EURASIP Journal on Applied Signal Processing 2001:2, 82–88

The Use of Adaptive Frame for Speech Recognition

Sam Kwong

Department of Computer Science, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong
Email: cssamk@cityu.edu.hk

Qianhua He

Department of Electronic Engineering, South China University of Technology, China
Email: eeqhhe@scut.edu.cn

Received 19 January 2001 and in revised form 18 May 2001

We propose an adaptive frame speech analysis scheme through dividing speech signal into stationary and dynamic region. Long frame analysis is used for stationary speech, and short frame analysis for dynamic speech. For computation convenience, the feature vector of short frame is designed to be identical to that of long frame. Two expressions are derived to represent the feature vector of short frames. Word recognition experiments on the TIMIT and NON-TIMIT with discrete Hidden Markov Model (HMM) and continuous density HMM showed that steady performance improvement could be achieved for open set testing. On the TIMIT database, adaptive frame length approach (AFL) reduces the error reduction rates from 4.47% to 11.21% and 4.54% to 9.58% for DHMM and CHMM, respectively. In the NON-TIMIT database, AFL also can reduce the error reduction rates from 1.91% to 11.55% and 2.63% to 9.5% for discrete hidden Markov model (DHMM) and continuous HMM (CHMM), respectively. These results proved the effectiveness of our proposed adaptive frame length feature extraction scheme especially for the open testing. In fact, this is a practical measurement for evaluating the performance of a speech recognition system.

Keywords and phrases: speech recognition, speech coding, adaptive frame, signal analysis.

1. INTRODUCTION

To date, the most successful speech recognition systems mainly use Hidden Markov Model (HMM) for acoustic modeling. HMM in fact dominates the continuous speech recognition field [1]. In order to improve the performance of speech recognition, a great deal of efforts had been made to study the training approaches for HMMs [2, 3, 4], or variations of the conventional HMM, such as the segment HMM [1], and the HMMs with state-conditioned second-order nonstationary [5]. In general, frame-based feature analysis for speech signals has been accepted as a very successful technique. In this method, time speech samples are blocked into frames of N samples, with adjacent frames separated by M samples. Then the spectral characteristic coefficients are calculated for each frame via some speech analysis methods (coding procedure), such as LPC, FFT analysis, Gabor expansion [6], or wavelets [7]. N is usually set to be the number of samples of 30–45 ms signal and M to be $N/3$, [8]. This procedure based on the assumption that speech signal could be considered as quasi-stationary if speech signal is examined over a sufficiently short period of time (between 5 and 100 ms). However, this is not true when the signal is measured over long periods of time (on the order of 0.2 seconds or more). For reducing the discontinuities associated with windowing, pitch synchronously speech processing may be utilized [9, 10]. This technique is mainly used for synthesis of speech and rate-reduction speech coding.

It is well known that the choice of analysis frame length is fundamental to any transform-based audio coding methods. A long transform length is most suitable for input signals whose spectrum remains stationary or varies slowly with time, such as the quasi-steady state voiced regions of speech. A long transform length provides greater frequency resolution. On the other hand, a shorter transform length, processing greater time resolution, is more desirable for signals that are changed rapidly in time. Therefore, the time versus frequency resolution tradeoff should be considered when selecting a transform frame length. The traditional approach to solve this dilemma is to select a single transform frame length that provides the best tradeoff of coding quality for both stationary and dynamic signals. For example, the typical value of N and M are 300 and 100 when the sampling rate is 6.67 kHz [8]. It is well known that there are areas where the speech signal is relatively constant and in other areas the signal may change rapidly in time. Therefore, this scheme does not provide good choices for both stationary and dynamic signals. However, in practice, this is the situation in extracting spectral features for speech recognition.

