# A STATISTICAL ACOUSTIC CONFUSABILITY METRIC BETWEEN HIDDEN MARKOV MODELS

*Hong You and Abeer Alwan*

Department of Electrical Engineering
University of California, Los Angeles, CA 90095
Email: hyou@icsl.ucla.edu, alwan@ee.ucla.edu

## ABSTRACT

With the wide application of Hidden Markov Models (HMMs) in speech recognition, a statistical acoustic confusability metric is of increasing importance to many components of a speech recognition system. Although distance metrics between HMMs have been studied in the past, they didn't include a way of accounting for speaking rate and durational variations. In order to account for the underlying speech signal's properties when computing such a metric between HMMs, we propose a dynamically-aligned Kullback Leibler (KL) divergence measurement and discuss a cost-efficient implementation of the metric. The proposed approach outperforms existing metrics in predicting phonemic confusions.

***Index Terms***— Hidden Markov Models, Speech recognition, Statistical acoustic confusability metric

## 1. INTRODUCTION

With the successful application of HMMs in speech recognition, it is increasingly important to be able to predict speech recognition performance without intensive tests which involve large amounts of testing data in various scenarios. Since automatic speech recognition (ASR) performance in HMM-based systems is largely determined by acoustic confusability between HMMs, a statistical acoustic confusability measurement is therefore desired. In addition, such a metric provides discriminative evaluation information of the system which is helpful for the development of discriminative training algorithms. The encapsulation of acoustic modeling confusability information provides compact and valuable input to improve state-of-the-art pronunciation modeling methods [1].

There are a few studies on developing HMM distance metrics [2, 3, 4, 5, 6]. In early studies, the stationary aspect of HMMs, i.e. the state dependent observation distribution, is taken into consideration. Various distance metrics, such as the Mahalanobis [5] and Euclidean distances [4], have been experimented with. Recently, both stationary and transient aspects of HMMs have been considered in HMM distance measurement by extending metrics between probability density functions to distance metrics between dynamic probability density functions [2, 3].

The dynamic characteristics of speech signals warrant that we normalize them before performing pattern comparison. Factors that require normalization include speaking rate and durational variations. Since a statistical acoustic confusability metric between HMMs, for example the KL divergence, computes the "distance" between two dynamic probabilistic distributions modeling speech signals, it is important that the impact of the speech signal's dynamic characteristics on the metric is minimized. Moreover, it is also necessary for an acoustic confusability metric to be computed based on comparing entire speech units instead of a partial comparison. In order to meet these challenges, state alignment is necessary. We propose a statistical acoustic confusability metric based on the KL divergence between HMMs as shown in Figure 1.

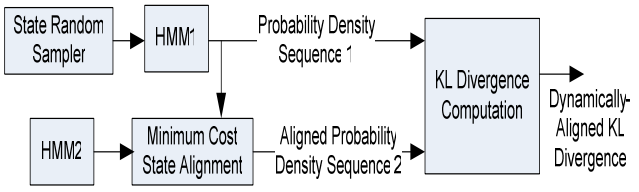


**Fig. 1**. Dynamically-Aligned KL Divergence between HMMs

The rest of the paper is organized as follows. We introduce a dynamically-aligned KL divergence metric in Section 2, and the implementation in Section 3. In Section 4, a comparative experiment is performed showing that the dynamically-aligned KL divergence outperforms a state-of-the-art HMM confusability metric. Section 5 provides a summary.

## 2. ACOUSTIC CONFUSABILITY METRIC

The computation of the KL divergence between HMMs, $\lambda^1$ and $\lambda^2$, for fixed state sequences, $S^1$ and $S^2$, is derived as in

[2].

$$D(\lambda^1, S^1||\lambda^2, S^2) = \sum_{i=1}^{N} D(\lambda^1, s_i^1||\lambda^2, s_i^2)$$
$$= \sum_{i=1}^{N} \int_{o_i} P(o_i|s_i^1) log \frac{P(o_i|s_i^1)}{P(o_i|s_i^2)} do_i$$

$P(o_i|s_i^1)$ and $P(o_i|s_i^2)$ are state-dependent observation distributions of HMMs $\lambda^1$ and $\lambda^2$ at states $s_i^1$ and $s_i^2$, respectively, assuming that the state sequences $S^1$ and $S^2$ are known and of the same length.

