ON-LINE ADAPTIVE LEARNING OF CDHMM PARAMETERS BASED ON MULTIPLE-STREAM PRIOR EVOLUTION AND POSTERIOR POOLING

Qiang HUO and Bin MA

Department of Computer Science and Information Systems,
The University of Hong Kong, Pokfulam Road, Hong Kong (e-mail: qhuo@csis.hku.hk)

10.21437/Eurospeech.1999-598

ABSTRACT

Based on the concept of *multiple-stream prior evolution and posterior pooling*, we propose a new incremental adaptive Bayesian learning framework for efficient on-line adaptation of the continuous density hidden Markov model (CDHMM) parameters. As a first step, we apply the affine transformations to the mean vectors of CDHMMs to control the evolution of their prior distribution. This new stream of prior distribution can be combined with another stream of prior distribution evolved without any constraints applied. In a series of comparative experiments on the task of continuous Mandarin speech recognition, we show that the new adaptation algorithm achieves a similar fast-adaptation performance as that of incremental MLLR (maximum likelihood linear regression) in the case of small amount of adaptation data, while maintains the good asymptotic convergence property as that of our previously proposed quasi-Bayes adaptation algorithms.

Keywords: speaker adaptation, on-line adaptation, adaptive learning, Bayesian approach, hidden Markov model

1. INTRODUCTION

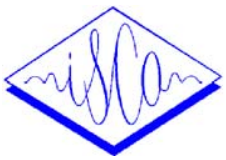
For a Gaussian-mixture continuous density HMM (CDHMM) based automatic speech recognition (ASR) system, adaptive learning of CDHMM parameters from adaptation/testing data provides a good way to reduce the possible acoustic mismatches between the training and testing conditions and thus to enhance the system performance robustness. A good CDHMM adaptation algorithm should have the characteristics of being *incremental*, *adaptive*, and *efficient*. In the past few years, we've developed several such on-line Bayesian adaptive learning algorithms [2, 3, 4]. For the approach in [2], we can adapt all of the CDHMM parameters related to the *observed* speech units in adaptation data. For the approach in [3], except for the covariance matrices, we can adapt not only all of the other CDHMM parameters of the *observed* speech units, but also the *mean vectors* of those *unseen* speech units by exploiting the *correlations* between different mean vectors. In practice, in order to avoid the over-smoothing and to reduce the memory requirement for storing all of the correlation coefficients, we usually disregard those weakly correlated means. Consequently, with the small amount of adaptation data, some of the mean vectors might be untouched. Other possibilities to enhance the adaptation efficiency have to be explored. Recently, we proposed a new adaptation framework [4] aiming at enhancing the efficiency further of the algorithms in [2, 3] in terms of both performance improvement and memory requirement. In this paper, we intend to introduce this new framework to a wider range of audience and present some new experimental results.

2. MULTIPLE-STREAM PRIOR EVOLUTION AND POSTERIOR POOLING FOR EFFICIENT HMM ADAPTATION

The key concept behind our approach is how to appropriately evolve the prior distribution of the CDHMM parameters. In a Bayesian framework, we intend to consider the uncertainty of the HMM parameters Λ by treating them as if they were random. Our prior knowledge about Λ is assumed to be summarized in a known joint *a priori* probability density function (pdf) $p(\Lambda|\varphi^{(0)})$ with *hyperparameters* $\varphi^{(0)}$, where $\Lambda \in \Omega$, Ω denotes an admissible region of the HMM parameter space. Such prior information may, for example, come from subject matter considerations and/or from previous experiences (training data). Let $\mathcal{X}_1^n = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$ be n independent sets of observation samples which are incrementally obtained and used to update our knowledge about Λ. Depending on the different assumptions, there are many ways to *evolve* $p(\Lambda)$. One way is to adopt the recursive Bayesian learning framework:

$$p(\Lambda|\mathcal{X}_1^n) = \frac{p(\mathcal{X}_n|\Lambda) \cdot p(\Lambda|\mathcal{X}_1^{n-1})}{\int_\Omega p(\mathcal{X}_n|\Lambda) \cdot p(\Lambda|\mathcal{X}_1^{n-1}) d\Lambda} \,. \qquad (1)$$

