AN ORTHOGONAL POLYNOMIAL REPRESENTATION OF SPEECH SIGNALS AND ITS PROBABILISTIC MODEL FOR TEXT INDEPENDENT SPEAKER VERIFICATION

Chi-Shi Liu[1,2], *Hsiao-Chuan Wang*[2], *Frank K. Soong*[3], *Chao-Shih Huang*[1]

[1] Telecommunication Labs. Ministry of Transportation and Communications, Taiwan, ROC.
[2] Dept. of Electrical Engineering, National Tsing Hua Univ., Taiwan, ROC.
[3] Speech Research Department, AT&T Bell Labs, USA.

ABSTRACT

In this paper, a segmental probabilistic model based on an orthogonal polynomial representation of speech signals is proposed. Unlike the conventional frame based probabilistic model, this segment based model concatenates the similar acoustic characteristics of consecutive frames into an acoustic segment and represents the segment by an orthogonal polynomial function. An algorithm which iteratively performs recognition and segmentation processes is proposed for estimating the parameters of the segment model. This segment model is applied in the text independent speaker verification. For a 20-speaker database, the experimental results show that the performance by using segment models is better than that by using the conventional frame based probabilistic model. The equal error rate can be reduced by 3.6% when the models are represented by 64-mixture density functions.

1. INTRODUCTION

Recently, Hidden Markov Model(HMM) was proposed for speaker recognition applications [2, 3, 4]. HMM is a method for modeling the speaker's acoustic space by probability. *Matsui* and *Furui* showed that the performance of speaker verification by HMM was better than that by the template modeling methods. However, for most of speaker recognition systems, the speaker model is a frame-based model[3, 4]. For a frame-based HMM, the observation probability in every state is obtained by a frame-based probability density function, but the observation probability density function at different states could be different. The observed frames in the same state should use the same probability density function to compute the observation probability. By using an N-state HMM to obtain the statistic of a given acoustic segment, we consider this acoustic segment as being piecewisely concatenated by N-state subsegments and its characteristic is obtained by a composition of state-by-state distribution. Furthermore, the statistics of frames over the same state are assumed to be independent in the frame-based HMM. This assumption is not good for speech signals since speech signals are short time stationary signals and the signals over the short-time state are dependent.

In this paper, we consider the speech signal as being composed of a sequence of stationary segments instead of frames. The states in HMM are segment-based states, not frame-based states. The spectral continuity over a segment is kept, and the probability distribution for this segment model is not longer a piecewise distribution over subsegments like the frame-based model at the same state. Moreover, the dependence between frames and the behavior of signal variation can be preserved.

Although there are good reasons to use a segment-based model for modeling speech signals, the methods of representing the segments, generating the segment model, and segmenting the speech signals should be proposed. The simplest way for generating a segment model is to segment speech signals every N frames[5], and then use the EM algorithm to generate the segment model. For this segment method, the training data and the size of model should be large enough to capture most of behavior of speech signals. The other method for generating a segment model was proposed by *Ostendorf etc*[6], and was successfully applied for speech recognition. Although good results are given, a lot of memories are needed for storing the parameters of segment models. In this paper, a new segment model is proposed for text independent speaker recognition. A segment, which is composed of L successive N-dimensional feature vectors, is considered to be a set of N trajectories whose lengths are equal to L. The previous study[4] showed that the performance depended on the acoustic resolution, *i.e.*, the number of states multiplied by the number of mixtures. We can maintain the acoustic resolution by using a single-state model with more mixtures. The observation probability density function of our segment model is described by a mixture of gaussian probability density functions, which are represented by three parameters, (A, U, C). Here A is the mixture means, U is the covariance matrix, and C is the weights on the mixtures. A mapping function $\mathcal{F}$ is used to map the mixture means into a time sequence of feature vectors. In this paper, the mapping function $\mathcal{F}$ is the orthogonal polynomial function. Then the mixture means are a set of the orthogonal polynomial coefficients. It is not easy to simultaneously estimate this model and the segment boundary. An iterative method to generate the segment-based speaker model is proposed in this paper. For the same database and experimental conditions, the segment-based model can perform better than the conventional frame-based models.

