2009 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS 2009) December 7-9, 2009

TP2-D-2

Gradient-Based Acoustic Features for Speech Recognition

Takashi Muroi* , Ryoichi Takashima* , Tetsuya Takiguchi† and Yasuo Ariki†
Department of Computer Science and Systems Engineering, Kobe University, Japan
* E-mail: {muroi, takashima}@me.cs.scitec.kobe-u.ac.jp
† E-mail: {takigu, ariki}@kobe-u.ac.jp

***Abstract*—This paper proposes a novel feature extraction method for speech recognition based on gradient features on a 2-D time-frequency matrix. Widely used MFCC features lack temporal dynamics. In addition, ΔMFCC is an indirect expression of temporal frequency changes. To extract the temporal dynamics more directly, we propose local gradient features in an area around a reference position. The gradient-based features were originally proposed as HOG (Histograms of Oriented Gradients) and applied to human body detection in image recognition. In this paper, we expand the application to include gradient-based acoustic features in speech recognition. The novel acoustic features were evaluated on a word-speech recognition task, and the results showed a significant improvement for clean speech and even for noisy speech when coupled with MFCC.**

I. Introduction

In speech recognition, MFCC (Mel-Frequency Cepstrum Coefficient), which is obtained by cosine transformation of a sub-band mel-frequency spectrum within a short time, is widely used. Due to the characteristic of a short-time spectrum, MFCC lacks temporal dynamic features. To overcome this defect, the regression coefficients of MFCC (ΔMFCC) are usually utilized [1], but they are an indirect expression of temporal frequency changes, such as formant transition or high-frequency plosive.

More direct expression of the temporal frequency changes will be a geometrical feature on a two-dimensional time-frequency local area (e.g. [2], [3], [4], [5], [6]). A typical narrowband magnitude spectrogram of speech displays several important and well known phenomena [7]. Conventionally, the acoustic features based on orientation patterns obtained from the spectrum pattern [8] were proposed, and utilizing these features improved the word recognition accuracy. We have proposed a geometrical feature computed within a 3-frame by 3-frequency band local area on the temporal frequency domain [9], [10]. In the image recognition fields, local features are commonly employed in a number of real-world applications, such as object recognition and image retrieval. In recent research, gradient-based features such as SIFT (Scale Invariant Feature Transform) [11] or HOG (Histograms of Oriented Gradients) [12] achieved high recognition performance on various image tasks. The work presented in this paper applies a gradient-based feature extraction method like SIFT (or HOG) to the speech recognition task and demonstrates the effectiveness of this method. In this paper, the effectiveness of the proposed features is verified through speech recognition experiments. To evaluate noise robustness of the proposed feature, two kinds of noise were added to the clean speech at -5≤ SNR ≤10 dB.

This paper is organized as follows. In Section 2, an extraction flow of the gradient-based features for speech recognition is described. In Section 3, the proposed method is described. In Section 4, the speech recognition experiment results are discussed.

II. Extraction flow of gradient-based features

Fig. 1 shows an extraction flow of gradient-based features. At first, speech waveforms are converted into the time-frequency domain using short-time Fourier transformation. At this point, a time sequence of short-time spectra (frames) is obtained. Next, a bilateral filter is applied to the time-frequency matrix for removing noise components and smoothing the spectra. Then local gradients are computed on the area around the 2-D reference position (time, frequency) obtained by grid sampling on the smoothed time-frequency matrix, forming the local gradient matrix. Finally, orientation histograms are computed on the local gradient matrix, and the gradient-based feature vector $\mathbf{X}$ is obtained.

In speech recognition, phoneme HMMs are first trained using the gradient-based features. Then, test speech data is converted into a sequence of gradient-based features, and word likelihood is computed using the trained phoneme HMMs.

