

New Model-Based HMM Distances with Applications to Run-Time ASR Error Estimation and Model Tuning

Chao-Shih Huang¹, Chin-Hui Lee² and Hsiao-Chuan Wang³

¹Advanced Research, Product Value Lab., Acer Inc., Tao-Yuan, Taiwan

²School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA

³Department of Electrical Engineering, National Tsing Hua University, Hsin-Chu, Taiwan
joseph_cs_huang@acer.com.tw, chl@ece.gatech.edu, hcwang@ee.nthu.edu.tw

Abstract

We propose a novel model-based HMM distance computation framework to estimate run-time recognition errors and adapt recognition parameters without the need of using any testing or adaptation data. The key idea is to use HMM distances between competing models to measure the confusability between phones in speech recognition. Starting with a set of simulated models in a given noise condition, the corresponding error rate could be estimated with a smooth approximation of the error count computed from the set of phone distances without using any testing data. By minimizing the estimated error between the desired and simulated models, the target model parameters could also be adjusted without using any adaptation data. Experimental results show that the word errors, estimated with the proposed framework, closely resemble the errors obtained by running actual recognition experiments on a large testing set in a number of adverse conditions. The adapted models also gave better recognition performances than those obtained with environment-matched models, especially in low signal-to-noise conditions.

1. Introduction

The most successful modeling approach to acoustic modeling for automatic speech recognition (ASR) is the use of hidden Markov model (HMM) to simultaneously characterize both the temporal and spectral variation of the speech signal [1]. Two families of HMM training algorithms have been adopted extensively. The first attempts to improve modeling accuracy in estimating the parameters of the signal distributions using the maximum likelihood (ML) [1] and maximum *a posteriori* (MAP) [2] objectives. The second tries to increase model discrimination power and enhance separation between competing models based on the minimum classification error (MCE) [3], maximum mutual information (MMI) [4] and minimum discrimination information (MDI) [5] training criteria. An in-depth review on recent advances in acoustic modeling can be found in [6]. The MCE approach, which is the key to formulating the techniques used in this study, is extensively surveyed in [7].

Since distances between competing models are good indicators of the quality of the set of trained models used in ASR, many approaches to computing HMM distances have been proposed. Although divergence-based distances, such as the well-known Kullback-Leibler information [8], could serve as a theoretical foundation for defining HMM distances, there are usually no closed-form solutions available when comparing two HMM's. The missing data nature with

multiple states and many mixture components and the time-varying property with the Markov assumption used in HMM characterization make it very difficult, if not impossible, to evaluate model-based distances even in very simple cases. Because of the generative nature with HMM, one way to approximate the HMM distance is to use Monte Carlo techniques to generate a collection of sample data and used them to compute sample-based likelihood differences and average them [9]. Other techniques have also been proposed (e.g. [10], [11]). Once HMM distances could be evaluated, many applications would follow. For example, distances between phone models have been used to predict word confusability [11] in defining speech recognition vocabularies.

In this paper, we propose a family of model-based HMM distances to compare a single HMM with a collection of competing HMM's. These new distances are derived from the sample-based misclassification measures often used in MCE training with the extension that the samples needed to evaluate the distance are replaced by an HMM. They are referred to as model-based discriminative HMM distances in this study. We show how these distances could be approximated with only HMM parameters without using any sample data. This nice property allows us to estimate run-time recognition error rates [12] dynamically without the need of using any testing data that are sometimes difficult to collect. The approximated error rates could be expressed in closed-form functions of a set of target HMM parameters and offer a mechanism to optimize these parameters in order to reduce the mismatches between HMM's trained in one acoustic condition and the desired HMM's needed in another adverse condition, without using any adaptation data.

We report on experimental results obtained with Japanese phone and word recognition in both clean and noisy conditions. Using only the HMM models learned from a training set of 12,000 phonetically balanced utterances, the proposed HMM distances estimate error rates that closely resemble to the error rates obtained by running actual recognition experiments. By tuning the target HMM parameters to minimize the estimated error rates during run-time, we achieved an additional error reduction over environmentally-matched models, especially in low signal-to-noise condition, without using any adaptation data.

2. Model-Based Discriminative HMM Distances

Given two probability densities, the Kullback-Leibler (KL) divergence [8] of $p(\cdot)$ with respect to $q(\cdot)$ is defined as

$$\mathcal{D}(p \parallel q) = \int \log(p(x)/q(x))p(x)dx \quad (1)$$

Closed-form expression of the KL divergence could only be obtained in a few cases, e.g. between multivariate Gaussians.