Starting the calculation of posterior pdf from $p(\Lambda|\varphi^{(0)})$, repeated use of equation (1) produces a sequence of densities $p(\Lambda|\mathcal{X}_1^1)$, $p(\Lambda|\mathcal{X}_1^2)$, and so forth. Because of the *missing-data* problem of CDHMM, there are some serious computational difficulties to directly implement this learning procedure [2]. Consequently, some approximations are needed in practice. One such approach called quasi-Bayes (QB) learning was developed in [2, 3]. Based on the concept of *density approximation*, the QB algorithm is designed to incrementally update the hyperparameters on the approximate posterior distribution. Actually, the *density approximation* also opens up the opportunity of appropriately manipulating the posterior distribution as we intend. For example, in order to make our Bayesian learning algorithms truly adaptive, we can introduce some *forgetting mechanisms*, namely



6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

exponential forgetting and *hyperparameter refreshing* as discussed in [2, 3] to adjust the contribution of previously observed sample utterances. Consequently, we will get a posterior distribution $p_{QB}(\Lambda|\mathcal{X}_1^n)$ which is different from the true posterior distribution $p_{true}(\Lambda|\mathcal{X}_1^n)$, but includes the appropriate information we want to learn from the observation data $\mathcal{X}_1^n$. In the following, we show how to use the concept of *prior evolution* and *density approximation* to derive a new *hybrid* adaptation algorithm.

In addition to the above method of prior evolution, we can also assume Λ to evolve in a more *constrained* way as $\Lambda^{(n)} = H_n(\Lambda^{(0)})$ where H_n represents a mapping from $\Lambda^{(0)}$ to $\Lambda^{(n)}$ and can be *incrementally* learned from the observation data $\mathcal{X}_1^n$. Then from $p(\Lambda|\varphi^{(0)})$, we can derive a new posterior distribution $p_{new}(\Lambda|\mathcal{X}_1^n) = p(\Lambda^{(n)})$. We can define the intended posterior distribution $p_{intend}(\Lambda|\mathcal{X}_1^n)$ as

$$p_{intend}(\Lambda|\mathcal{X}_1^n) = \epsilon \times p_{QB}(\Lambda|\mathcal{X}_1^n) + (1-\epsilon) \times p_{new}(\Lambda|\mathcal{X}_1^n) \quad (2)$$

where $0 \leq \epsilon \leq 1$ is a parameter to control the *relative importance* of the above two mixture components. By using the concept of the *density approximation* again here, we can use another distribution, $p(\Lambda|\varphi^{(n)})$, with the same parametric form as $p(\Lambda|\varphi^{(0)})$ but different hyperparameters $\varphi^{(n)}$, to approximate the $p_{intend}(\Lambda|\mathcal{X}_1^n)$. Then we can continue the Bayesian learning process by using Eq. (1). The MAP (maximum *a posteriori*) estimate of the CDHMM parameters derived from the evolving prior distribution can be used to update the speech recognition system.

The above procedure is general enough to extend to the case of *multiple-stream prior evolution and posterior pooling* when more than two streams are considered. It also provides a good tool to exploit respectively the different *knowledge sources* in an appropriate way. Such information can be incorporated into the existing system so that the system can be continuously adapted to the new adaptation data and/or evolve in a desired way. In the next section, as a first step, we consider applying affine transformations to the mean vectors of CDHMM to control their evolution.

