International Symposium on Chinese Spoken Language Processing (*ISCSLP'98*), ASR-B3, pp.164-168, Dec.7-9, 1998, Singapore

THE SIMILARITY MEASURE AMONG ACOUSTIC MODELS AND ITS TWO APPLICATIONS

WU Jian, ZHENG Fang, WU Wen-Hu, FANG Di-Tang
Speech Lab., Dept. of Computer Science and Technology,
Tsinghua Univ., Beijing
Tel: +86-10-6277 2001, FAX: +86-10-6277 2001, E-mail: jwu@sp.cs.tsinghua.edu.cn

ABSTRACT

The distance measure of two stochastic processes is a key problem in the processing of stochastic signals. In speech recognition, the distance between two basic recognition models can provide the information about the relation and the difference of these two units. In fact, the distance measure can depict the model's availability. We can improve the hit rate of recognition results by adjusting the distance between basic unit models. In recent years, many definitions have been put forward for calculating the exact value of the distance between stochastic processes. We also have developed a simplified distance measure based on CDCPM (Center-Distance Continuous Probability Model) which is an improved version of CHMM (Continuous Hidden Markov Model). And since CDN (Center-Distance Normal) distribution is derived from the normal distribution, the definition can be extended to other types of acoustic models such as Segmental HMM easily. In this paper, we will focus on this simplified definition of distance measure and propose two examples applied to continuous speech recognition. And the experiment result shows it preserve very good performance without additory computation.

1. INTRODUCTION TO CDCPM

Dominant acoustic models in speech recognition (SR) are HMMs, including continuous mixture density HMMs [Bahl 90, Rabiner 85, Juang 85b] with full covariance matrices or diagonal covariance matrices, semi-continuous HMMs [Huang 89], and VQ-based discrete HMMs [Rabiner 83].

A Continuous HMM (CHMM) is represented by state transition probability matrix A, observation probability density function (PDF) matrix B and initial probability distribution vector π. Many algorithms are developed to estimate these HMM parameters, such as Baum-Welch [Baum 72], EM (Expectation and Maximization) [Dempster 77], MMIE (Maximum Mutual Information Estimation) [Bahl 86], and MAP (Maximum a Posterior) [Gauvain 92]. Also many algorithms are developed for recognition, such as Viterbi algorithm [Viterbi 67] and Frame Synchronous Search algorithm [Lee 89].

But many reports [Juang 85a, Lyu 98] show that the transition probability matrix A in traditional HMM is not as useful as observation probability density function matrix B. In our research work on Mandarin isolated syllable recognition, we find that the state decoding process can be separated from the pattern match process. According to this, we developed a kind of new model that discards the matrix A to reduce the complexity in training process. Furthermore, in order to reduce more computational complexity in recognition process, we choose CDN distribution rather than Gaussian distribution to represent the observation PDF.

The final results provide a recognition rate of 52.25% (control experiment on CHMM gives 50.93%) at a speed of about 20 times faster than CHMM, which shows this simplified model has a good behavior in Mandarin isolated syllable recognition.

1.1. The CDCPM

In CDCPM, each utterance of a given PLU (Phone Like Unit) is divided into N states before training, and the feature vectors in each state are modeled by an observation probability distribution $b_i(O_t)$, which is composed of several Center-Distance Normal Distribution. Below, we shall present the concept of CDCPM in a nutshell.

1.1.1 Center-Distance Normal Distribution

The PDF of a normal random variable ξ with mean value μ_x and standard deviation σ_x is as follows,

$$N(x;\mu_x,\sigma_x)=\frac{1}{\sqrt{2\pi}\,\sigma_x}\exp\left\{-(x-\mu_x)^2/2\sigma_x^2\right\} \qquad x\in(-\infty,\infty) \tag{1}$$

Define a new random variable $\eta=|\xi-\mu_x|$, we have the PDF of η as

$$p(y;\sigma_x)=\frac{2}{\sqrt{2\pi}\,\sigma_x}\exp(-y^2/2\sigma_x^2),\quad y\ge 0 \tag{2}$$

In fact, η is the distance between a normal variable ξ and its mean value μ_x, thus the defined distribution is referred to as Center-Distance Normal (CDN) distribution.

