COMPARISON OF LARGE MARGIN TRAINING TO OTHER DISCRIMINATIVE METHODS FOR PHONETIC RECOGNITION BY HIDDEN MARKOV MODELS

Fei Sha[*]

Computer Science Division
University of California
527 Soda Hall
Berkeley, CA 94720-1776

Lawrence K. Saul[†]

Department of Computer Science and Engineering
University of California (San Diego)
9500 Gilman Drive, Mail Code 0404
La Jolla, CA 92093-0404

ABSTRACT

In this paper we compare three frameworks for discriminative training of continuous-density hidden Markov models (CD-HMMs). Specifically, we compare two popular frameworks, based on conditional maximum likelihood (CML) and minimum classification error (MCE), to a new framework based on margin maximization. Unlike CML and MCE, our formulation of large margin training explicitly penalizes incorrect decodings by an amount proportional to the number of mislabeled hidden states. It also leads to a convex optimization over the parameter space of CD-HMMs, thus avoiding the problem of spurious local minima. We used discriminatively trained CD-HMMs from all three frameworks to build phonetic recognizers on the TIMIT speech corpus. The different recognizers employed exactly the same acoustic front end and hidden state space, thus enabling us to isolate the effect of different cost functions, parameterizations, and numerical optimizations. Experimentally, we find that our framework for large margin training yields significantly lower error rates than both CML and MCE training.

Index Terms— speech recognition, discriminative training, MMI, MCE, large margin, phoneme recognition

1. INTRODUCTION

Most modern speech recognizers are built from continuous-density hidden Markov models (CD-HMMs). The hidden states in these CD-HMMs model different phonemes or sub-phonetic elements, while the observations model cepstral feature vectors. Distributions of cepstral feature vectors are most often represented by Gaussian mixture models (GMMs). The accuracy of the recognizer depends critically on the careful estimation of GMM parameters.

In this paper, we present a systematic comparison of several leading frameworks for parameter estimation in CD-HMMs. These frameworks include a recently proposed scheme based on the goal of margin maximization [1, 2], an idea that has been widely applied in the field of machine learning. We compare the objective function and learning algorithm in this framework for *large margin* training to those of other traditional approaches for parameter estimation in CD-HMMs. The most basic of these traditional approaches involves maximum likelihood (ML) estimation. Mainly, however, we focus on competing discriminative methods in which parameters are estimated directly to maximize the conditional likelihood [3, 4] or minimize the classification error rate [5]. Though not as straightforward to implement as ML estimation, discriminative methods yield much lower error rates on most tasks in automatic speech recognition (ASR).

We investigate salient differences between CML, MCE, and large margin training through carefully designed experiments on the TIMIT speech corpus [6]. Though much smaller than typical corpora used for large vocabulary ASR, the TIMIT corpus provides an apt benchmark for evaluating the intrinsic merits of different frameworks for discriminative training. We compare the phonetic error rates on the TIMIT corpus from multiple systems trained with different parameterizations, initial conditions, and learning algorithms. All other aspects of these systems, however, were held fixed. In particular, the different systems employed exactly the same acoustic front end and model architectures (e.g., monophone CD-HMMs with full Gaussian covariance matrices). From the results of these experiments, we are able to tease apart the significant factors that differentiate competing methods for discriminative training.

The paper is organized as follows. In section 2, we review CD-HMMs as well as several different methods for parameter estimation, including our own recent formulation of large margin training [1, 2]. In section 3, we compare the performance of phonetic recognizers trained in all these different frameworks. Finally, in section 4, we conclude with a brief discussion of future directions for research.