The other parts of this paper are organized as follows. Section 2 describes our new segmental probabilistic model and

0-7803-2431-5/95 $4.00

discusses the approach to derive this segment model. Databases and experiments are given in Section 3. Section 4 gives a brief summary to our proposed method.

2. SEGMENTAL PROBABILISTIC MODEL(SPM)

SPM is a segment based model. The smallest unit for SPM is an acoustic segment. Several successive frames with similar characteristics are concatenated into an acoustic segment. Thus, for a given speech signal, we need to partition this speech signal into several acoustic segments according to some criterion. The number of segments, the length for each segment and the SPM are simultaneously determined in the training phase. In the following sections, we discuss the representation of SPM, the optimal criterion for generating the SPM, and the way of partitioning speech signals according to the given criterion.

2.1. REPRESENTATION OF SEGMENTS BY THE ORTHOGONAL POLYNOMIAL FUNCTION

Since SPM is a segment-based model, the parameters in this model should be mapped into a time sequence of vectors, and then the likelihood between a time sequence of given vectors and mapping vectors could be computed. Given a set of orthogonal coefficients, $\{a_0, \cdots, a_r, \cdots, a_R\}$, we can use the following formula to regenerate a time sequence of L-length feature vectors, $X_l = \{x(1), \cdots, x(l), \cdots, x(L)\}$. This mapping formula is given by

$$X_l = \mathcal{F}(A; L; R), \tag{1}$$

with the column vector $x(l)$ equal to

$$x(l) = \sum_{r=0}^{R} a_r \phi_r^L(l), \quad for \quad l = 1, \cdots, L, \tag{2}$$

where $\mathcal{F}$ is the orthogonal polynomial function whose input arguments are a set of orthogonal coefficients A, the segment length L and the degree of the orthogonal polynomial function, R. $\phi_r^L(l)$ is a polynomial of degree r. The dimension of a feature vector $x(l)$ is assumed to be d. This assumption is also used for all following sections.

2.2. FORMULATION OF THE SPM

A SPM is represented by

$$\Lambda = \{c_m, A_{R,m}, U_m | m = 1, \cdots, M\},$$

where $A_{R,m} = [a_{0,m}, \cdots, a_{r,m}, \cdots, a_{R,m}]$ is a set of orthogonal coefficients which are used to generate segment mean according to eqns.(1) and (2), $a_{r,m}$ is an orthogonal coefficient vector for an orthogonal polynomial of degree r, U_m is the $d \times d$ dimensional covariance matrix, c_m is the mixture weight and M is the total number of mixtures. For a set of signal feature vectors, $X = \{x(1), \cdots, x(t), \cdots, x(T)\}$, the log-likelihood for this signal X is given by

$$\log P(X|\Lambda) = \max_{B} \sum_{j=0}^{J-1} \log P(X_j|\Lambda, B), \tag{3}$$

where B is a possible segment boundary in the set

$$\{b_0, \cdots, b_j, \cdots, b_J | b_j \in [b_{j-1} + 1, T], \quad for \quad j = 1, \cdots, J \quad with \quad b_0 = 0 \quad and \quad b_J = T\}, \tag{4}$$

J is the number of partitioned segments in accordance with B, $X_j = \{x(b_j + 1), \cdots, x(b_{j+1})\}$ is the jth segment, $\log P(X_j|\Lambda, B)$ is defined as

$$\log P(X_j|\Lambda, B) = \log \sum_{m=1}^{M} c_m P(X_j|A_{R,m}, U_m, B), \tag{5}$$

where

$$P(X_j|A_{R,m}, U_m, B) = \tag{6}$$

$$\prod_{t=b_j+1}^{b_{j+1}} (2\pi)^{-\frac{d}{2}} |U_m|^{-\frac{1}{2}} \exp[-\frac{1}{2} o_m^j(t)^T U_m^{-1} o_m^j(t)], \tag{7}$$

$$o_m^j(t) = (x(t) - \sum_{r=0}^{R} a_{r,m} \phi_r^{(b_{j+1}-b_j)}(t - b_j)). \tag{8}$$

Eqn.(3) illustrates that the maximum log-likelihood $\log P(X|\Lambda)$ is obtained by choosing the optimal segment boundary from all possible segment boundaries of eqn.(4). The dynamic programming algorithm[8] can be used to find the optimal segment boundary for a given speech signal.