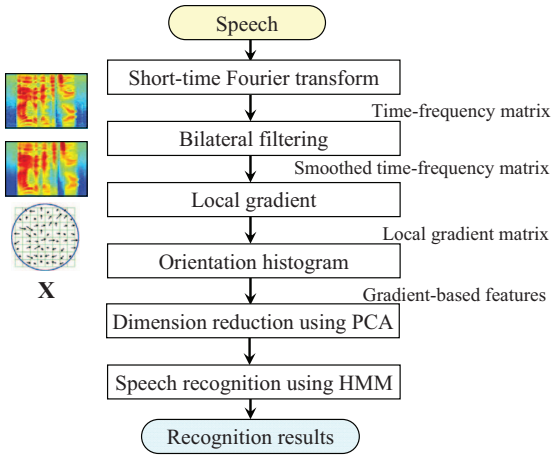


Fig. 1. Extraction flow of our gradient-based acoustic features

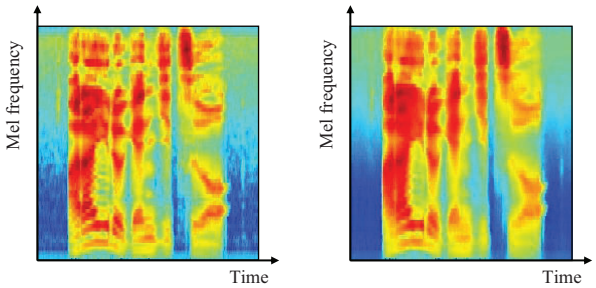
978-1-4244-5016-9/09/$25.00 ©2009 IEEE

III. Gradient-based features

A. Bilateral filtering

The spectrogram of the speech signal is composed of mel-frequency filter-bank outputs after short-time Fourier transform. In order to remove noise components contained in the spectrogram and to enhance the global changes (such as formant transition), a bilateral filter [13] was applied, which smoothed the spectrograms while preserving edges by means of a non-linear combination of the nearby values of the power spectrum.

Fig. 2 shows the mel-frequency filter-bank outputs and the filtered time-frequency matrix of the speech spoken by a Japanese male speaker under clean conditions.



(a) Filter-bank outputs (b) After bilateral filtering

Fig. 2. Bilateral filter

B. Local feature descriptor

The local feature descriptor is obtained at the reference position as the orientation histogram of the local gradient features computed on the smoothed time-frequency matrix.

1) Local gradient features: We compute local gradient features on the smoothed time-frequency matrix after bilateral filtering. Fig. 3 illustrates the local feature gradients. The gradient magnitudes and orientations are extracted around the reference position, using the power spectrum of the bilateral filtered mel-scale time-frequency matrix. At each position (t, f), the gradient magnitude $m(t, f)$ and orientation $\theta(t, f)$ are computed as follows:

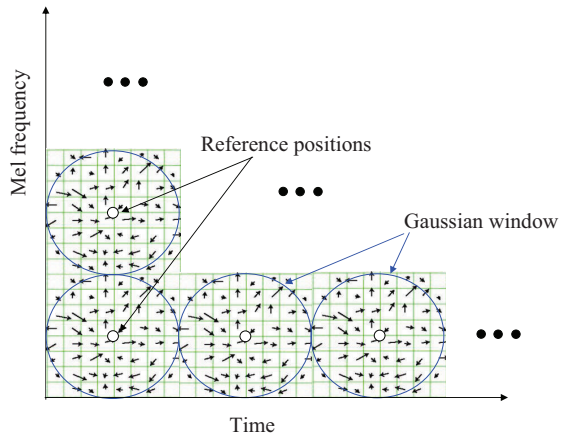


Fig. 3. Local gradient features in the time-frequency domain

$$m(t,f) = \sqrt{d_t(t,f)^2 + d_f(t,f)^2} \quad (1)$$

$$\theta(t,f) = tan^{-1}\frac{d_f(t,f)}{d_t(t,f)} \quad (2)$$

$$\begin{cases} d_t(t,f) = r(t+1,f) - r(t-1,f) \\ d_f(t,f) = r(t,f+1) - r(t,f-1) \end{cases} \quad (3)$$

where $r(t, f)$ is the power spectrum at the position (t, f) on the smoothed time-frequency matrix composed of time t and frequency f. The gradient magnitudes are weighted by a two-dimensional Gaussian circular window. (An 8×8 window was used in this paper.) The purpose of this Gaussian window is to put less emphasis on gradients that are far from the center.

2) Orientation histogram: The orientation histogram was calculated from local gradient features in a 4×4 area, where we divided an 8×8 area into four separate areas as shown in Fig. 4. Each orientation histogram was calculated from the gradient magnitude $m(t, f)$ and orientation $\theta(t, f)$ as follows:

$$h_l(\theta') = \sum_x \sum_y m(x,y)\delta[\theta', \theta(x,y)] \quad (4)$$

$$\theta' = \{0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°\} \quad (5)$$

where (x, y) is a grid point in each (4×4) area, and l = 1, 2, 3, 4 is the index of each area. In this paper, we quantize $\theta(t, f)$ to the 8 discrete orientation levels θ' as shown in Equation (5). The δ function in Equation (4) is the Kronecker delta, and it returns 1 if the quantized $\theta(x, y)$ is equal to θ'.

