Bi-Spectral Acoustic Features for Robust Speech Recognition

Kazuo ONOE^{†a)}, *Nonmember*, Shoei SATO[†], Shinichi HOMMA[†], *Members*, Akio KOBAYASHI[†], *Nonmember*, Toru IMAI[†], and Tohru TAKAGI[†], *Members*

SUMMARY The extraction of acoustic features for robust speech recognition is very important for improving its performance in realistic environments. The bi-spectrum based on the Fourier transformation of the third-order cumulants expresses the non-Gaussianity and the phase information of the speech signal, showing the dependency between frequency components. In this letter, we propose a method of extracting short-time bi-spectral acoustic features with averaging features in a single frame. Merged with the conventional Mel frequency cepstral coefficients (MFCC) based on the power spectrum by the principal component analysis (PCA), the proposed features gave a 6.9% relative lower word error rate in Japanese broadcast news transcription experiments.

key words: bi-spectrum, non-Gaussianity, phase information, speech recognition

1. Introduction

We are studying a large vocabulary continuous speech recognition (LVCSR) system [1] for real-time subtitling of live TV programs. It has already been used in some news broadcasts for anchorpersons' read speech. To expand the subtitling to the whole news programs, it is necessary to recognize not only read speech but also speech in various situations such as reports from noisy environments or spontaneous commentaries. In this letter, we propose a new feature extraction method based on bi-spectral analysis for robust speech recognition under noisy conditions, by which conventional power spectrum-based features are influenced. The bi-spectrum obtained from the Fourier transformation of the third-order cumulants expresses the non-Gaussianity and the phase information related to the dependency of the frequency components of the speech signal, whereas the conventional features of Mel frequency cepstral coefficients (MFCCs) are obtained from the power spectrum that is the Fourier transformation of the second-order cumulants. As the bi-spectral features and MFCCs are extracted from mutually independent statistics, their complementally merged features are expected to be more robust for speech recognition. The short-time bi-spectral acoustic features are obtained by averaging features in a single frame and are merged with MFCCs by using principle component analysis (PCA).

After describing the proposed acoustic features,

Manuscript received July 3, 2007.

Manuscript revised September 10, 2007.

[†]The authors are with NHK Science and Technical Research Laboratories, Tokyo, 157-8510 Japan.

a) E-mail: onoe.k-ec@nhk.or.jp

DOI: 10.1093/ietisy/e91-d.3.631

we describe experimental speech recognition results for Japanese TV news.

2. Bi-Spectral Acoustic Features

2.1 Power Spectrum and Bi-Spectrum

The usefulness of the bi-spectrum has been known for a long time [2], [3]. There are many reports of its effectiveness in the field of noise suppression [4] and voice activity detection [5]. The bi-spectrum represents the non-Gaussianity of an observed signal, because it becomes zero if the signal is Gaussian. It also provides phase information in speech by extracting correlations between frequency components based on their phase relations [3]. This property becomes more valuable if a non-linear process is assumed as a speech production model [6], [7], which generates a frequency component even not included in the source signal [3]. The non-Gaussianity and the phase information related to the dependency of the frequency components cannot be gotten from the power spectrum. For the speech recognition under noisy conditions, the bi-spectrum does not suffer from Gaussianity included in the noise. Moreover, the bi-spectrum gives additional information about the correlations between frequency components over the conventional acoustic features based on the power spectrum. The acoustic features derived from the power spectrum and the bi-spectrum are theoretically independent, and their combination is expected to give more valuable information for speech recognition.

If $x(n)$ is a discrete time signal with a zero-mean, the second-order cumulants known as the autocorrelation and the third-order cumulants are defined as

$$c_2(\tau_1) = E\{x(n) \cdot x(n + \tau_1)\}; \quad (1)$$

$$c_3(\tau_1, \tau_2) = E\{x(n) \cdot x(n + \tau_1) \cdot x(n + \tau_2)\}, \quad (2)$$

where τ_1 and τ_2 are lags. Note that the parameter dimension of the cumulants increases with their order. The power spectrum and the bi-spectrum are defined with a discrete Fourier transformation (DFT):

$$P(f_1) = DFT(c_2(\tau_1)); \quad (3)$$

$$B(f_1, f_2) = DFT(c_3(\tau_1, \tau_2)). \quad (4)$$

Let $x(n)$ be a periodic signal with window size N , i.e., $x(n) = x(n + N)$. Then, the Fourier series $X(f)$, power spectrum $P(f)$ and bi-spectrum $B(f_1, f_2)$ are expressed as follows,

Fig. 1 Magnitude bi-spectral feature of vowel /o/.

$$X(f) = \sum_{n=0}^{N-1} x(n) e^{-j \frac{2\pi}{N} f n}; \quad (5)$$

$$P(f) = \frac{1}{N} X(f) X^*(f); \quad (6)$$

$$B(f_1, f_2) = \frac{1}{N} X(f_1) X(f_2) X^*(f_1 + f_2), \quad (7)$$

where X^* means the complex conjugate of X . The power spectrum expresses each frequency component f independently, but the bi-spectrum is a measure showing the dependency of three frequency components $(f_1, f_2, f_1 + f_2)$.