### 2.1. HMM State Alignment Constraints

We introduce nonlinear state alignment in order to account for variations observed in speech signals, and define our statistical acoustic confusability metric based on it. A dynamically-aligned KL divergence metric between HMMs is defined as:

$$D^a(\lambda^1, \lambda^2) = E_{S^1,S^2}(D^a(\lambda^1, S^1||\lambda^2, S^2))$$
$$D^a(\lambda^1, S^1||\lambda^2, S^2) = D(\lambda^1, S^1||\lambda^2, \phi(S^2||S^1))$$

where $\phi(S^2||S^1)$ is an aligned state sequence of $S^2$ given $S^1$ and $E_{S^1,S^2}$ is the expectation over random variables $S^1$ and $S^2$. For the HMM state alignment to be meaningful in terms of normalizing duration and speaking rate variation, these state alignment constraints are necessary:

1. ***End-pointing non-emitting state alignment constraint:*** This eliminates the possibility of aligning a emitting state with an non-emitting state, therefore guarantees that only global alignment paths are valid.

2. ***State transition monotonicity constraint/state transition probability constraint:*** This constraint implies that state alignment needs to satisfy the state transition topology of an HMM.

### 2.2. Dynamically-Aligned KL Divergence Between HMMs

The choice of a state-alignment function reflects the level of linguistic knowledge we want to include in the acoustic confusability metric. However, using linguistic knowledge to quantitatively compare different phonemic units is an unsolved problem. Hence, a minimum cost criterion has been a natural choice for many speech pattern comparison tasks. In our case, we choose the alignment function, $\phi(S^2||S^1)$, that minimizes the KL divergence between HMMs for a given state sequence, $S^1$.

$$D^{min-a}(\lambda^1, S^1||\lambda^2, S^2) = min_\phi D(\lambda^1, S^1||\lambda^2, \phi(S^2||S^1))$$

Using this alignment scheme, the dynamically-aligned KL divergence between HMMs $\lambda^1$ and $\lambda^2$ is defined as

$$D^{min-a}(\lambda^1, \lambda^2) = E_{S^1,S^2}(min_\phi D(\lambda^1, S^1||\lambda^2, \phi(S^2||S^1)))$$

In the rest of the paper, the dynamically-aligned KL divergence between HMMs with a minimum cost alignment scheme, i.e. $D^{min-a}(\lambda^1, \lambda^2)$, is used.

The minimum cost alignment function guarantees that the dynamically-aligned KL divergence between identical HMM models is always zero. Furthermore, for HMMs with almost identical observation distributions sequentially but different state transitions, the dynamically-aligned KL divergence is very close to zero, indicating that the acoustic confusability between the models is high after normalizing variations in dynamic characteristics. Finally, when HMMs $\lambda^1$ and $\lambda^2$ are models of different phonemic units, the minimum cost state alignment function implies that the valid dynamic probabilistic distribution space, $\phi(S^2||S^1)$, is searched and the minimum KL distance is selected.

Figure 2 illustrates the concept. $D^{min-a}(\lambda^1, \lambda^2)$ is the minimum distance from $(\lambda^1, S^1)$ to the set $(\lambda^2, \phi(S^2||S^1))$. As shown, the dynamic probability distribution $(\lambda^2, S^2)$ may not be a valid dynamic probability distribution for searching the minimum distance, i.e. $((\lambda^2, S^2) \notin (\lambda^2, \phi(S^2||S^1))$ can happen.

