EURASIP Journal on Applied Signal Processing 2004:3, 340–346

Using Mel-Frequency Cepstral Coefficients in Missing Data Technique

Zhang Jun

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China

School of Electronic and Communication Engineering, South China University of Technology, Guangzhou 510640, China
Email: zhj_angun@sina.com.cn

Sam Kwong

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China
Email: cssamk@cityu.edu.hk

Wei Gang

School of Electronic and Communication Engineering, South China University of Technology, Guangzhou 510640, China
Email: ecgwei@scut.edu.cn

Qingyang Hong

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China
Email: qyhong@cs.cityu.edu.hk

Received 19 February 2003; Revised 16 June 2003; Recommended for Publication by Mukund Padmanabhan

Filter bank is the most common feature being employed in the research of the marginalisation approaches for robust speech recognition due to its simplicity in detecting the unreliable data in the frequency domain. In this paper, we propose a hybrid approach based on the marginalisation and the soft decision techniques that make use of the Mel-frequency cepstral coefficients (MFCCs) instead of filter bank coefficients. A new technique for estimating the reliability of each cepstral component is also presented. Experimental results show the effectiveness of the proposed approaches.

Keywords and phrases: MFCC, missing data techniques, robust speech recognition.

1. INTRODUCTION

In spite of many years of efforts, the robustness of speech recognition in the noisy environment is still a fundamental unsolved issue in today's automatic speech recognition (ASR) systems. Recently, missing data theory [1, 2, 3, 4] is proposed as an operationalization to improve the robustness of the ASR decoding process. Experimental results show that it can significantly restore the ASR performance with little prior assumptions made about the characteristics of the environment noises. However, most of the previous marginalisation approaches are only derived and tested for the filter bank features due to the convenience of detecting the unreliable data in the frequency domain. Most often, cepstral features are the parameterisation of choice for many speech recognition applications. For example, the Mel-frequency cepstral coefficient (MFCC) [5] representation of speech is probably the most commonly used representation in speech recognition and recently being standardized for the distributed speech recognition (DSR) [6]. Generally, cepstral features are more compactible, discriminable, and most importantly, nearly decorrelated such that they allow the diagonal covariance to be used by the hidden Markov models (HMMs) effectively. Therefore, they can usually provide higher baseline performance over filter bank features. Applying missing data techniques to cepstral features is obviously attractive and natural.

Unfortunately, while decorrelating, the cepstral transform also smears localized spectral uncertainty over global cepstral uncertainty. This defect dose not only bring the difficulty to the detection of the unreliable cepstral components but also seems to contradict the basic assumption of missing data theory that some part of the feature vector should be untainted by the noise [4]. However, when the distortions are not too severe, there will be some cepstral components that are less affected and can provide correct discrimination

information while using the clean speech models. If we regard these components as reliable data, then the marginalisation approach should also be applied to the cepstral features. Its performance will depend on how severely the noise distorts the cepstral feature. Fortunately, it can be seen that even the full band features that smear distortions over the entire vector are much more affected by band-limited noises than those features that localize the spectral distortions, they do perform well in many full band noises. This phenomenon is also reported in [7, 8, 9]. It means that in many cases, the full band features are not more affected by the noise than the subband ones. Therefore, it can be expected that the cepstral marginalisation will also perform well under such situations.

To implement the cepstral marginalisation approach, we propose a new technique to evaluate the reliability of each feature component in the Mel-cepstrum domain. Two criteria for detecting the reliable cepstral components are presented and combined together to form a more accurate joint decision. Then the marginalisation approach is applied to the MFCCs by using this combined criterion. Based on the proposed cepstral marginalisation approach, a cepstral soft decision approach is also developed to further improve the robustness of the MFCC recognizer.

2. CEPSTRAL MARGINALISATION

2.1. Detection of the reliable cepstral features

The major difficulty of the cepstral marginalisation is how to determine the reliable/unreliable components of the speech data. In this paper, we propose two ways to estimate the influence of noises on the cepstral component. One is based on the speech enhancement and the other is based on a noise mask model. By setting the threshold, a criterion for selecting the reliable data can be obtained from each method. After that, we combine these two criteria together and propose a soft technique to determine the final reliable/unreliable decision for each cepstral component.