The local feature descriptor was formed as a vector containing the four orientation histograms. In the experiments, the local feature descriptor was composed of 8 orientations × 4 areas = 32 elements at each reference position.

Finally, the gradient-based feature vector $\mathbf{X}$ was obtained by packing the local feature descriptors at each frame as shown in Fig. 5. In our experiments, we calculated 64 mel-frequency bands, and eight reference positions were set at each frame by grid sampling (without overlap areas). Therefore, the number of dimensions of the vector $\mathbf{X}$ is 256 (32 elements × 8 descriptors). Because this is so high, the HMMs used in the speech recognition may be estimated inaccurately and unstably. To solve this problem, PCA (Principal Component Analysis) was used to reduce the dimension effectively.

IV. Word-recognition experiments

A. Experimental setup

In order to confirm the efficiency of the proposed method, the speech data were extracted from the A-set of the ATR Japanese database, and the noise data were extracted from the CENSREC-1-C database [14]. The total number of speakers was ten (5 males and 5 females). The training data were composed of 2,620 utterances per speaker, and 1,000 clean or noise-added utterances were used for testing each speaker. The

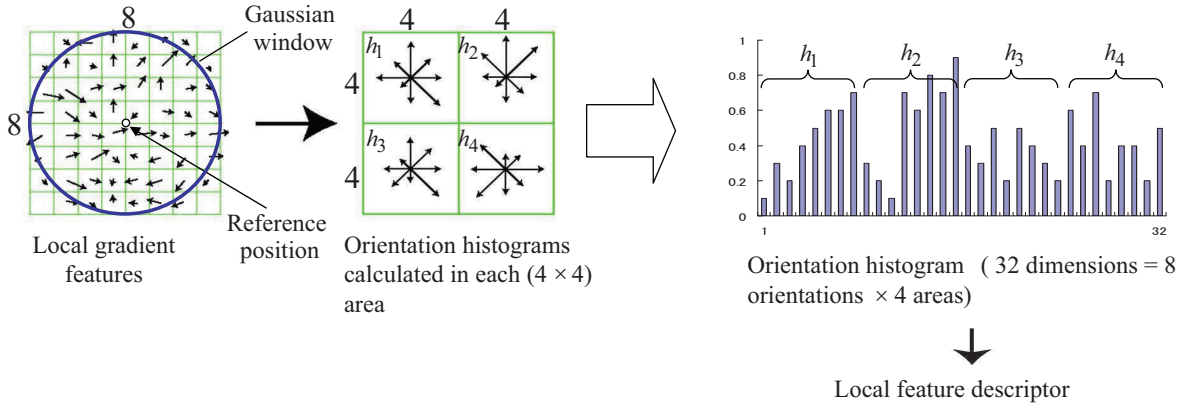


Fig. 4. Orientation histogram

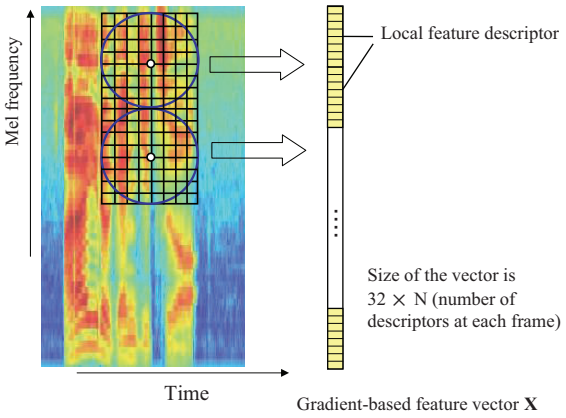


Fig. 5. Gradient-based feature vector obtained by packing the local feature descriptors at a frame

