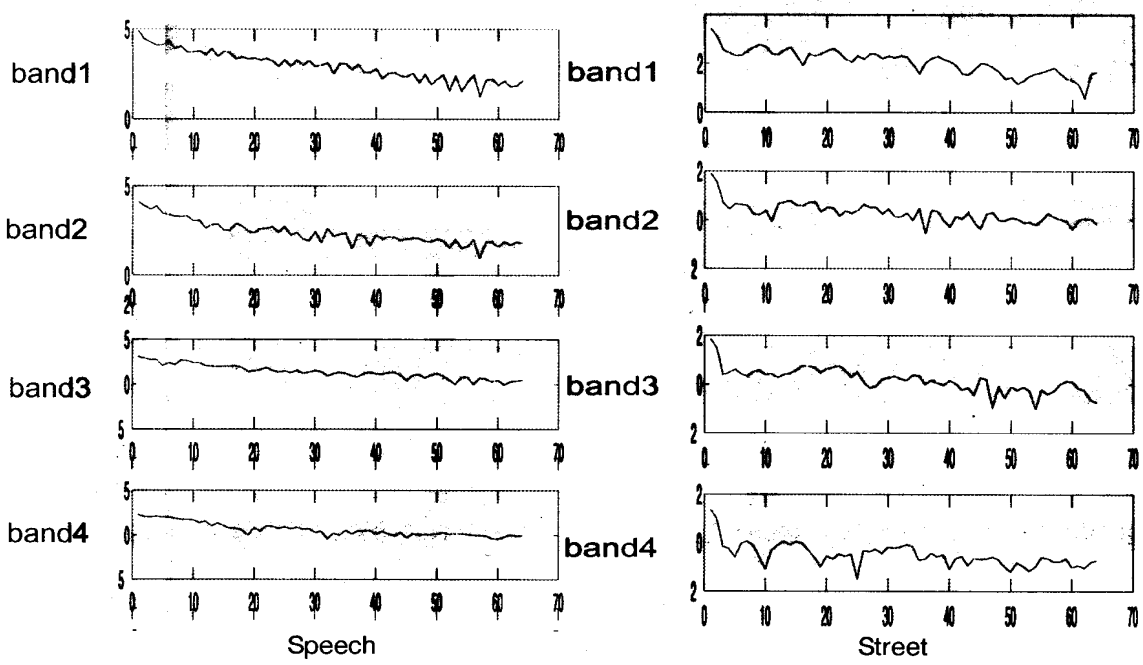

Fig. 4.- Modulation spectra of Speech (left) and Street noise (right) (horizontal axes are the modulation frequency in Hz).

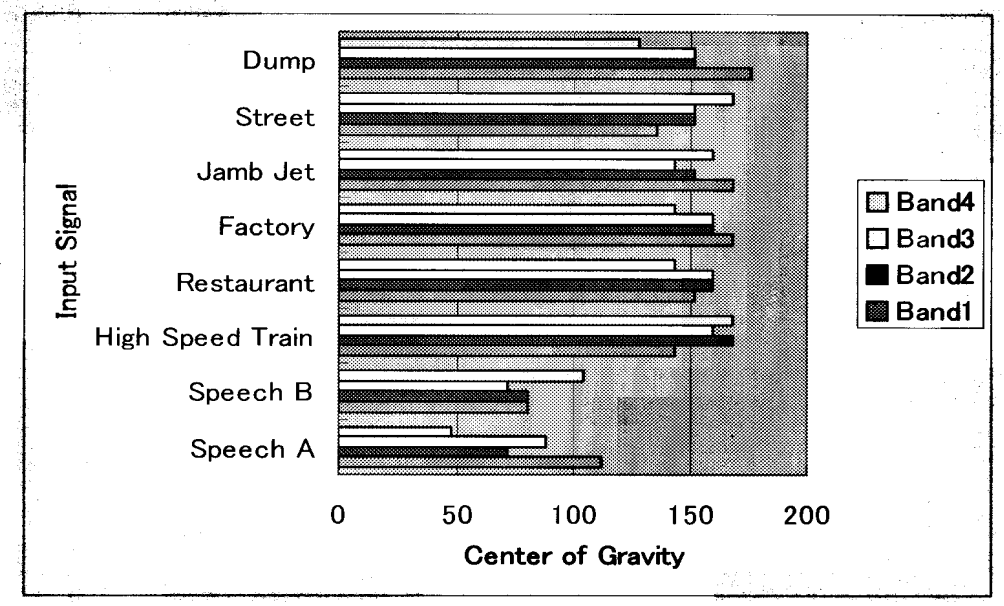

Fig. 5.- Position of the center of gravity (in ms).

Finally, we obtained the modulation cepstrum by taking the inverse FFT of the modulation spectrum. We obtained a distribution reflecting the statistical feature by accumulating at all signal continuation times because the proposed processing is performed for every frame. We computed the center-of-gravity point of this distribution. Fig. 5 shows the value of the center of gravity point searched for. The numerical value in Fig. 5 is the position of the center of gravity of the distribution pattern (0 – 200 ms in frequency domain).

DISCUSSION

From Fig. 3 one may surmise that speech and environmental noises have unique characteristics for each band in the time sequence. Fig. 4 shows that the modulation spectra resulting from the FFT of the energy contours for each band have their own distinctive features. One could further point out the manifest difference in pattern among accumulating distributions of modulation spectra in a short period. The difference between speech and environmental noise was obvious from Fig. 5, which shows the point of the center of gravity obtained in accumulated distributions. Since the point of the center of gravity on the frequency band has a critical boundary of 125 ms, this tells us that if the point of the center of gravity is above this critical boundary, we can identify an acoustic signal as environmental noise. And if the point is lower than that, it is identified as speech signal.

CONCLUSION

We introduced a new concept, the modulation cepstrum, and applied it to two speech utterances and six environmental noises. We investigated the center of gravity of accumulated distributions of the modulation cepstrum. By comparing the positions of the centers of gravity, we successfully discriminated between speech and environmental noise. We conclude that this feature parameter is effective for distinguishing speech from environmental noise.

For future work, we would like to investigate the features more specifically, not only to discriminate speech from environmental noises but also to differentiate environmental noises.

BIBLIOGRAPHICAL REFERENCES

[1] L. Mauuary and J. Monne, "Speech/non-speech detection for voice response systems," *Proc. Eurospeech*, Berlin, Germany, pp.1097-1100, 1993.
[2] J. D. Hoyt, H. Wechsler, "Detection of human speech in structured noise," *Proc. IEEE ICASSP*, Vol. 2, pp.237-239, 1994.
[3] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Process.*, Vol. 2, No. 4, pp. 578-589, 1994.
[4] T. Arai, M. Pavel, H. Hermansky and C. Avendano, "Syllable intelligibility for temporally filtered LPC cepstral trajectories," *J. Acoust. Soc. Am.*, Vol. 105, No.5, pp. 2783-2791, 1999.
[5] T. Arai, M. Takahashi and N. Kanedera, "On the important modulation frequency bands of speech for human speaker recognition," *Proc. ICSLP*, Vol. 3, pp. 774-777, 2000.
[6] 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 Communication*, Vol. 28, pp. 43-55, 1999.