China-Japan Joint Conf. on Acoustics, 191-194, 2002

MODULATION CEPSTRUM DISCRIMINATING BETWEEN SPEECH AND ENVIRONMENTAL NOISE

Tooru Miyoshi, Takahito Goto, Takaaki Doi, Taeko Ishida, Takayuki Arai and Yuji Murahara
Dept. of Electrical and Electronics Engineering, Sophia University, Tokyo, Japan

ABSTRACT

We introduce "the modulation cepstrum," a novel representation of an acoustic signal used to distinguish speech signal and environmental noises. The modulation cepstrum was computed by taking the inverse Fourier transform of the logarithmic modulation spectrum, a spectral representation of the temporal dynamics of a sub-band. We calculated the center of gravity of accumulated the modulation cepstrum for eight seconds of an acoustic signal as an index. The experimental result showed that this index enabled us to discriminate between speech and noise signals.

1.INTRODUCTION

Environmental noises often degrade the performance of a system, and sometimes they are harmful to human life. Currently, many studies have been done to obtain refined speech recognition algorithms that are robust against respective kinds of noises [1]. We also need to find a parameter that reveals the properties pertaining to noise and speech. One of the more robust techniques of speech recognition, is a feature called RASTA-PLP, which takes the properties of human auditory system into consideration [2]. RASTA filtering is a processing for modulation spectrum, which is a spectral representation of the temporal dynamics of the bandpassed signal. It is known to have important information for speech and speaker recognition [3][4]. In particular, for speech recognition, the improvement of the recognition rate has been achieved by using components of a 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 a modulation cepstrum. By this method, we will see that we can reveal prominent differences between environmental noises and speech signals, and we confirm that this technique is efficient.

2.METHOD

First, we 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 the 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 a new term, “the modulation cepstrum,” which we 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.

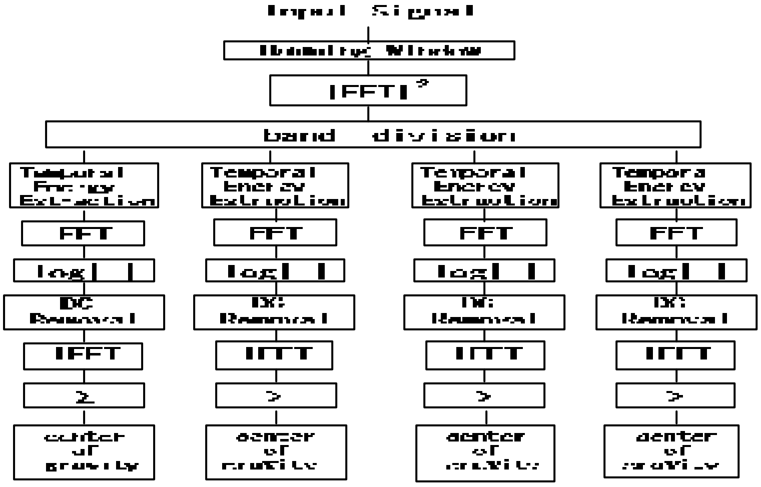

Fig. 1. Block Diagram

An 8000-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: 0-500Hz, Band 2: 500-1000Hz, Band 3: 1000-2000Hz, and Band 4: 2000-4000Hz). 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 1024-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 the distribution pattern.

3.EXPERIMENTAL RESULTS

We used six types of environmental noises from the “Ambient Noise Database for Telephonometry 1996” distributed by NTT Advance Technology and we also used the “Multilingual Speech Corpus” for speech data, from the Department of Eastern and Western Linguistic Culture, Tsukuba University in Japan. Six types of environmental and speech sound sources are shown in Table1.

Table1: 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 a 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 male/female, Chinese male/female, English male/female

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, which is a logarithmic spectral representation of the temporal dynamics for each frequency band. Then we calculated the modulation cepstrum that was given by IFFT of the modulation spectrum for each frame. We obtained the distribution that reflected the statistical feature by accumulating the cepstra of each frame for the signal duration. We defined the point of the center-of-gravity obtained by distribution as the feature parameter. Fig. 3 shows the center of gravity point. The value on the horizontal axis ranges from 0-200 ms in quefrency.