Another alternative to solve this dilemma is to apply the variable frame rate (VFR) analysis techniques. Those are well-established methods for data reduction in speech coding and have been used in various speech recognition systems [11]. Ponting and Peeling [12] gave the first demonstration and showed that VFR could successfully improve the performance of speech recognition. Based on a similarity measure between two feature vectors, the VFR algorithm is designed to retain all the input feature vectors when they are changing rapidly and remove a large number of vectors when they are relatively constant. VFR results in considerable saving of computational load by reducing the number of frames processed. Specifically, if $D(i, j)$ is the distance between the previously selected frame j and the current frame i, and the threshold is T, then the rule is to select frame i as the next output frame if $D(i, j) \geq T$, [12]. Furthermore, Besacier et al. [13] used the frame pruning (a variation of VFR) for speaker identification, and significant improvement was achieved on the TIMIT and NON-TIMIT databases. An evident shortcoming of the VFR is that the discarded feature vectors have no contribution to the recognition procedure.

For speech signals, some regions are relatively constant (stationary speech) while others may change rapidly in time (dynamic speech). Under a constraint of fixed (average) frame rate, a practical approach is to code the dynamic speech with higher time resolution by analyzing it with shorter frame, and code stationary speech in higher frequency resolution by analyzing it with longer frame.

This paper proposes an adaptive frame length selection approach (AFL) that can adapt to the frequency/time resolutions of the transform depending upon the spectral and temporal characteristics of the signal being processed. During the feature extraction process, each speech frame is examined at different time scales. If a transient is detected in the second half of a speech frame, then this frame will be analyzed with short frame (a normal frame is divided into two short frame of same length in the experiments). Thus, more accurate speech coding can be achieved by using the adaptive frame length method. It is expected that this accurate coding of speech could result in better recognition performance, which are confirmed with the experiments carried out in this study.

This paper is organized as follows. Transient detection method and the analysis approaches for transient frame are described in Section 2, which is the major contribution of this work. In Section 3, experimental results on TIMIT and NON-TIMIT databases are presented to demonstrate the effectiveness of the AFL on recognition performance. Finally, we summarize our findings in Section 4.

2. THE METHOD

For applying the adaptive frame length analysis in speech recognition, the first task is to segment speech signal into quasi-stationary and dynamic intervals. Then the spectral feature vectors of quasi-stationary signal are computed with normal frames (N samples) and dynamic signal with short frames. For practical reasons, the length of short frame is set to be half of a normal frame, that is, a normal frame is divided into two short frames. A frame is analyzed in short or normal frame based on whether a transient could be detected in that particular frame or not. If a transient is found, we call it as transient frame. The transient detection algorithm is presented first, and then two expressions are suggested for the feature vectors of transient frame.

2.1. *Transient detection*

Transients are detected for each normal frame in order to decide whether a frame is analyzed in two short frames or not. The pre-emphasized version of the speech signal is examined for an increase in energy from one subframe time-segment to the next. Subframes are examined in different time scales. If a transient is detected in the second half of a speech frame, that frame is then analyzed using short frames.

Transient detection could be done in three steps:

(1) segmentation of the frame into subframes in different time scales,
(2) peak amplitude detection for each subframe segment,
(3) threshold comparison. The transient detector will output a flag to indicate whether this frame is analyzed in short for normal frame.

(1) Frame segmentation: normal frame of N pre-emphasized samples are segmented into a hierarchical tree in which level 1 represents two $N/2$ length frames and level 2 is four segments of length $N/4$.

(2) Peak detection: the sample with the largest magnitude is identified for each segment on every level of the hierarchical tree. Let $P[j][k]$ be the peaks of kth segment of level j, where $j = 1, 2$ is the hierarchical level number, k is the segment number within level j.

(3) Threshold comparison: the relative peak levels of successive segments on each level of the hierarchical tree are checked. If any of the following inequalities is true, then transient is found in current frame:

$$P[1][2] * T[1] > P[1][1] \tag{1}$$

or

$$P[2][k] * T[2] > P[2][k-1], \quad k = 3, 4, \tag{2}$$

where $T[j]$ is the pre-defined threshold for level j. Since there is no perfect theoretical way for determining a good setting for $T[j]$, these thresholds are determined experimentally. We do it by investigating the effect of varying the values of $T[j]$ and pick the one that gave the best test performance of speech recognition for the experiments. For the 8 bit NON-TIMIT database, the optimal setting is $T[1] = 0.2$ and $T[2] = 0.15$. For the 16 bit TIMIT database, the optimal setting is $T[1] = 0.1$ and $T[2] = 0.075$.