3. CONSTRAINED PRIOR EVOLUTION OF THE MEAN VECTORS BY LINEAR TRANSFORMATIONS

Suppose there are M speech units in the recognizer, each being modeled by a Gaussian mixture CDHMM. Consider a collection of such M CDHMM's $\Lambda = \{\lambda_q\}_{q=1,\cdots,M}$, where $\lambda_q = (\pi^{(q)}, A^{(q)}, \theta^{(q)})$ denotes the set of parameters of the q-th N-state CDHMM used to characterize the q-th speech unit, of which, $\pi^{(q)}$ is the initial state distribution, $A^{(q)} = [a_{ij}^{(q)}]$ is the transition probability matrix, and $\theta^{(q)}$ is the parameter vector composed of mixture parameters $\theta_i^{(q)} = \{\omega_{ik}^{(q)}, m_{ik}^{(q)}, \Sigma_{ik}^{(q)}\}$ for each state i with the state observation density being a mixture of multivariate Gaussian pdf's: $p(\mathbf{x}|\theta_i^{(q)}) = \sum_{k=1}^{K} \omega_{ik}^{(q)} \mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$, where the mixture coefficients $\omega_{ik}^{(q)}$'s satisfy the constraint $\sum_{k=1}^{K} \omega_{ik}^{(q)} = 1$, and $\mathcal{N}(\mathbf{x}|m_{ik}^{(q)}, \Sigma_{ik}^{(q)})$ is the k-th normal mixand with $m_{ik}^{(q)}$ being the D-dimensional mean vector and $\Sigma_{ik}^{(q)}$ being the $D \times D$ covariance matrix with its d-th diagonal element being $\sigma_{ik}^{(q)2}(d)$. For notational convenience, it is assumed that all the state observation pdf's have the same number of mixture components.

In this study, we only consider the case of CDHMMs in which the covariance matrices are specified. We define the parameter vector $\mathbf{m}$ to be the collection of the mean vectors of all the Gaussian mixture components of CDHMMs and denoted simply by an operator "vec" as $\mathbf{m} = vec\{m_{ik}^{(q)}\}$. We also define another operator "$block-diag$" to denote a block diagonal matrix, e.g., $\Xi = block-diag\{\Sigma_{ik}^{(q)}\}$, with each diagonal block element to be also a matrix, e.g., $\Sigma_{ik}^{(q)}$. Further denote $\lambda_q' = (\pi_i^{(q)}, a_{ij}^{(q)}, \omega_{ik}^{(q)})$. The initial prior pdf of Λ is assumed to be: $g(\Lambda) = g(\mathbf{m}) \prod_{q=1}^{M} g(\lambda_q')$, where $g(\lambda_q')$ takes the special form of a matrix beta pdf with sets of positive hyperparameters of $\{\eta_i^{(q)}\}, \{\eta_{ij}^{(q)}\}, \{\nu_{ik}^{(q)}\}$, and $g(\mathbf{m}) = \mathcal{N}(\mathbf{m}|\boldsymbol{\mu}(0), \mathbf{U}(0))$ has a joint normal pdf with mean vector $\boldsymbol{\mu}(0) = vec\{\mu_{ik}^{(q)}(0)\}$ and covariance matrix $\mathbf{U}(0)$ [2, 3]. In this study, we only consider the case of multiple stream prior evolution for mean vectors of CDHMMs. All of the other HMM parameters will evolve in a single stream as described in [2, 3].

In the new stream of prior evolution, at time instant n, we assume that mean vectors $m_{ik}^{(q)}(n)$'s have been evolved from the original mean vectors $m_{ik}^{(q)}(0)$'s by linear transformations as follows:

$$m_{ik}^{(q)}(n) = A_{c_1(i,k,q)}^{(n)} m_{ik}^{(q)}(0) + b_{c_2(i,k,q)}^{(n)}$$

