Comparison of Different Implementations of MFCC

ZHENG Fang (郑 方), ZHANG Guoliang (张国亮) and SONG Zhanjiang (宋战江)

Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology, Tsinghua University, Beijing 100084, P.R. China

E-mail: fzheng@sp.cs.tsinghua.edu.cn

Received October 15, 1999; revised February 23, 2001.

Abstract The performance of the Mel-Frequency Cepstrum Coefficients (MFCC) may be affected by (1) the number of filters, (2) the shape of filters, (3) the way in which filters are spaced, and (4) the way in which the power spectrum is warped. In this paper, several comparison experiments are done to find a best implementation. The traditional MFCC calculation excludes the 0th coefficient for the reason that it is regarded as somewhat unreliable. According to the analysis and experiments, the authors find that it can be regarded as the generalized frequency band energy (FBE) and is hence useful, which results in the FBE-MFCC. The authors also propose a better analysis, namely the auto-regressive analysis, on the frame energy, which outperforms its 1st and/or 2nd order differential derivatives. Experiments with the "863" Speech Database show that, compared with the traditional MFCC with its corresponding auto-regressive analysis coefficients, the FBE-MFCC and the frame energy with their corresponding auto-regressive analysis coefficients form the best combination, reducing the Chinese syllable error rate (CSER) by about 10%, while the FBE-MFCC with the corresponding auto-regressive analysis coefficients reduces CSER by 2.5%. Comparison experiments are also done with a quite casual Chinese speech database, named Chinese Annotated Spontaneous Speech (CASS) corpus. The FBE-MFCC can reduce the error rate by about 2.9% on an average.

Keywords MFCC, frequency band energy, auto-regressive analysis, generalized initial/final

1 Introduction

The extraction and selection of the best parametric representation of acoustic signals are important tasks in the design of any speech recognition system. It significantly affects the recognition performance. A compact representation would be provided by a set of mel-frequency cepstrum coefficients (MFCC), which are the results of a cosine transform of the real logarithm of the short-term energy spectrum expressed on a mel-frequency scale[1]. The MFCCs are proved more efficient[2]. The calculation of the MFCC includes the following steps.

(1) The discrete Fourier transform (DFT) turns the windowed speech segment into the frequency domain and the short-term power spectrum $P(f)$ is obtained.

(2) The spectrum $P(f)$ is warped along its frequency axis f (in hertz) into the mel-frequency axis as $P(M)$, where M is the mel-frequency, using Eq.(1)[3,4]. This is to approximately reflect the human's ear perception.

$$M(f) = 2595\log_{10}(1 + f/700) \tag{1}$$

(3) The resulted warped power spectrum is then convolved with the triangular band-pass filter $P(M)$ into $\theta(M)$. The convolution with the relatively broad critical-band masking curves $\psi(M)$ significantly reduces the spectral resolution of $\theta(M)$ in comparison with the original $P(f)$, which allows for the down sampling of $\theta(M)$. The discrete convolution of $\psi(M)$ with $\theta(M)$ yields samples of the critical-band power spectrum as $\theta(M_k)$ $(k = 1, \dots, K)$ in (2), where Ω_k's are linearly spaced in the mel-scale. Then K outputs $X(k) = \ln(\theta(M_k))$ $(k = 1, \dots, K)$ are obtained. The K filters in the implementation of discrete convolution are simulated as shown in Fig.1(a). In the implementation,

$\theta(M_k)$ is the average rather than the sum.

$$\theta(M_k) = \sum_{M} P(M - M_k)\psi(M), \quad k = 1, \ldots, K \tag{2}$$

(4) The MFCC is computed using (3) and often $D \ll K$ because of the compression ability of MFCC.

$$\mathrm{MFCC}(d) = \sum_{k=1}^{K} X_k \cos\left[d(k - 0.5)\frac{\pi}{K}\right], \quad d = 1, \ldots, D \tag{3}$$

Accordingly, more detailed and deeper research could be done to tune the implementation of the MFCC.

