TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 05/23 pp158-161
Volume 10, Number 2, April 2005

Improved MFCC-Based Feature for Robust Speaker Identification*

WU Zunjing (吴尊敬), CAO Zhigang (曹志刚)**

State Key Laboratory on Microwave and Digital Communications,
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Abstract: The Mel-frequency cepstral coefficient (MFCC) is the most widely used feature in speech and speaker recognition. However, MFCC is very sensitive to noise interference, which tends to drastically degrade the performance of recognition systems because of the mismatches between training and testing. In this paper, the logarithmic transformation in the standard MFCC analysis is replaced by a combined function to improve the noisy sensitivity. The proposed feature extraction process is also combined with speech enhancement methods, such as spectral subtraction and median-filter to further suppress the noise. Experiments show that the proposed robust MFCC-based feature significantly reduces the recognition error rate over a wide signal-to-noise ratio range.

Key words: Mel-frequency cepstral coefficient (MFCC); robust speaker identification; feature extraction

Introduction

Speaker recognition systems based on the Mel-frequency cepstral coefficient (MFCC) feature can provide a high recognition rate for clean speech, but its performance degrades dramatically in noisy environments. Noise interference introduces mismatches between the pre-trained acoustic models and the input features. The techniques for robust speaker recognition can be classified into three categories based on the working spaces. The first category pre-processes the noisy speech signal to obtain a better estimation of clean speech, such as the traditional spectral subtraction (SS)[1] and minimum mean-square error estimation[2] methods. The second category focuses on robust feature representation of the speech, e.g., RASTA processing[3], cepstral mean subtraction[4], and adaptive component weighting[5]. In the third category, acoustic model parameters are adapted to match the noisy speech, such as the parallel model combination[6] method.

This paper presents a method combining robust representations in the feature space and speech enhancement in signal space. The main objective in the feature space is to extract the acoustic features that give an accurate description of the stochastic characteristics of the input speech. The features are expected to be robust to the noise and can be modeled by the hidden Markov model or Gaussion mixture model distribution. The standard MFCC feature analysis is very successful except when the noise is present. Because the log function in the MFCC generation is very sensitive to noise, a combined piecewise function was developed which combines a power law function and the log function (referred as PL) to replace the log function to improve the MFCC robustness. In signal space, the classical SS method effectively suppresses noise to improve the signal-to-noise (SNR) of the input speech. Most modulation energy of speech is at around 4 Hz, so a median-filter (MF) for filter-bank energies on time trajectory can be used to suppress spectral components that change more quickly than the speech. SS and MF methods can be combined together with newly

Received: 2003-10-01; revised: 2004-04-16
* Supported by the National Natural Science Foundation of China (No. 6007201)
** To whom correspondence should be addressed.
E-mail: caozg-dee@tsinghua.edu.cn; Tel: 86-10-62771026

proposed transformation function to further increase the system robustness.

1 Compression Function

The standard MFCC analysis consists of five steps[7]:

1) Pre-process the input speech and detect the end-points;

2) Perform an fast Fourier transform on the input speech signal;

3) Calculate the Mel-frequency bank energies by integrating the spectral energy coefficients within triangular frequency bins arranged uniformly on the Mel-frequency scale;

4) Perform the discrete cosine transform on the logarithm of the filter-bank energies;

5) Append first order differentials.

In step 3, the log transformation nonlinearly compresses the filter-bank energies in accord with a human auditory response. However, the log transformation is sensitive to noise because its slope is very steep when the energy is low, so serious mismatches are introduced in low-energy log filter-banks. The effects of the large variation induced by the log transformation can be seen clearly in Fig. 1.