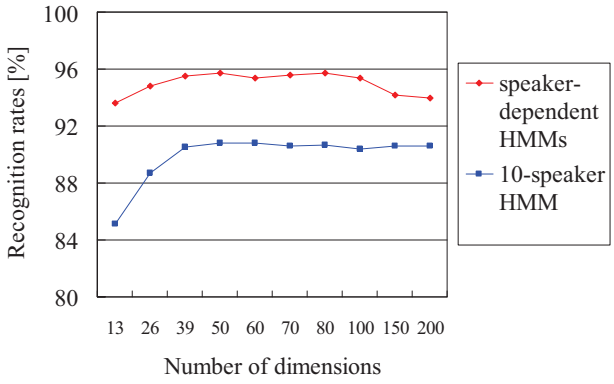


Fig. 6. Results of word recognition as a function of the number of feature dimensions after PCA

speech waveforms were transformed into the time-frequency matrix using short-time Fourier transformation with 25-ms frame width and 10-ms frame shift. Next, a 64-channel mel-frequency filter-bank analysis was performed on the above components. To evaluate our method with real noise data, two kinds of noise data (restaurant and street) were used from the CENSREC-1-C database [14], which were recorded in -5≤ SNR ≤10 dB environments. Word recognition rates were computed by averaging the results from the ten speakers.

B. Experimental results

The gradient-based feature vector **X** obtained in this paper equals the 256-dimensional vector (32 elements × 8 descriptors). It is a large-dimension vector, so PCA (Principal Component Analysis) is used to reduce the dimension effectively. Fig. 6 shows the word recognition results as a function of the number of feature dimensions using PCA, where the speaker-dependent HMMs and the HMM trained using ten speakers were used. The highest recognition rate for both HMMs was obtained when there were 50 dimensions of the gradient-based features vector **X**. So, the following experiments used the 50-dimension gradient-based features and were carried out using the HMM trained using ten speakers.

The results of word recognition using the gradient-based features, compared with the recognition results using MFCC are shown in Fig. 7. The recognition rates using the gradient-based features improved to 90.8%, 72.4%, and 67.6% for clean, restaurant, and street environments, respectively, due to accumulation of the direct expression of temporal features.

Since the gradient-based features showed a high speech recognition rate using a single feature for each kind of noise, we examined combinations with MFCC and Δ MFCC. The feature combination is based on a stream weighting method that concatenated each feature vector by equally weighting the respective feature. As shown in Fig. 8, compared with the results of MFCC and ΔMFCC only, the gradient-based features (HOG) improved the recognition rates by 3.8% for both types of noise after the HOG features were combined with MFCC and Δ MFCC. This indicates that the gradient-based features contain information that can improve the recognition rate obtained by MFCC and ΔMFCC combination.

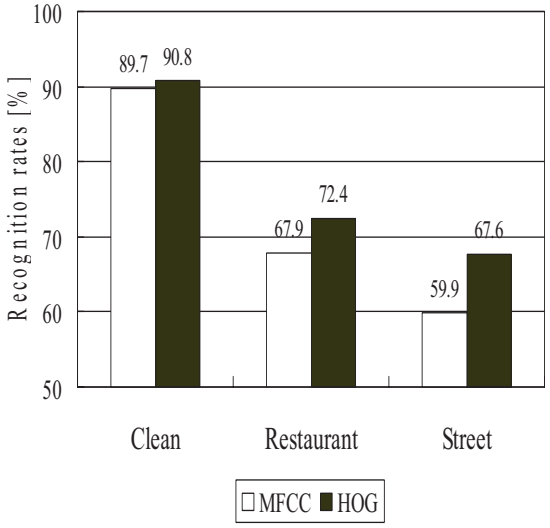


Fig. 7. Results of word recognition when using each single feature

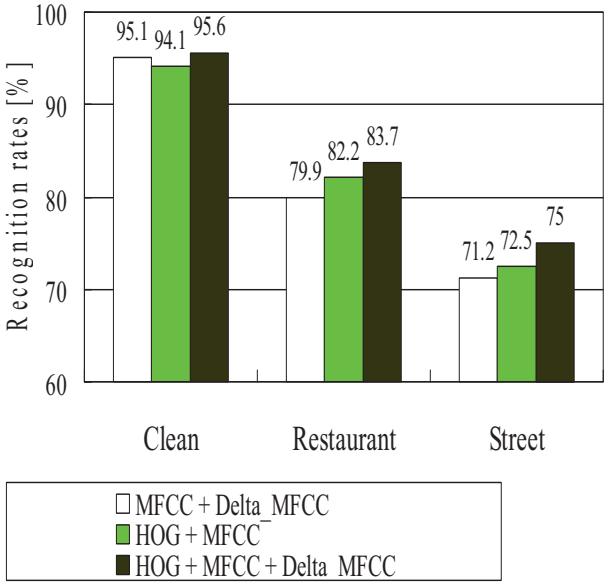


Fig. 8. Results of word recognition when using combined features

V. Conclusion

We described a new feature extraction method based on gradient orientations of the power spectrum on a time-frequency matrix. Experimental comparisons with MFCC in noisy environments corrupted by two kinds of noise have suggested that the proposed feature offers better encoding of temporal features and is more noise robust than MFCC.

