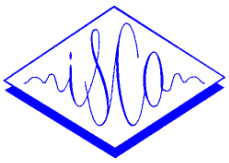

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

SPEAKER RECOGNITION WITH TEMPORAL TRANSITION MODELS

Haizhou Li Jean-Paul Haton Jian Su Yifan Gong
CRIN-CNRS/INRIA-Lorraine, Campus Scientifique, BP 239, 54506 Vandoeuvre-lès-Nancy, France
E-mail: hzli,jph,gong@loria.fr, eesujian@cpccux0.cityu.edu.hk

ABSTRACT

In this paper, a temporal transition model (TTM) of speech is proposed for speaker recognition and verification. The TTM is introduced to encode the short time dynamics of speech. The issues on the model building, the distance measures and the implementation are addressed. A set of experiments were conducted based on TTM, which gave a 98.9% recognition rate and 99.5% verification rate on a database of 72 French speakers. The fact is confirmed that temporal dynamics of utterance encodes well speaker specificity.

1 INTRODUCTION

The research issue of temporal dynamics of speech has attracted more and more attention in speech recognition. Applications such as SM[1] and STM[2] in continuous speech recognition demonstrated its successes. Recently several studies on the speaker specificity of spectral dynamics have been carried out. They show that the dynamics carries discriminative characteristics of speakers[3, 4]. An AR-Vector model was proposed by Montacie[3] to capture the spectral dynamics in text-free speaker recognition experiment, which is costly in computation. Other approaches such as nonlinear interpolation of vectors by Gong[4] employed the correlation between successive vectors in speech feature sequence. In this paper, we use the maximum likelihood estimation algorithm in preparing speaker dependent statistical models, the Gaussian speaker model[5]. Then a temporal transition model (TTM) is constructed based on the speaker model. Some TTM distance measures are also discussed. Finally, a set of experiments on speaker recognition and verification is carried out to show the effectiveness of the model.

2 SPEAKER DATA MODELING

Let $X = \{x_0, \ldots, x_t, \ldots, x_T\}$ be a set of data from speaker s. If the *a priori* probability of category ω_j, $Pr(\omega_j)$, and the category-conditional pdf of drawing sample x_t from category ω_j, $f(x_t/\omega_j)$ are known, where ω_j is the parameters describing the category distribution, the probability of drawing the sample data x_t from a mixture of J-pdf, Ω_s, can be given as follows

$$f(x_t/\Omega_s) = \sum_{j=1}^{J} f(x_t/\omega_j) Pr(\omega_j) \qquad (1)$$

where $f(x_t/\omega_j)$ is called component density and $Pr(\omega_j)$ the mixture coefficient, subject to the constraint

$$\sum_{j=1}^{J} Pr(\omega_j) = 1. \qquad (2)$$

It is known that the *a posteriori* probability is given as

$$f(\omega_j/x_t) = \frac{f(x_t/\omega_j) Pr(\omega_j)}{\sum_{j'=1}^{J} f(x_t/\omega_{j'}) Pr(\omega_{j'})} \qquad (3)$$

In this paper, we have

$$f(x_t/\omega_j) = N(x_t; m_j, \Sigma_j) \qquad (4)$$

as Gaussian density.

Given a set of training data, it is possible to unsupervisedly cluster all sample data from a speaker into a certain number of states, which are also referred to as mixture components ω_j in Eq.(1). The Expectation-Maximization algorithm is one of the most popular algorithms for the maximum likelihood estimation. When we consider a speech utterance as a sequence of articulatory evolution, a state could be associated with an articulatory configuration. The observation probability of mixture density could be interpreted as the occurrence probability of certain articulatory configuration in an evolution. The speaker Gaussian mixture model here serves as the underlying definition of state for temporal transition models.

3 TEMPORAL TRANSITION MODEL

The TTM is motivated by two ideas. One is that most of pattern classification approaches are devoted to

static patterns, or the patterns with fixed size of components. The problem will become simpler if speech dynamics could be translated into a static model. Another idea is that a temporal transition within a small time slot carries speaker-specific information[3, 4].