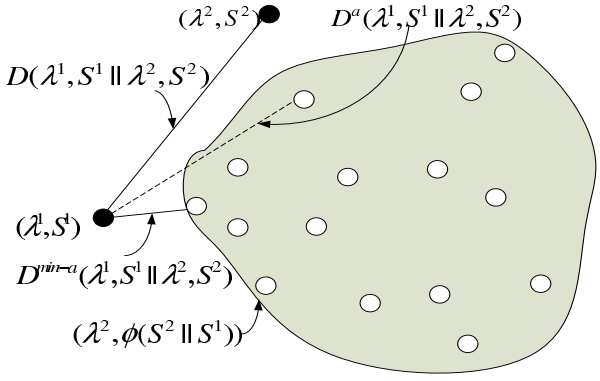


**Fig. 2**. Dynamically-aligned KL divergence illustration

It is easy to show that the minimum cost state-alignment function $min_\phi D(\lambda^1, S^1||\lambda^2, \phi(S^2||S^1))$ is independent of $S^2$. Therefore, the definition of dynamically-aligned KL divergence can be further simplified as

$$D^{min-a}(\lambda^1, \lambda^2) = E_{S^1}(min_X D(\lambda^1, S^1||\lambda^2, X))$$

where $min_X D(\lambda^1, S^1||\lambda^2, X)$ is the minimum KL distance between the dynamic probabilistic distribution generated by HMM $\lambda^1$ with $S^1$ as state sequence and the dynamic probabilistic distribution set that can be generated with HMM $\lambda^2$ using state sequence set $X$.

The mapping from $S^2$ to the state sequence that achieves the minimum KL distance is a nonlinear state-alignment function which minimizes the KL divergences between HMMs given current state sequences and hence minimizes the effects of dynamic behavior variations on the metric.

$D^{min-a}(\lambda^1, \lambda^2)$ is the average minimum KL distance between the models, which is averaged over the state transition automaton of HMM $\lambda^1$. Therefore, the dynamically-aligned KL divergence, $D^{min-a}(\lambda^1, \lambda^2)$, is an asymmetric divergence metric.

Compared with the average divergence distance metric (ADD) [2], $D^{min-a}(\lambda^1, \lambda^2)$ applies minimum cost nonlinear state alignment concept to normalize the speech signal's variations due to dynamic characteristics. Moreover, we enforce two state-alignment constraints so that speech unit distance computation is based on comparing the whole unit. This eliminates the possibility of partial comparison which may occur when computing the ADD metric. A speech unit could be a phoneme, triphone, word, etc.

## 3. IMPLEMENTATION

Assume HMM $\lambda^1$ has $M$ states, HMM $\lambda^2$ has $N$ states, the KL divergence between non-emitting and emitting states is infinite, and the KL divergence between two non-emitting states is zero. With these assumptions, dynamic programming is applied to implement the search problem in two steps.

***Initialization step***: Compute $D(\lambda^1, s_i^1 || \lambda^2, s_j^2)$ for $s_i^1 = 1, ..., M$ and $s_j^2 = 1, ..., N$. The KL divergence between Gaussian mixture models can be analytically approximated using the method proposed in [7]. Represented in vector matrix form, we define the state divergence matrix $V$, an $M$x$N$ matrix, as $V(s_i^1, s_j^2) = D(\lambda^1, s_i^1 || \lambda^2, s_j^2)$.

***Search step***: Assume $S^1 = [s_1^1, ..., s_L^1]$. We use $N_c(i, s_j^2)$ to represent the DP node KL divergence cost at position $(i, s_j^2)$, and use $P_c(i, s_j^2)$ to denote the DP path KL divergence cost from initial position to position $(i, s_j^2)$. The DP update of node cost and path cost is

$N_c(i, s_j^2) = D(\lambda_1, s_i^1 || \lambda_2, s_j^2) = V(s_i^1, s_j^2)$

For emitting node $(i, s_j^2)$

$P_c(i, s_j^2) = N_c(i, s_j^2) + min\{P_c(i-1, s_j^2), P_c(i-1, s_{j-1}^2)\}$

For non-emitting node $(i, s_j^2)$

$P_c(i, s_j^2) = N_c(i, s_j^2) + P_c(i-1, s_{j-1}^2)$

Therefore,

$$min_X D(\lambda^1, S^1 || \lambda^2, X) = P_c(L, N)$$

Finally, $D^{min-a}(\lambda^1, \lambda^2)$ is numerically approximated by averaging over a sample set which is generated by a Monte-Carlo simulation using a sample set of 200.