Another example of an HMM distance, in the MCE formulation, is defined between a target discriminant function and a collection of competing discriminants, and evaluated as a sample-based misclassification function for a given sample X , as follows:

$$d_i(X, \Lambda) = -g(X, \lambda_i) + \ln \left\{ \frac{1}{N} \sum_{j, j \neq i} \exp[\eta \cdot g(X, \lambda_j)] \right\}^{1/\eta}, \quad (2)$$

where $g(X, \theta)$ is the log likelihood of observing X for HMM with parameter θ , and Λ is the set of all HMM parameters, η is a positive weighting constant, and N is the number of competing HMM's. By replacing X , with a density, $p(\cdot)$, that could have generated X , we can now define a corresponding model-based misclassification function as

$$\mathcal{M}_i(\Lambda) = -g(p, \lambda_i) + \ln \left\{ \frac{1}{N} \sum_{j, j \neq i} \exp[\eta \cdot g(p, \lambda_j)] \right\}^{1/\eta}. \quad (3)$$

where $g(p, \lambda_i)$ is a model-based discriminant function. Using the KL divergence defined in Eq. (1), the model-based discriminant, $g(p, \lambda_i) = -\mathcal{D}(p(x) \| p(\lambda_i))$, could in principle be evaluated and therefore the model-based distance in Eq. (3) could be computed. Here are some examples with Gaussian mixture models (GMM's) and HMM's with Gaussian mixture state densities.

2.1. Distance between Gaussian Mixture Models

A GMM density is expressed as a weighted sum of Gaussian densities. We assume that the similarity between two classes, modeled by GMM's, is estimated by a weighted sum of all pairs of the mixture components from the two classes, or two mixture Gaussian states, i and j . This is evaluated as

$$\mathcal{B}(s_i, s_j) = \sum_{m=1}^{Q_i} c_{im} \sum_{n=1}^{Q_j} c_{jn} \exp(-\mathcal{D}(p_{im} \| p_{jn})), \quad (4)$$

where Q_i and Q_j are the numbers of the Gaussian mixture components of the states, respectively, and c_{im} and c_{jn} are the respective mixture weights of the m -th Gaussian component p_{im} of state i and the n -th Gaussian component p_{jn} of state j . The divergence-based discriminant function for GMM is then given by $g(p_i, \lambda_j) = \log\{\mathcal{B}(s_i, s_j)\}$.

2.2. Distance between Phones and Words

Let $S = \{s_1, s_2, \dots, s_{J_1}\} \in \lambda_1$ and $S' = \{s'_1, s'_2, \dots, s'_{J_2}\} \in \lambda_2$ be the state sequences of two HMM's modeling phones or words in ASR, with J_1 and J_2 being the respective lengths of the state sequences. In our approach, the Viterbi algorithm [1] is used to find the most likely state sequence, $\bar{s} = \{\bar{s}_i\}_{i=1}^T$, $T = \sum_{k=1}^J t_k$, with λ_2 with respect to the sequence generated by λ_1 , where t_k is the duration of state k of the λ_1 . State duration models for each phone estimated from the training set are used in sequence generation. The similarity is calculated with

$$\ln\{\mathcal{H}(\lambda_1, \lambda_2)\} = \frac{1}{T} \sum_{i=1}^J \sum_{t=1}^{t_i} \ln\{a_{\bar{s}_{t-1}, \bar{s}_t, \lambda_2} \mathcal{B}(s_i, \lambda_1, \bar{s}_t)\}, \quad (5)$$

where $a_{\bar{s}_{t-1}, \bar{s}_t, \lambda_2}$ is the probability of transition from state \bar{s}_{t-1} to state \bar{s}_t and $\tau = \sum_{j=1}^{J-1} t_j + t$. Therefore, in this case we could assume $g(p_i, \lambda_j) = \ln\{\mathcal{H}(\lambda_i, \lambda_j)\}$.