The above procedure is designed for detecting the transients where the energy of signal is increasing. If the transient happening for the signal energy is decreasing, then a decrease in energy from one subframe time-segment to the next is examined by the following inequalities:

$$P[1][2] * T[1] < P[1][1] \tag{3}$$

or

$$P[2][k] * T[2] < P[2][k-1], \quad k = 3, 4. \tag{4}$$

Experiments were designed to demonstrate effect of the two cases:

(i) only the transients with increasing energy are considered,
(ii) both the transients with increasing and decreasing energy are considered. The results are discussed in Section 3.

2.2. *Feature representation of transient frames*

What is the feature vector of a transient frame? The following two factors are considered:

(1) the feature extracted from a short frame (half of a normal frame),
(2) the feature representation of the transient frame should be identical to that of a normal frame.

Otherwise, extra computation cost is needed to deal with the difference between the feature vectors of the transient and the normal frames.

A straightforward implementation is to analyze each half-frame (short frame) separately, and a feature vector with the same dimension is computed. In other words, two feature vectors represent a transient frame. Of course, these two vectors have higher time resolution against the vector of normal frame (15 ms versus 30 ms if 30 ms window is used). Experimental results demonstrated that this simple representation could give substantial improvement to the performance of speech recognition system. It is also evident that this schedule increased the frame rate as well as the computation cost for both training and recognition. Therefore, we designed to use another feature extracting scheme that provides the same frame rate as the fixed frame length schedule. Let D be the dimension size of the feature vector for the normal frame. Each short frame is analyzed separately and produces $D/2$ feature coefficients. And the two vectors of $D/2$ coefficients are interleaved together on a coefficient-by-coefficient basis to form a single vector of D coefficients. Let $e_1, e_2, \ldots, e_{D/2}$ be the feature vector of the first half subframe, and $f_1, f_2, \ldots, f_{D/2}$ be the feature vector of the second half subframe, these two vectors are interleaved to form a vector with D coefficients: $e_1, f_1, e_2, f_2, \ldots, e_{D/2}, f_{D/2}$. Then the transient frame could be treated in the same way for the normal frames. In Section 3, experimental results showed that this feature extraction approach has comparable performance to the previous feature extraction schedule, but does not cause any extra computation complexity.

3. EXPERIMENTAL RESULTS

Two sets of experiments were investigated, one is to recognize 21 English words and the other is to recognize 31 Chinese words. Standard maximum likelihood training was used to estimate the HMMs' parameters, and Viterbi algorithm was adopted to give the recognition performance.

3.1. *The databases*

The first database was extracted from TIMIT corpus. TIMIT [14] corpus contains a total of 6300 utterances. Ten utterances spoken by each of 630 speakers from 8 major dialect regions of the United States, which has been designed to provide speech data for the acquisition of acoustic-phonetic knowledge and for the development and evaluation of automatic speech recognition system. TIMIT used 16K sampling rate and 16 bit precision. In TIMIT, each of the two dialect "shibboleth" sentences labeled as sa1 and sa2 has 630 tokens separately. The sentences are

(1) She had your dark suit in greasy wash water all year. (sa1)
(2) Don't ask me to carry an oily rag like that. (sa2)

We extracted the tokens of the above 21 words from all sentences under the test group of TIMIT corpus. But the word "an" has much less utterances than other words. In order to make its utterances be comparable to that of other words, some tokens of "an" in the DR1 subdirectory of TRAIN group were extracted. Table 1 lists the number of utterances of each word used for training and test. Totally, 3698 utterances were used. For each word, the first 100 utterances were used for HMM parameter estimation, and the left were used for open test (speaker-independent, totally 1598 utterances for 21 words).

Table 1: Distribution of words' utterances.

she	had	your	dark	suit	in	greasy	wash	water	all	year
185	183	185	171	168	185	168	168	170	185	177
don't	ask	me	to	carry	an	oily	rag	like	that	
180	169	185	185	169	165	168	168	179	185	

As mentioned before, two expressions could be used to represent the feature vector of a transient frame, and transient could be detected in energy increasing mode and in energy decreasing mode. There are four combinations with different expression and different transient detection scope, which results in 4 sets of spectral feature vectors for TIMIT. Including the basic feature sets without transient detection, 8 sets of spectral feature vectors (FS1 to FS8) are extracted:

(1) FS1: features were extracted with fixed frame length, no transient was detected. The speech was analyzed in 30 ms frame to produce a 24-dimensional feature vector through the standard LPC analysis [8]. The feature vector consists of 12 weighted cepstrum coefficients and 12 delta-cepstrum coefficients. And adjacent frames had 10 ms step.