Assume that the noise is added in the time domain. Let $c_y(i)$ and $c_x(i)$ denote the ith MFCC components of the noisy speech and the clean speech, respectively, where $1 \le i \le I$, I is the dimension of the MFCC vector. Then $c_y(i)$ can be expressed as follows:

$$c_y(i) = c_x(i) - c_n(i), \tag{1}$$

where $c_n(i)$ can be viewed as the noise in the cepstrum domain. If $c_n(i)$ can be estimated, then the impact of the noise to the clean feature can also be determined. Let $Y(j)$, $X(j)$, and $N(j)$ denote the jth filter bank outputs of the linear power spectra of the noisy speech, clean speech, and noise, respectively. Then $c_n(i)$ can be expressed as

$$\begin{aligned} c_n(i) = c_x(i) - c_y(i) &= \sum_{j=0}^{J-1} a_{ij}\left(\log\left(X(j)\right) - \log\left(Y(j)\right)\right) \\ &= \sum_{j=0}^{J-1} a_{ij} \log \frac{X(j)}{Y(j)}, \end{aligned} \tag{2}$$

where $0 \le j \le J-1$, J is the number of filter bank channels, and a_{ij} is the DCT coefficients. Using some kinds of the enhancement techniques like the spectral subtraction, $X(j)$ can be estimated, so the estimation of $c_n(i)$ can be given by the following:

$$\hat{c}_n(i) = \sum_{j=0}^{J-1} a_{ij} \log \frac{\hat{X}(j)}{Y(j)}, \tag{3}$$

where $\hat{X}(j)$ and $\hat{c}_n(i)$ denote the estimation of $X(j)$ and $c_n(i)$. When $\hat{c}_n(i)$ is larger than a given threshold, $c_y(i)$ can be regarded as unreliable. So the first criterion for choosing a reliable component can be given by the following:

$$\left|c_y(i)\right| > \beta_1 \left|\hat{c}_n(i)\right|. \tag{4}$$

Obviously, speech enhancement algorithms cannot always give accurate estimations of the clean features, especially when the SNR is low. It can be seen that an unreliable component with a small $\hat{c}_n(i)$, which is caused by the inaccuracy of the enhancement, cannot be detected using (4). To overcome this defect, we propose to use another method to estimate the influence of the noises. For additive noises, $c_y(i)$ can be expressed as

$$c_y(i) = \sum_{j=0}^{J-1} a_{ij} \log\left(Y(j)\right) = \sum_{j=0}^{J-1} a_{ij} \log\left(X(j) + N(j)\right). \tag{5}$$

Assume that either the clean speech or the noise will dominate in each filter bank channel and the channel output can be approximated to the dominating one. For each channel, a threshold can be applied to determine which signal is dominating. Then $Y(j)$ can be expressed as

$$Y(j) \approx \begin{cases} X(j), & Y(j) > \alpha\hat{N}(j), \\ N(j), & Y(j) \le \alpha\hat{N}(j), \end{cases} \tag{6}$$

where $\hat{N}(j)$ is the estimation of the noise, α is an empirical threshold factor which can be determined in the experiment. Substituting (6) in (5), we have the following:

$$c_y(i) \approx \sum_{j,Y(j)>\alpha\hat{N}(j)} a_{ij} \log\left(X(j)\right) + \sum_{j,Y(j)\le\alpha\hat{N}(j)} a_{ij} \log\left(N(j)\right). \tag{7}$$

According to (7), another criterion for choosing the reliable components can be given by

$$\sum_{j,Y(j)>\alpha\hat{N}(j)} \left|a_{ij} \log\left(Y(j)\right)\right| > \beta_2 \left(\sum_{j,Y(j)\le\alpha\hat{N}(j)} \left| a_{ij} \log\left(Y(j)\right)\right| \right). \tag{8}$$

Combining (8) with (4), the unreliable components with a small $\hat{c}_n(i)$ can also be detected. It is more accurate to use a joint decision than an individual one. We can simply adopt

an "and" operation to achieve such a decision, that is, a component will be considered as reliable when conditions (4) and (8) are satisfied.