Since SPM is a segment model, the model parameters and the segment boundary should be simultaneously estimated. Thus, in the training phase, the optimal criterion is to find the optimal segment boundary B_{op} and the model parameters Λ_{op} such that the log-likelihood $\log P(X|\Lambda_{op}, B_{op})$ is maximum for the given training feature vectors, $X = \{x(1), \cdots, x(T)\}$. $\log P(X|\Lambda_{op}, B_{op})$ is defined as

$$\log P(X|\Lambda_{op}, B_{op}) = \max_{B,\Lambda} \sum_{j=0}^{J-1} \log P(X_j|\Lambda, B), \tag{9}$$

where B is a possible segment boundary in the set of eqn.(4), $\log P(X_j|\Lambda, B)$ is defined as eqn.(5) and J is the number of partitioned segments. It is not an easy task to directly solve eqn.(9) for obtaining the segment boundary and the segment model. In this paper, we propose an iterative algorithm to solve eqn.(9). The procedures are estimation of the segment model using the known boundary, and then estimation of the segment boundary using the known segment model. These two estimation procedures are iterated until this algorithm is converged. We depict this algorithm as follows.

[Iterative Algorithm :]

(1) **Initialization :**

initially guess B^0, and set i to 1.

The initial segment boundary is obtained by the maximum likelihood segmentation method[7] except the gaussian mean is represented by the orthogonal polynomials.

(2) **Reestimation of the segment model using the known segment boundary :**

The new segment model Λ^i for the known segment boundary B^{i-1} is obtained by maximizing the following log-likelihood function,

$$\log P(X|\Lambda^i, B^{i-1}) = \max_{\Lambda'} \sum_{j=0}^{j^{i-1}-1} \log P(X_j|\Lambda', B^{i-1}). \quad (10)$$

The EM algorithm can be used to solve the above equation.

(3) Reestimation of the segment boundary using the known segment model :

The new segment boundary B^i for the known segment model Λ^i are determined by maximizing the following log-likelihood function,

$$\log P(X|\Lambda^i, B^i) = \max_{B} \sum_{j} \log P(X_j|\Lambda^i, B). \quad (11)$$

The above equation can be solved by the dynamic programming algorithm[8].

(4) Termination : If it is converged, then stop; otherwise set i to $i+1$ and go to step 2.

It can be shown that the sequence of the log-likelihoods is increased with the iteration time. For the ith iteration of the segment model Λ^i, the log-likelihood $\log P(X|\Lambda^i, B^{i-1})$ is greater than $\log P(X|\Lambda^{i-1}, B^{i-1})$ since Λ^i is the optimal model for the segment boundary B^{i-1}. Similarly, the log-likelihood $\log P(X|\Lambda^i, B^i)$ is greater than $\log P(X|\Lambda^i, B^{i-1})$ since the segment boundary B^i are the optimal boundary for the segment model Λ^i. Thus, we have the relation,

$$\log P(X|\Lambda^{i-1}, B^{i-1}) \leq \log P(X|\Lambda^i, B^{i-1}) \leq \log P(X|\Lambda^i, B^i),$$

at the ith iteration. By extending this relation to other iterations, we have

$$\log P(X|\Lambda^0, B^0) \leq \cdots \leq \log P(X|\Lambda^i, B^i) \leq \cdots$$

This relation indicates the increase of log-likelihood by the above iterative algorithm. Speaker models are obtained by the above training procedure. In the verification phase, the scores of the test utterances for a claimed speaker model Λ are computed, and then compared with the threshold associated with this speaker model to verify the identity of a claimed speaker. The *posteriori* equal error rate is used to measure the system performance.

