Improved HMM Training and Scoring Strategies with Application to Accent Classification

Levent Arslan, John H.L. Hansen
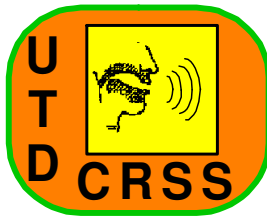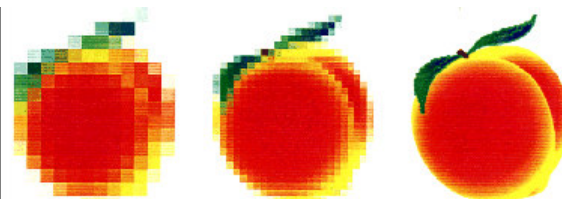
Center for Robust Speech Systems

Erik Jonsson School of Engineering & Computer Science
Department of Electrical Engineering
The University of Texas at Dallas
P.O. Box 830688, EC33
Richardson, TX 75083-0688
972 – 883 – 2910 (Phone)   972 – 883 - 2710 (Fax)
John.Hansen@utdallas.edu (email)

The 1996 International Conference on
Acoustics, Speech,
and Signal Processing

Improved HMM Training and Scoring Strategies with Application to Accent Classification

Levent M. Arslan   and   John H.L. Hansen

Robust Speech Processing Laboratory, Duke University
Box 90291, Durham, North Carolina 27708-0291
http://www.ee.duke.edu/Research/Speech   larslan@ee.duke.edu   jhlh@ee.duke.edu

ABSTRACT

In this study we propose two methods to improve HMM speech recognition performance. The first method employs an adjustment in the training stage, whereas the second method employs it in the scoring stage. It is well known that speech recognition system performance increases when the amount of labeled training data is large. However, due to factors such as inaccurate phonetic labeling, end-point detection, and voiced-unvoiced decisions, the labeling procedure can be prone to errors. In this study, we propose a selective hidden Markov Model (HMM) training procedure in order to reduce the adverse influence of atypical training data on the generated models. To demonstrate its usefulness, selective training is applied to the problem of accent classification, resulting in a 9.4% improvement in classification error rate.

The second goal is to improve HMM scoring performance. The objective of HMM training algorithms is to maximize the probability over the training tokens for each model. However, this does not guarantee a minimized error rate across the entire model set. Typically, biases in the confusion matrices can be observed. We propose a method for estimating the bias from input training data, and incorporating it into the general scoring algorithm. Using this technique, a 9.8% improvement is achieved in accent classification error rate.

1. INTRODUCTION

For speech recognition, the two most popular training strategies are hidden Markov models [13], and artificial neural networks (ANN) [11]. In general, HMM's are preferred over neural networks, because their implementations are simpler and faster. Moreover, HMM performance has been shown to be slightly better than neural networks. Recently, HMM-ANN hybrids were proposed which combine both modelling strategies in order to improve performance [2]. Recent developments in speech recognition technology have resulted in many speech recognition systems achieving acceptable error rates. However, the performance of these systems depends heavily on the complexity of the task vocabulary as well as available training data. For example, distinguishing between words such as "white" and "wide" is clearly more difficult than distinguishing between "hot" and "destination". Studies have been conducted that attempt to increase the separability among similar speech patterns. Linear discriminant analysis [6] is a method of transforming and scaling variables to improve classification performance. It was first successfully applied to speech recognition by Hunt [4] in Independent Mel-scale Linear Discriminant Analysis (IMELDA). Various studies have noted improvements in sub-word recognition using this technique [5, 7]. Further refinements to this technique by Ayer [3] resulted in its use for whole-word recognition, and by Parris [12] to incorporate state specific mixture densities. Juang proposed another technique in order to minimize the number of errors in the training set [10] by weighing the feature set, resulting in improvement in the highly confusable e-set.

In this paper, the problem of accent classification [1] is taken under the framework of distinguishing among confusable speech patterns, since the goal is to distinguish among different pronunciations of the same utterance. We propose modifications to the forward-backward training and Viterbi scoring algorithms. In Sec. 2, we describe the selective training method. In Sec. 3, the model bias removal procedure is developed and employed during scoring of the HMM's. In Sec. 4, the accent database used in our experiments is summarized. Sec. 5 presents the experiments conducted, and discusses the results. Finally, conclusions and future work are presented.