2.2. Detection of the reliable delta cepstral coefficients

In traditional ASRs, the time derivatives are usually added to the static parameters to enhance the recognizer performance. The marginalisation approach can also be applied to these coefficients. In the filter bank marginalisation, one solution to this problem is called the "strict mask" [10]. It treats the derivatives as missing if any of the features involved in their calculations are missing. The strict mask is sufficient for filter bank features because the reliable features tend to be clustered into time-frequency blocks. However, it may not be feasible for cepstral features since the missing mask pattern is more random. Applying the strict mask will cause the sparseness of the reliable derivatives, thus, we propose to use another way to detect the reliable derivatives. It is also based on the combination of the enhancement and noise mask methods that are described in Section 2.1.

Usually, the delta coefficients can be calculated using the following expression:

$$\Delta c(i) = \frac{\sum_{t=-T}^{T} t c(i+t)}{\sum_{t=-T}^{T} t^2}. \tag{9}$$

The noise of the delta cepstral coefficients can be expressed as

$$\begin{aligned}\Delta c_n(i) &= \Delta c_x(i) - \Delta c_y(i) \\ &= \frac{\sum_{t=-T}^{T} t\left(c_x(i+t) - c_y(i+t)\right)}{\sum_{t=-T}^{T} t^2} \\ &= \frac{\sum_{t=-T}^{T} t c_n(i+t)}{\sum_{t=-T}^{T} t^2}.\end{aligned} \tag{10}$$

When the cepstral noise $\hat{c}_n(i)$ is estimated using the enhancement, $\Delta\hat{c}_n(i)$ can be given as

$$\Delta\hat{c}_n(i) = \frac{\sum_{t=-T}^{T} t\hat{c}_n(i+t)}{\sum_{t=-T}^{T} t^2}. \tag{11}$$

So, one criterion for choosing the reliable delta cepstral components can be given by

$$\left|\Delta c_y(i)\right| > \beta_3 \left|\Delta\hat{c}_n(i)\right|. \tag{12}$$

On the other hand, with the noise mask approximation, $\Delta c_y(i)$ can be expressed as

$$\begin{aligned}\Delta c_y(i) \approx \frac{1}{\sum_{t=-T}^{T} t^2}\Bigg[&\sum_{t=-T}^{T}\sum_{j,Y(j)>\alpha\hat{N}(j)} ta_{ij}\log\left(X(j+t)\right) \\ &+ \sum_{t=-T}^{T}\sum_{j,Y(j)\le\alpha\hat{N}(j)} ta_{ij}\log\left(N(j+t)\right)\Bigg].\end{aligned} \tag{13}$$

So, another criterion for choosing the reliable delta cepstral components can be given by

$$\begin{aligned}&\sum_{t=-T}^{T}\sum_{j,Y(j)>\alpha\hat{N}(j)} \left|ta_{ij}\log\left(Y(j+t)\right)\right| \\ &\quad > \beta_4 \sum_{t=-T}^{T}\sum_{j,Y(j)\le\alpha\hat{N}(j)} \left|ta_{ij}\log\left(Y(j+t)\right)\right|.\end{aligned} \tag{14}$$

Combining these two criteria, a delta cepstral component can be decided as reliable when conditions (12) and (14) are satisfied.

2.3. Marginalisation

Using (4), (8) and (14), the reliable cepstral and delta cepstral components can be picked out from the whole feature vectors. For the continuous density HMM (CDHMM) recognition system with diagonal-only covariance, the marginalised probability of observations can be given by

$$p(\mathbf{x}|C_m) = \sum_{n=1}^{N} w_{mn} \prod_{i,\text{reliable}} \mathbf{N}(x_i, \mu_{mn}(i), \sigma_{mn}^2(i)), \tag{15}$$

where $\mathbf{x}$ is the observation vector, C_m is the mth state of the HMM model, w_{mn} is the weight factor associated with the nth Gaussian component of the state C_m, and μ_{mn} and σ_{mn}^2 are the mean and variance of the Gaussian PDF.