On the other hand, in many ASR systems, the 0th coefficient of the MFCC cepstrum is ignored due to its unreliability[3]. As a matter of fact, the 0th coefficient can be regarded as a collection of average energies of all frequency bands in the signal that is being analyzed. We will prove by experiments in this paper that this analysis is reasonable.

The energy information is another very important feature in ASR. Basically, the commonly used energy-related features include the frame energy and the first order and/or second order time derivatives. Many experiments have shown that the system performance can be improved when the energy information is added as another model feature in addition to the cepstrum[5]. Our previous experiments on the performance of the combination of the cepstrum and its time derivatives show that the auto-regressive analysis[6] outperforms the 1st/2nd order differential analysis[7]. This suggests the use of the auto-regressive analysis on the energy. In this paper, we will come to the conclusion that the auto-regressive analysis of the energy is better than the first/second order differential analysis.

In this paper, several experiments are designed and done step by step to compare the effects of several different implementations and those of the ways how the energy information is integrated.

2 Experiment Condition

The standard Mandarin database used here is the ‘863’ Database. Digitized speech at 16kHz is pre-emphasized using a simple first-order digital filter $H(z) = 1 - 0.95z^{-1}$, and then blocked into frames of 32ms (512 sampling points) in length spaced every 16ms (256 sampling points). The D-order (where $D = 16$) cepstral analysis is performed to every Hamming-windowed frames and the auto-regression analysis (ARA) is performed to every 5 adjacent frames[6]. The cepstral coefficients and their auto-regression coefficients form the basic features for the automatic speech recognition systems. It is divided into training and testing parts. The training set covers 180,063 Chinese syllable samples of 30 men’s utterances while the testing set covers 70,462 Chinese syllable samples of 8 men’s utterances.

A kind of Segmental Probability Model has been proposed based on the desertion of the HMM probability transition matrix named mixed Gaussian continuous probability model (MGCPM) in our previous paper[8].

In this experiment, the 6-state 8-mixture based MGCPMs are adopted to model the 397 toneless Chinese syllables as the speech recognition units (SRUs).

3 Step-by-Step Experiments

In this section, we will give the designs and the results of the step-by-step experiments on Mel-frequency cepstrum analysis. To be brief, we define $F^{\&aRa}$ as Feature F and its auto-regressive analysis coefficients, and D^1F/D^2F as its 1st/2nd order time differential derivative. We denote the traditional MFCC defined in Section 3 by MFCC0.

Our previous comparison of the combination of the MFCC and the derivatives shows that MFCC plus its auto-regressive coefficients performs better than MFCC plus its 1st or 2nd order differential MFCC. Therefore, the MFCC0$^{\&aRa}$ is adopted as the baseline in this paper.

3.1 Comparisons of MFCC Implementation

According to the MFCC calculation, the performance of MFCC may be affected by: (1) the number of the filters, (2) the shape of the filters, (3) the way in which the filters are spaced, overlapped or not, or (4) the way in which the power spectrum is warped. In order to find which factors are more important, we design several comparison experiments.

Table 1 gives the results of different numbers of filters. The recognizer reaches the maximal performance at the filter number $K = 35$. Too few or too many filters do not result in better accuracy. In this case, each filter covers about 158 Mels. Hereafter, if not specifically stated, the number of filters is chosen to be $K = 35$.

Table 1. Different Numbers of Overlapped Triangular Filters

No. of filters (MFCC0$^{\&aRa}$)	Top 1	Top 5	Top 10
25	67.39	91.56	95.57
30	67.73	91.72	95.66
35	**68.01**	**91.79**	**95.77**
40	67.84	91.92	95.82
45	67.86	91.81	95.74

Traditionally, the filters are triangular. As a matter of fact, rectangular filters can be alternatives. And in PLP analysis[9], Hermansky adopts a particular shape of the critical-band curve given by (4), as illustrated in Fig.1(c).

