A DTW-BASED DISSIMILARITY MEASURE FOR LEFT-TO-RIGHT HIDDEN MARKOV MODELS AND ITS APPLICATION TO WORD CONFUSABILITY ANALYSIS

Qiang HUO and Wei LI

Department of Computer Science, The University of Hong Kong, Pokfulam Road, Hong Kong
(Email: qhuo@cs.hku.hk, pwli@cs.hku.hk)

ABSTRACT

We propose a dynamic time-warping (DTW) based distortion measure for measuring the dissimilarity between pairs of left-to-right continuous density hidden Markov models with state observation densities being mixture of Gaussians. The local distortion score required in DTW is defined as an approximate Kullback-Leibler divergence (KLD) between two Gaussian mixture models (GMMs). Several approximate KLDs are studied and compared for pairs of GMMs with different properties, and one of them is identified for being used in our DTW-based HMM dissimilarity measure. In an experiment of identifying automatically the subsets of confusable Putonghua (Mandarin Chinese) syllables, it is observed that the result based on the proposed HMM dissimilarity measure is highly consistent with the one based on syllable recognition confusion matrix obtained on a testing data set.

Index Terms: speech recognition, dissimilarity measure, hidden Markov model, Kullback-Leibler divergence.

1. INTRODUCTION

After many years research, left-to-right (LR) Gaussian mixture continuous density hidden Markov model (CDHMM) remains predominant as a speech modeling technique in automatic speech recognition (ASR) area. How to measure the dissimilarity of two given CDHMMs without running recognition experiments has been an important research topic for several decades due to its potential applications in a variety of contexts in ASR (e.g., [4, 8, 15, 14, 17, 16, 1, 2]). In a pioneering work [6], Juang and Rabiner proposed to use Kullback-Leibler divergence (KLD) as a dissimilarity measure of two HMMs with arbitrary observation densities, but only experimental results on discrete HMMs (DHMMs) were reported. A similar approach was applied to measuring KLD for pairs of LR-CDHMMs in e.g., [14]. In order to calculate the KL divergence, a Monte Carlo (MC) simulation procedure is used to generate a large number of observation sequences from the HMMs being measured. Alternatively, some training data has to be used for calculating KLD as reported in [8]. It is not practical to use the above procedures in ASR applications that require a quick response and/or have no access to training data. To address this issue, an upper bound of the KLD for ergodic DHMMs and CDHMMs was proposed in [3]. It was extended recently to calculate the upper bound of the KLD [13] and a so-called Average Divergence Distance (ADD) [12] for LR-CDHMMs. Other heuristic approaches to measuring efficiently the dissimilarity between HMMs were also reported for both DHMMs (e.g., [4]) and LR-CDHMMs (e.g., [15, 17, 16, 1, 2]).

Similar to the applications in [11, 15, 17, 1, 2], we've also been working on a problem of how to identify automatically the subsets of confusable words in the vocabulary of a given ASR task without doing recognition experiments on some testing data. By considering the following two facts:

- Most of the current state-of-the-art ASR systems are still based on a so-called "Beads-on-a-String" notion that a word is composed of a sequence of phone segments, and each phone segment (or a basic speech unit) is modeled by an LR-CDHMM;
- The transition probabilities play a less important role in LR-CDHMM in comparison with that of state observation densities;

we propose to represent a specific pronunciation of a word as a sequence of state-dependent Gaussian mixture models (GMMs). Such a representation would be a reasonable approximation. Therefore, the problem of measuring the confusability of two vocabulary words can be cast as a problem of measuring the dissimilarity of two sequences of GMMs with possibly different lengths. Such a problem can be solved efficiently by using a dynamic time-warping (DTW) approach as described in detail in e.g., [10]. The key technical issue then becomes how to measure the dissimilarity of pairs of GMMs that can be used as the local distortion scores required in DTW. The KLD of GMMs offers a good theoretical answer. Because no closed form solution exists for calculating the KLD of GMMs, we have studied several approximate KLDs proposed originally in [5] for pairs of GMMs with different "topography" properties [9]. It is the purpose of this paper to report our study on this topic.