The bi-spectrum has useful characteristics in the processing of random signals.

- The cumulant spectrum of an order more than two becomes zero if the signal is Gaussian. Therefore, the non-zero cumulant spectrum provides a measure of non-Gaussian components.
- The bi-spectrum of white noise is flat because each frequency component is independent.
- The bi-spectrum shows the dependency of three frequency components $(f_1, f_2, f_1 + f_2)$ based on their phase relations.

The magnitude bi-spectrum from one frame of the vowel /o/ is shown in Fig. 1 as an example. A highly correlated pair of frequencies (f_1, f_2) gives a white grid because of the pitch frequency and its overtone elements. Due to the symmetry and the complex conjugate, the unique values of the bi-spectrum are only within the triangular region enclosed by the dotted line in Fig. 1.

2.2 Bi-Spectral Acoustic Features for Speech Recognition

The bi-spectrum is generally a complex function and is two-dimensional, and this results in too many parameters for speech recognition. The variance of the bi-spectrum estimator is greater than the variance of the power spectrum estimator. It also becomes much larger if the bi-spectrum is processed from just one frame. Therefore, it is necessary to average the bi-spectral features in the triangular region for

Fig. 2 Mean average direction of fixed f_3 .

smoothing inside a single frame.

Moreno [8] calculates such a mean value in the direction with the frequency f_1 fixed and gets one-dimensional bi-spectral features. The features are supposed to be inputs of the Mel-frequency filter banks to obtain MFCC-like acoustic features. A better result is reported with the features other than MFCCs in isolated word recognition experiments under noisy condition.

We propose a way of averaging that takes the mean values in the direction such $f_1 + f_2 = f_3$ is fixed for one-dimensional features. The features are expressed as

$$\hat{B}(f_3) = 1/F_3 \sum_{f_2=1}^{F_3} |B(f_3 - f_2, f_2)|^{1/3}, \quad (8)$$

where F_3 is a number of elements of f_3 . As shown in Figs. 1 and 2, the direction of fixed f_3 represents highly correlated bi-spectral features due to harmonic overtones on the basis of the pitch of voiced speech as well as the direction of fixed f_1 . In our preliminary experiment [9], this way of averaging was better than the way that kept f_1 fixed. $\hat{B}(f_3)$ is calculated from the cubic root of the absolute value of the bi-spectrum so that the dynamic range would be almost the same as the magnitude spectrum of the MFCC analysis. The bi-spectral acoustic features are obtained by taking $\hat{B}(f_3)$ as the inputs of the Mel-filter banks and carrying out the subsequent procedure in the same way as MFCC.

As the bi-spectral features and MFCCs are extracted from mutually independent statistics, their merged features should be more robust for speech recognition. Therefore, these features (12 dimensions of the bi-spectral features and 12 dimensions of MFCCs) are merged into one feature vector (24 dimensions in total) with a PCA transformation for de-correlation, and only the top-12 transformed dimensions are used for compression.

Samples of the Mel-frequency filter bank outputs of the MFCC and the bi-spectral features with fixed f_3 are shown in Fig. 3, where a male reporter's Japanese utterance with background noise of the audience in a stadium is analyzed. Noise components can be observed in the filter bank outputs of the MFCC, which are represented in the top left rectangular region enclosed by the dotted line, but they are much less in the filter bank outputs of the bi-spectral features. On the other hand, features supposed to be related to non-Gaussianity or dependency of frequency components appear in the filter bank outputs of the bi-spectral features represented in the bottom right region enclosed by the dotted line. Such differences in bi-spectral features from the MFCCs should be helpful to give additional information. In fact,

Fig. 3 Mel-frequency filter bank outputs of (a) the MFCC and (b) the bi-spectral features with fixed f_3 .

the proposed bi-spectral features helped to correctly recognize the above sample utterance in the following recognition experiment in which conventional MFCCs alone were not enough to recognize it.

3. Experiments

3.1 Experimental Setup

To evaluate the proposed acoustic features, we prepared an evaluation set of 643 sentences and 8,399 words from NHK's Japanese broadcast news programs. They were field reports and spontaneous commentaries uttered by male announcers and reporters. The signal to noise ratio (SNR) of the field reports subset distributed from 0 to 51 dB with background noise such as car traffic, a helicopter, rainstorm, or audience in a stadium.

The training data for the acoustic models of triphone three-state Hidden Markov Models (HMMs) were also collected from NHK news programs recorded in multiple acoustic conditions. They were 60,000 sentences (200 hours in total) uttered by male speakers. The speech data were digitized at 16 kHz sampling with pre-emphasis and a Hamming window (25 ms width and 10 ms shift). The number of Mel-frequency filter banks was 24. The data were analyzed into 39-dimensional feature parameters (12-dimensional features of MFCC, bi-spectrum, or their merged ones with the log-energy and their first- and second-order regression coefficients).

The language models were bigrams and trigrams with a lexical size of 60 K, and they were adapted to the latest news programs. The LVCSR system was a two-pass decoder

Table 1 Recognition results (WER).