## 4. EXPERIMENTS

The dynamically-aligned KL divergence between HMMs is experimented with using monophone HMMs trained and tested with the TIMIT database. Front-end speech features are Mel frequency cepstral coefficients and their first and second derivatives. 40 English monophones are trained using the training set of TIMIT. A left-to-right HMM topology is employed to model monophones with 2 non-emission states and 3 emission states for each model. There are three Gaussian mixtures in each state observation distribution. All the Gaussian mixtures have diagonal covariance matrices. Monophone recognition experiment is performed using the testing set of TIMIT, which generates the ASR monophone confusion matrix used in the following comparison.

We compute the dynamically-aligned KL divergences between monophone HMM pairs as shown by the gray scale image in Figure 3, and used them to test the metric's ability to predict phoneme confusion pattern and confusion likelihood. As a comparison baseline, we compute a state-of-the-art statistical acoustic confusability metric, the ADD metric[2], between HMMs in the same manner.

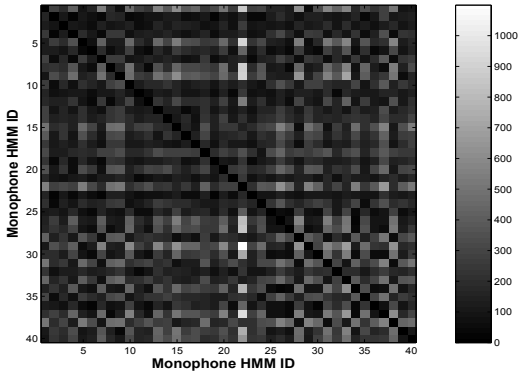


**Fig. 3**. Dynamically-aligned KL divergence measurements between monophone HMM pairs in gray scale

Cross correlations between the metric confusability values and the actual confusion matrix values are computed. The average cross correlation is computed by cross correlating each row of the confusion matrix with each row of a confusability measurement, and averaging over the row-wise cross correlation results. A smaller HMM-based distance metric, such as the dynamically-aligned KL divergence or the ADD measurement, corresponds to more acoustic confusability between two speech units. In these situations, it is more likely to observe a higher confusion likelihood for the corresponding entry in the confusion matrix. Therefore, the normalized cross-correlation is usually negative, and a value closer to $-1$ indicates a better prediction ability of the HMM-based acoustic distance metric.

In addition, a simple segmental linear function is used, shown by Figure 4, to first process the computed acoustic confusability measurement. The motivation of such a function is to adjust a two-class comparison measurement to predict a multi-class classification task. For the segmental linear function used, the parameter $C$ is the size of the most confusable

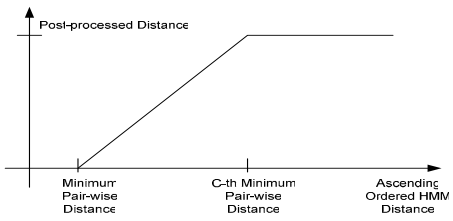


**Fig. 4**. Processing function

set. When applying any acoustic confusability metric to related ASR areas, such as pronunciation modeling, a processing function as the one used here is also necessary.