The rest of the paper is organized as follows. In Section 2, we describe our DTW-based procedure for measuring the dissimilarity between pairs of LR-CDHMMs. In Section 3, we compare, via simulation experiments, two approximation methods with the one based on MC simulation for calculating the KLD of GMMs. As a result, one of them is identified for being used in our DTW-based HMM dissimilarity measure. In Section 4, we report experiments on how to use the proposed HMM dissimilarity measure for identifying automatically the subsets of confusable Putonghua syllables. Finally, we conclude the paper in Section 5.

2. A DTW-BASED PROCEDURE FOR MEASURING DISSIMILARITY OF TWO LR-CDHMMS

Consider two LR-CDHMMs, $\{\lambda_i, i = 1, 2\}$, whose states from left-to-right are denoted as $\{q_1^{(i)}, \cdots, q_{N_i}^{(i)}; i = 1, 2\}$, where N_i is

This research was supported by a grant from HKU's Seed Funding Programme for Basic Research. Dr. Donglai Zhu's contribution in training the triphone models used in this study is also gratefully acknowledged.

the number of states for λ_i. For each state $q_s^{(i)}$, its observation density is a GMM denoted as $p(Y|q_s^{(i)}) = \sum_{m=1}^{M_s^{(i)}} c_{sm}^{(i)} \mathcal{N}(Y; \mu_{sm}^{(i)}, \Sigma_{sm}^{(i)})$ where $M_s^{(i)}$ is the number of Gaussian components for the sth state of λ_i; $c_{sm}^{(i)}$'s are nonnegative Gaussian mixture weights with constraint $\sum_{m=1}^{M_s^{(i)}} c_{sm}^{(i)} = 1$; $\mathcal{N}(Y; \mu_{sm}^{(i)}, \Sigma_{sm}^{(i)})$ is a normal distribution with a D-dimensional mean vector $[\mu_{sm1}^{(i)}, \cdots, \mu_{smD}^{(i)}]^{Tr}$ and a $D \times D$ diagonal covariance matrix $\Sigma_{sm}^{(i)} = diag\{(\sigma_{sm1}^{(i)})^2, \cdots, (\sigma_{smD}^{(i)})^2\}$.

As described in [10], by introducing two warping functions

$$j = \phi_1(t) \quad t = 1, \cdots, T \ ,$$
$$k = \phi_2(t) \quad t = 1, \cdots, T \ ,$$

where $j = 1, \cdots, N_1$, $k = 1, \cdots, N_2$, and T is the length of a warping path, we propose to use the following DTW procedure to measure the dissimilarity of two LR-CDHMMs, λ_1 and λ_2:

Step 1: Initialization

$$D_A(1,1) = d(q_1^{(1)}, q_1^{(2)}) \tag{1}$$

where $d(q_1^{(1)}, q_1^{(2)})$ is the dissimilarity measure between the GMM for state $q_1^{(1)}$ and the GMM for state $q_1^{(2)}$.

Step 2: Recursion

For $1 \le j \le N_1$, $1 \le k \le N_2$ such that j and k stay within the allowable grid defined by the following local continuity constraints:

$$\phi_1(t+1) - \phi_1(t) \le 1 \ ,$$
$$\phi_2(t+1) - \phi_2(t) \le 1 \ ;$$

compute

$$D_A(j,k) = \min \left\{ \begin{array}{c} D_A(j-1,k) + d(q_j^{(1)}, q_k^{(2)}), \\ D_A(j-1,k-1) + d(q_j^{(1)}, q_k^{(2)}), \\ D_A(j,k-1) + d(q_j^{(1)}, q_k^{(2)}) \end{array} \right\}; \tag{2}$$

where $d(q_j^{(1)}, q_k^{(2)})$ is the dissimilarity measure between the GMM for state $q_j^{(1)}$ and the GMM for state $q_k^{(2)}$. Consequently, only values of (j,k) that can be reached from $(1,1)$ and can end ultimately at (N_1, N_2) are evaluated in the above recursion step.

Step 3: Termination

$$D_{HMM}(\lambda_1, \lambda_2) = \frac{D_A(N_1, N_2)}{T_o} \ , \tag{3}$$

where T_o is the length of the optimal warping path identified at the end of the dynamic programming recursion. $D_{HMM}(\lambda_1, \lambda_2)$ calculated as the above is defined to be the dissimilarity score between λ_1 and λ_2.