2. SELECTIVE TRAINING

In speech recognition, the generation of accurate models for speech units is essential in achieving high performance. In order to generate accurate models, one often needs a substantial amount of training data. However, it is not always trivial to collect sufficient data, depending on the application. For example, most speaker-dependent continuous systems are expected to operate with a minimal amount of training data. Even when there is sufficient training data available, the data labeling is often prone to errors (inaccurate phonetic labels, end-point detection, voiced-unvoiced decision, etc.). These errors may cause modeling inaccuracies in the HMM training phase.

In Fig. 1, a scatter plot of a set of 2-dimensional feature vectors is shown. The aim here is to distinguish between two classes of data, labeled A and B. We would like to generate statistical models for the two classes based on the training data shown in the figure. The common approach to this problem is to represent the two classes with 2-dimensional Gaussian densities, with means and variances

computed from corresponding class samples. Following this approach, the means for classes A and B are found to be m_{Ai} and m_{Bi} respectively. The corresponding variances are represented by dashed ellipses. In the figure, an outlier exists for both classes. These outliers bias the models, and would result in test errors for the models that were just generated. Of course, if they were labeled correctly to begin with, it would have been better to leave the models as they are. However, if the labeling procedure is prone to errors, it may be better to exclude these outliers in the training process to estimate more accurate models. When the outliers are excluded, the means for the two classes shift to m_{Af} and m_{Bf}, and the new variances are represented with the solid ellipses. It is clear that the new model statistics can characterize the data better excluding outliers.

This approach can be extended to the training of HMM's from labeled data. For the application of foreign accent classification, the labeling of the data may be extremely unreliable depending on the level of accent each speaker possesses. For example, an utterance from a non-native speaker may be used in the training of accented models, even though the speaker had perfect pronunciation. One approach to the solution of this problem would be to generate initial models from bootstrap training data assuming that they are correctly labeled. Then, the outliers in the training data can be identified by testing the training data using these initial models. Finally, the models could be re-trained using the same training data excluding the outliers. Another approach is to weigh the training tokens according to their relative match to the initial models and their degree of dissimilarity from the rest of the models. In this case, the likelihood ratio can be used in calculating weights. The weights for the training tokens can be calculated as

$$w_{ik} = \frac{P(X_{ik}|\lambda_i)}{(\prod_{j=1, j\neq i}^{N} P(X_{ik}|\lambda_j))^{1/(N-1)}}$$

where λ_i is the ith word model, which consists of the mean vectors μ, covariance matrices Σ, mixture coeffient matrix C, and state transition matrix A, and X_{ik} is the kth training token of the ith word. In terms of log-probabilities the weight expression becomes

$$ln(w_{ik}) = ln(P(X_{ik}|\lambda_i)) - \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} ln(P(X_{ik}|\lambda_j)).$$

In our experiments, we also used a relaxation parameter ν to control the dynamic range of weight adjustment,

$$w_{ik} = e^{\nu(ln(P(X_{ik}|\lambda_i)) - \frac{1}{N-1} \sum_{j=1, j\neq i}^{N} ln(P(X_{ik}|\lambda_j)))}$$

As ν takes on larger values, the influence of the outliers on the generated models is reduced. A special case of selective HMM training is when no weight adjustment is applied (i.e., when $\nu = 0$), which corresponds to traditional HMM training. The Forward-Backward re-estimation equations are adjusted to take into account the new set of weights for

the training tokens. The new set of equations becomes:
State transition matrix entries:

$$\bar{a}_{ij} = \frac{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N-1} \gamma_n(i,j)}{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N-1} \sum_j \gamma_n(i,j)}$$

where R is the total number of training observation sequences, w_r is the weight assigned to r^{th} observation sequence, P_r is the total probability of rth observation sequence given the model, and $\gamma_n(i,j)$ is the probability of making a transition from state i to state j at time n. The remainder of the parameter re-estimation equations are as follows:
Mixture coefficients:

$$\bar{c}_{jk} = \frac{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \xi_n(j,k)}{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \sum_{k=1}^{L} \xi_n(j,k)}$$

Mean vector entries:

$$\bar{\mu}_{jk} = \frac{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \xi_n(j,k) x_n}{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \xi_n(j,k)}$$