3. Model-Based Error Estimation

Error estimation is an important topic studied extensively in pattern recognition. Accurate error estimation makes it easy for a researcher to predict the performance of a system without the need of collecting a large testing database, which is sometimes expensive to manage. However, there is very little done in ASR to estimate recognition errors using only the given set of trained HMM's because time-warping is needed [12]. Here we apply the proposed HMM distances discussed to perform error estimation for isolated phone and word recognition. Due to the added complexity to handle multiple strings, error estimation for continuous speech is more involved and will not be investigated. In the current study, the overall error rate could be estimated by

$$\mathcal{L}(\Lambda) = \sum_{i=1}^N E(c_i) P(c_i), \quad (6)$$

where $E(c_i) \in [0,1]$ is the estimated error of the i^{th} class and $P(c_i)$ is the *a priori* class probability. We propose the use of $\ell(\mathcal{M}_i(\Lambda))$ in Eq. (3) to approximate $E(c_i)$, with $\ell(\cdot)$ being a smoothed 0-1 function, like a sigmoid, that converts any distance to an approximated error count as follows [7],

$$\ell(\mathcal{M}_i(\Lambda)) = \frac{1}{1 + \exp(-\gamma \mathcal{M}_i(\Lambda) + \beta)}, \quad (\gamma > 0) \quad (7)$$

where β and γ are location and slope constants. When $\mathcal{M}_i(\Lambda)$ is much smaller than zero, it is clear a correct classification is implied and little loss is incurred. Depending on the model-based discriminant evaluated in Eqs. (4) and (5), the values of the quantity in Eq. (7) vary. We are now ready to estimate the overall error rate by approximating the error in Eq. (6) as

$$\mathcal{L}(\Lambda) \approx \sum_{i=1}^N \ell(\mathcal{M}_i(\Lambda)) P(c_i). \quad (8)$$

4. Model-Based Parameter Tuning

By expressing the model-based discriminant, $g(\cdot)$, in Eq. (3) as a function of two sets of parameters, the loss function in Eq. (8) could be extended to predict the error rate at mismatched conditions. We could also perform MCE training to tune HMM parameters. Given two sets of HMM's, Λ_1 and Λ_2 which could be trained in the same or under different conditions, we have

$$\mathcal{L}(\Lambda_1, \Lambda_2) = \sum_{r=1}^N \ell(\mathcal{M}_r(\Lambda_1, \Lambda_2)) P(c_r). \quad (9)$$

Then we can define a performance degradation measure \mathcal{L}_p as

$$\mathcal{L}_p(\Lambda^{(n)}, \Lambda^{(c)}) = \sum_{r=1}^N \{\ell(\mathcal{M}_r(\Lambda^{(n)}, \Lambda^{(n)})) - \ell(\mathcal{M}_r(\Lambda^{(c)}, \Lambda^{(c)}))\} P(c_r), \quad (10)$$

where $\Lambda^{(c)}$ and $\Lambda^{(n)}$ are the clean model and the tuned model for matching the testing condition, respectively. The quantity \mathcal{L}_p has a property that the error rate in clean condition is usually the smallest, i.e., $\ell(\mathcal{M}_r(\Lambda^{(n)}, \Lambda^{(n)})) \geq \ell(\mathcal{M}_r(\Lambda^{(c)}, \Lambda^{(c)}))$ for $\forall r$.

Another useful measure not considered in the MCE formulation is modeling accuracy. Following the same model-based scenario, we consider a loss with respect to environment-matched models

$$\mathcal{L}_m(\Lambda^{(n)}, \Lambda^{(0)}) = \sum_{r=1}^N \{1 - \ell(\mathcal{R}_r(\Lambda^{(n)}, \Lambda^{(0)}))\} P(c_r), \quad (11)$$

where $\mathcal{R}_r(\Lambda^{(n)}, \Lambda^{(0)})$ is the loss in modeling accuracy for class r defined by

$$\mathcal{R}_r(\Lambda^{(n)}, \Lambda^{(0)}) = \frac{1}{J_r} \sum_{k=1}^{J_r} \log \left[\sum_{m=1}^{M_k} c_{km} \exp(-\mathcal{D}(p_{km}^{(n)} \| p_{km}^{(0)})) \right], \quad (12)$$

where J_r and M_k are the numbers of states and mixtures of the k^{th} state for class r , respectively. Thus, the objective function of the proposed algorithm jointly considers the losses in recognition performance \mathcal{L}_p and modeling accuracy \mathcal{L}_m as follows,

$$\mathcal{L}(\Lambda^{(n)}, \Lambda^{(0)}, \Lambda^{(c)}) = \alpha \mathcal{L}_m(\Lambda^{(n)}, \Lambda^{(c)}) + (1 - \alpha) \mathcal{L}_p(\Lambda^{(n)}, \Lambda^{(0)}), \quad (13)$$

where α is a positive weighting factor. The generalized probabilistic descend (GPD) algorithm [7] could now be used to minimize the overall loss iteratively.