In the above procedure, a key technical issue is how to define the local dissimilarity score $d(q_j^{(1)}, q_k^{(2)})$. The following KLD of GMMs offers a good theoretical answer since the KLD is the average discrimination information per observation between two hypotheses modeled as random variables:

$$KL(p(Y|q_j^{(1)}), p(Y|q_k^{(2)})) = \int p(Y|q_j^{(1)}) \log \frac{p(Y|q_j^{(1)})}{p(Y|q_k^{(2)})} dY. \tag{4}$$

Unfortunately, there is no closed form solution for the above integration. Therefore some approximations are required. Among many choices, the methods proposed in [5] attracted our special attention. We therefore conducted a comparative study via simulation experiments with a hope to identify the most appropriate one for being used in the above DTW procedure. In the following section, we report the result of our study on this sub-topic.

3. APPROXIMATE KLDS OF TWO GMMS

To simply the notation, let's consider two GMMs, whose parameters are denoted as $\theta_i = \{c_m^{(i)}, \mu_m^{(i)}, \Sigma_m^{(i)}\}$ for $i = 1, 2$ respectively. The number of Gaussian components are denoted as $M^{(1)}$ and $M^{(2)}$ accordingly. The first method we studied is to use Monte-Carlo (MC) simulations to approximate the KLD of two GMMs as follows:

$$KL(p(Y|\theta_1), p(Y|\theta_2)) \approx \frac{1}{N_{sim}} \sum_{t=1}^{N_{sim}} \log \frac{p(Y_t|\theta_1)}{p(Y_t|\theta_2)} \ , \tag{5}$$

where Y_t are sampled from $p(Y_t|\theta_1)$ and N_{sim} is the number of samples simulated. In our experiments, $N_{sim} = 1000$.

The second method we studied is a so-called *matching based approximation* method proposed in [5]. The approximate KLD of two GMMs is calculated as follows:

$$KL(p(Y|\theta_1), p(Y|\theta_2)) \approx \sum_{m=1}^{M^{(1)}} c_m^{(1)} [\log \frac{c_m^{(1)}}{c_{\pi(m)}^{(2)}} + KL(\mathcal{N}(Y; \mu_m^{(1)}, \Sigma_m^{(1)}), \mathcal{N}(Y; \mu_{\pi(m)}^{(2)}, \Sigma_{\pi(m)}^{(2)}))] \ , \tag{6}$$

with the following matching function $\pi(m)$:

$$\pi(m) = \arg\min_n \{-\log c_n^{(2)} + KL(\mathcal{N}(Y; \mu_m^{(1)}, \Sigma_m^{(1)}), \mathcal{N}(Y; \mu_n^{(2)}, \Sigma_n^{(2)}))\}, \tag{7}$$

where $KL(\mathcal{N}(Y; \mu_m^{(1)}, \Sigma_m^{(1)}), \mathcal{N}(Y; \mu_n^{(2)}, \Sigma_n^{(2)}))$ is the KLD of two Gaussians [5].

The third method we studied is a so-called *unscented transformation (UT) based approximation* method proposed also in [5]. The approximate KLD of two GMMs is calculated as follows:

$$KL(p(Y|\theta_1), p(Y|\theta_2)) \approx \sum_{m=1}^{M^{(1)}} \frac{c_m^{(1)}}{2D} \sum_{t=1}^{2D} \log \frac{p(Y_t|\theta_1)}{p(Y_t|\theta_2)} \ , \tag{8}$$

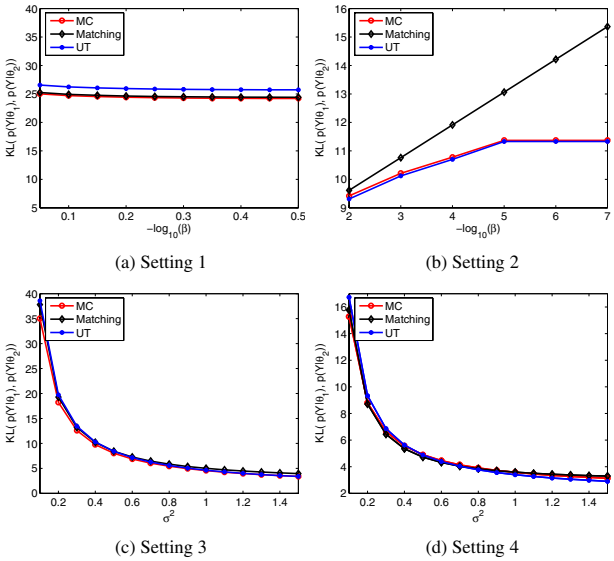
where Y_t's are $2D$ "sigma points" selected according to $p(Y|\theta_1)$ as described in [5]. Other choices for generating "sigma points" are also provided in [7], but we have not tried them out yet.