$$\Psi(B) = \begin{cases} 0, & \text{for } B < -1.3 \\ 10^{2.5(B+0.5)}, & \text{for } -1.3 \le B \le -0.5 \\ 1, & \text{for } -0.5 < B < +0.5 \\ 10^{-1.0(B-0.5)}, & \text{for } +0.5 \le B \le +2.5 \\ 0, & \text{for } +2.5 < B \end{cases} \tag{4}$$

where B is the warped frequency in Bark. This piece-wise shape for the simulated critical-band-masking curve is an approximation to the asymmetric masking curve of Schroeder[4]. It is a rather crude approximation of what is known about the shape of auditory filters. It exploits Zwicker's[10] proposal that the shape of auditory filters is approximately constant on the Bark scale. The filter skirts are truncated at −40dB. From this point forward in this paper, this curve is referred to as the *Schroeder* curve.

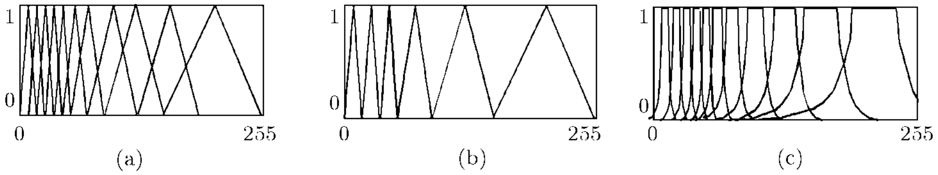

Fig.1. The band-pass filters used in MFCC calculation (horizontal axis: DFT frequency sampling point).

This experiment compares the effects of the above 3 different shapes of the critical-band filters, triangular, rectangular and *Schroeder* curve. The results are given in Table 2. We do not see too much difference.

Table 2. Different Filter Shapes and Frequency Warping

Features $(^{\&aRa})$		Top 1	Top 5	Top 10
Warping	*Filter Shape*			
`MEL`	`XTRI`	`68.01`	`91.79`	`95.77`
MEL	TRI	66.35	91.21	95.42
MEL	**XRECT**	**68.38**	**92.14**	**95.91**
MEL	RECT	66.36	91.18	95.37
BARK	XTRI	67.61	91.56	95.57
BARK	TRI	66.99	91.38	95.43
BARK	XRECT	67.59	91.53	95.57
BARK	RECT	67.00	91.35	95.53
BARK	SCHROEDER	67.25	91.53	95.51

Note: In this table, XTRI stands for crossed/overlapped triangular filters while TRI for non-overlapped, and XRECT for overlapped rectangular filters while RECT for non-overlapped.

In the traditional MFCC calculation, the mel-scale is used to warp the power spectrum, as approximately described in Eq.(1). In the PLP technique, the spectrum $P(f)$ is warped along its frequency axis f into the Bark frequency B according to Eq.(5)[9,11]. This particular Bark-hertz transformation is due to [4]. This gives us an alternative for the shape of critical-band filters, resulting in the Bark-frequency Cepstral Coefficient (BFCC). From the comparison results listed in Table 2, MFCC is better than BFCC.

$$B(f) = 6 \ln \left\{ \frac{f}{600} + \sqrt{\left(\frac{f}{600}\right)^2 + 1} \right\} \tag{5}$$

In our experiments, filters can be either overlapped (see Fig.1(a) as an example) or side by side (see Fig.1(b) as an example), except for the *Schroeder* filters that are always overlapped due to their design purpose. Each of them has the equal width in the warped frequency axis. In the overlapped scheme, any two adjacent filters will overlap half the width with each other. The experimental results are also listed in Table 2.

The experimental results given in this section lead to the conclusion that the differences between these scales (Bark or Mel), filter shapes (triangular, rectangular or *Schroeder*) are not very significant. But whether the filters are overlapped or not makes a big difference. Overlapped filters always achieve higher hit rate. To clarify, we refer to the traditional mel-frequency cepstrum with 35 overlapped triangular filters as MFCC0 as stated in Section 4.