2. PARAMETER ESTIMATION IN HMMS

CD-HMMs define a joint probability distribution over a hidden state sequence $S = \{s_1, s_2, \ldots, s_T\}$ and an observed output sequence $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$, given by

$$\log p(\boldsymbol{X}, S) = \sum_t \left[\log p(s_t|s_{t-1}) + \log p(\boldsymbol{x}_t|s_t)\right]. \quad (1)$$

For ASR, the hidden states s_t and observed outputs $\boldsymbol{x}_t$ denote phonetic labels and acoustic feature vectors, respectively, and the distributions $p(\boldsymbol{x}_t|s_t)$ are typically modeled by multivariate GMMs:

$$p(\boldsymbol{x}_t|s_t = j) = \sum_{m=1}^{M} \omega_{jm} \mathcal{N}(\boldsymbol{x}_t; \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}). \quad (2)$$

In eq. (2), we have used $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ to denote the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, while the constant M denotes the number of mixture components per GMM. The mixture weights ω_{jm} in eq. (2) are constrained to be nonnegative and normalized: $\sum_m \omega_{jm} = 1$ for all states j.

[*]Part of the work was performed at University of Pennsylvania.

[†]This work is supported by the National Science Foundation under Grant Number 0238323 and the UCSD FWGrid Project, NSF Research Infrastructure Grant number NSF EIA-0303622.

Let $\boldsymbol{\theta}$ denote all the model parameters including transition probabilities, mixture weights, mean vectors, and covariance matrices. The goal of parameter estimation in CD-HMMs is to compute the optimal $\boldsymbol{\theta}^*$ (with respect to a particular measure of optimality), given N pairs of observation and target label sequences $\{\boldsymbol{X}_n, Y_n\}_{n=1}^N$.

In what follows, we review the optimizations for well-known frameworks based on maximum likelihood (ML), conditional maximum likelihood (CML), and minimum classification error (MCE). We also review our most recently proposed framework for large margin training [2].

2.1. Maximum likelihood estimation

The simplest approach to parameter estimation in CD-HMMs maximizes the joint likelihood of output and label sequences. The corresponding estimator is given by

$$\boldsymbol{\theta}^{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log p(\boldsymbol{X}_n, Y_n) \tag{3}$$

For transition probabilities, ML estimates in this setting are obtained from simple counts (assuming the training corpus provides phonetic label sequences). For GMM parameters, the EM algorithm provides iterative update rules that converge monotonically to local stationary points of the likelihood. The main attraction of the EM algorithm is that no free parameters need to be tuned for its convergence.

2.2. Conditional maximum likelihood

CD-HMMs provide transcriptions of unlabeled speech by inferring the hidden label sequence Y with the highest posterior probability: $Y = \arg\max_S p(S|\boldsymbol{X})$. The CML estimator in CD-HMMs directly attempts to maximize the probability that this inference returns the correct transcription. Thus, it optimizes the conditional likelihood:

$$\boldsymbol{\theta}^{\mathrm{CML}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log p(Y_n|\boldsymbol{X}_n). \tag{4}$$

In CML training, the parameters must be adjusted to increase the likelihood gap between correct labelings Y_n and incorrect labelings S. This can be seen more explicitly by rewriting eq. (4) as:

$$\boldsymbol{\theta}^{\mathrm{CML}} = \arg\max_{\boldsymbol{\theta}} \left[\log p(\boldsymbol{X}_n, Y_n) - \log \sum_S p(\boldsymbol{X}_n, S) \right]. \tag{5}$$

The CML estimator in eq. (4) is closely related to the maximum mutual information (MMI) estimator [7, 8], given by:

$$\boldsymbol{\theta}^{\mathrm{MMI}} = \arg\max_{\boldsymbol{\theta}} \sum_n \log \frac{p(\boldsymbol{X}_n, Y_n)}{p(\boldsymbol{X}_n)p(Y_n)}. \tag{6}$$

Note that eqs. (4) and (6) yield identical estimators in the setting where the (language model) probabilities $p(Y_n)$ are held fixed.

2.3. Minimum classification error

MCE training is based on minimizing the number of sequence misclassifications. The number of such misclassifications is given by:

$$\mathcal{N}_{\mathrm{err}} = \sum_n \mathrm{sign}\left[-\log p(\boldsymbol{X}_n, Y_n) + \max_{S \neq Y_n} \log p(\boldsymbol{X}_n, S)\right] \tag{7}$$

where $\mathrm{sign}[z] = 1$ for $z > 0$ and $\mathrm{sign}[z] = 0$ for $z \leq 0$. To minimize eq. (7), the parameters must be adjusted to maintain a likelihood gap between the correct labeling and all competing labelings. Unlike CML training, however, the size of the gap in eq. (7) does not matter, as long as it is finite.