To compare the above three methods, experiments are conducted for four pairs of GMMs with settings of parameters shown in Table 1, where $D = 2$, each GMM has two Gaussian components, MN_1 and MN_2 are the number of modes in two GMMs respectively. As demonstrated in e.g., [9], the number of modes is not necessarily the same as the number of components in GMM. The above experimental design was inspired by the work in [9] such that each pair of GMMs has different "topography" properties. Therefore an interesting comparison can be made, as shown in Fig. 1, to understand the behaviors of the above three approximate KLD methods under the above different conditions.

Table 1. Settings of parameters of four pairs of GMMs.

Setting No.	GMMs	$\left(c_m^{(i)}, \mu_m^{(i)}, \Sigma_m^{(i)}\right)$	Remarks
1	$i=1$ $m=1$	$\left(0.5, \begin{bmatrix}3\\3\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	$MN_1 = 2$
	$i=1$ $m=2$	$\left(0.5, \begin{bmatrix}-2\\-2\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	
	$i=2$ $m=1$	$\left(\beta, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1&0\\0&0.07\end{bmatrix}\right)$	$MN_2 = 1,2,3$ for $\beta \in [0.05, 0.5]$
	$i=2$ $m=2$	$\left(1-\beta, \begin{bmatrix}1\\1\end{bmatrix}, \begin{bmatrix}0.07&0\\0&1\end{bmatrix}\right)$	
2	$i=1$ $m=1$	$\left(0.5, \begin{bmatrix}-2\\4\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	$MN_1 = 2$
	$i=1$ $m=2$	$\left(0.5, \begin{bmatrix}2\\4\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	
	$i=2$ $m=1$	$\left(\beta, \begin{bmatrix}2\\0\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	Simulate $\beta \dashrightarrow 0$
	$i=2$ $m=2$	$\left(1-\beta, \begin{bmatrix}-2\\0\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	
3	$i=1$ $m=1$	$\left(0.5, \begin{bmatrix}-3\\-3\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	$MN_1 = 2$
	$i=1$ $m=2$	$\left(0.5, \begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	
	$i=2$ $m=1$	$\left(0.5, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2&0\\0&\sigma^2\end{bmatrix}\right)$	$MN_2 = 3,2,1$ for $\sigma^2 \in [0.1, 1.5]$
	$i=2$ $m=2$	$0.5, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}\sigma^2&0\\0&2\end{bmatrix}\right)$	
4	$i=1$ $m=1$	$\left(0.5, \begin{bmatrix}2\\6\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	$MN_1 = 2$
	$i=1$ $m=2$	$\left(0.5, \begin{bmatrix}4\\4\end{bmatrix}, \begin{bmatrix}1&0\\0&1\end{bmatrix}\right)$	
	$i=2$ $m=1$	$\left(0.5, \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}2&0\\0&\sigma^2\end{bmatrix}\right)$	$MN_2 = 3,2,1$ for $\sigma^2 \in [0.1, 1.5]$
	$i=2$ $m=2$	$0.5, \begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}\sigma^2&0\\0&2\end{bmatrix}\right)$	

In *Setting 1* as listed in Table 1, it is noted that the number of modes in the second GMM changes from 1 to 3 when the mixture weight varies from 0.05 to 0.5. In *Settings 3 and 4*, the number of modes in the second GMM changes from 3 to 1 when the variance σ^2 varies from 0.1 to 1.5. The difference between *Settings 3 and 4* is that the matching results are different when the matching based approach is used. In *Setting 3*, each component in GMM θ_1 is matched to a different component in GMM θ_2. However in *Setting 4*, both components in GMM θ_1 are matched to the same component in GMM θ_2. From Fig. 1, we make the following observations: 1) In *Settings 3 and 4*, both matching-based and UT-based KLD measures are a good approximation to the KLD calculated by using the MC method; 2) In *Setting 1*, all the three approximate KLDs follow the same trend when the mixture weight β changes; 3) In *Setting 2*, when the mixture weight is close to 0 or 1, the UT-based method can give a more accurate approximation to the KLD than the matching based method. As for the computational complexity, among three methods, the matching-based method is the most efficient, followed by the UT-based method. The MC method is the most expensive one. According to the above simulation results, we decided to use the matching-based KLD to serve as the local distortion score $d(q_j^{(1)}, q_k^{(2)})$ required in our DTW-based procedure described in Section 2.