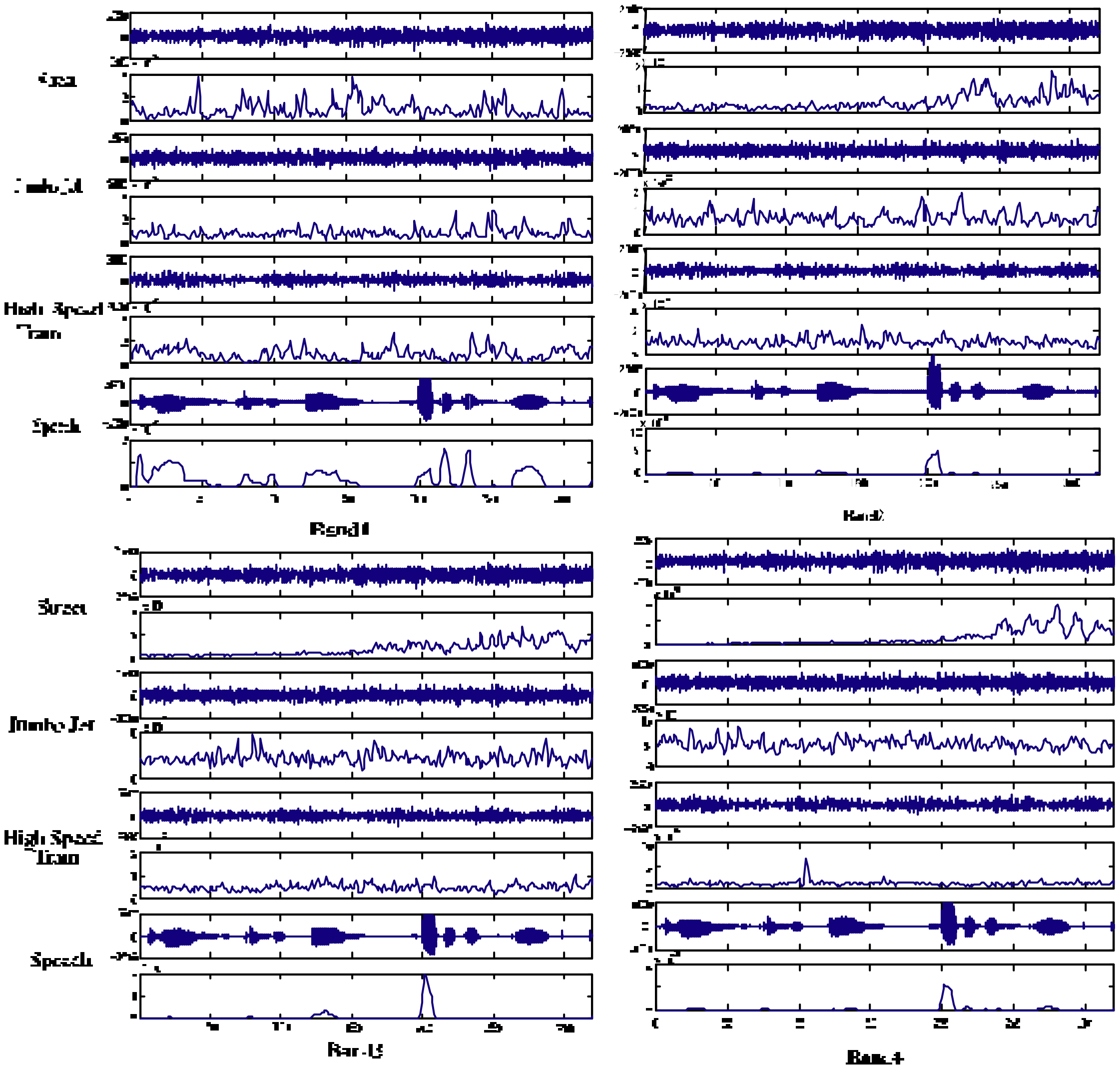

Fig. 2. Comparison between source signal and intensity contours

DISUCUSSION

According to Fig. 2, we found that speech and environmental noise have their own characteristics in the waveforms for each band energy. The modulation spectrum also has its own characteristic, but it is mainly different in its shape in the frequency domain. So it is difficult for us to get a quantitative index from the modulation spectrum. In order to quantify the difference in shape of the modulation spectrum, we calculated the center of gravity of accumulated modulation cepstrum because we assumed that the distribution pattern of the modulation cepstrum would reflect the difference in shape of the modulation spectrum. As we expected, we could find manifest difference in the pattern among accumulated distributions of each modulation cepstrum. Fig. 3 shows that the center of gravity of each signal reflects the difference between speech sounds and environmental noises. We can say that the point of the center of gravity had a critical boundary of 130 ms, which does not change with language or gender. If the point of the center of gravity was above that critical boundary, we can recognize the acoustic signal as an environmental noise. If not, it was a speech sound.

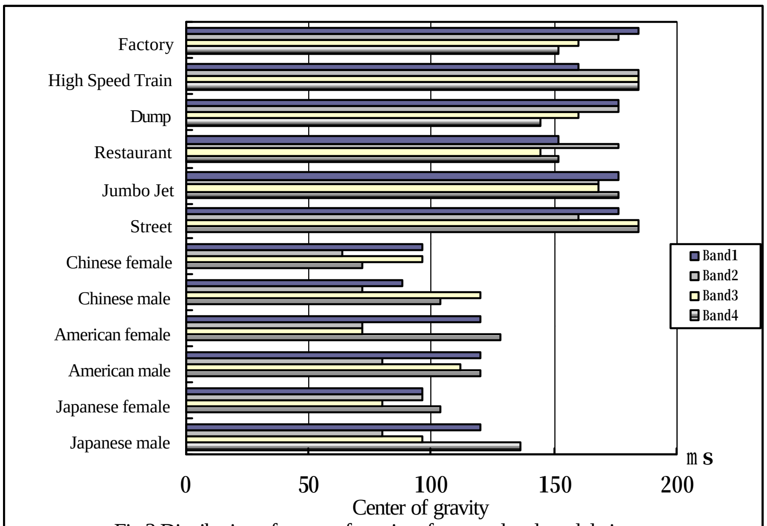

Fig.3 Distribution of center of gravity of accumulated modulation cepstrum

CONCLUSIONS

We introduced a new concept, the 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.

References

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

[2] H. Hermansky and N. Morgan, "RASTA processing of speech", IEEE Trans. Speech and Audio Processing, **Vol.2, No.4, pp.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., **Vol.105, No.5, pp.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, **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 Communication, **Vol.28, pp.43-55** (1999)