3. SOFT DECISION

3.1. Noisy speech model

Due to the cepstral transformation, even a little noise that exists in some frequency bands will affect all the feature components. So, in a noisy environment, each cepstral component will always have a portion of the noise in the clean speech. Obviously, it is more sensible to adjust the weight of each component according to its influence level than using a binary decision of reliable or unreliable.

Given a noisy observation, the components that are less affected by the noise will have distributions close to the clean ones while those severely affected will be more uncertain and might have much different characteristics. According to [4], the distribution of a noisy observation can be modeled as a weighed sum of a known distribution that is obtained during the training process and an unknown distribution for the uncertain data. We model the noisy speech in a similar way. While using the diagonal-only covariance, the probability of a noisy observation can be given by

$$p(\mathbf{x}|C_m) = \sum_{n=1}^{N} w_{mn} \prod_{i=1}^{I} \left(\varepsilon_i p_1(x_i|C_m, n) + (1-\varepsilon_i) p_2(x_i)\right), \tag{16}$$

where $p_1(x_i|C_m, n)$ denotes the clean distribution as

$$p_1(x_i|C_m, n) = N(x_i, \mu_{mn}(i), \sigma_{mn}^2(i)), \tag{17}$$

and where $p_2(x_i)$ denotes the distribution of uncertain data. When no prior knowledge about this distribution is available, it can be assumed that the uncertain data will have a uniform distribution in the range of values observed during training as

$$p_2(x_i) = \frac{1}{x_{i,\max} - x_{i,\min}}, \tag{18}$$

where $x_{i,\max}$ and $x_{i,\min}$ are the maximum and minimum values of the ith component observed in the training data.

In the acoustic backing-off approach, ε_i refers to the prior probability of observing a reliable datum and needs to be determined in advance. It is obviously that this assumption is not suitable for real world applications. Instead of setting up a static value in advance, we adjust ε_i according to the noise level of each cepstral component. These levels are estimated using the two methods described in Section 2.

3.2. Weights adjustment

Let ε_i and ε_i' denote the weights for the ith cepstral and delta cepstral components, respectively. Using the enhancement method, we can adjust them by

$$\begin{aligned} \varepsilon_{1i} &= \frac{|\hat{c}_x(i)|}{|\hat{c}_x(i)| + \gamma_1 |\hat{c}_n(i)|}, \\ \varepsilon'_{1i} &= \frac{|\Delta\hat{c}_x(i)|}{|\Delta\hat{c}_x(i)| + \gamma_2 |\Delta\hat{c}_n(i)|}. \end{aligned} \tag{19}$$

Using the noise mask method, the weights can be adjusted as

$$\begin{aligned} &\varepsilon_{2i} \\ &= \frac{\sum_{j,Y(j)>\alpha\hat{N}(j)} |a_{ij} \log(Y(j))|}{\sum_{j,Y(j)>\alpha\hat{N}(j)} |a_{ij} \log(Y(j))| + \gamma_3 \sum_{j,Y(j)\le\alpha\hat{N}(j)} |a_{ij} \log(Y(j))|}, \\ &\varepsilon'_{2i} \\ &= \left[\sum_{t=-T}^{T} \sum_{j,Y(j)>\alpha\hat{N}(j)} |ta_{ij} \log(Y(j+t))|\right] \\ &\quad \times \left[\sum_{t=-T}^{T} \sum_{j,Y(j)>\alpha\hat{N}(j)} |ta_{ij} \log(Y(j+t))| \right. \\ &\qquad \left. + \gamma_4 \sum_{t=-T}^{T} \sum_{j,Y(j)\le\alpha\hat{N}(j)} |ta_{ij} \log(Y(j+t))|\right]^{-1}. \end{aligned} \tag{20}$$