3.1 State transition

Considering a time slot of Q vectors

$$\mathbf{x}_t = \{x_t, \ldots, x_{t+Q-1}\} \tag{5}$$

a Q-dimensional space could be spanned by the *a posteriori* state transition probabilities of $p(\omega_{d_1}, \ldots, \omega_{d_Q}/\mathbf{x}_t)$. For each $\mathbf{x}_t$, we have a Q dimensional array with J elements in each dimension, which is called a state transition probability array and is considered to be related to the articulatory characteristics of speakers. For simplicity, let

$$p_{jt} = p(\omega_j/x_t), \tag{6}$$

therefore a set of probability $\{p_{0t}, ..., p_{Jt}\}$ could be obtained for each speech observation x_t. Suppose that the occurrences of successive speech vectors are independent, hence

$$Pr(\omega_{d_1}, \ldots, \omega_{d_Q}/\mathbf{x}_t) = p_{d_1 t} \times \cdots \times p_{d_Q t+Q-1} \tag{7}$$

Since we have

$$\sum_{d_q=1}^{J} p_{d_1 t} = 1 \tag{8}$$

it is noted that the summation of all elements in a transition probability array is equal to one.

$$\sum_{d_1=1}^{J} \cdots \sum_{d_Q=1}^{J} p_{d_1 t} \times \cdots \times p_{d_Q t+Q-1} = 1 \tag{9}$$

Taking the average of all transition probability arrays over the short term speech sessions X, i.e. $T-Q+1$ slots, along all Q dimensions by

$$\theta_{d_1 \ldots d_Q} = \frac{\sum_{t=1}^{T-Q+1} p_{d_1 t} \times \cdots \times p_{d_Q t+Q-1}}{T-Q+1} \tag{10}$$

a TTM with J^Q elements is obtained. There are $(T-Q+1)$ transition probability arrays for T successive feature vectors because of the endpoint effect. One can easily verify

$$\sum_{d_1=1}^{J} \cdots \sum_{d_Q=1}^{J} \theta_{d_1 \ldots d_Q} = 1 \tag{11}$$

3.2 Distance measures

After building a TTM for a given speech session based on the speaker specific Gaussian mixture models, the distance measures are required to evaluate the differences between TTMs. To deal with two probability distribution models, one possibility is to use the information divergence (ID) which is usually used in measuring the difference between distributions[6, 1]. Given two TTM $A = \{a_{d_1 \ldots d_Q}\}$ and $B = \{b_{d_1 \ldots d_Q}\}$, The ID distance is given as

$$D(A,B) = \sum_{d_1=1}^{J} \cdots \sum_{d_Q=1}^{J} a_{d_1 \ldots d_Q} \log \frac{a_{d_1 \ldots d_Q}}{b_{d_1 \ldots d_Q}} \tag{12}$$

We know that (cf. Appendix 1),

$$D(A,B) \geq 0 \tag{13}$$

but it is asymmetric:

$$D(A,B) \neq D(B,A). \tag{14}$$

An alternative is to combine two of them as

$$D^*(A,B) = [D(A,B) + D(B,A)]/2 \tag{15}$$

and give a symmetric one with more computational load. However, if the same order of A and B, or the order of reference patterns and test patterns on the distance measuring, is kept, one can still obtain consistent results with $D(A,B)$.

Now we consider the Q dimensional array of TTM as an one dimensional vector for simplicity by letting $\theta_i = \theta_{d_1 \ldots d_Q}$ and $I = J^Q$, where

$$i = \{\{\{d_1 \times J + d_2\} \times J\} + \cdots + d_Q\} \tag{16}$$

in the cardinal set $[1, I]$. We have a TTM vector of $\Theta = \{\theta_i\}$. Eq.(11) is rewritten as

$$\sum_{i=1}^{I} \theta_i = 1. \tag{17}$$

In the case of multiple reference pattern matching, one is usually asked to find the centroid in a pattern cluster. A centroid Θ^c in a cluster of K templates, could be obtained by minimizing the objective function

$$f(\theta^c) = \sum_{k=1}^{K} \sum_{i=1}^{I} \theta_i^c \log \frac{\theta_i^c}{\theta_i^k} \tag{18}$$