3. DATABASES AND EXPERIMENTS

3.1. DATABASES

The database[1] used in the following experiments consists of 20,000 isolated digit utterances recorded by 100 speakers, 50 males and 50 females. The utterances were recorded over dialed-up local telephone lines. Each speaker was asked to utter 200 digits, 20 repetitions of each digit, in five recording sessions over a period of two months. In each recording session, the speakers were prompted to utter four complete sets of the digits with random order. The first 20 speakers in 100-speaker database are used for the experiments. The first 80 utterances of each speaker are used for training and the rest 120 utterances are used for testing. The average of equal error rates is obtained by alternatively choosing one of 20 speakers as a claimed speaker and then taking the average of these equal error rates.

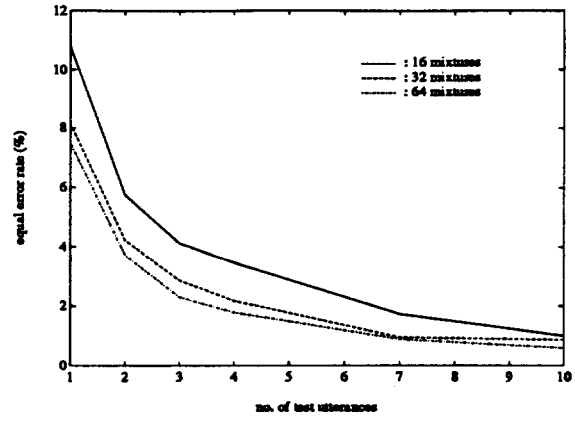

Figure 1: The equal error rates of SPM for different digital lengths and mixture numbers.

All of utterances are bandpass filtered from 200 to 3200 Hz and sampled at 6.67 kHz. The digitized speech signal is preemphasized using the filter, $H(z) = 1 - 0.95z^{-1}$. Nine autocorrelation coefficients are calculated over 45 msec after a Hamming windowing and shifted by every 15 msec. These autocorrelation coefficients are used to calculate 8 cepstral coefficients. Diagonal covariance matrix is used for all probabilistic models.

3.2. EXPERIMENTS

There are several experiments to be done. The first part of experiments is to evaluate the performance of the SPM for speaker verification by varying the mixture number and the degree of the orthogonal polynomial. The second part of experiments is to compare the performance of the SPM with that of the conventional frame based probabilistic model.

A. Evaluating the performance of the SPM :

The mixture number in SPM represents that the number of acoustic segments is used for modeling speaker's characteristics. The larger the mixture number is, the higher the spectral resolution for speaker model can be obtained. To evaluate the mixture number on the performance of speaker verification, the first experiment is given to examine the performance of the speaker verification system for different mixture numbers. In the all following experiments the searching window size is set to 9 for the training and the verification phases.

Fig.1 shows the equal error rates of the SPM for different digital lengths and mixture numbers. The degree of the orthogonal polynomial used in this experiment is set to 3. The results show that the equal error rate is reduced as the digital length or the mixture number are increased. As the mixture number is beyond 32, the improvement becomes saturated. This indicates that the mixture number equal to 32 is good enough to represent the speaker's characteristics. The degree of the orthogonal polynomial affects the accuracy of the SPM. The smallest length of a parti-

tioned segment is determined by the degree of the orthogonal polynomials. For an r-degree orthogonal polynomial to be used, the smallest length for any partitioned segment is $(r+1)$. Thus, the type of the basic segment for the SPM and its characteristics depends on the degree of the orthogonal polynomials. Besides, computation time and memory storage are affected by the degree.

Table 1 shows the influence of the degree of the orthogonal polynomials on the performance of speaker verification. The test digital length used in this experiment is equal to 1. The results show that the best performance is in the degree equal to 3 and the improvement from the 0th degree to the first degree is the greatest one. Moreover, as the degree is greater than 1, the improvement is small. We also find that as the degree is greater than 3, the accuracy rate is decreased. For higher degree orthogonal polynomials, we need a larger training database to obtain a more reliable model which is robust to test utterances. However, for lower orthogonal polynomials, the accuracy of the SPM is worse. This is just a tradeoff problem. A better choice is to use the third degree orthogonal polynomial.