By calculating the mean value μ_y of CDN variable η in Eq. (2), we have

$$\mu_y = \sqrt{\frac{2}{\pi}}\sigma_x \quad \text{or}$$
$$\sigma_x = \sqrt{\frac{\pi}{2}}\mu_y \tag{3}$$

Substituting Eq. (3) into (2), the PDF can be rewritten as

$$p(y;\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2/\pi\mu_y^2),\ y \geq 0. \tag{4}$$

D-dimensional case is similar to mono-dimensional case. Denote the (weighted) Euclidean distance between a *D*-dimensional normal vector $\vec{\xi}$ and its mean value vector $\vec{\mu}_{\mathbf{x}}$ by another random variable $\eta = y(\vec{\xi},\vec{\mu}_x)$. Assume η is a CDN variable, then its CDN pseudo-PDF (**PPDF**) is

$$N_{CD}(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y) = \frac{2}{\pi\mu_y}\exp(-y^2(\vec{x},\vec{\mu}_{\mathbf{x}})/\pi\mu_y^2) \tag{5}$$

Strictly speaking, $N_{CD}(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y)$ is the PDF of $y(\vec{\xi},\vec{\mu}_{\mathbf{x}})$ instead of that of $\vec{\xi}$, it is just for convenience and comparison. For simplification, Eq. (5) is called a CDN PPDF while Eq. (4) a CDN PDF.

1.1.2 CDCPM

According to the definition of CDN, we can denote the form of CDCPM with N states and M CDN densities each state for a given PLU as follows:

$$\lambda : \{b_i(\cdot),\quad 1 \leq i \leq N\} \tag{6}$$

where $b_i(O_t) = \max_{1\leq m\leq M} N_{CD}(\vec{x};\vec{\mu}_{\mathbf{x}},\mu_y)$

So the score of an utterance S = O_1 O_2 …O_T matching again the CDCPM model θ is

$$Score\{\mathbf{O}|\lambda\} = \prod_{t=1}^{T} b_{s_t^{(ML)}}(\vec{\mathbf{o}}_t) \tag{7}$$

1.2. The Similarity Measure among CDCPMs

Compared to traditional HMM, CDCPM not only discards the transition probability matrix and the initial probability vector in traditional HMM but also considers that the PDF of every state doesn't observe the Mixed Gauss Distribution but the Mixed CDN Distribution. This change makes the training and recognition process simpler and more efficient. But it also makes the classic definition of distance between two HMMs useless. Based on CDCPM, we define a kind of distance to measure the difference of two acoustic models.

1.2.1 The distance definition for CDN distribution

Consider the CDN PDF described in Eq. (4), by shifting this function by *d* along *y* axis and then unfolding the CDN PDF function, we get the corresponding normal PDF as follows:

$$p_0(y,\mu_y) = \frac{1}{\pi\mu_y}\exp\left(-y^2/\pi\mu_y^2\right)$$
$$y \in (-\infty,\infty) \tag{8}$$

So the distance between two CDNs can be mapped into a distance between two corresponding Gauss Distributions, which is shown as in Fig.1.

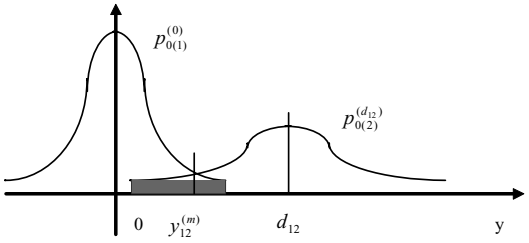

Fig.1 The distance between two Gaussian Distribution
*The area in shadow represents the distance of these two distributions

1.2.2 The distance definition for CDCPMs

The distance measure between two CDCPMs is based on the distance measure for the CDN distribution. For CDCPMs let

$$\Lambda_k = \left\{N_{CD(knm)} \middle| 1 \leq n \leq N, 1 \leq m \leq M\right\}$$
$$k = 1,2 \tag{9}$$

denote two CDCPMs. Define

$$D(\Lambda_1,\Lambda_2) = \frac{1}{N}\sum_{n=1}^{N}\left[\min_{1\leq m\leq M} D(N_{CD(1nm)}, N_{CD(2nm)})\right] \tag{10}$$

as the CDCPM model distance measure.

According to the definition hereinbefore, we know that the computation of the distance doesn't require the real speech signal data. Its result comes directly from CDCPM models' parameters. So it will be quick and effective in actual training and recognition process. Because CDN distribution is derived from the normal distribution, the definition we proposed could be extended to other types of acoustic model based on normal distribution such as Segmental HMM. Furthermore, although the original form of the distance is asymmetric, we can get a symmetry distance matrix by a simple transform. We have done many experiments to make use of the feature of our acoustic model trained from the distance matrix. The next two sections are the examples about it and the experiment results show that it has a good performance in Mandarin Syllable-Based Speech Recognition.