These weights can also be combined together to improve the performance. We calculate the combined weights by

$$\begin{aligned} \varepsilon_i &= \min(\varepsilon_{1i}, \varepsilon_{2i}), \\ \varepsilon'_i &= \min(\varepsilon'_{1i}, \varepsilon'_{2i}). \end{aligned} \tag{21}$$

4. EXPERIMENTS

Clean speech data for training and testing are taken from the TI46 speaker-dependent isolated word corpus. Digits 0–9 spoken by all male speakers are used. There are 26 utterances of each digit from each speaker: 10 of these utterances are designated as training tokens and the other 16 are designated as testing tokens. Speech data are sampled at 12500 Hz and linearly quantified with 12 bits. Four noises from the NOISEX-92 [11] database with distinct characteristics: white noise, F16 noise, pink noise, and factory noise, are artificially added to the clean speeches with different SNRs.

Each digit is modeled by an HMM which composes of five no-skip straight-through emitting states. Each state has three diagonal Gaussian mixtures. Both filter bank coefficients and MFCCs are used in the experiments. Input speeches are segmented into overlapping frames with 25 milliseconds length and 10 milliseconds shift. Twenty triangular filters are uniformly distributed on a Mel-frequency scale and their log energy outputs form the 20-dimension filter bank coefficients. Twelve MFCCs are computed using DCT transformation on these filter bank coefficients. The delta coefficients are computed and appended to the basic acoustic vectors in the front-end. We use the HTK tools 3.0 [12] for both the feature extraction and the HMM model training.

4.1. Evaluation of the proposed approaches

The performance of the proposed approaches is evaluated with the four types of noises. For the cepstral marginalisation and soft decision approaches, a simple nonadaptive linear spectral subtraction in (22) is employed as an enhancement preprocess:

$$\hat{X}(j) = \max(Y(j) - \hat{N}(j), \lambda Y(j)), \tag{22}$$

where λ is the flooring factor, which is set to 0.05 in the experiments. The first 20 frames of noisy speeches are assumed to be the noises. Their average power spectra are used to estimate $\hat{N}(j)$. We empirically set α, β_1–β_4, and γ_1–γ_4 to 1.0. The HTK recognition process is modified according to (15) and (16) to implement the marginalisation and soft decision approaches.

Table 1 shows the average recognition rates of the baseline MFCC recognizer and the proposed approaches. For comparison, the results of the spectral subtraction (SS), cepstral mean subtraction (CMS), and filter bank marginalisation with SNR criterion plus strict mask are also listed in the table. Here, "MG" refers to marginalisation and "SD" refers to soft decision.

Both the SS and CMS gain improvements over the baseline performance. It can be seen that the cepstral mean subtraction is less effective for additive noises than the spectral subtraction. This is probably because the CMS is mainly designed to cope with the stationary convolution distortions. Both the proposed approaches and the filter bank marginalisation show significant improvements over these two techniques. Comparing with the filter bank marginalisation, the cepstral marginalisation gives higher average recognition rates for the four types of noises. It is worse for the

Table 1: Average recognition rates of various techniques for the four types of noises.

	Clean	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
MFCC_D	100.00	99.59	97.64	87.91	58.61	25.78	11.50
MGCC_D+SS	100.00	99.50	98.54	92.17	75.24	49.25	22.54
MFCC_D+CMS	100.00	98.89	97.90	89.85	64.65	29.71	12.09
FBANK_D+SS+MG	99.92	99.90	99.73	98.83	88.29	64.42	35.71
MFCC_D+SS+MG	100.00	99.94	99.84	98.61	93.36	75.45	42.79
MFCC_D+SS+SD	100.00	99.90	99.79	99.53	98.22	90.02	51.23

Table 2

(a) Average recognition rates of the cepstral marginalisation approaches with different criteria for the four types of noises.

	Clean	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
SS	100.00	99.50	98.54	92.17	75.24	49.25	22.54
Criterion 1	100.00	99.82	99.71	98.28	91.99	71.06	36.40
Criterion 2	100.00	99.71	99.06	95.43	81.10	51.92	21.22
Combined criterion	100.00	99.94	99.84	98.61	93.36	75.45	42.79

(b) Average recognition rates of the cepstral SD approaches with different weights for the four types of noises.