(2) FS2: transient detection with energy increasing indicated whether a frame was a transient frame. The frames without transient were analyzed normally as for FS1. Frames with transient were divided into two short frames. Each short frame of the transient frame was represented with a vector of 12 elements. These two 12-order vectors were then interleaved together on a coefficient-by-coefficient basis and formed a single vector of 24 elements.

(3) FS3: features were extracted with transient detection of increasing energy. But each short frame of a transient frame was represented with a vector of 24 elements. Two normal feature vectors represent one transient frame. The difference between FS3 and FS2 is the representation of the transient frame.

(4) FS4: transients with energy increasing or decreasing were detected. A transient frame was represented with two 12-order vectors of its two short frames. These two 12-order vectors were interleaved together on a coefficient-by-coefficient basis to form a single vector of 24 coefficients. The difference between FS4 and FS2 is the scope of transient detection.

(5) FS5: transients with energy increasing or decreasing were detected. But a transient frame was represented with two 24-order vectors of its two short frames. Thus, two normal feature vectors represent a transient frame. The difference between FS5 and FS4 is the representation of the transient frame.

(6) FS6: all frames are represented with the interleaved half-spectral resolution feature vectors.

(7) FS7: the retained features of VFR are coded with the original 15 ms frame analysis.

(8) FS8: the retained features of VFR are coded with original 20 ms feature analysis.

Continuous density hidden Markov models (CHMM) and discrete hidden Markov models (DHMM) were separately trained and tested on the above 5 feature sets.

A NON-TIMIT database contains 31 Chinese words was used, which includes digits (0–9), words for number operation: addition, subtraction, multiplication, division, etc. This was an informal database, which consisted of 96 utterances of each word, spoken by the authors through microphone, digitized with sound blaster card in 8 K sampling rate and 8 bit precision. Just gave a secondary investigation of the effect of adaptive frame length analysis in speech recognition. 60 of the 96 utterances were used for HMM estimation, and the other 36 were used for open test. The same analysis procedure was carried on the NON-TIMIT database as we did on the TIMIT database to get 5 different feature sets.

3.2. HMM topology and parameter initialization

Standard HMM topology was used in the experiments, the structure parameters of the HMM are

- Left-to-right HMM with six states were used for both CHMM and DHMM.
- Two Gaussian mixtures were used in each state of CHMM.
- 256 distinct vectors comprise the codebook of DHMM.

Model parameters were initialized with the labeled training data in a uniform distribution way, that is, each utterance in the training feature set was segmented into partitions with equal length and attached them to each state of the model in the order from left to right. By counting observations in each state, the parameter initialization of DHMM is straightforward. But it is little complex to initialize the mixture parameters of CHMM, which was completed in the following way. If there are L vectors in the ith state, the L vectors are first clustered into P classes (P is the number of mixtures in each state) with K-mean cluster algorithm [8]. The center vector and the standard deviation are then computed for each class and then used as the initial value of the mixture Gaussian density parameters.

3.3. Experiments

The following experiments are done for this paper:

(1) Experiments with continuous density hidden Markov model on 5 feature sets of TIMIT database. 2100 tokens of the 21 English words were recognized in the closed test (test with the HMM training data), and 1598 utterances were recognized in the open test. Both the recognition error rates and phoneme classification rates are listed in the first line of Tables 2 and 5, where FS1–FS5 were defined in Section 3.1. Considering the characteristics of TIMIT corpus, these results are for speaker-independent.

(2) Experiments with discrete HMM on 5 feature sets of TIMIT database. The recognition error rates and phoneme classification rates are listed in the second line of Tables 2 and 5. These results are also for speaker-independent.