for $D(A,B)$ and,

$$f(\theta^c) = \sum_{k=1}^{K} \sum_{i=1}^{I} \theta_i^k \log \frac{\theta_i^k}{\theta_i^c} \tag{19}$$

for $D(B,A)$ subject to the constraint of $\sum_{i=1}^{I} \theta_i^c = 1$ according to the Min-max theory. The centroids for

the two distances under discussion could be derived as Eq.(20) for $D(A,B)$ and Eq.(21) for $D(B,A)$ (cf. Appendix 2). Unfortunately, it is not straightforward to give a centroid with the sysmetric measure $D^*(A,B)$.

$$\theta_i^c = \frac{\sqrt[K]{\prod_k \theta_i^k}}{\sum_i \sqrt[K]{\prod_k \theta_i^k}} \tag{20}$$

$$\theta_i^c = (1/K)\sum_k \theta_i^k \tag{21}$$

4 EXPERIMENTS

A database from 72 French speakers was used in the experiments where 36 male and 36 female speakers were involved. It consists of three-word French phrases[1] naturally uttered five times by each speaker. Using one of the five sessions as the test data and others as the training data, rotating the orders resulted in five assessment sets. The results are reported as the average of the five sets. A 12th mel-scale cepstral analysis was performed each 10ms time interval with 32ms Hamming window over each session. All silences and stops in the sessions were eliminated since they are not discriminative. When modeling the mixture model for a speaker, an eight-mixture Gaussian pdfs was trained ($J = 8$) with the EM algorithm. TTMs with different state dependency Q were built for each session.

4.1 Speaker recognition

Three experiments were carried out in the recognition test. In Table 1, C is referred to as the case where TTM centroids for $D(A,B)$ or $D(B,A)$ are used. NC means no centroid is used and all four TTM references are evaluated in each recognition trial for each speaker. It is noted that the Gaussian pdfs computation is time consuming. We have proposed a tied model and MMI learning algorithm[5] for tying the mixtures across speakers into a sharing Gaussian kernel so that the computation of pdfs can be significantly reduced. Speakers are only discriminated by the mixture coefficients. Tying pdfs will surely reduce the discriminability of the system. A compensation is made by increasing the number of sharing components. The experiments with different mixture number are reported in Table 1, referred to as $J = 32, 64, 128$.

4.2 Speaker verification

All experiments for speaker recognition are carried out for speaker verification as well. It is known that a normalized distance outperforms unnormalized one by introducing a group of speaker as the background

[1] *cent beaux papis*

	Q=1	Q=2	Q=3
NC	84.8/85.3	98.9/97.2	98.7/97.5
C	85.9/85.5	98.5/97.1	98.5/97.7
	J=32	J=64	J=128
C/Q=2	70.6/77.4	96.5/94.4	98.2/97.2

Table 1: Recognition accuracies (%) with $D(A,B)/D(B,A)$

speakers[7]. We use a new cohort selection approach to group 24 cohort speakers for each speaker as presented in [7]. The cohort speakers are excluded as imposters to avoid the unexpected optimistically biasing of the results. The verification accuracy is defined as the average of the *true-accept* (true claim goes with acceptance) and *false-reject* (false claim goes with rejection) rates. Only $D(A,B)$ distance is evaluated in verification experiment since it is shown to be more effective than $D(B,A)$.

	Q=1	Q=2	Q=3
NC	90.8	99.4	99.5
C	92.8	98.9	98.9
	J=32	J=64	J=128
C/Q=2	83.6	98.5	99.3

Table 2: Speaker verification accuracies (%)

5 DISCUSSION

We have developed a speaker model for speaker recognition and verification. The two ideas presented in section 3 have been implemented. A 98.9% recognition accuracy and 99.5% verification accuracy are obtained as the best results in the experiments.

According to the modeling procedure, we know that a larger Q means to have an average of transition probability over a longer period, which may lead to an ambigious description of short term dynamics. By observing the results for Q=2 and Q=3, one notes that the two cases perform almost the same. Other than the accuracy, one factor that we should consider is the computational complexity. When Q increases, the components in the model will augment exponentially and introduce more computation. When $Q = 1$, actually, no speech dynamics is taken into account. We conclude that TTM with Q=2 is a reasonable model setting for the data base.