where $A_{c_1(i,k,q)}^{(n)}$ is a $D \times D$ matrix and $b_{c_2(i,k,q)}^{(n)}$ is a D-dimensional bias vector. These transformations can be shared by different mean vectors in a very flexible way. $c_1(i,k,q)$ and $c_2(i,k,q)$ represent the class indexs which are the results of two mappings from distinct mixture component labels to the shared transformation class labels. For simplicity, we only study the case of $c_1(i,k,q) = c_2(i,k,q) = c(i,k,q)$ here. As in [5], we use a hierarchical regression tree to define the above mappings by attaching to each node of the tree a distinct transformation. These transformations can be incrementally estimated from $\mathcal{X}_1^n$ by using an approximate maximum likelihood approach described in [5, 1]. In evolving $m_{ik}^{(q)}(n)$'s, the transformation $\{A_{c(i,k,q)}^{(n)}, b_{c(i,k,q)}^{(n)}\}$ will be chosen by traversing the above tree to make sure the most detailed yet reliably estimated one be used. With the above constraints applied to the $m_{ik}^{(q)}$'s, the new stream of prior pdf will evolve as follows:

$$p_{new}(\mathbf{m}|\mathcal{X}_1^n) = \mathcal{N}(\mathbf{m}|\boldsymbol{\mu}_{new}(n), \mathbf{U}_{new}(n))$$

where

$$\boldsymbol{\mu}_{new}(n) = \mathcal{A}(n) \cdot \boldsymbol{\mu}(0) + \mathcal{B}(n)$$
$$\mathbf{U}_{new}(n) = \mathcal{A}(n) \cdot \mathbf{U}(0) \cdot \mathcal{A}^t(n)$$

with $\mathcal{A}(n) = block-diag\{A_{c(i,k,q)}^{(n)}\}$ and $\mathcal{B}(n) = vec\{b_{c(i,k,q)}^{(n)}\}$. Another stream of prior pdf will evolve as described in [2, 3] as follows:

$$p_{QB}(\mathbf{m}|\mathcal{X}_1^n) = \mathcal{N}(\mathbf{m}|\boldsymbol{\mu}_{QB}(n), \mathbf{U}_{QB}(n))$$

By pooling $p_{QB}(\mathbf{m}|\mathcal{X}_1^n)$ and $p_{new}(\mathbf{m}|\mathcal{X}_1^n)$ together as described in Eq.(2), we obtain a mixture of Gaussian pdf's:

$$p_{intend}(\mathbf{m}|\mathcal{X}_1^n) = \epsilon \times p_{QB}(\mathbf{m}|\mathcal{X}_1^n) + (1-\epsilon) \times p_{new}(\mathbf{m}|\mathcal{X}_1^n)$$

We can now use another Gaussian pdf $\mathcal{N}(\mathbf{m}|\boldsymbol{\mu}(n), \mathbf{U}(n))$ to approximate the above pdf $p_{intend}(\mathbf{m}|\mathcal{X}_1^n)$ under the criterion of minimizing the *Kullback-Leibler directed divergence* of $\mathcal{N}(\mathbf{m}|\boldsymbol{\mu}(n), \mathbf{U}(n))$ from $p_{intend}(\mathbf{m}|\mathcal{X}_1^n)$. It can be derived that

$$\begin{aligned}\boldsymbol{\mu}(n) &= \epsilon\boldsymbol{\mu}_{QB}(n) + (1-\epsilon)\boldsymbol{\mu}_{new}(n) \\ \mathbf{U}(n) &= \epsilon\mathbf{U}_{QB}(n) + (1-\epsilon)\mathbf{U}_{new}(n) + \\ &\quad \epsilon(1-\epsilon)(\boldsymbol{\mu}_{QB}(n) - \boldsymbol{\mu}_{new}(n))(\boldsymbol{\mu}_{QB}(n) - \boldsymbol{\mu}_{new}(n))^t\end{aligned}$$

The weight ϵ $(0 \leq \epsilon \leq 1)$ is a monotonically increasing function of the amount of available adaptation data so that a good asymptotic convergence can be achieved.