	All (643 sent.)	Field reports (395 sent.)	Spontaneous commentaries (231 sent.)
MFCC	7.2%	7.0%	11.7%
Bi-spectrum (f_1)	8.5%	8.5%	13.7%
Bi-spectrum (f_3)	7.9%	7.4%	12.9%
MFCC+ bi-spectrum (f_1)	6.9%	6.6%	11.5%
MFCC+ bi-spectrum (f_3)	6.7%	6.3%	11.3%

that performed the N-best Viterbi search in the first pass with the triphone-HMMs and bigram, and rescored the N-best sentences by trigram in the second pass.

3.2 Results

Table 1 shows the recognition results for the conventional MFCCs, bi-spectral features with fixed f_1 and f_3 , and their merged features. The word error rate (WER) of the bi-spectral features alone with fixed f_3 was 0.7% worse than that of MFCC, but it was better than that of the bi-spectral features with fixed f_1 . The higher WER of the individual bi-spectral features suggests that their representation of non-Gaussianity and the dependency of frequency components was not good enough for speech recognition, unlike power spectrum-based features which were directly related to the spectral envelope. However, the bi-spectral features with fixed f_3 merged by PCA obtained a 0.5% (absolute) or 6.9% (relative) gain compared with MFCC for all the evaluation data. The improvement was 0.7% (absolute) or 10.0% (relative) for the field reports subset. The WER reduction for all the evaluation data and the field reports subset were significant in a paired difference t -test [10] at a significance level of 0.05, but it was not significant for the spontaneous commentaries subset. Bi-spectral features with fixed f_1 , however, did not show significant improvement in the t -test even after merging with MFCC. The difference in the averaging direction between f_1 and f_3 , where a higher f_3 gave more averaged samples, resulted in the difference in the recognition performance between them.

One may feel that it is strange that the bi-spectral features which individually yielded a higher WER obtained the best result after they were merged with the MFCCs by PCA. However, it is theoretically possible that new independent features could contribute to reducing the probability of error, if these new features provide any additional information over that of inadequate features, as long as they meet the complexity requirements of a classifier such as for its modeling or training samples [11]. The power spectrum and the bi-spectrum are mutually independent statistics, and the bi-spectral features provide additional information not included in MFCCs. It is considered that the effect of the additional information such as non-Gaussianity and the phase information related to the dependency of frequency

components appeared most in the field reports subset under noisy conditions in the experiment.

4. Conclusions

We described a bi-spectral analysis as a new way of feature averaging and PCA-based merging with the MFCCs. In broadcast news transcription experiments, the bi-spectral features alone did not reduce the word error rate, but the complementary merged features of PCA reduced the word error rate by 6.9% (relative) for all the evaluation data and 10.0% for the field reports subset.

In the future, we will extend our research on the bi-spectral acoustic features for more effective merging with various complimentary acoustic features and for better bi-spectral feature averaging in a single frame and over a number of frames.

References

- [1] A. Ando, T. Imai, A. Kobayashi, S. Homma, J. Goto, N. Seiyama, T. Mishima, T. Kobayakawa, S. Sato, K. Onoe, H. Segi, A. Imai, A. Matsui, A. Nakamura, H. Tanaka, T. Takagi, E. Miyasaka, and H. Isono, "A simultaneous subtitling system for broadcast news programs with a speech recognizer," *IEICE Trans. Inf. & Syst.*, vol.E86-D, no.1, pp.15–25, Jan. 2003.
- [2] M. Rosenblatt and J.W. Van Ness, "Estimation of the bispectrum," *Ann. Math. Statist.*, vol.36, no.4, pp.1120–1136, 1965.
- [3] C.L. Nikias and J.M. Mendel, "Signal processing with higher-order spectra," *IEEE Signal Process. Mag.*, vol.10, pp.10–37, 1993.
- [4] R. Fulchiero and A.S. Spanias, "Speech enhancement using the bi-spectrum," *ICASSP*, vol.4, pp.488–491, 1993.
- [5] J.M. Górriz, J. Ramírez, J.C. Segura, C.G. Puntonet, and L. García, "Effective speech/pause discrimination using an integrated bispectrum likelihood ratio test," *ICASSP*, vol.1, pp.801–804, 2006.
- [6] J. Turunen, J.T. Tantt, and P. Laula, "Hammerstein model for speech coding," *EURASIP Journal on Applied Sig. Proc.*, no.12, pp.1238–1249, 2003.
- [7] J.W.A. Fackrell, *Bispectral analysis of speech signal*, Ph. D. Thesis, University of Edinburgh, 1996.
- [8] A. Moreno and M. Rutllan, "Integrated polispectrum on speech recognition," *ICSLP*, vol.2, pp.1281–1284, 1996.
- [9] K. Onoe, S. Sato, S. Homma, A. Kobayashi, T. Imai, and T. Takagi, "A study of bi-spectrum features for speech recognition," *Proc. Spring Meeting of the ASJ*, (28), 1-P-4, 2007.
- [10] C.P. Cox, *A Handbook of Introductory Statistical Methods*, pp.46–50, John Wiley & Sons, 1987.
- [11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification Second ed.*, pp.107–111, John Wiley & Sons, 2001.