The nondifferentiability of the sign and $\max$ functions in eq. (7) makes it difficult to minimize the misclassification error directly. Thus, MCE training [9] adopts the surrogate cost function:

$$\mathcal{N}_{\mathrm{err}} \approx \sum_n \sigma\left(-\log p(\boldsymbol{X}_n, Y_n) + \log\left[\frac{1}{C}\sum_{S \neq Y_n} e^{\eta \log p(\boldsymbol{X}_n, S)}\right]^{\frac{1}{\eta}}\right). \tag{8}$$

In this approximation, a sigmoid function $\sigma(z) = (1 + e^{-\alpha z})^{-1}$ replaces the sign function $\mathrm{sign}[z]$, and a softmax function (parameterized by η) replaces the original $\max$. The parameters α and η in this approximation must be set by heuristics. The sum in the second term is taken over the top C competing label sequences.

2.4. Large margin training

Recently, we proposed a new framework for discriminative training of CD-HMMS based on the idea of margin maximization [1, 2]. Our framework has two salient features: (i) it attempts to separate the likelihoods of correct versus incorrect label sequences by margins proportional to the number of mislabeled states [10]; (ii) the required optimization is convex, thus avoiding the pitfall of spurious local minima. These features also distinguish our approach to large margin training of CD-HMMs from other recent formulations [11].

We start by reviewing the discriminant functions in large margin CD-HMMs [1, 2]. These parameterized functions of observations $\boldsymbol{X}$ and states S take a form analogous to the log-probability in eq. (1). In particular, we define

$$\mathcal{D}(\boldsymbol{X}, S) = \sum_t \left[\lambda(s_{t-1}, s_t) + \rho(\boldsymbol{x}_t, s_t)\right] \tag{9}$$

in terms of state-state transition scores $\lambda(s_{t-1}, s_t)$ and state-output emission scores $\rho(\boldsymbol{x}_t, s_t)$. Unlike eq. (1), however, eq. (9) does not assume that the transition scores $\lambda(s_{t-1}, s_t)$ are derived from the logarithm of normalized probabilities. Likewise, the emission scores $\rho(\boldsymbol{x}_t, s_t)$ in eq. (9) are parameterized by sums of *unnormalized* Gaussian distributions:

$$\rho(\boldsymbol{x}_t, s_t = j) = \log \sum_m e^{-(\boldsymbol{x}_t - \boldsymbol{\mu}_{jm})^{\mathrm{T}} \boldsymbol{\Sigma}_{jm}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_{jm}) - \theta_{jm}}, \tag{10}$$

where the nonnegative scalar parameter $\theta_{jm} \geq 0$ is entirely independent of $\boldsymbol{\Sigma}_{jm}$ (as opposed to being related to its log-determinant).

To obtain a convex optimization for large margin training, we further reparameterize the emission score in eq. (10). In particular, we express each mixture component's parameters $\{\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}, \theta_{jm}\}$ as elements of the following matrix:

$$\boldsymbol{\Phi}_{jm} = \begin{bmatrix} \boldsymbol{\Sigma}_{jm}^{-1} & -\boldsymbol{\Sigma}_{jm}^{-1}\boldsymbol{\mu}_{jm} \\ -\boldsymbol{\mu}_{jm}^{\mathrm{T}}\boldsymbol{\Sigma}_{jm}^{-1} & \boldsymbol{\mu}_{jm}^{\mathrm{T}}\boldsymbol{\Sigma}_{jm}^{-1}\boldsymbol{\mu}_{jm} + \theta_{jm} \end{bmatrix}. \tag{11}$$