	Clean	25 dB	20 dB	15 dB	10 dB	5 dB	0 dB
Weight 1	100.00	99.80	99.69	99.44	97.56	84.40	42.36
Weight 2	100.00	99.88	99.74	99.08	94.91	79.52	43.53
Combined weight	100.00	99.90	99.79	99.53	98.22	90.02	51.23

white noise, slightly better for the F16 noise and pink noise, and significantly better for the factory noise. The cepstral SD approach is superior to both marginalisation approaches for all types of noises. These results confirm our prediction that the cepstral marginalisation can work well for many kinds of full band noises, and also show the effectiveness of the SD approach.

4.2. Combination of the criteria and the weights

To show the effectiveness of our combined criteria for the cepstral marginalisation, Table 2a lists the average recognition rates of different criteria for the four types of noises. Here, criterion 1 refers to the criteria shown in (4) and (12), criterion 2 is from (8) and (14), and the combined criterion is from (4), (8) and (14). The results of the SS are also listed in the table.

It can be seen that the recognition rates are improved whenever the marginalisation approaches are applied with criterion 1, criterion 2, or the combined criterion. For individual criteria, criterion 1 gives better performance than criterion 2. This is probably because criterion 1 is more closely related to the enhancement preprocess. Nevertheless, the combined criterion is able to achieve the highest recognition rates. Thus, it can conclude that the joint decision is more accurate than the individual one.

The average recognition rates of the cepstral SD approach with the individual or combined weights are also listed in the Table 2b. Here, weight 1 is used from (19), weight 2 is derived from the (20), and the combined weight refers to (21). As the combined criterion does in the cepstral marginalisation, the combined weight also gives the best performance in the cepstral SD approaches.

4.3. Influence of different types of noises to the cepstral feature

One of the major factors that affect the performance of the marginalisation and SD approaches is how severely the noises distort the features. If we consider the effect of cepstral distortions to be additive, the normalized mean square error (NMSE) can be used to evaluate the distortion level of a cepstral component [13]. To show the impacts of different types of noises to the MFCCs, we compute the NMSE between the corresponding components of the clean and noisy MFCCs when the SNR is 10 dB. The results are listed in Table 3.

As can be seen, the four types of full band noises distort all the MFCC components. For the white noise and pink noise, C1 are the mostly affected. For the F16 noise, C9 and C10 are much more affected than the other components. Obviously, the additive noises in the time domain cause the signal to be distorted in the cepstrum domain. The level of distortions depends both on the level of noises and the clean speech. The results in Table 3 show the trend that the noises with flat spectra will distort the lowest cepstral component most. The noises with energies that concentrate on some frequency bands will give particular distortions to some cepstral components. Due to the nonstationary property of factory noise, it is hard to analysis its impact through the NMSE. But the result shows that C1 and the higher-order coefficients are more affected. Among the four types of noises, the NMSE of

Table 3: NMSE of the 12 MFCCs for the four types of noises when the SNR is 10 dB.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12
White	2.53	0.79	0.96	0.52	0.94	0.47	0.96	0.72	0.97	0.89	0.80	1.07
F16	1.02	0.47	0.60	0.41	0.65	0.44	0.80	0.67	2.09	2.15	0.73	0.85
Pink	1.24	0.53	0.70	0.41	0.75	0.43	0.77	0.67	0.88	0.80	0.76	0.88
Factory	0.85	0.52	0.59	0.38	0.72	0.42	0.75	0.66	0.88	0.79	0.69	0.97

white noise is the largest. This phenomenon explains why the cepstral marginalisation approach performs worse under the white noise condition.

5. CONCLUSION

In this paper, we propose the new cepstral marginalisation and cepstral soft decision approaches for the MFCCs. In the experiments on the TI46 speaker-dependent isolated word corpus and four types of noises from the NOISEX-92 database, it shows that the proposed approach can efficiently improve the performance of the MFCC recognizer and give higher average recognition rates than the filter bank marginalisation. It shows that the marginalisation approach that is applied to the features rather than filter bank representations can also perform well when these features are not too severely affected by the environment noises. The cepstral soft decision approach gives the best performance in the experiments. It is believed that further improvement can be gained when the weights are determined in a more precise manner.