2. A QUICK ALGOTITHM

In the recent typical SR (speech recognition) systems, the basic idea of recognition is that an incoming speech signal S matches against a large number of competing models，and then identify the first top n PLUs with highest scores as the candidates. This method has its shortcoming: many match scores will be discarded once the highest score is identified. For traditional HMM, the cost of working out the match score is too great. This makes it difficult to get the possible candidates in real time. In order to solve this problem without too much cost, we developed a quick algorithm that makes use of the similarity measure between stochastic processes.

2.1. Equivalence Model Classes (EMC)

Let $M_i(i=1,\cdots,N)$ be the N competing PLU models. Depend on the distance definition between stochastic processes, we get a matrix $D=\{d_{ij}\}$ in size $N\times N$, where $d_{ij}=d(M_i,M_j)$ represents the distance between models M_i and M_j. Studying the value of each element in this matrix, we can find an interesting phenomenon: if the distance between two models M_i and M_j is smaller than a predefined threshold ε, the row vectors $\vec{d}_{i\bullet}$ and $\vec{d}_{j\bullet}$ also are strongly comparable. It is because the acoustic similarity of these two PLUs. So we can carry out a new "distance" matrix $E=\{e_{ij}\}$ directly from distance matrix D, where $e_{ij}=1-\rho(\vec{d}_{i\bullet},\vec{d}_{j\bullet})$ and $\rho(\cdot,\cdot)$ is the cross-correlation function.

Since matrix E reflects not only the distance between PLUs but also the relation with other PLUs, it represents the similarity between PLU models more exactly than the matrix D. From this matrix we can easily get K equivalence model classes $\mathrm{MC_i}\ (i=1,\cdots,K)$ by a simple cluster algorithm. Every class consists of T_i confusing PLU models $\mathrm{MC_{ij}}\ (j=1,\cdots,T_i)$, which satisfies

$$MC_i\cap MC_j=\varnothing\ \forall i\neq j \qquad (11)$$

and

$$\bigcup_{i=1}^{K}MC_i=\{M_j\,|\,j=1,\cdots,N\} \qquad (12)$$

According to the total probability formula, the probability $P(s|M_i)$ can be described as follows,

$$P(s|M_i)=\sum_{j=1}^{K}P(s|MC_j,M_i)\cdot P(MC_j|M_i) \qquad (13)$$

Suppose $M_i\in MC_k\ (1\le k\le K)$, Eq.(13) can be simplified ulteriorly (Since the utterance of M_i is not similar to the PLUs in $MC_j\ (j\neq k)$, the probability $P(MC_j|M_i)$ is approximately equal to $\delta(j,k)$). The simplified result is

$$P(s|M_i)\approx P(s|MC_k,M_i) \qquad (14)$$

Let $\overline{MC_i}$ be the delegate PLU model of class MC_i, Eq.(14) can be changed to Eq.(15)

$$P(s\,|\,M_i)\approx P(s\,|\,\overline{MC_k}) \qquad (15)$$

2.2. Quick Algorithm

Known the approximate scores of utterance matches against each PLUs, we can discard many impossible candidate and then identify the most likely PLUs by calculating the exact match scores against a few remaining similar PLUs. Through this method, we get the same recognition results in low computational cost. The idiographic algorithm is as follows,

Step1. Calculate the matrix E, then classify all the PLUs into **K** EMCs and select the delegate PLU of every model class;
Step2. For each utterance **S = O_1 O_2 …O_T**, work out the approximate match scores against all the N PLU models;
Step3. Sort the score list and prune the PLUs in the last **(K-C)** model class candidates;
Step4. Calculate the exact value of match scores against the models in rest C model classes;
Step5. Take the candidate result as the final result of this algorithm.

2.3. Experiments and Results

In this part, the speech database used in all the following experiments and some front-end signal processing performed on the database are described. And we will show the test results using this new quick algorithm.