Our framework for large margin training optimizes the matrices $\boldsymbol{\Phi}_{jm}$, as opposed to the conventional GMM parameters $\{\boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}, \theta_{jm}\}$. Since the matrix $\boldsymbol{\Sigma}_{jm}$ is positive definite and the scalar θ_{jm} is nonnegative, we also require the matrix $\boldsymbol{\Phi}_{jm}$ to be positive semidefinite (as denoted by the constraint $\boldsymbol{\Phi}_{jm} \succ 0$). With this reparameterization, the emission score in eq. (10) can be written as:

$$\rho(\boldsymbol{x}_t, s_t = j) = \log \sum_m e^{-\boldsymbol{z}_t^{\mathrm{T}} \boldsymbol{\Phi}_{jm} \boldsymbol{z}_t} \quad \text{where } \boldsymbol{z}_t = \begin{bmatrix} \boldsymbol{x}_t \\ 1 \end{bmatrix}. \tag{12}$$

Note that this score is convex in the elements of the matrices $\mathbf{\Phi}_{jm}$.

For large margin training of CD-HMMs, we seek parameters that separate the discriminant functions for correct and incorrect label sequences. Specifically, for each joint observation-label sequence $(\boldsymbol{X}_n, Y_n)$ in the training set, we seek parameters such that:

$$\mathcal{D}(\boldsymbol{X}_n, Y_n) - \mathcal{D}(\boldsymbol{X}_n, S) \;\geq\; \mathcal{H}(Y_n, S), \quad \forall\, S \neq Y_n \tag{13}$$

where $\mathcal{H}(Y_n, S)$ denotes the *Hamming distance* between the two label sequences [10]. Note how this constraint requires the log-likelihood gap between the target sequence Y_n and each incorrect decoding S to scale in proportion to the number of mislabeled states.

Eq. (13) actually specifies an exponentially large number of constraints, one for each alternative label sequence S. We can fold all these constraints into a single constraint by writing:

$$-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \max_{S \neq Y_n} \{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)\} \;\leq\; 0. \tag{14}$$

In the same spirit as the MCE derivation for eq. (8), we obtain a more tractable (i.e., differentiable) expression by replacing the $\max$ function in eq. (14) with a "softmax" upper bound:

$$-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \log \sum_{S \neq Y_n} e^{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)} \;\leq\; 0. \tag{15}$$

The exponential terms in eq. (15) can be summed efficiently using a modification of the standard forward-backward procedure.

While we would like to find parameters that satisfy the large margin constraint in eq. (15) for all training sequences $\{\boldsymbol{X}_n, Y_n\}_{n=1}^N$, in general this is not possible. For such "infeasible" scenarios, we instead compute the parameters that minimize the total amount by which these constraints are violated:

$$\min \sum_n \left[-\mathcal{D}(\boldsymbol{X}_n, Y_n) + \log \sum_{S \neq Y_n} e^{\mathcal{H}(Y_n, S) + \mathcal{D}(\boldsymbol{X}_n, S)} \right]_+ . \tag{16}$$

The "+" subscript in eq. (16) denotes the hinge function: $[z]_+ = z$ if $z > 0$ and $[z]_+ = 0$ if $z \leq 0$. The optimization of eq. (16) is performed subject to the positive semidefinite constraints $\mathbf{\Phi}_{jm} \succ 0$. We can further simplify the optimization by assuming that each emission score $\rho(\boldsymbol{x}_t, y_t)$ in the first term is dominated by the contribution from a single (pre-specified) Gaussian mixture component. In this case, the overall optimization is convex; see [2] for further details.

3. EMPIRICAL COMPARISON OF METHODS

We experimented with the discriminative frameworks in section 2 (as well as several variants of these frameworks) to explore the effects of different parameterizations, initializations, and cost functions.

3.1. Setup

CD-HMMs were evaluated on the task of phonetic recognition [12]—namely, mapping speech utterances to sequences of phonemes, as opposed to higher-level units, such as words. Phonetic label sequences of test utterances were inferred using Viterbi decoding. Note that for the Viterbi decoding of test utterances, we did not make any use of manually time-aligned phonetic transcriptions; in particular, we did not assume that the boundaries between phonetic segments were correctly located prior to decoding. This distinguishes the task of phonetic recognition, considered in this paper, from the simpler task of phonetic classification [13] considered in our earlier work [1].