5. Experimental Results

We use two databases, ASJ-CSC (The ASJ Continuous Speech Corpus) and JSDC (Japanese Speech Data Corpus), for performance evaluation. These corpora were recorded with desktop microphones and sampled at 16 kHz with 16-bit quantization. The training set contains 12,000 ASJ-CSC utterances spoken by 50 speakers. The test sets vary. ASJ-CSC is used for phone recognition and JSDC for word recognition.

Feature extraction is performed as follows: for each 30 ms frame (with 12 ms overlap), a 25-dimensional feature vector (12 cepstrum, 12 delta cepstrum, and delta log energy) was extracted using 512-point FFT with a mel-scale filterbank. 35 context-independent Japanese monophone HMM's were trained with three states per model and 16 Gaussian mixture components per state. White noise samples from the NOISEX-92 database were used to simulate the additive noise in this study.

5.1. Error Estimation with Isolated Phone Recognition

Using Eq. (5) to evaluate the distances between two phones, the model-based discriminant function in Eq. (6) could be computed. The phone error is then estimated with Eq. (8) and an error rate of 25.2% was achieved as shown in Table 1.

To validate error estimation, we use a testing database for phone recognition, with a set of 2,500 ASJ-CSC utterances, spoken by 14 speakers who are different from those in the training set. Only substitution errors are considered to simulate isolated phone recognition. To avoid insertion and deletion errors, the test utterances were first segmented by forced Viterbi alignment. The experimental results in Table 1 indicate that the recognized phone error of 25.8% is close to the estimated error of 25.2%.

5.2. Error Estimation with Isolated Word Recognition

The JSDC city name sub-corpus is used for isolated word recognition testing. It consists of a vocabulary of 100 Japanese cities with phonetically rich names, and with a total of 5,000 utterances spoken by another group of 50 talkers. A subset containing 2000 utterances spoken by 20 talkers were tested here. The experimental results are shown in the bottom row of Table 1. Again, the experimental word error rate (1.7%) is close to the experimental one (1.8%) obtained with Eq. (8).

Table 1. A comparison of error rate estimation.

	Sample-based error	Model-based error
Phone error rate	25.8%	25.2%
Word error rate	1.7%	1.8%

5.3. Error Estimation with Word Recognition in Noise

The above error estimation algorithm can be used to predict recognition errors in adverse condition, either with clean models or environment-adapted models. Using artificially simulated noisy speech data and the phone and state segmentation obtained in clean environment, all the HMM parameters can be re-estimated, at a specific signal-to-noise (SNR) level, to create a collection of environment-adapted models [13]. Therefore the same procedure used in Section 5.2 can now be adopted for error estimation. The results are shown in Figure 1.

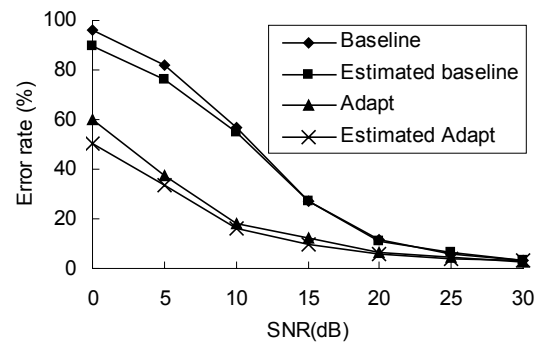


Figure 1 The experimental and estimated error rates of the baseline (clean) and adapted models in different conditions

Four performance curves are plotted, two each for estimated and experimental error rates, respectively. The legends, "Baseline" and "Adapt", denote the error rates of recognizing the noise-corrupted speech or estimating error rates by using clean and environment-adapted models,

respectively. It is interesting to note that the estimated error rates are close to the experimental ones in all noisy conditions tested. For lower SNR, the estimated error rates were underestimated. Due to the use of the sigmoid function in Eq. (7), some outliers were smoothed out, causing the estimate of the loss function in Eq. (8) to be lower.

5.4. Model Tuning with Word Recognition in Noise

To tune the parameter at each specific SNR level, the proposed model tuning algorithm (MTA) starts from a set of environment-adapted models and then adjusts the HMM parameters iteratively using the GPD algorithm. The experimental results are shown in Table 2, where "MTA" denotes the word error rate obtained by running isolated word recognition experiments on the 2,000 utterances from the JSDC city name testing set using the models trained with the proposed model tuning algorithm. Different values of the weighting factor α in Eq. (12) are experimented. It is shown that the MTA achieves slightly better performances over the environment-adapted when both modeling accuracy and model discrimination power are equally considered ($\alpha = 0.5$).