(3) Experiments with continuous density HMM on 5 feature sets of NON-TIMIT database. 1860 tokens of the 31 Chinese words were recognized in the closed test, and 1126 utterances were recognized in the open test. The recognition error rates are listed in the first line of Table 3. These results are for speaker-dependent.

(4) Experiments with discrete HMM on 5 feature sets of NON-TIMIT database. The recognition error rates are listed in the second line of Table 3. These results are for speaker-dependent.

3.4. Experimental results

We have the following observations based on the results listed in Tables 2 and 3.

(1) For the open test, the proposed adaptive frame length analysis approach could generate steady improvement in all cases. It does not depend on what kinds of HMMs, databases or feature sets are used. With the TIMIT database, the adaptive frame length (AFL) gave an error reduction from 4.54% to 9.58% for CHMM, and from 4.47% to 11.21% for DHMM. With the NON-TIMIT database, AFL gave an error reduction from 2.63% to 9.5% for CHMM, and from 1.91% to 11.55% for DHMM.

(2) For the closed test, we do not know definite conclusions here. The influence of AFL to system performance depends on the feature set used. It showed that the AFL may be benefit from some feature sets or harm others. For example,

Table 2: The results (error rate %) on TIMIT database.

	Closed set test					Open set test				
	FS1	FS2	FS3	FS4	FS5	FS1	FS2	FS3	FS4	FS5
CHMM	5.07	4.87	4.6	5.07	4.53	12.11	11.21	11.04	10.95	11.56
DHMM	2.94	2.97	2.94	2.2	2.63	16.32	14.49	15.59	15.24	15.06

Table 3: The results (error rate %) on NON-TIMIT database.

	Closed set test					Open set test				
	FS1	FS2	FS3	FS4	FS5	FS1	FS2	FS3	FS4	FS5
CHMM	4.85	4.73	4.87	4.19	4.52	6.42	6.25	5.81	5.99	6.17
DHMM	1.41	1.28	1.07	1.35	1.43	9.44	8.53	9.14	9.26	8.35

it gave DHMM an error reduction of 25.17% (from 2.94% to 2.2%) on the FS4 of TIMIT. However, for the FS2 of TIMIT, it gave DHMM an error increase of 1.02% (from 2.94% to 2.97%). Overall speaking, AFL results in some improvement of recognition performance.

(3) DHMM gave better results on the closed test, but worse results on the open test than that of the CHMM. For example, within the TIMIT database, the lowest closed error rate of CHMM (4.53%) is higher than the highest closed error rate of DHMM (2.97%). But the highest open error rate of CHMM (12.11%) is lower than the lowest open error rate of DHMM (14.49%). This conclusion is true for NON-TIMIT database. It shows that the CHMM is more robust than DHMM and it could provide better matches between the testing and training conditions. Therefore, CHMM is preferable for high-performance speech recognition systems.

(4) It is also observed that the feature set benefiting most for open test usually has the lowest improvement in the closed test. For instance, with CHMM-based recognizer, FS4 of TIMIT gave most improvement of open test performance (9.58%), but the least improvement of closed test performance (0.0%). With DHMM-based recognizer, FS2 of TIMIT gave the highest error reduction in the open test (11.21%), however, the lowest in the close test (−1.02%). What does this mean? One simple reason may be that FS4 (FS2) of TIMIT gives least mismatch between training and testing environments if English speech is assumed to be generated by CHMM (DHMM).

We also observe that even though FS3/FS5 used more coefficients to represent a transient frame than the FS2/FS4 did, it does not always provide better performance. It might be due to the reason that hidden Markov model is only an approximating model for speech. Different feature vectors fit HMM at certain different degree of angles, and may result in different modeling performance. At one extreme, the perfect representation of speech may not give the best performance for the HMM-based recognizer. This is the reason why variable frame rate, [12], could improve the performance of speech recognizer. There, it is evident that the retained vectors could not give better representation of speech than the original vector set does.

Table 4: Results (error rate %) on TIMIT database for comparison.