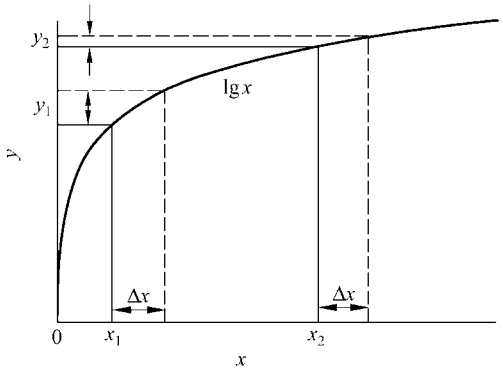

Fig. 1 Mismatches induced by log transformation

Assume that a noise signal Δx is added to x at x_1 or x_2. The mismatch after the log transformation with the noise Δx is Δy_1 for x_1 and Δy_2 for x_2. Clearly, noise influence on the log filter-banks at low energies is much larger than high-energy banks due to the log function effect. The log filter-bank energies of clean speech and noisy speech in white noise environment are shown in Fig. 2. The mismatches between the clean speech and the noisy speech are very large for low-energy banks on the ordinate.

Since the low energy banks tend to be affected by the noise, a method was developed to increase the compression ratio for these banks. The lower segment of the log function was replaced by a power function as:

$$f_{\mathrm{PL}}(x)=\begin{cases}\lambda x^{\frac{1}{\lambda}}/C^{\frac{1}{\lambda}} & x\le C;\\ \lg x+\lambda-\lg C, & x>C.\end{cases}$$

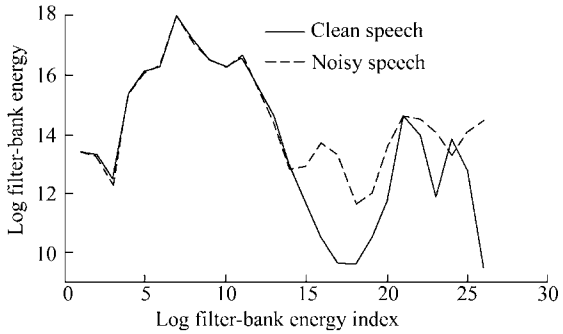

Fig. 2 Log filter-bank energies of clean and noisy speech

Although the function is piecewise, it is still continuous and smooth. This function ($f_{\mathrm{PL}}(x)$) combines a power function and the log function, where C is the noise masking level and λ is the compression coefficient. However, some detail information is lost by the scheme. C and λ must be chosen carefully to make a trade-off between the compression increase and the information loss. C should be chosen according to the noise level with the best value about equal to the Mel-frequency bank energies of the background noise. λ should be selected to reduce the effects of noise while reserving as much speech information as possible to provide high clean recognition accuracy. The filter-bank energies using the combined transformation with $C=5\times10^6$ and $\lambda=2$ are shown in Fig. 3. The mismatch is much less in Fig. 3 for low-energy banks.

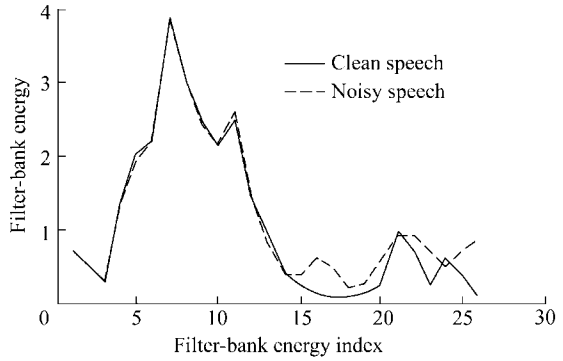

Fig. 3 Filter-bank energies using the combined transformation

When C and λ equal different values, the functional forms are illustrated in Fig. 4 for several values of C and λ. When $\lambda \to +\infty$, the function approximates the log function, so the log function is a special case of this combined function. Several different functions with different parameters were studied to choose the optimal function.