3.2 Integrating Energy Information

Researches have proved that the energy information, as well as the differential derivatives, is useful. In this section, we compare two different kinds of energy information, the frame energy (FE) and the frequency band energy (FBE). Because the log energy is better than the energy itself[5], we give mainly the results for the log energy related experiments.

Given a frame of speech $s(n)$, $1 \le n \le N$, the frame energy (FE) can be defined and calculated as

$$\mathrm{FE} = \sum_{n=1}^{N} |s(n)|, \quad \text{or} \quad \mathrm{FE} = \sqrt{\sum_{n=1}^{N} s^2(n)} \tag{6}$$

In the traditional MFCC calculation using Eq.(3), the first dimension is eliminated[3]. Taking $d = 0$, we have

$$\mathrm{MFCC}(0) = \sum_{k=1}^{K} X_k = \ln \prod_{k=1}^{K} \theta(M) \triangleq 2 \ln \prod_{k=1}^{K} E_k^{(g)}, \tag{7}$$

where $E_k^{(g)} = \sqrt{\theta(M_k)}$, and $\theta(M_k)$ is the output of the k-th filter. If the critical-band filter has the rectangular shape, $\theta(M_k)$ is the average power energy in the k-th frequency band. So for any kind of critical band filter, $E_k^{(g)}$ can be regarded as the generalized frequency band energy (FBE). Compared with the frame energy, FBE contains more information. It contains energy information of several different sub-bands of the whole frequency band. Based on the analysis, we have reasons to think that FBE should be included.

Because the logarithm has the compression function, MFCC(0) is more sensitive to $E_k^{(g)}$ in low-valued region and less sensitive in high-valued region than the original product of energies. This is similar to ear's hearing characteristics. Based on this analysis, we change (3) into (3′).

$$\mathrm{MFCC}(d) = \sum_{k=1}^{K} X_k \cos\left[d(k-0.5)\frac{\pi}{K}\right], \quad d = 0, \ldots, D \tag{3'}$$

The resulted MFCC calculated by (3′) is referred to as the FBE-MFCC, denoted by MFCC1 hereafter. According to the calculation of the DFT of the given speech frame $s(n)$, $1 \le n \le N$, we have

$$\mathrm{DFT}(0) = \sum_{n=1}^{N} s(n) \tag{8}$$

which is the direct current component of the signal. The experiment on the DC component is just for the purpose of comparison with the energy.

The experiment is designed to compare and find out which kind of information is better: (1) FE — the frame energy, (2) LnFE — the logarithm of frame energy, (3) LnDFT0 — the logarithm of DFT(0), or (4) FBE — the frequency band energy. The experimental results are listed in Table 3.

Table 3. The Energy/DC Component Information

Features ($^{\&aRa}$)	Top 1	Top 5	Top 10
`MFCC0`	`68.01`	`91.79`	`95.77`
MFCC0 + LnDFT0	68.14	92.07	95.88
MFCC0 + FE	69.52	92.50	96.11
MFCC0 + LnFE	70.46	92.99	96.35
MFCC0 + FBE (i.e., MFCC1)	**70.51**	**92.96**	**96.39**

From Table 3, the integration with any of the four items is better than the original MFCC (i.e., MFCC0), but the FBE is the most useful one. The reason why the FBE is better is that FBE includes energy information of several frequency sub-bands while (log) frame energy or log DFT(0) includes only part of them.

What should be mentioned is that in Table 3 all features are used together with their auto-regressive analysis coefficients. There is no exception to the frame energy. In the previous research, the most frequently used frame energy information includes the frame energy itself, the logarithm of frame energy, and/or the 1st/2nd differential (log) frame energy.

Our comparison of the combinations of the MFCC and the time derivatives shows that MFCC plus its auto-regressive coefficients performs better than MFCC plus its 1st or 2nd order differential MFCC (refer to Section 4). This suggests the definition of the auto-regressive frame energy as

$$\mathrm{ARE}(t) = G \cdot \sum_{n=-n_0}^{n=n_0} nE(t), \tag{9}$$

where G is a gain constant that is the same as that in the auto-regressive analysis for MFCC. The result listed in Table 4 is the best evidence for the use of the auto-regressive analysis of the frame energy.