Covariance matrix entries:

$$\bar{\Sigma}_{jk} = \frac{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \xi_n(j,k)(x_n - \mu_{jk})(x_n - \mu_{jk})'}{\sum_{r=1}^{R} \frac{w_r}{P_r} \sum_{n=1}^{N} \xi_n(j,k)}$$

where $\xi_n(j,k)$ is the probability of being in state j at time frame n with the kth mixture component accounting for x_n, i.e.,

$$\xi_n(j,k) = \left[\frac{\alpha_n(j)\beta_n(j)}{\sum_{j=1}^{M} \alpha_n(j)\beta_n(j)} \right] \left[\frac{c_{jk} f_{jk}(x_n)}{\sum_{m=1}^{L} c_{jm} f_{jm}(x_n)} \right]$$

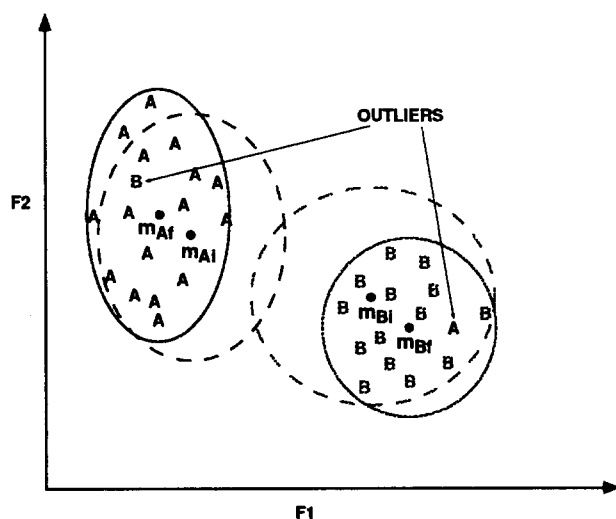## SELECTIVE TRAINING ILLUSTRATION

Figure 1: An illustration of the adverse influence of outliers when creating statistical models.

590

3. MODEL BIAS REMOVAL

Most errors in speech recognition are due to confusable word pairs in the vocabulary set such as "white" vs. "wide", or subword units such as "f" vs. "s". When the confusion matrices for these systems are analyzed, we observed biases towards one of the words in the confusable pairs. For example, in a study on speech under stress [8], the model for "wide" was favored by all the "wide" utterances, as well as most "white" utterances. Our proposed method is to balance the biases in the confusion matrices with the constraint of maximizing the separation between confusable word pair models or minimizing the error rate in the training set.
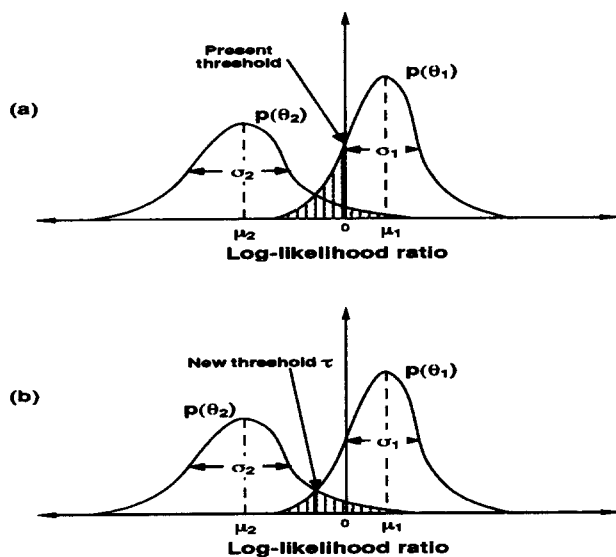
BIAS ESTIMATION BETWEEN CONFUSABLE PAIR

Figure 2: An illustration of model bias removal between confusable word pairs.

Normally, when an observation vector sequence X is to be classified as one of the two models λ_1 and λ_2, we calculate $P(X|\lambda_1)$ and $P(X|\lambda_2)$, and select the model which results in the higher likelihood. The decision criterion can be formulated as

$$Choose \quad \lambda_1 \qquad if \quad \theta > 0$$
$$Choose \quad \lambda_2 \qquad otherwise$$

where θ is the log-likelihood ratio $ln(\frac{P(X|\lambda_1)}{P(X|\lambda_2)})$. Although this rule is mathematically correct and simple, at times there may be a bias towards one of the models. This is especially true for speech recognition systems. Most speech recognition systems generate models with the criterion of maximizing the probability over the training tokens of each word. However, this does not guarantee the minimization of the error rate across the entire vocabulary. Our proposed solution to this problem is to make the adjustment in the scoring procedure *instead* of transforming the feature vector or modifying the training procedure as is the case with most other confusion discriminant algorithms. Here, updating the threshold (which is normally 0) in the decision