In the following experiments, as a first step, we ignore the correlations between $m_{ik}^{(q)}$'s, i.e., $\mathbf{U}(0)$ is assumed to be a diagonal covariance matrix. Then, the above formulas can be greatly simplified by treating the evolution of the individual $m_{ik}^{(q)}$ separately. After $u_{ik}^{(q)}(n)$ is calculated, we set all off-diagonal elements to be 0 so that the QB evolution will be started again from a new prior in a consistent way. In this case, we can also use a different $\epsilon_{ik}^{(q)}$ for each $m_{ik}^{(q)}$. In this study, the following sigmoid function is adopted:

$$\epsilon_{ik}^{(q)} = \frac{1}{1 + exp(-\alpha c_{ik}^{(q)} + \beta)} \tag{3}$$

where $c_{ik}^{(q)}$ is the related accumulated "EM count", $\alpha > 0$ and $\beta > 0$ are two control parameters.

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

To examine the viability and the efficacy of the proposed method, a series of experiments for continuous speech recognition of Putonghua (Mandarin Chinese) are performed. The database we used is the HKU96 Putonghua Corpus developed in our laboratory. The HKU96 corpus consists of a total of 20 native Putonghua speakers, 10 females and 10 males, each speaking: (1) all Putonghua syllables in all tones at least once, (2) 11 words of 2 to 4 syllables, (3) 16 digit strings of 4 to 7 digits, (4) 3 sentences of 7 rhymed syllables with /a/, /i/ and /u/ endings respectively, and (5) hundreds of sentences with verbalized punctuation from newspaper text. All speech recording were made in a quite room with a single National Cardioid Dynamic Microphone. Speech was digitized using a Sound Blaster 16 ASP A/D card plugged into a 486 PC at 16-bit accuracy and with a sampling rate of 16KHz. We used 18224 sentences (about 23 hours of raw speech) from 18 speakers (9 females and 9 males) for training. Other two speakers (1 female and 1 male) are used for speaker-independent (SI) testing and speaker adaptation. For testing data, we randomly choose 378 sentences (about 25 minutes of raw speech which includes 4122 syllables or 10351 phones) from the female speaker, and 215 sentences (about 12 minutes of raw speech which includes 2362 syllables or 5788 phones) from the male speaker. The remaining sentences from those two speakers are used for adaptation.

Input speech was first pre-emphasized by a fixed first-order system, $1 - 0.97z^{-1}$, and then grouped into frames of 25ms with a frame shift of 10ms. For each frame, a Hamming window was applied followed by the computation of 12 MFCC's. 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 (CMS) is applied for acoustic normalization both in training and testing.

The baseline system is a speaker independent, cross-syllable-triphone, decision-tree-based tied-state system and is trained by using the HTK2.1 toolkit [6]. The adopted context-independent (CI) phone set consists of 36 phones plus silence. With this phone set definition, there are 8022 triphones in Putonghua. Among them, 5594 triphones are observed in our training data set, with only 4760 triphones each appearing at least 3 times. Each phone is modeled by a left-to-right three-emitting-state Gaussian-mixture CDHMM without state skipping. Each state has 4 Gaussian mixture components with each component having a diagonal covariance matrix. A special three-state CDHMM is also used for silence modeling.

The recognition task is the recognition of 410 Putonghua *base syllables* disregarding tones. The recognition network enforces silence at the start and end of sentences and allows optional silences between syllables. As for syllable language model, a uniform grammar with a syllable perplexity of 411 (i.e., each syllable can be followed by any of the 410 base syllables and silence) is used. All the recognition experiments are performed with the search engine provided by HTK2.1 toolkit.

4.2. Speaker Adaptation Results

In building the baseline system, about 150 linguistic questions are used in decision-tree construction and the relevant thresholds for stopping criterion are adjusted to generate 3019 tied states. For this system, an averaged syllable accuracy of 75.8% over 2 testing speakers is achieved.

Starting from the baseline system, we performed supervised incremental speaker adaptation experiments on two testing speakers by using four different methods, namely,

(1) the incremental MLLR adaptation method in [5, 1],

(2) the incremental QB adaptation method without correlation in [2],

(3) the incremental QB adaptation method with correlation in [3], and

(4) the new adaptation method which includes the evolution of two streams of prior pdf's, i.e., QB without correlation and linear transformation constrained prior.