Table 4. Effects of the Differential Derivatives of Frame Energy

Features (besides MFCC0$^{\&aRa}$)	Top 1	Top 5	Top 10
LnFE + D^1LnFE	68.87	92.35	96.00
LnFE + D^2LnFE	69.23	92.53	96.05
LnFE + D^1LnFE + D^2LnFE	69.43	92.55	96.12
LnFE$^{\&aRa}$	70.46	92.99	96.35

3.3 Combining Frame Energy and FBE-MFCC

Now that the FBE-MFCC is the best among all kinds of combinations of the traditional MFCC (MFCC0) and the frame energy as well as the frame energy's derivatives, there arises a question: what about combining the frame energy and its derivatives into the FBE-MFCC? The results are given in Table 5. The conclusion is as expected. The log frame energy and its ARA coefficients are the best to be integrated into the FBE-MFCC and the corresponding ARA coefficients. As a supplement of Table 2 where the XRECT is better than the XTRI, an additional comparison is done between the two kinds of filter shapes, and the results are given in Table 6. We can see again that the differences between the filter shapes are not so significant.

Table 5. Combining the Frame Energy Information into FBE-MFCC (i.e., MFCC1)

Features (besides MFCC1$^{\&aRa}$)	Top 1	Top 5	Top 10
None	70.51	92.96	96.39
LnFE + D^1LnFE	70.41	92.65	96.28
LnFE + D^2LnFE	70.79	92.84	96.29
LnFE + D^1LnFE + D^2LnFE	70.97	92.85	96.40
LnDFT0$^{\&aRa}$	68.92	92.06	95.94
FE$^{\&aRa}$	70.27	92.68	96.28
LnFE$^{\&aRa}$	**71.19**	**92.98**	**96.41**

Table 6. Additional Comparison on the Filter Shape

Features ($^{\&aRa}$)	Top 1	Top 5	Top 10
MFCC1 + LnFE [35 XTRI]	**71.19**	**92.98**	**96.41**
MFCC1 + LnFE [35 XRECT]	71.17	92.99	96.47

3.4 A Uniform Calculation of the FBE-MFCC

According to the above discussion, the proposed FBE-MFCC calculation has a uniform programming form. Firstly, the FBE item is inserted by modifying the index range of the DCT outputs. Secondly, the calculation of the frame energy is performed during the windowing and pre-emphasizing of the speech segment (i.e., frame). All these coefficients form the set of FBE-MFCC coefficients. Because the calculation of the frame energy is integrated into the calculation of the FBE-MFCC, the frame energy seems to be hidden in the application of the programming.

The auto-regressive analysis can then be performed on the FBE-MFCC.

4 Application to CASS Corpus

Experiments are also done on a spontaneous Chinese speech database, named Chinese Annotated Spontaneous Speech (CASS) corpus. It is created to collect samples of most phonetic variations in Mandarin spontaneous speech due to pronunciation effects, including allophonic changes, phoneme reduction, phoneme deletion and insertion, as well as duration changes[15]. The CASS corpus was transcribed into a five-level annotation.

- *Character Level.* Canonical sentences (known as word/character sequences) are transcribed.
- *Toned Pinyin* (*or Syllable*) *Level.* The canonical toned pinyin transcription is generated.
- *Initial/Final Level.* This semi-syllable level's transcription only includes the time boundaries for each (observed) surface form initial/final.
- *SAMPA-C Level.* This level contains the observed pronunciation in SAMPA-C[12,13], a labeling set of machine-readable IPA symbols adapted for Chinese from the Speech Assessment Methods Phonetic Alphabet (SAMPA).
- *Miscellaneous Level.* Several labels related to spontaneous phenomenon are used to independently annotate the spoken discourse phenomena, including modal/exclamation, noise, silence, murmur/unclear, lengthening, breathing, disfluency, coughing, laughing, lip smack, crying, non-Chinese, and uncertain segments.