(a) Setting 1 (b) Setting 2

(c) Setting 3 (d) Setting 4

Fig. 1. A comparison of three approximate KLD measures of GMMs under four experimental settings.

4. CONFUSABILITY ANALYSIS OF PUTONGHUA SYLLABLES USING HMM DISSIMILARITY MEASURE

In order to verify the effectiveness of the proposed HMM dissimilarity measure for identifying automatically the subsets of confusable words in the vocabulary of a given ASR task, we take 410 Putonghua base syllables disregarding tones as our vocabulary. The basic speech units are triphones considering both the within-syllable and cross-syllable contextual dependencies. The context-independent (CI) phone set consists of 36 phones plus silence. Each triphone is modeled by a three-emitting-state LR-CDHMM without state skipping. Each state has 8 Gaussian mixture components with each component having a diagonal covariance matrix. A special three-state CDHMM is also used for silence modeling. The 39-dimensional feature vector used in this study consists of 12 MFCC's and log-scaled energy normalized by the peak of the individual sentence, plus their first and second order derivatives. Sentence-based cepstral mean subtraction is applied for acoustic normalization both in training and testing. A speaker independent, decision-tree-based tied-state HMM system with 3001 tied states was trained by using HTK3.0 toolkit. The training data consists of 153796 sentences (about 180 hours) from 216 male and 114 female speakers extracted from four Putonghua corpora, namely HKU96, HKU99, 863 and MSRA.

Using the above trained models, we conduct an isolated syllable recognition experiment by using 33021 utterances from 10 male and 10 female speakers extracted from isolated syllable part of HKU96 Putonghua corpus. An averaged syllable recognition accuracy of 60.9% is achieved. At the same time, a syllable confusion matrix SCM_R is also obtained, with the (i,j)th element $SCM_R[i,j]$ being the percentage of the ith syllable mis-recognized as the jth syllable. From SCM_R, a symmetric syllable dissimilar-

ity matrix SIM_R can be defined with the (i, j)th element calculated as

$$SIM_R[i, j] = -\log[\max(SCM_R(i, j), SCM_R(j, i))]. \quad (9)$$

Using the DTW-based HMM dissimilarity measure, a model based symmetric syllable dissimilarity matrix SIM_H can be obtained with the (i, j)th element calculated as

$$SIM_H[i, j] = (D_{HMM}(\lambda_i, \lambda_j) + D_{HMM}(\lambda_j, \lambda_i))/2, \quad (10)$$

where λ_i, λ_j are the LR-CDHMMs for the ith and jth syllables respectively.

The following agglomerative hierarchical clustering procedure is then used for identifying the subsets of confusable syllables:

1. **begin: initialize** $threshold$, $c \leftarrow n$, $D_i \leftarrow \{syl_i\}$, where syl_i is the ith syllable; $i = 1, \cdots, n$; and $n = 410$ is the total number of syllables in the vocabulary.
2. **do:** $c \leftarrow c - 1$; find the nearest clusters, say D_i and D_j, with the smallest dissimilarity $d(D_i, D_j)$; merge D_i and D_j.
3. **until:** all $d(D_i, D_j) > threshold$; $i < j$; $i, j = 1, \cdots, c$.
4. **return:** c clusters.

In the above procedure, c is the number of clusters. The dissimilarity between a pair of syllable clusters is calculated as

$$d(D_i, D_j) = \min_{\{syl^i \in D_i, syl^j \in D_j\}} SIM(syl^i, syl^j), \quad (11)$$

where $SIM(syl^i, syl^j)$ is the dissimilarity between the syllables syl^i and syl^j, which takes the value of the corresponding element in SIM_R or SIM_H respectively.

Under the settings of $threshold = 1$ for SIM_R-based clustering and $threshold = 10$ for SIM_H-based clustering, the final numbers of clusters are 108 and 90 respectively. For each cluster $G_R(i)$ obtained by the SIM_R-based clustering, a corresponding cluster $G_H(j)$ obtained by the SIM_H-based clustering is selected with the largest measure, $CP(G_R(i), G_H(j))$, defined as follows:

$$\frac{syllable\ number\ in\ both\ G_R(i)\ and\ G_H(j)}{syllable\ number\ in\ G_R(i)}. \quad (12)$$

Some examples are illustrated in Table 2. The agreement between pairs of clusters is very high. Among 108 pairs of clusters, 27 pairs have $CP = 1$ and 48 pairs have a CP greater than 0.5.