	Closed set test			Open set test		
	FS6	FS7	FS8	FS6	FS7	FS8
CHMM	11.76	5.47	5.54	16.19	11.35	11.27
DHMM	4.37	2.88	2.97	18.04	14.22	14.06

Another question arises that if the improvement of AFL comes from the interleaved half-spectral resolution feature used in the FS2/FS4 for dynamic intervals of speech. In order to confirm this, another feature set (FS6) of the TIMIT database was extracted, where all frames were represented with interleaved half-spectral resolution features. The results are listed in Table 4. The CHMM-based recognizer gave 88.24% accuracy for the training data set and 83.81% for the open test set. From Table 2, the CHMM-based recognizer trained with the standard feature set FS1 gave a closed recognition rate of 94.93%, and an open recognition rate of 87.89%. For DHMM-based recognizer, the closed correctness is 95.63%, and the open correctness is 81.96%. But the worst closed correctness in Table 2 is 97.03%, and worst open correctness is 83.68%. These results absolutely confirmed that the interleaved half-spectral resolution feature gave a poor recognition performance than the method used in this work.

We also would like to know the relationship between the performance of the proposed adaptive frame length and variable frame rate (VFR) [12]. In order to give a comparison, two experiments were investigated on the TIMIT database. The comparison was carried on the base of the same computational load of AFL. In the first experiment, speech signal was analyzed with 15 ms frame, and then variable frame rate analysis was adopted to select the retained vectors. By adjusting the threshold T, we made the retained feature vector set (FS7) give the same average data rate of feature set FS2 or FS4. CHMM-based recognizer gave a closed error rate of 5.47%, and an open correctness of 11.35%. DHMM-based recognizer gave a closed error rate of 2.88%, and an open correctness of 14.22%. In the second experiment, speech signal was analyzed with 20 ms frame, and then variable frame rate analysis was adopted to select the retained vectors. By

Table 5: The results (phoneme classification rate %) on TIMIT database.

	Closed set test					Open set test				
	FS1	FS2	FS3	FS4	FS5	FS1	FS2	FS3	FS4	FS5
CHMM	67.95	68.52	69.29	68.02	69.46	60.37	63.19	64.01	64.83	62.53
DHMM	69.49	69.30	69.53	74.42	71.55	54.19	60.02	56.58	57.63	58.2

Table 6: Results (phoneme classification rate %) on TIMIT database for comparison.

	Closed set test			Open set test		
	FS6	FS7	FS8	FS6	FS7	FS8
CHMM	61.47	67.63	67.28	56.23	62.9	63.1
DHMM	68.14	70.13	69.32	52.78	60.77	61.04

adjusting the threshold T, we made the retained feature vector set (FS8) give the same average data rate of feature set FS2 or FS4. CHMM-based recognizer gave a closed error rate of 5.54%, and an open correctness of 11.27%. DHMM-based recognizer gave a closed error rate of 2.91%, and an open correctness of 14.06%. All these results are listed in Table 4. The results of the phoneme classification are represented in Table 6.

Comparing those with the results listed in Tables 2 and 5, it indicated that the VFR did better than AFL to the open performance of DHMM-based recognizer, but worse to the performance of CHMM-based recognizer. Therefore, VFR and AFL may be complementary to each other, each does the best for different acoustic model. It can also be seen that the VFR pays more attention to the important part (features) of speech pattern by lowering the redundant information. On the other hand, the DHMM uses some representative features in the future space to represent speech pattern. When the left speech partition in the VFR matches the representative features well, then the VFR may make the DHMM-based recogniser provide better performance. Contrarily, the AFL pays more attention to the key part (features) of speech pattern through representing it in a more precise manner, and the CHMM gives a complete probability description to the speech feature space. Therefore, AFL may be more suitable for the CHMM-based speech recogniser.

4. CONCLUSIONS

In this paper, we have presented a method for transient frame detection. Two expressions are proposed for transient frame detection and the proposed speech analysis scheme is simple. Word recognition experiments on the TIMIT and NTIMIT with continuous density HMM and discrete HMM showed that steady performance improvement could be achieved for open set testing. On the TIMIT database, adaptive frame length approach (AFL) gave an error reduction of 4.47% to 11.21% and 4.54% to 9.58% for DHMM and CHMM, respectively. On the NTIMIT database, AFL gave DHMM an error reduction of 1.91% to 11.55% and 2.63% to 9.5% for CHMM. These results gave preliminary evidence for the effectiveness of the proposed adaptive frame length for speech analysis. Especially for the open testing, this is the practical usage for evaluating how a speech recognition system works.