4.1 Generalized Initials/Finals (GIFs)

In spontaneous speech, there are two kinds of differences between the canonical initials/finals (IFs) and their surface forms if the deletion and insertion are not considered. One is the sound change from one IF to a SAMPA-C sequence close to its canonical IF, such as nasalization, centralization, voiceless, voiced, rounding, syllabic, pharyngealization, and aspiration. We refer to the surface form of an IF as its *generalized IF* (GIF). Obviously, the IFs are special GIFs. The other is the phone change directly from one IF to another quite different IF or GIF, for example, initial '*zh*' may be changed into '*z*' or voiced '*z*'.

To model the sound variability when the semi-syllable level units are SRUs, the first thing to do is to choose and define the GIF set. The canonical IF set consists of 21 initials and 38 finals, totally 59 IFs. By searching in the CASS corpus, we initially obtain a GIF set of over 140 possible SAMPA-C sequences (pronunciations) of IFs. Two examples are given in Table 7. However, some of them occur for only a couple of times which can be regarded as least frequently observed sound variability forms, therefore they are merged into the most similar canonical IFs. Finally we have 86 GIFs.

Table 7. Examples for IFs and Their Possible Pronunciations in SAMPA-C Format

IF		**Comments**
Pinyin	**SAMPA-C**	
z	/ts/	Canonical
z	/ts_v/	Voiced
z	/ts`/	Changed to '*zh*'
z	/ts`_v/	Changed to voiced '*zh*'
e	/7/	Canonical
e	/7`/	Retroflexed, or changed to '*er*'
e	/@/	Changed to /@/ (a GIF)

These well-chosen GIFs are taken as SRUs. In order to better model the spontaneous speech, additional garbage models are also built for breathing, coughing, crying, disfluency, laughing, lengthening, modal, murmur, non-Chinese, smacking, noise, and silence.

4.2 Experiment on CASS

The CASS corpus is divided into two parts. The first part is the training set with about 3.0 hours' spontaneous speech data and the second is the testing set with about 15 minutes' spontaneous speech data. The HTK is used for both the training and testing. A 3-state 16-gaussian HMM is used to model each GIF. The feature extraction frame size is 32ms with 16ms overlap between any two adjacent frames.

Experimental results in Table 8 include (1) GIF-X: IF comparison without the syllable lexicon constraint; (2) GIF-S: IF comparison with the syllable lexicon constraint; and (3) SYL-S: syllable comparison with the syllable lexicon constraint. The listed figures are error rate reduction based on the percent correct $\%Cor = \%Hit = Hit/Num * 100\% = (Num - Del - Sub)/Num * 100\%$ and the percent accuracy $\%Acc = (Hit - Ins)/Num * 100\% = (Num - Del - Sub - Ins)/Num * 100\%$[14] with MFCC0$^{(aRa)}$ as the baseline, where *Num* is the total number of GIFs in reference transcriptions, and *Hit*, *Del*, *Sub* and *Ins* indicate numbers of hits, deletion errors, substitution errors and insertion errors respectively.

Table 8. Comparison on Features

Features ($^{\&aRa}$)	**GIF-X**		**GIF-S**		**SYL-S**	
	EC↓	EA↓	EC↓	EA↓	EC↓	EA↓
MFCC0+FEB=MFCC1	3.1	2.4	4.0	2.8	2.7	2.3
MFCC0+LnFE	3.7	3.0	4.9	3.0	3.1	2.3
MFCC0+FBE+LnFE=MFCC1+LnFE	2.3	1.9	2.9	1.6	−0.8	−1.1

Note: EC↓ means the error rate reduction based on %*Cor*, while EA↓ based on %*Acc*.

From this table, we can see that for a casual speech database with quite different channels, the FBE-MFCC outperforms the traditional MFCC; it can reduce the error rate by about 2.9% on an average. But for a situation under the adverse environment, the performance of the method integrating the FBE and the frame energy is different from that for the 863 Database. For 863 Database, MFCC0+FBE+LnFE is better than both MFCC0+FBE and MFCC0+LnFE, and any of the above combinations is better than MFCC0. On the contrary, the integration of frame energy information may reduce the performance. Nevertheless, the fact remains that FBE-MFCC is always better than traditional MFCC.