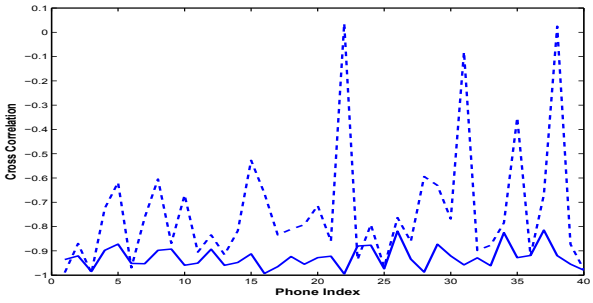


**Fig. 5**. Cross-correlation values between a HMM distance metric and the ASR phone recognition confusion pattern with processing function $C = 5$. Dotted line: ADD metric. Solid line: Dynamically-aligned KL divergence metric.

**Table 1**. *Average cross-correlation values between HMM divergence metrics and phoneme confusion scores*

| Cross-correlation | C=40 | $C = 10$ | $C = 5$ | $C = 2$ |
|---|---|---|---|---|
| ADD | -0.2833 | -0.6559 | -0.7357 | -0.7641 |
| Proposed metric | -0.3848 | -0.8122 | -0.9266 | -0.9833 |

In Figure 5, cross-correlations between the processed confusability metric and phone confusion pattern for each phoneme are shown. Peaks in the cross correlation curve correspond to less accurate predictions of phoneme confusions. It is interesting to note that the top 4 peaks of the ADD metric correspond to the vowels oy, aw, aa, and ey. Three of these are dipthongs with dynamic characteristics which may have resulted in the higher values. For the majority of phonemes, we observe improved phone confusion pattern prediction of the proposed metric over the ADD metric. In addition, we show in Table 1 the average cross-correlation coefficients obtained by the dynamically-aligned KL divergence and the ADD metric. Improvement in predicting the confusion pattern and confusion likelihood using the dynamically-aligned KL divergence measurement is observed, which tends to monotonically increase as the parameter of the processing function decreases.

## 5. SUMMARY

We propose a dynamically-aligned KL divergence metric between HMMs. Dynamic state alignment is necessary for computing HMM distance measurement due to the dynamic characteristics of speech signals. When compared with the ADD metric[2], the dynamically-aligned KL divergence can indicate the confusability between HMMs in a more accurate way. Key differences between the proposed metric and the ADD metric[2] are: First, the speech signal's duration and speaking rate variations are minimized with a nonlinear state alignment scheme. Second, a minimum cost state alignment is applied to reduce the computational cost from $O(N^2)$ to $O(N)$, where $N$ is the size of the sample set. Finally, our experiment implies that using a processing function for the divergence measurement is necessary for it to be successfully applied in several applications, such as pronunciation modeling.

## 6. ACKNOWLEDGEMENTS

This work is supported in part by the NSF and by a Fellowship from the Radcliffe Institute of Advanced study to Abeer Alwan. We thank Jorge Silva for insightful discussions.

## 7. REFERENCES

[1] E. Fosler-Lussier, I. Amdal, and H. J Kuo, "A framework for predicting speech recognition errors," *Speech Communication*, vol. 46, pp. 153–170, 2005.

[2] J. Silva and S. Narayanan, "Average divergence distance as a statistical discrimination measure for Hidden Markov Models," *IEEE Transaction on Speech and Audio Processing*, vol. 14, pp. 890–906, May 2006.

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

[4] M. Falkhausen, H. Reininger, and D. Wolf, "Calculation of distance measures between Hidden Markov Models," *Eurospeech*, pp. 1487–1490, 1995.

[5] M.-Y. Tsai and L.-S. Lee, "Pronunciation variations based on acoustic phonemic distance measures with applications examples of Mandarin Chinese," *ASRU*, pp. 117–121, 2003.

[6] M. Vihola, M. Harju, P. Salmela, J. Suontausta, and J. Savela, "Two dissimilarity measures for HMM and their application in phoneme model clustering," *ICASSP*, pp. 933–936, 2002.

[7] M. N. Do, "Fast approximation of Kullback Leibler distance for dependence trees and Hidden Markov Models," *IEEE Signal Processing Letters*, vol. 10, pp. 115–118, Apr. 2003.