Experimental results showed that variable frame rate, [12], is more suitable for discrete HMM-based recognizer, and adaptive frame length analysis is more suitable for CHMM-based recognizer.

5. ACKNOWLEDGEMENT

This work was supported in part by the City University of Hong Kong Strategic grant 7001059 and the National Natural Science Foundation of China project 69881001.

REFERENCES

[1] M. Ostendorf, V. D. Vassilios, and A. K. Owen, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. On Speech and Audio processing*, vol. 4, no. 5, pp. 360–378, 1996.

[2] J. F. Jason and B. M. John, "On adaptive HMM state estimation," *IEEE Trans. On Signal processing*, vol. 46, no. 2, pp. 475–486, 1998.

[3] S. Kwong, Q. H. He, K. F. Man, and K. S. Tang, "A maximum model distance approach for HMM-based speech recognition," *Pattern Recognition*, vol. 31, no. 3, pp. 219–229, 1998.

[4] B.-H. Juang, W. Chou, and C.-H. Lee, "Minimum classification error rate methods for speech recognition," *IEEE Trans. On Speech and Audio processing*, vol. 5, no. 3, pp. 257–265, 1997.

[5] L. Deng and C. Rathinavelu, "A Markov model containing state-conditioned second-order non-stationarity: application to speech recognition," *Computer Speech and Language*, vol. 9, pp. 63–86, 1995.

[6] H. Feichtinger and T. Stronmas, *Gabor Analysis and Algorithms: Theory and Applications*, Birkhaüser, 1998.

[7] I. Daubechies, "The wavelet transform, time-frequency localization, and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, 1990.

[8] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*, Chapter 3. PTR Prentice-Hall, 1993.

[9] G. Evangelista, "Pitch-synchronous wavelet representation of speech and music signals," *IEEE Trans. On Signal Processing*, vol. 41, no. 12, pp. 3313–3330, 1993 Special Issue on Wavelets and Signal Processing.

[10] H. Yang, S.-N. Koh, and P. Sivaprakasapillai, "Pitch synchronous multi-band (PSMB) speech coding," in *ICASSP-95*, 1995, vol. 1, pp. 516–519.

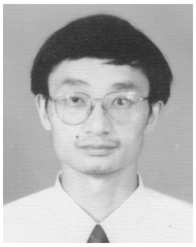
[11] Y. L. Chow, M. O. Dunham, and O. A. Kimball, "BYBLOS: The BBN continuous speech recognition system," in *Proc. of the IEEE ICASSP*, Dallas, 1987, pp. 89–92.

[12] K. M. Ponting and S. M. Peeling, "The use of variable frame rate analysis in speech recognition," *Computer Speech and Language*, vol. 5, pp. 169–179, 1991.

[13] L. Besacier and J. F. Bonastre, "Frame pruning for automatic speaker identification," in *Signal Processing IX: Theories and Applications*, 1998, vol. 1, pp. 367–370.

[14] W. Fisher, V. Zue, J. Bernsten, and D. Pallet, "An acoustic-phonetic database," *JASA*, vol. suppl. A 81(S92), 1986.

Sam Kwong received his B.Sc. degree and M.Sc. degree 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. 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 Scientific staff where he worked on both the DMS-100 voice network and the DPN-100 data network project. 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. His research interests are in Genetic Algorithms, Speech Processing and Recognition, Data Compression and Networking.



Qianhua He received the B.Sc. degree in physics from Hunan Normal University in 1987, the M.Sc. degree in medical instrument engineering from Xi'an Jiaotong University in 1990, and the Ph.D. degree in communication engineering from South China University of Technology, in 1993. Since 1993, he has been at the Institute of Radio and Auto-control of South China University of Technology. And he has been an associate professor since 1997. His research interests include speech recognition, speaker recognition, and natural language processing, optimal algorithm design, such as genetic algorithm and neural networks, microcomputer application in industry, encoding methods of multiple channel audio. From May 1994 to March 1996, July 1998 to June 1999, and May 2000 to July 2000, he worked with the department of computer science of the City University of Hong Kong.