5 Summary

From the step-by-step design and implementation of the experiments, we come to the following conclusions:

(1) The MFCC(0), i.e., the frequency band energy (FBE) information, is useful to be included in the MFCC, referred to as FBE-MFCC in this paper to be distinguished from the traditional MFCC, no matter the speech environment is of high-quality or is adverse.

(2) For high-quality environments (for example similar to that for the 863 Database), the combination of the FBE-MFCC and the frame energy (FE) with their auto-regressive analysis coefficients is the best one; while for adverse environments (for instance similar to that for CASS corpus), the combination of the MFCC and the frame energy with their auto-regressive analysis coefficients is the best one.

(3) The integration of both the frame energy and the FBE into the MFCC should be treated differently for different applications. In some cases, the integration of the FE information and the FBE information achieves the best performance, while not in other cases.

(4) The uniform calculation makes the programming and application of the feature extraction simpler and more straightforward and it can provide an option for different applications.

References

[1] Pols L C W. Spectral analysis and identification of Dutch vowels in monosyllabic words [dissertation]. Free University, Amsterdam, The Netherlands, 1966.

[2] Davis S B, Mermelstein P. Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. ASSP*, Aug., 1980.

[3] Picone J W. Signal modeling techniques in speech recognition. In *Proceedings of the IEEE*, 1993, 81(9): 1215–1247.

[4] Schroeder M R. Recognition of complex acoustic signals. *Life Science Research Report,* Bullock T H (ed.), Abakon Verlag, Berlin, 1997, 55: 323–328.

[5] Huang X D, Acero A, Alleva F *et al.* From SPHINX-II to WHISPER — Making Speech Recognition Usable. *Automatic Speech and Speaker Recognition: Advanced Topics.* Lee C H, Soong F K, Paliwal K K (eds.), USA: Kluwer Academic Publishers, 1996, pp.481–508.

[6] Furui S. Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoust., Speech, and Signal Processing*, Feb., 1986, 34(1): 52–59.

[7] Zheng F. Studies on speaker-independent continuous digit recognition methods and Chinese speech corpus [thesis]. Department of Computer Science & Technology, Tsinghua University, June, 1992.

[8] Zheng F, Mou X-L, Wu W-H *et al.* On the embedded multiple-model scoring scheme for speech recognition. *International Symposium on Chinese Spoken Language Processing* (*ISCSLP'98*), Singapore, Dec. 7–9, 1998, ASR-A3: 49–53.

[9] Hermansky Hynek. Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.,* April, 1990, 87(4): 1738–1752.

[10] Zwicker E. Masking and psychological excitation as consequences of ear's frequency analysis. In *Frequency Analysis and Periodicity Detection in Hearing*, Plomp R, Smoorenburg G F (eds.), Sijthoff Leyden, The Netherlands, 1970.

[11] Zwicker E. Subdivision of the audible frequency range into critical bands. *J. Acoust. Soc. Am.*, Feb., 1961, 33.

[12] Chen X-X, Li A-J *et al.* An application of SAMPA-C for standard Chinese. In *International Conference on Spoken Language Processing* (*ICSLP*'2000), Oct. 16–20, 2000, Beijing.

[13] Li A-J, Chen X-X, Sun G *et al.* The phonetic labeling on read and spontaneous discourse corpora. In *International Conference on Spoken Language Processing* (*ICSLP*'2000), Oct. 16–20, 2000, Beijing.

[14] Young S, Kershaw D, Odell J *et al.* The HTK Book, Version 2.2, Entropic Ltd., 1999.

[15] Li A-J, Zheng F, Byrne W, Fung P *et al.* CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In *International Conference on Spoken Language Processing* (*ICSLP*'2000), Oct. 16–20, 2000, Beijing.