The speech database is a Mandarin Continuous Speech Database recorded by 38 men. Each speaker uttered one set of sentences in a continuous mode. The database contains 250,657 syllables totally. We used 180,065 of them to train the CDCPMs' parameters and distance matrix. And the remaining part is used for testing. All the recorded materials were obtained in an officelike laboratory environment through a close-talk noise-canceling microphone. They are digitized with a sampling frequency of 16KHz. The filtered speech was taken by a 32ms Hamming window. And then the cepstral coefficients derived from LPC of order 16 were extracted for each 16ms window shift. The acoustic model used for experiments is the CDCPM with 6 states and 16 mixtures each state.

Table 1 shows the result on both CDCPM using traditional algorithm and CDCPM using modified recognition algorithm. EMC means the CDCPM using quick algorithm, while NORMAL means CDCPM using traditional algorithm. The parameter K and C are the same meaning as hereinbefore. The parameter M means

the mixtures number each state used in the CDCPM.

Table 1 Test Result List of Quick Algorithm (%)

	Top n	1	2	3	4	5
1	EMC (*K*=56,*C*=10)	49	65	72	76	79
2	EMC (*K*=63,*C*=10)	49	65	73	77	79
3	EMC (*K*=56,*C*=15)	50	67	75	79	82
4	NORMAL (*M*=8)	40	56	64	70	74
5	NORMAL (*M*=16)	52	70	78	83	86

The results list in table 1 show that this quick algorithm has a good reduction in the time complexities without too much reduction in hit rate (only from 52% down to 50% for the first candidate and from 86% down to 82% for the first five candidates). And we can learn the movement tendency when parameters K and C were changed.

2.4. Discussion

From the description of quick algorithm, we can easily know the complexity of new algorithm is only $\left(\frac{C}{K}+\frac{K}{N}-\frac{C}{N}\right)$ times of the complexity of traditional algorithm. Choosing proper value of K and C, the complexity can be reduced by almost 50%. This means that after adopting the quick algorithm, the computational complexity of recognition using CDCPM with 16 mixtures every state will reduce to the same level as that using CDCPM with 8 mixtures every state. The results in table 1 (row 3 and row 5) show that the hit rate using complex models and quick algorithm is higher than that using simple models but traditional algorithm. With this property, we can choose more detailed acoustic models to make the hit rate of recognition results higher under the same level of time and memory complexity, and the experiment results have proved the efficiency of this method.

3. A NEW CLASSIFIER

3.1. The Principle of New Classifier

The traditional Bayesian classifier that has been used frequently has its advantage mathematically, but its performance depends crucially on how well the class distributions are separated. So once the form of distribution we choose doesn't accord with the real distribution, the classified result will be elusory. One of the solution is to utilize more information besides the data distribution, such as the relation among PLUs. Speaker recognition has slathered this idea. Here we also will develop a kind of new classifier based on the distance measure of the acoustic model, which makes use of the correlation between the PLU models.

Let $SCORE(s_j|M_i)$ denote the score of an utterance s_j of PLU M_j matching against M_i.

$I_k=\{I_{ki}|i=1,\cdots,T_k\}\ (1\le k\le K)$ denotes the subscript set of model class MC_k.

$\overrightarrow{SC_{kj}}=(SC_{kj}^1,SC_{kj}^2,\cdots,SC_{kj}^{T_k})$ denotes the vector consists of match score between s_j and the model in model class MC_k, where $SC_{kj}^i=SCORE(s_j|MC_{ki})$.

We suppose $\overrightarrow{SC_{kj}}$ satisfies a certain kind of probability distribution. In order to simplifier the computation, we define

$$p(\overrightarrow{SC_{kj}}|\theta)=\sum_{i=1}^{M}\lambda_i p(\overrightarrow{SC_{kj}}|\theta_i) \tag{16}$$

where M is the number of mixtures in this model, $p(\overrightarrow{SC_{kj}}|\theta_i)$ is the PDF of the mixture, λ_i is the weighted coefficient which satisfies $\sum_{j=1}^{M}\lambda_i=1,\lambda_i\ge 0$, and θ_i is the parameter of the single mixture, which is composed by mean vector $\overrightarrow{\mu_i}$ and covariance matrix R_i. The formula of $p(\overrightarrow{SC_{kj}}|\theta_i)$ can be defined as:

$$p(\overrightarrow{SC_{kj}}\mid\theta_i)=(2\pi)^{-d/2}\mid R_i\mid^{-1/2}\cdot \exp\left(-\frac{1}{2}(\overrightarrow{SC_{kj}}-\overrightarrow{\mu_i})^T R_i^{-1}(\overrightarrow{SC_{kj}}-\overrightarrow{\mu_i})\right) \tag{17}$$

3.2. Experiments and Results

Two experiments in different manners using this classifier have been conducted, and the test result is list in Table 2 and discussed below. The basic acoustic model is also CDCPM with 6 states and 16 mixtures each states. The probability model that is used in this classifier has 2 mixtures for each PLU. PUC is the experiments that use this classifier to prune some candidate after traditional recognition process. And RUC is the experiments that use this classifier to replace the traditional classifier.