ACKNOWLEDGMENT

This work was supported by the City University Strategic Grant 7001416 and 7001488.

REFERENCES

[1] M. Cooke, A. Morris, and P. Green, "Missing data techniques for robust speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '97)*, vol. 2, pp. 863–866, Munich, Germany, April 1997.

[2] A. Morris, M. Cooke, and P. Green, "Some solution to the missing feature problem in data classification, with application to noise robust ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, vol. 2, pp. 737–740, Seattle, Wash, USA, May 1998.

[3] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2001.

[4] J. de Veth, B. Cranen, and L. Boves, "Acoustic backing-off as an implementation of missing feature theory," *Speech Communication*, vol. 34, no. 3, pp. 247–265, 2001.

[5] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, 1980.

[6] ETSI ES 201 108 V 0.08, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-End Feature Extraction Algorithm; Compression Algorithm," 1999.

[7] S. Okawa, E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing (ICASSP '98)*, pp. 641–644, Seattle, Wash, USA, May 1998.

[8] A. Morris, A. Hagen, H. Glotin, and H. Bourlard, "Multi-stream adaptive evidence combination for noise robust ASR," *Speech Communication*, vol. 34, no. 1-2, pp. 25–40, 2001.

[9] R. Hariharan, I. Kiss, and O. Viikki, "Noise robust speech parameterization using multiresolution feature extraction," *IEEE Trans. Speech, and Audio Processing*, vol. 9, no. 8, pp. 856–865, 2001.

[10] J. P. Barker, L. Josifovski, M. P. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP '00)*, vol. 1, pp. 373–376, Beijing, China, October 2000.

[11] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.0)*, Cambridge University Technical Services, Cambridge, UK, 2000.

[13] J. Huerta and R. Stern, "Speech recognition from GSM codec parameters," in *Proc. International Conference on Spoken Language Processing (ICSLP '98)*, vol. 4, pp. 1463–1466, Sydney, Australia, November 1998.

Zhang Jun was born in Guangdong province, China, in 1975. He received his B.S and M.S. degrees from Zhong Shan University, China, in 1997 and 2000, respectively, and his Ph.D. degree from the South China University of Technology, China, in 2003, all in electronic and communication engineering. He worked as a Research Assistant in the City University of Hong Kong from August 2002 to May 2003. He is currently in the School of Electronic and Communication Engineering, South China University of Technology. His research interests include robust speech recognition and low bit rate speech coding.

Sam Kwong received his B.S. and M.S. degrees in electrical engineering from The State University of New York at Buffalo, USA and University of Waterloo, Canada, in 1983 and 1985, respectively. In 1996, he obtained his Ph.D. degree from the University of Hagen, Germany. From 1985 to 1987, he was a Diagnostic Engineer with the Control Data Canada where he designed the diagnostic software to detect the manufacture faults of the VLSI chips in the Cyber 430 machine. He later joined the Bell Northern Research Canada as a member of the scientific staff. In 1990, he joined the City University of Hong Kong as a Lecturer in the Department of Electronic Engineering. He is currently an Associate Professor in the Department of Computer Science.

Wei Gang was born in January 1963. He received the B.S., M.S., and Ph.D. degrees in 1984, 1987, and 1990, respectively, from Tsinghua University and South China University of Technology. He was a visiting scholar to the University of Southern California from June 1997 to June 1998. He is currently a Professor at the School of Electronic and Communication Engineering, South China University of Technology. He is a committee member of the National Natural Science Foundation of China. His research interests are signal processing and personal communications.

Qingyang Hong received his M.S. degree of Computer Science from Xiamen University in 2001. Currently, he is a Ph.D. student in the Department of Computer Science at City University of Hong Kong. His research direction is statistical speech and speaker recognition.