In future research, we will investigate some optimal numbers for the Gaussian window size, the area size of orientation histogram, and the discrete orientation level. Then we will verify the effectiveness of the method for other types of noises. We will also investigate combinations of speech enhancement or model adaptation techniques (e.g. [15], [16], [17]).

References

[1] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum," IEEE Transactions on Acoustic, Speech, and Signal Processing, vol. 34, pp. 52-59, 1986.

[2] T. Nitta, "Feature Extraction for Speech Recognition Based on Orthogonal Acoustic- feature Planes and LDA," Proc. of ICASSP, pp. 421-424, 1999.

[3] S. Y. Zhao and N. Morgan, "Multi-Stream Spectro-Temporal Features for Robust Speech Recognition," Proc. of Interspeech, pp. 898-901, 2008.

[4] B. T. Meyer and B. Kollmeier, "Optimization and Evaluation of Gabor feature sets for ASR," Proc. of Interspeech, pp. 906-909, 2008.

[5] S.-N. Tsai; L.-S Lee, "Improved robust features for speech recognition by integrating time-frequency principal components (TFPC) and histogram equalization (HEQ)," Proc. of ASRU, pp. 297-302, 2003.

[6] Y. Li and D. Wang, "Musical Sound Separation Using Pitch-Based Labeling and Binary Time-Frequency Masking," Proc. of ICASSP, pp. 173-176, 2008.

[7] K. Schutte and J. Glass, "Speech Recognition with Localized Time-Frequency Pattern Detectors," Proc. of ASRU, pp. 341-344, 2007.

[8] H. Matumura, R. Oka, K. Kogure, and Y. Kojima, "Speaker-Independent Spoken Word Recognition by Using the Orientation Patterns Obtained from the Vector Field of Spectrum Pattern," Transactions of IEICE, Vol. 72-D-II, No. 4 , pp. 487-498, 1989.

[9] Y. Ariki, S. Kato, and T. Takiguchi, "Phoneme Recognition Based on Fisher Weight Map to Higher-Order Local Auto-Correlation," Proc. of Interspeech, pp. 377-380, 2006.

[10] T. Muroi, T. Takiguchi, and Y. Ariki, "Speaker Independent Phoneme Recognition Based on Fisher Weight Map," Proc. of the 2nd International Conference on Multimedia and Ubiquitous Engineering, pp. 253-257, 2008.

[11] D. Lowe, "Distinctive image feature from scale invariant keypoints," Proc. of International Journal of Conference on ComputerVision (IJCV), 60(2), pp. 91-110, 2004.

[12] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 720-723, 2007.

[13] C. Tomasi, R. Manduchi, "Bilateral Filtering for Gray and Color Images," Proc. of International Conference on Computer Vision, pp. 829-846, 1998.

[14] N. Kitaoka, K. Yamamoto, T. Kusamizu, S. Nakagawa, T. Yamada, S. Tsuge, C. Miyajima, T. Nishiura, M. Nakayama, Y. Denda, M. Fujimoto, T. Takiguchi, S. Tamura, S. Kuroiwa, K. Takeda, and S. Nakamura, "Development of VAD evaluation framework CENSREC-1-C and investigation of relationship between VAD and speech recognition performance," Proc. of ASRU, pp. 607-612, 2007.

[15] D. Yu, L. Deng, J. Droppo, J. Wu, Y. Gong, and A. Acero, "A Minimum-Mean-Square-Error Noise Reduction Algorithm on Melfrequency Cepstra for Robust Speech Recognition," Proc. of ICASSP, pp. 4041-4044, 2008.

[16] R. Gomez, T. Toda, H. Saruwatari, K. Shikano, "Improving Rapid Unsupervised Speaker Adaptation Based on HMM Sufficient Statistics," Proc. of ICASSP, pp. 1001-1004, 2006.

[17] A. Betkowska, K. Shinoda, and S. Furui, "Speech Recognition Using FHMMs Robust Against Nonstationary Noise," Proc. of ICASSP, pp. 1029-1032, 2007.

[18] K. Yu, M. J. F. Gales, and P. C. Woodland, "Unsupervised Discriminative Adaptation Using Discriminative Mapping Transforms," Proc. of ICASSP, pp. 4273-4276, 2008.