Table 2 Test Result List of New Classifier (%)

	Top n	1	2	3	4	5
1	PUC	54	75	82	87	89
2	RUC	49	68	79	84	86
3	NORMAL	52	70	78	83	86

This classifier uses a kind of new evaluation standard, which emphasizes not only the real distribution of every unit but also the relationship of them, to represent the

between-sample variability and between-model variability. From the result, we come to a conclusion that this new classifier can not only remedy the error of normal classifier, but also achieve a good performance when adopted solely. An important thing that should be announced here is that the final results are very dependent on the selection of equivalence model class.

4. CONCLUSION AND EXPECTATION

The results discussed above are preliminary research on similarity measure among acoustic models. In future research, we will try to get further improvement on the description ability of the distance measure and other corresponding work. The direction can be as follows:
1) To make the distance measure among acoustic models more explicit;
2) To make use of more information about the distance between PLUs to improve the rejection ability of new classifier.

REFERENCE

[1] **Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, R.L., (1986)** "*Maximum mutual information estimation of hidden Markov model parameters for speech recognition,*" Proc. ICASSP-86, 49-52, Apr. 1986

[2] **Bahl, L.R., Brown, P.F., de Souza, P.V., Mercer, K.L., (1990)** "Speech Recognition with Continuous-parameter Hidden Markov Models," *Readings in Speech Recognition*, edited by Alex Waibel & Kai-Fu Lee, 1990, pp.332-339

[3] **Baum, L.E., (1972)** "An Inequality and Associated Maximization Technique in Statistical Estimation of Probabilistic Functions of Markov Processes," *Inequalities*, 3, 1972

[4] **Dempster, A.P., Laird, N.M., Rubin, D.B., (1977)** "*Maximum likelihood from incomplete data via the EM algorithm*," Proc. R. Stat. Soc. B., 39(1):1-38, 1977

[5] **Gauvain, J.-L., Lee, C.-H., (1992)** "*Improved acoustic modeling with Bayesian learning,*" Proc. ICASSP-92, 1:481-485, 1992.

[6] **Huang, X.D., Jack, M.A., (1989)** "Semi-Continuous Hidden Markov Models for Speech Signals," *Computer Speech and Language (1989)*, 3:239-251

[7] **Juang, B.-H., Rabiner, L.R., (1985a)** "A probabilistic distance measure for hidden Markov models," *AT&T Tech. J.,* 64(2): 391-408, Feb. 1985

[8] **Juang, B.-H., Rabiner, L.R., (1985b)** "Mixture autoregressive hidden Markov Models for speech signals," *IEEE Trans. on Acoust., Speech, and Signal Processing*, vol. ASSP-33, 1985, pp.1404-1413

[9] **Lee, C.-H., Rabiner, L.R., (1989)** "A Frame-Synchronous Network Search Algorithm for Connected Word Recognition," *IEEE Trans. on ASSP*, Vol. 37, No.11, Nov.1989, pp.1649-1658

[10] **Lyu, R.-Y., Hong, I-C., Shen, J.-L., Lee, M.-Y., Lee, L.-S., (1998) "**Isolated Mandarin Base-Syllable Recognition Based upon the Segmental Probability Model**",** *IEEE Trans. On SAP*, Vol. 6, No. 3, May 1998, pp. 293-299

[11] **Rabiner, L.R., Levinson, S.E., Sondhi, M.M., (1983)** "On the application of vector quantization and hidden Markov models to speaker-independent, isolated word recognition," *Bell Syst. Tech. J.*, vol.62, pp.1075-1105

[12] **Rabiner, L.R., Juang, B.-H., Levinson, S.E., Sondhi, M.M., (1985)** "Recognition of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities," *AT&T Technical Journal*, vol. 64, No.6, July-August 1985, pp.1211-1234

[13] **Viterbi, A.J., (1967)** "Error Bounds for Convolutional Codes and An Asymptotically Optimum Decoding Algorithm," *IEEE Trans. on IT-13(2)*, Apr., 1967