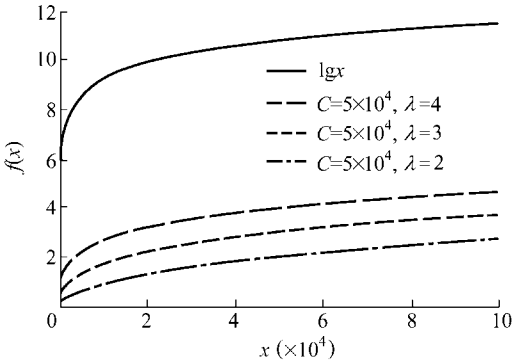

Fig. 4 Transformation functions corresponding to different parameters *C* and λ

2 Combining Proposed Function with SS and MF

Classic spectral subtraction is a simple and effective speech enhancement method. The median-filter for filter-bank energies on the time trajectory suppresses the spectral components that change more quickly than the speech. Therefore, a combination of these methods with the proposed transformation function $f_{\mathrm{PL}}(x)$ (referred to here as SS_MF_PL) was used to extract the robust feature. The standard MFCC analysis described in Section 1 was modified to include these methods in the feature extraction process. In step 1, the SS method was added with the noise estimation based on the first 20 frames and the results of the endpoint detection. After step 3, a median-filter was used to smooth the filter bank energies and to restrain the high-frequency components, with the filter bank energies then normalized by the integral area of the corresponding triangular frequency bin. The normalization seeks to normalize the noise in every filter bank to facilitate selection of one noise mask C for all banks. Step 4 used the transformation function $f_{\mathrm{pl}}(x)$ to replace the log function.

3 Experimental Results

The system was evaluated with a 26-dimensional MFCC_0_D and a 32-mixture Gaussion mixture model. A set of 30 speakers was selected from the dr6' region of the TIMIT database, including 16 females and 14 males. Each speaker provided 10 sentences lasting about 3 s each with 8 sentences used to build the Gaussian mixture models and the other 2 sentences used for the test. Silent sentences were removed by endpoint detection based on the frame energies. Gaussian white noise from the Noisex-92 database was added to test sentences artificially at different SNR levels of 0 dB, 5 dB, 10 dB, 20 dB, and clean speech.

Two tests were carried out with the first test evaluating the performances of the baseline system and enhanced systems using the SS and SS_MF (combination of SS and MF) methods, the second evaluating the combination method described in Section 2 with various values of C and λ.

3.1 Test 1

Results from the first set of experiments for baseline, SS and SS_MF methods are illustrated in Fig. 5. The identification accuracy for clean speech was 99.7% for the baseline system and 99.5% for the SS enhanced system. With the SS_MF system, the identification accuracy was reduced to 97.2% for clean speech. For noisy speech with SNR from 0 to 20 dB, the results in Fig. 5 show that the baseline system performance degraded rapidly with the SS method providing improved system performance in noisy environments and the SS_MF method providing better performance in the low SNR ranges.

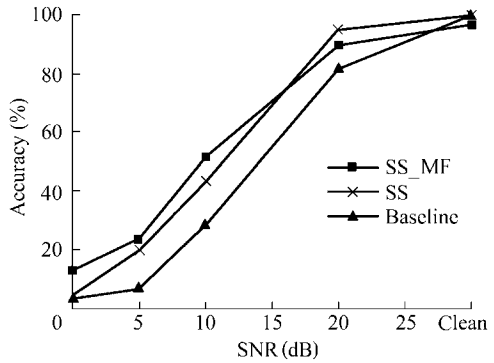

Fig. 5 Performance of baseline, SS, and SS_MF methods

3.2 Test 2

The second set of tests evaluated the combined SS_PL_MF method with various values of C and λ. C

was chosen to be 10^5, 10^6, 10^7, and 10^8 with λ equal to 2 or 3. The results are shown in Figs. 6 and 7.

The results in Figs. 6 and 7 show that the present improved system very effectively improves the system performance in very noisy environments compared with the SS and SS_MF methods. However, the system performance is dependent on the values of C and λ. Increasing the value of C increases the identification accuracy for low SNR while degrades the identification accuracy for high SNR because of the competing influences of the lost of speech information and the noise reduction. In general, the combined system with λ =2 performs better than with λ =3.