rule above, will result in minimization of the total number of errors. The new rule is expressed as:

$$Choose \quad \lambda_1 \qquad if \quad \theta > \tau$$
$$Choose \quad \lambda_2 \qquad otherwise$$

where τ is determined from the training data as follows. Let X_1 and X_2 denote the training observation vector sets for the confusable pair $word_1$ and $word_2$ respectively. The models generated from the training set are denoted as λ_1 and λ_2. Next, define log-likelihood ratio functions θ_1 and θ_2 as

$$\theta_{1n} = ln(\frac{P(X_{1n}|\lambda_1)}{P(X_{1n}|\lambda_2)}) \quad \theta_{2n} = ln(\frac{P(X_{2n}|\lambda_1)}{P(X_{2n}|\lambda_2)}) \quad n = 1, ..., N$$

where N is the number of training tokens for $word_1$ and $word_2$.

In Fig. 2, the probability distribution functions of θ_1 and θ_2 are plotted for a hypothetical case. In the top graph, it can be seen that model 2 is selected in most of the erroneous decisions. In the lower graph of Fig. 2, the estimate of the new threshold τ is shown as the point where the two distributions intersect. This point minimizes the size of the shaded area, which is the total error region. If $p(\theta_1) = N(\mu_1, \Sigma_1)$, and $p(\theta_2) = N(\mu_2, \Sigma_2)$, then the decision boundary τ can be computed from the equation

$$\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{-\frac{(\tau-\mu_1)^2}{2\sigma_1^2}} = \frac{1}{\sqrt{2\pi\sigma_2^2}}e^{-\frac{(\tau-\mu_2)^2}{2\sigma_2^2}}$$

$$\frac{\sigma_2}{\sigma_1} = \frac{e^{-\frac{(\tau-\mu_1)^2}{2\sigma_2^2}}}{e^{-\frac{(\tau-\mu_1)^2}{2\sigma_1^2}}}.$$

Taking the natural logarithm of both sides, results in,

$$ln(\sigma_2) - ln(\sigma_1) = -\frac{(\tau - \mu_1)^2}{2\sigma_2^2} + \frac{(\tau - \mu_1)^2}{2\sigma_1^2}.$$

If we rearrange the terms on both sides, the equation reduces to a second order polynomial of the form, $A\tau^2 + B\tau + C = 0$ where

$$A = \sigma_1^2 - \sigma_2^2$$
$$B = 2\sigma_2^2\mu_1 - 2\sigma_1^2\mu_2$$
$$C = \sigma_1^2\mu_2^2 - \sigma_2^2\mu_1^2 + 2\sigma_1^2\sigma_2^2(ln(\sigma_2) - ln(\sigma_1))$$

The root of the polynomial which lies between μ_1 and μ_2 is the new decision boundary.

4. ACCENT DATABASE

In order to investigate accent, a vocabulary of words was established which contains accent sensitive phonemes or phoneme combinations [1, 9]. Vocabulary choice was based on a literature review of language education of American English as a second language. The data corpus was collected using a head-mounted microphone, from speakers among the general Duke University community. The test vocabulary consists of twenty isolated words (sample words

591

include: aluminum, thirty, bringing, target, bird). Available speech includes neutral American English, and English under the following accents: German, Chinese, Turkish, French, Persian, Spanish, Italian, Hindi, Romanian, Japanese, Russian, and others. For the studies conducted here, we focus on American English speech from twenty-seven speakers across the following accents: Turkish, Chinese and German.

5. EVALUATIONS

We applied our proposed method for selective training to foreign accent classification. Four accent types (neutral, Turkish, German, Chinese) were considered in the experiments. By using the selective training method, the average classification rate using single words among 4 accents improved from 62.8% to 66.3% for an open test set (a 9.4% error reduction).