M	ML	CML	MCE	Margin
1	40.1%	36.4%	35.6%	31.2%
2	36.5%	34.6%	34.5%	30.8%
4	34.7%	32.8%	32.4%	29.8%
8	32.7%	31.5%	30.9%	28.2%

Table 1. Phonetic error rates from differently trained CD-HMMs, with M mixture components per GMM. See text for details.

For each test utterance, Viterbi decoding yielded a frame-by-frame phonetic transcription with one label for each analysis window of speech. We matched the label sequences from Viterbi decoding against ground truth phonetic transcriptions (obtained manually) and used dynamic programming to compute the minimum string edit distance for each utterance. We report error rates as the number of insertion, deletion, and substitution errors over the entire corpus, normalized by the total string length.

All CD-HMMs were trained and tested on utterances from the TIMIT speech corpus [6]. We followed standard practices in preparing the training, development, and test sets [12, 13]. Our recognizers employed standard front ends and model architectures. Acoustic feature vectors were derived from 13-dimensional mel-frequency cepstral coefficients (MFCCs) and their first and second derivatives. MFCCs were computed on 25 ms analysis windows with 10 ms of overlap between consecutive windows. Each CD-HMM recognizer had 48 states, one for each of 48 broad phonetic categories and transcription markers (e.g., silence). For evaluation, these 48 labels were further simplified to 39 phonetic categories, following the convention in [12]. To vary the model size, we experimented with different numbers of mixture components per state, ranging from 1 to 8 in each GMM. For each mixture component, we estimated a full covariance matrix from the training corpus; there was no parameter-tying across different states or mixture components.

We used gradient-based numerical optimizations for CML, MCE, and large margin training. For CML training, we used conjugate gradient descent; for MCE training, we used steepest gradient descent (which worked better); for margin maximization, we used a combination of conjugate gradient and projected subgradient descent, as described in previous work [1, 2]. For CML training, we obtained competitive results from conjugate gradient descent and did not experiment with the extended Baum-Welch algorithm [14].

3.2. Experimental results

Table 1 shows the error rates of different CD-HMMs trained by ML, CML, MCE, and margin maximization. Here, M denotes the number of mixture components per state (in each GMM). As expected, all the discriminatively trained CD-HMMs yield significant improvements over the baseline CD-HMMs trained by ML. On this particular task, the results show that MCE does slightly better than CML, while the largest relative improvements are obtained by large margin training (by a factor of two or more). Using MMI on this task, Kapadia et al [14] reported larger relative reductions in error rates than we have observed for CML (though not better performance in absolute terms). It is difficult to compare our findings directly to theirs, however, since their ML and MMI recognizers used different front ends and numerical optimizations than those in our work.

Several possible factors might explain the better performance of CD-HMMs trained by margin maximization. These include: (i) the relaxation of Gaussian normalization constraints by the parameterization in eq. (11), yielding more flexible models, (ii) the convexity of our margin-based cost function eq. (16), which ensures that its

m	CML	Unnormalized	Reinitialized	Reweighted
1	36.4%	36.0%	32.6%	33.6%
2	34.6%	36.3%	31.7%	32.8%
4	32.8%	33.6%	31.2%	32.8%
8	31.5%	31.6%	28.9%	31.0%

Table 2. Phonetic error rates from CD-HMMs trained by CML and three variants of CML. See text for details.

optimization (unlike those for CML and MCE) does not suffer from spurious local minima, and (iii) the closer tracking of phonetic error rates by the margin-based cost function, which penalizes incorrect decodings in direct proportion to their Hamming distance from the target label sequence. We conducted several experiments with variants of CML and MCE training in an attempt to determine which of these factors (if any) played a significant role.

Some preliminary results are reported in Table 2, for CD-HMMs trained by three variants of CML. In the first variant, we relaxed the normalization constraints on the mixture weights and log-determinant prefactors of the GMMs. In the second variant, we initialized the GMMs from a different starting point; in particular, instead of baseline GMMs trained by ML estimation, we used large margin GMMs that had been trained for segment-based phonetic classification [1]. Finally, in the third variant, we maximized a reweighted version of the conditional likelihood:

$$\max \sum_n \log p(\boldsymbol{X}_n, Y_n) - \log \sum_S e^{\mathcal{H}(Y_n,S)+\log p(\boldsymbol{X}_n,S)}. \quad (17)$$

The reweighting in eq. (17) penalizes incorrect decodings in proportion to their Hamming distance from the target label sequence, analogous to the cost function of eq. (16) for large margin training.