It should be noted that the proposed adaptation algorithm does not need any adaptation data, a desirable feature with the model-based performance measures used in Eqs. (11)-(13). It is also noted that an additional 9% word error reduction over the environment-adapted models is observed at the 0dB SNR level, a side benefit with HMM parameter tuning based on minimizing model-based loss functions.

Table 2 Word error rate by running word recognition using different sets of models in various noisy conditions. The reference word accuracy in clean condition is 98.3%.

	Baseline (Clean)	Adapt (Matched)	MTA (α)				
			0.1	0.3	0.5	0.7	0.9
30 dB	3.0	2.9	3.0	2.9	2.9	2.9	2.9
25 dB	6.1	4.3	4.5	4.3	4.2	4.2	4.3
20 dB	11.3	6.7	6.6	6.2	6.4	6.4	6.7
15 dB	27.3	12.0	11.8	11.1	11.1	11.2	11.9
10 dB	56.7	18.1	18.4	16.7	16.8	17.0	17.7
5 dB	82.2	37.2	35.3	34.5	34.3	34.3	36.5
0 dB	96.3	60.3	57.2	55.5	54.9	55.6	58.1

6. Summary

In this paper, we proposed a new family of model-based HMM distances. They are used to measure separation between a target model and a set of competing models. By imposing a smoothed 0-1 function on the HMM distance, it could then be used to estimate the errors of comparing the corresponding HMM with competing HMM's. Since most phones and words in speech recognition are modeled by HMM's, the proposed distances can be used to predict recognition error rates without the need of collecting a large testing database. Experimental results show that the estimated error rate is close to the experimental one in both isolated phone and word recognition. The proposed error estimation scheme also works well in predicting isolated word recognition performances in noisy conditions. As a side benefit, the proposed algorithm could also serve as an

objective measure for model parameter adjustment without using any adaptation data. Starting from a set of environment-adapted models, the model tuning algorithm improves the recognition performance, especially in low SNR conditions. Our future work includes the extension of error estimation to continuous speech recognition.

7. Acknowledgements

This research was partially supported by the National Science Council, Taiwan, ROC, under Grant No. NSC-91-2219-E-007-017. The first author also thanks Philips Speech Processing, APAC, for supporting most of the work carried out in this study.

8. References

- [1] L. R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, Vol. 77, pp. 257-286, Feb. 1989.
- [2] J.-L. Gauvain and C.-H. Lee, "Maximum A Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Trans. on Speech and Audio Proc.*, Vol. 2, No. 2, pp. 291-298, April 1994.
- [3] B.-H. Juang, W. Chou and C.-H. Lee, "Discriminative Methods for Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, Vol. 5, No. 3, pp. 257-265, May 1997.
- [4] L.R. Bahl, P.F. Brown, P.V. de Souza and R.L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition", *Proc. ICASSP-86*, pp. 49-52, Tokyo, 1986.
- [5] Y. Ephraim, A. Dembo, and L. R. Rabiner, "A minimum discrimination information approach for hidden Markov modeling", *IEEE Trans. Information Theory*, Vol. 35, No. 5, pp. 1001-1013, September 1989.
- [6] C.-H. Lee and Q. Huo, "On Adaptive Decision Rules and Decision Parameter Adaptation for Automatic Speech Recognition," *Proceedings of the IEEE*, Vol. 88, No. 8, pp. 1241-1269, August 2000.
- [7] S. Katagiri, B.-H. Juang and C.-H. Lee, "Pattern Recognition Using A Generalized Probabilistic Descent Method," *Proceedings of the IEEE*, Vol. 86, No. 11, pp. 2345-2373, Nov. 1998.
- [8] S. Kullback, "Information Theory and Statistics", New York, Dover, 1968.
- [9] 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, Feb. 1985.
- [10] M. Faulhansen, H. Reininger and D. Wolf, "Calculation of Distance Measures between Hidden Markov Models," *Proc. EuroSpeech95*, 1995.
- [11] B. T. Tan, Y. Gu and T. Thomas, "Word Confusability Measures for Vocabulary Selection in Speech Recognition," *Proc. 1999 IEEE Workshop on ASRU*, Snow Bird, UT, 1999.
- [12] C.-S. Huang, H. -C. Wang, and C. -H. Lee, "A Study of Model-based Error Rate Estimation for Automatic Speech Recognition", to appear in *IEEE Trans. on Speech and Audio Processing*, 2003.
- [13] C.-S. Huang, H. -C. Wang, and C. -H. Lee, "An SNR-incremental stochastic matching algorithm for noisy speech recognition", *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 866-973, 2001.