Table 1: The equal error rates(%) for different degrees of orthogonal polynomials.

degree of orthogonal polynomials	no. of test utterances		
	1	4	7
0	11.33	3.40	1.87
1	8.52	2.33	1.63
2	7.79	1.82	1.05
3	7.49	1.78	0.86
4	7.92	1.90	1.19

B. Comparing the performance of the SPM with that of the conventional frame based probabilistic model :

Frame based probabilistic Model(FBPM) is one of good models used for speaker verification[4]. In this experiment, FBPM is used as a baseline model. The degree of the orthogonal polynomial for the SPM is set to 3. The mixture number for FBPM and SPM is set to 64. The searching window size is constrainted to 9.

Table 2 shows the performance of the FBPM and the SPM for different digital lengths and mixture numbers. The results clearly show that the performance of the SPM is better than that of the FBPM. For the less mixture number and digital length, the equal error rate by the SPM is much less than that by the FBPM. As compared with the results in Table 1, the performance of the FBPM is still worse than those of SPM with the degree of the orthogonal polynomial greater than 0.

4. SUMMARY

In this paper, a segmental probabilistic model based on the orthogonal polynomial representation of speech signals was proposed. The spectral dependence and the spectral continuity over the intra-segment are captured into this model. An iterative algorithm was proposed to estimate the SPM.

Table 2: The equal error rates(%) for SPM and FBPM

model type	no. of mixtures	no. of test utterances		
		1	4	7
SPM	16	10.79	3.46	1.72
	32	8.15	2.17	0.95
	64	7.49	1.78	0.86
FBPM	16	15.74	6.61	4.10
	32	12.44	4.48	2.72
	64	11.08	3.18	1.89

Experiments showed that the performance of the SPM was better than that of the conventional FBPM. For the less mixture number, the improvement by the SPM is greater than that by FBPM. The degree of the orthogonal polynomial used for the SPM would affect the performance of the SPM. The results showed that the best degree was set to 3.

5. ACKNOWLEDGEMENT

Authors would like to thank AT&T Bell Laboratories for supplying the database. We also thank Dr. J.T. Wang, Dr. I.C. Jou, Dr. B.S. Jeng and our colleagues for supporting this research.

6. REFERENCES

[1] F. K. Soong, A. E. Rosenberg and L. R. Rabiner and B. H. Juang, "A vector quantization approach to speaker recognition", *AT&T Technical Journal*, vol. 66, pp. 14-26, Mar./Apr., 1987.

[2] A. E. Rosenberg, C-H. Lee and F. K. Soong, "Sub-word unit talker verification using hidden markov models", *Proc. ICASSP-90*, vol. 1, pp. 269-272, Apr., 1990.

[3] N. Tishby, "On the application of mixture AR hidden Markov models to text independent speaker recognition", *IEEE Tran. Acoust., Speech, Signal Processing*, vol. ASSP-39, pp. 563-570, Mar., 1991.

[4] T. Matsui and S. Furui, "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/continuous HMMs", *Proc. ICASSP-92*, Vol. 2, pp. 157-160, Mar., 1992.

[5] B-H. Juang and F. K. Soong, "Speaker recognition based on source coding approaches", *Proc. ICASSP-90*, Vol.2, pp. 613-616, Apr., 1990.

[6] M. Ostendorf and S. Roukos, "A stochastic segment model for phoneme-based continuous speech recognition", *IEEE Tran. Acoust., Speech, Signal Processing*, Vol. ASSP-37, pp. 1857-1869, Dec., 1989.

[7] T. Svendsen and F. K. Soong, "On the automatic segmentation of speech signals", *Proc. ICASSP-87*, Vol. 1, pp. 77-80, Apr., 1987.

[8] H. Ney, "The use of a one-stage dynamic programming algorithm for connected word recognition," *IEEE Tran. Acoust., Speech, Signal Processing*, Vol. ASSP-32, pp.263-271, Apr., 1984.