The experimental results on these variants of CML training reveal several interesting findings. First, the unnormalized GMMs performed slightly worse than the normalized GMMs, possibly due to overfitting. It seems that in the absence of margin-based criteria, the extra degrees of freedom in unnormalized GMMs help to maximize the conditional likelihood in ways that are not correlated with the phonetic error rate. Second, with better initializations, the CD-HMMs trained by CML approached the performance of CD-HMMs trained by margin maximization. This result highlights a significant drawback of optimizations, such as CML and MCE, that are not convex and depend on initial conditions. Third, we observed that the reweighted conditional likelihood in eq. (17) led to improved performance for smaller models. This positive effect diminished for larger models, however, perhaps due to the increased difficulty of non-convex global optimization in larger parameter spaces. Finally, though not reported here, we also experimented with corresponding variants of MCE training, obtaining similar results.

4. CONCLUSION

In this paper we have compared large margin training in CD-HMMs to two other leading frameworks for discriminative training. On the task of phonetic recognition, we observed that our formulation of large margin training achieved significantly better performance than either CML or MCE training. Follow-up experiments suggested two possible reasons for this better performance: the convexity of the optimization for large margin training (versus those for CML and MCE training), and the penalizing of incorrect decodings in direct proportion to the number of mislabeled states. In future research, we are interested in applying large margin training to large vocabulary ASR, where both CML and MCE training have already demonstrated significant reductions in word error rates [7, 8, 9].

5. REFERENCES

[1] F. Sha and L. K. Saul, "Large margin Gaussian mixture modeling for phonetic classification and recognition," in *Proceedings of ICASSP 2006*, Toulouse, France, 2006, pp. 265–268.

[2] F. Sha and L. K. Saul, "Large margin hidden Markov models for for automatic speech recognition," in *Advances in Neural Information Processing Systems 19*, B. Schölkopf, J.C. Platt, and T. Hofmann, Eds., Cambridge, MA, 2007, MIT Press.

[3] A. Nádas, "A decision-theoretic formulation of a training problem in speech recognition and a comparison of training by unconditional versus conditional maximum likelihood," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 31, no. 4, pp. 814–817, 1983.

[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," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, 1986, pp. 49–52.

[5] B.-H. Juang and S. Katagiri, "Discriminative learning for minimum error classification," *IEEE Trans. Sig. Proc.*, vol. 40, no. 12, pp. 3043–3054, 1992.

[6] L. F. Lamel, R. H. Kassel, and S. Seneff, "Speech database development: design and analsysis of the acoustic-phonetic corpus," in *Proceedings of the DARPA Speech Recognition Workshop*, L. S. Baumann, Ed., 1986, pp. 100–109.

[7] V. Valtchev, J. J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, pp. 303–314, 1997.

[8] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Computer Speech and Language*, vol. 16, pp. 25–47, 2002.

[9] E. McDermott, T. J. Hazen, J. Le Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large vocabulary speech recognition using minimum classification error," *IEEE Trans. Speech and Audio Processing*, Jan. 2007.

[10] B. Taskar, C. Guestrin, and D. Koller, "Max-margin markov networks," in *Advances in Neural Information Processing Systems (NIPS 16)*, Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, Eds., pp. 25–32. MIT Press, Cambridge, MA, 2004.

[11] Xinwei Li, Hui Jiang, and Chaojun Liu, "Large margin HMMs for speech recognition," in *Proceedings of ICASSP 2005*, Philadelphia, 2005, pp. 513–516.

[12] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, 1989.

[13] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *Proceedings of Eurospeech 97*, Greece, 1997, pp. 401–404.

[14] S. Kapadia, V. Valtchev, and S. J. Young, "MMI training for continuous phoneme recognition on the TIMIT database," in *Proc. of ICASSP 93*, Minneapolis, MN, 1993, vol. 2, pp. 491–494.