5. CONCLUSION

In this paper, a new DTW-based distortion measure is proposed for measuring the dissimilarity between pairs of LR-CDHMMs. Its effectiveness has been confirmed in the above experiments. We've also been using this new measure in other ASR applications. We will report those studies elsewhere.

6. REFERENCES

[1] J. Anguita, S. Peillon, J. Hernando, and A. Bramoulle, "Word confusability prediction in automatic speech recognition," *Proc. ICSLP-2004*, Jeju Island, Korea, 2004, pp.1489-1492.

[2] G. Bouwman, B. Cranen and L. Boves, "Predicting word correct rate from acoustic and linguistic confusability," *Proc. ICSLP-2004*, 2004, pp.1481-1484.

Table 2. An illustration of several pairs of confusable syllable subsets identified by using information derived from syllable recognition confusion matrix and the proposed dissimilarity measure of syllable LR-CDHMMs.

Recognition Confusion Matrix	Dissimilarity Measures
jian jie	jian jie
liu miu niu yo you	liu miu niu yo you
bi bin bing di ding	bi bin bing di
duan dui dun zhui	dui dun
cuo huo kuo po tuo	chuo cuo huo po tuo
chu cu hu pu tu	chu cu tong tu
bei dei fei gei	bei dei ei gei
gua guai guan guang guo	gua guang guo wa wang
lin ling mi min ming	ding li lin ling
ai an ang ao en	a an ang er
lian lie mian nian nie	lian lie lue nue yue

[3] M. N. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Processing Letters*, Vol. 10, No. 4, pp.115-118, 2003.

[4] M. Falkhausen, H. Reininger and D. Wolf, "Calculation of distance measures between hidden Markov models," *Proc. Eurospeech-1995*, Madrid, 1995, pp.1487-1490.

[5] J. Goldberger, S. Gordon and H. Greenspan, "An efficient image similarity measure based on approximations of KL-Divergence between two Gaussian mixtures," *Proc. ICCV-2003*, Nice, 2003, pp.487-493.

[6] B.-H. Juang and L. R. Rabiner, "A probabilistic distance measure for hidden Markov models," *AT&T Technical Journal*, Vol. 64, No. 2, pp.391-408, 1985.

[7] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, Vol. 92, No. 3, pp.401-422, 2004.

[8] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," *Proc. ICSLP-1996*, Philadelphia, 1996, pp.2195-2198.

[9] S. Ray and B. G. Lindsay, "The topography of multivariate normal mixtures," *The Annals of Statistics*, Vol. 33, No. 5, pp.2042-2065, 2005.

[10] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[11] D. B. Roe and M. D. Riley, "Prediction of word confusabilities for speech recognition," *Proc. ICSLP-1994*, Yokohama, Japan, 1994, pp.227-230.

[12] J. Silva and S. Narayanan, "A statistical discrimination measure for hidden Markov models based on divergence," *Proc. ICSLP-2004*, Jeju Island, Korea, 2004, pp.657-660.

[13] J. Silva and S. Narayanan, "An upper bound for the Kullback-Leibler divergence for left-to-right transient hidden Markov models," submitted to *IEEE Trans. on Information Theory*, June 2005

[14] R. Singh, B. Raj, and R. M. Stern, "Structured redefinition of sound units by merging and splitting for improved speech recognition," *Proc. ICSLP-2000*, Beijing China, 2000, pp.151-154.

[15] B.-T. Tan, Y. Gu and T. Thomas, "Word confusability measures for vocabulary selection in speech recognition," *Proc. ASRU-1999*, Keystone, 1999, pp.185-188.

[16] M.-Y. Tsai and L.-S. Lee, "Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on Mandarin Chinese," *Proc. ASRU-2003*, Virgin Islands, 2003, pp.117-122.

[17] M. Vihola, M. Harju, P. Salmela, J. Suontausta and J. Savela, "Two dissimilarity measures for HMMs and their application in phoneme model clustering," *Proc. ICASSP-2002*, Orlando, 2002, pp.933-936.