Next, the proposed bias removal method is applied to the confusable word pairs in the SUSAS speech under stress database [8]. Speech spoken under stressed style include loud, angry, fast, slow, soft, clear, and Lombard effect conditions. We considered 9 such word pairs, and concentrated our efforts on distinguishing only between confusable word pairs. The decision boundary was estimated for each word pair from the speech of 6 speakers (12 tokens each). In the test set, two different speakers, and 10 different types of stress conditions from all speakers in the database was included. The error rate was found to be 5.89% using the Viterbi algorithm, and a threshold of 0 for the likelihood ratio. When the model bias estimation method was used to shift the thresholds for the confusable word pairs, the error rate was reduced to 5.16% (a 12.36% improvement). A total of 3024 words was used in the test set.

Next, the optimum boundary decision method is applied to foreign accent classification. The method was used to distinguish between neutral American accent and Turkish accent in a 20-word database. The standard Viterbi algorithm without bias removal resulted in an error rate of 3.8% in the closed set. Using the estimated bias shift between the accents for each word, the error rate was reduced to 1.6% (a 61.1% reduction from original). For the test set, the error rate was reduced from 14.3% to 12.9% (a 9.8% reduction from the original). The results are illustrated in Table 1. Here, it is noted that the error rate in the closed set improves substantially, with positive, but not as substantial improvement for the test set. This can be explained by the fact that the optimum boundary decision method concentrates highly on minimizing the error rate on the closed set.

6. CONCLUSION

In this paper, two new techniques have been proposed to improve the performance of speech recognition algorithms. The first method was based on selective training, where outliers are given reduced weight in the training phase. The second method was based on a bias removal procedure where a new likelihood decision boundary is estimated in order to balance the overall error in the input vocabulary

Error rates in classification of Turkish accent versus neutral accent		
Scoring method	Standard	After bias removal
Closed set	3.8 %	1.6 %
Open test set	14.3 %	12.9 %

Table 1: Accent classification performance improvement by bias removal over closed and test sets.

set. These methods were evaluated for an accent classification task. Both methods resulted in consistent improvement in classification rate. Finally, while the proposed methods were evaluated using isolated word HMM's, these methods methods can easily be applied to other speech research areas such as continuous speech recognition and speaker ID systems.

References

[1] L. M. Arslan and J.H.L. Hansen. "Language Accent Classification in American English". submitted to *Speech Communications*, August 1995.

[2] Yoshua Bengio, Renato De Mori, Giovanni Flammia, and Ralf Kompe. Global Optimization of a Neural Network–Hidden Markov Model Hybrid. *IEEE Trans. Neural Net.*, 3(2):252–9, March 1992.

[3] C.M. Ayer et al. "A Discriminatively Derived Linear Transform for Improved Speech Recognition". In *Proc. Eurospeech*, Berlin, 1993.

[4] M.J. Hunt et al. "An investigation of PLP and IMELDA Acoustic Representations and of their Potential for Combination". In *Proc. IEEE ICASSP*, Toronto, 1991.

[5] X. Aubert et al. "Continuous Mixture Densities and Linear Discriminant Analysis for Improved Context Dependent Acoustic Models". In *Proc. IEEE ICASSP*, Minneapolis, 1993.

[6] R.A. Fisher. "The Use of Multiple Measures in Taxonomic Problems". *Contr. to Math. Stats.*, pages 32.179–32.188, 1950.

[7] R. Haeb-Umbach and H. Ney. "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition". In *Proc. IEEE ICASSP*, San Fransisco, 1992.

[8] J.H.L. Hansen. "Morphological Constrained Enhancement with Adaptive Cepstral Compensation (MCE-ACC) for Speech Recognition in Noise and Lombard Effect". *IEEE Trans. on Speech & Audio Proc.*, 2(4):598–614, October 1994.

[9] J.H.L. Hansen and L.M. Arslan. "Foreign Accent Classification using Source Generator Based Prosodic Features". In *Proc. IEEE ICASSP*, pages 836–839, Detroit, 1995.

[10] B.H. Juang and S Katagiri. "Discriminative Learning for Minimum Error Classification". *IEEE Trans. on Signal Processing*, 40:3043–3054, 1992.

[11] David P. Morgan and Christopher L. Scofield. *Neural Networks and Speech Processing*. Kluwer, 1991.

[12] E.S. Parris and M.J. Carey. "Estimating Linear Discriminant Parameters for Continuous Density Hidden Markov Models". In *Proc. ICSLP*, Yokohama, 1994.

[13] L.R. Rabiner and B.H. Juang. "An introduction to hidden Markov models". *IEEE ASSP Magazine*, pages 4–16, 1986.