A Appendix 1

We could prove the ID distance in Eq.(12) not less than zero by introducing one augmented objective

function with a Lagrange factor λ,

$$f(a_i) = \sum_{i=1}^{I} a_i \log \frac{a_i}{b_i} + \lambda(\sum_i a_i - 1) \quad (22)$$

Equating the partial derivative of $f(a_i)$ to zero with respect to a_i, we have

$$a_i/b_i = e^{-(1+\lambda)} \quad (23)$$

Applying $\sum a_i = 1$ and $\sum b_i = 1$, we come up with $\lambda = -1$ and $f(a_i)$ has its only critical point at $a_i = b_i$. Since

$$\frac{\partial^2 f(a_i)}{\partial^2 a_i} = \frac{1}{a_i} > 0 \quad (24)$$

It is concluded that $f(a_i)$ obtains its minimum of zero when

$$a_i = b_i. \quad (25)$$

Q.E.D

B Appendix 2

The proof for Eq.(20) and (21) comes from the following augmented objective functions by introducing a Lagrange factor λ,

B.1 Case: $D(A, B)$

$$f(\theta_i^c) = \sum_{k=1}^{K}\sum_{i=1}^{I} \theta_i^c \log \frac{\theta_i^c}{\theta_i^k} + \lambda(\sum_i \theta_i^c - 1) \quad (26)$$

Equating the partial derivative of $f(\theta_i^c)$ to zero with respect to θ_i^c

$$\theta_i^c = \sqrt[K]{\prod_k \theta_i^k} e^{-\frac{K+\lambda}{K}} \quad (27)$$

could be derived. Substituting θ_i^c into $\sum_i \theta_i^c = 1$, we have

$$e^{\frac{K+\lambda}{K}} = \sum_{i'=1}^{I} \sqrt[K]{\prod_k \theta_{i'}^k} \quad (28)$$

Substituting Eq.(28) into Eq.(27), the result is obtained.

$$\theta_i^c = \frac{\sqrt[K]{\prod_k \theta_i^k}}{\sum_{i'=1}^{I} \sqrt[K]{\prod_k \theta_{i'}^k}} \quad (29)$$

B.2 Case: $D(B, A)$

$$f(\theta_i^c) = \sum_{k=1}^{K}\sum_{i=1}^{I} \theta_i^k \log \frac{\theta_i^k}{\theta_i^c} + \lambda(\sum_i \theta_i^c - 1) \quad (30)$$

Equating the partial derivative of $f(\theta_i^c)$ to zero with respect to θ_i^c

$$\theta_i^c = \frac{1}{\lambda}\sum_k \theta_i^k \quad (31)$$

could be derived. Substituting θ_i^c into $\sum_i \theta_i^c = 1$, we have $\lambda = K$ so that

$$\theta_i^c = \frac{1}{K}\sum_k \theta_i^k \quad (32)$$

REFERENCES

[1] C. Chan and H. Li. Level building with static models in connected digits recognition. In *International Conference on Computer Architecture & Digital Signal Processing*, Hong Kong, Oct. 1989.

[2] Y. Gong and J.-P. Haton. Stochastic trajectory modeling for speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, volume I, pages 57–60, Adelaide, Australia, April 1994.

[3] C. Montacie and J.-L. Le Floch. Discriminant AR-vector models for free-text speaker verification. In *Proceedings of European Conference on Speech Technology*, pages 21–23, Berlin, Germany, 1993.

[4] Y. Gong and J.-P. Haton. Non-linear vectorial interpolation for speaker recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, USA, March 1992.

[5] H. Li, J.-P. Haton, and Y. Gong. On MMI learning of Gaussian mixture for speaker models. In *Proceedings of European Conference on Speech Technology*, Madrid, Spain, 1995.

[6] R. Ash. *Information theory.* Inter-Science Publisher, 1967.

[7] K. T. Ng, H. Li, J.-P. Haton, and J. Su. Nonparametric distance measures of speaker verification. *IEE Electronic Letters*, 31(9), 1995.