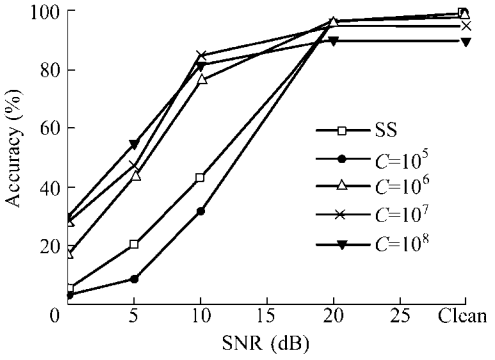

Fig. 6 Performance of SS_PL_MF system for λ =2

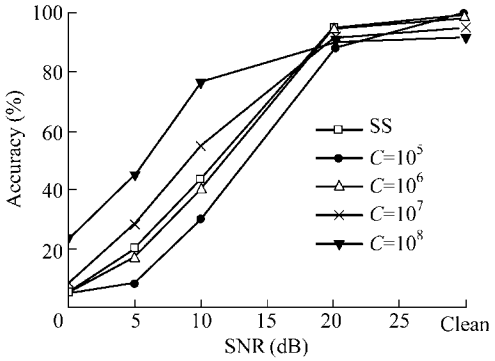

Fig. 7 Performance of SS_PL_MF system for λ =3

As described above, C should be chosen according to the noise level with the best value about equal to the Mel-frequency bank energies of the background noise. In the current tests, the noise filter-bank energies were close to 10^8 for SNR equal to 0 dB, 10^7 for 10 dB, and 10^6 for 20 dB. For λ =2, C=10^8 achieved the best performance for 0 dB and 5 dB, while C=10^7 was better for 10 dB and C=10^6 was better for 20 dB, which validates that C should be equal to the Mel-frequency bank energies of the background noise.

Considering the overall performance, λ =2, especially for C=10^7, performed the best because it kept high recognition accuracy with high SNR while improving the system performance at low SNR. The system performance with C=10^7 and λ=2 was close to the performance of the optimal system using different C for different noise levels. Thus, in practice the system can use either optimal values of C and λ for different background noise levels or just use C=10^7 and λ =2 directly without choosing C and λ to simplify the application.

4 Conclusions

An improved MFCC-based feature identification method was developed, which combines a piecewise transformation function with SS and MF. Test results using the TIMIT database show that the identification accuracy for the baseline system degrades rapidly in noisy environments and that the improved MFCC-based feature method more effectively improves system performance than the SS method. The system parameters can be optimized based on the background noise level. λ =2 gave the best overall performance as a trade-off between preserving speech information and suppressing the noise.

References

[1] Lim J S, Oppenheim A V. Enhancement and bandwidth compression of noisy speech. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing. Washington, DC: IEEE Press, 1979: 1586-1604.

[2] Ephraim Y. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Trans. on Acoustics Speech and Signal Proc.*, 1985, **33**(2): 443-445.

[3] Hermansky H, Morgan N. RASTA processing of speech. *IEEE Trans. on Speech Audio Proc.*, 1994, **2**(4): 578-589.

[4] Mammone R J, Zhang X Y, Ranmachandran R P. Robust speaker recognition: A feature-based approach. *IEEE Signal Processing Magazine,* 1996, **3**(5): 58-71.

[5] Assaleh K T, Mammone R J. New LP-derived features for speaker identification. *IEEE Trans. on Speech and Audio Proc.*, 1994, **2**(4): 630-637.

[6] Gales M J F, Young S J. Robust continuous speech recognition using parallel model combination. *IEEE Trans. on Speech and Audio Proc.*, 1996, **4**(5): 352-359.

[7] Young S. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Proc. Magazine*, 1996, **13**(5): 45-57.