In the experiments, the regression tree is built and then fixed for all of the Gaussian mixture components of the CDHMMs in the baseline recognition system by using a divisive Gaussian distribution clustering method with a distortion measure being the symmetric divergence measure between two Gaussian distributions. In MLLR and hybrid

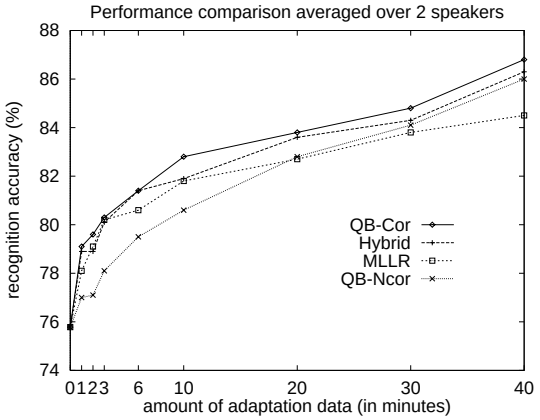


Figure 1: Performance (syllable accuracy in %) comparison averaged over 2 testing speakers as a function of amount (in minutes) of available adaptation data per speaker for four on-line adaptation methods: QB with correlation (QB-Cor), new hybrid method (Hybrid), MLLR, QB without correlation (QB-Ncor)

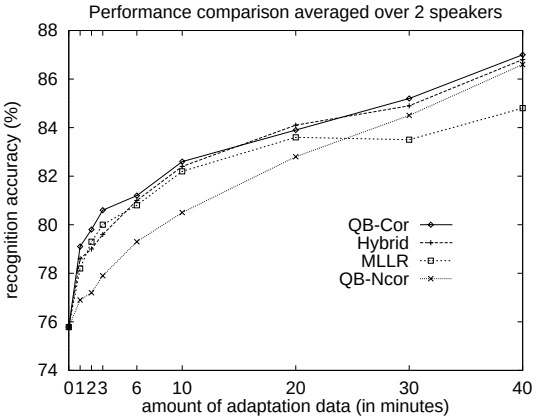


Figure 2: A similar performance comparison as in Fig. 1 by running the adaptation algorithms in batch mode

adaptation, different number (utmost 256) of affine transformations are adaptively chosen based on the amount of available adaptation data. In [4], full affine transformations are used. In this study, an improved performance is achieved by using block-diagonal transformations of each having three blocks corresponding to static features, their delta, and delta-delta versions. The two control parameters in Eq.(3) are chosen as $\alpha = 1.0, \beta = 1.0$. In QB adaptation, the prior is evolved sentence by sentence. However, in incremental MLLR estimation of the affine transformations, an updating interval of 30 seconds of speech is used. In QB adaptation of correlated CDHMMs, the correlation neighborhood size is chosen to be 8 (see explanation in [3]).

Figure 1 shows the performance (syllable accuracy in %) comparison averaged over 2 testing speakers as a function of amount of available adaptation data (in terms of minutes of raw speech) among the above four adaptation methods. The experimental results confirm our expectation, i.e., the new hybrid algorithm

- achieves a similar fast-adaptation performance as that of incremental MLLR in the case of small amount of adaptation data, while
- maintains the good asymptotic convergence property as that of the original QB algorithm.

In the current experiments, we observed that the QB adaptation method with correlation achieves the best overall performance. Figure 2 shows a similar performance comparison by running the above four adaptation algorithms in batch mode. The same conclusion can be drawn.

5. DISCUSSION AND CONCLUSION

In this study, we propose a new incremental adaptive Bayesian learning framework for efficient on-line adaptation of the CDHMM parameters. In a series of comparative experiments, we show that the new method has a better behavior as desired for a good adaptation algorithm than the methods of on-line MLLR and QB adaptation without correlation. In conclusion, the QB adaptation algorithm with correlation and the new adaptation method proposed in this paper are good candidates for efficient adaptation of CDHMM parameters. The former usually requires more memory than the latter. The new framework of *multiple-stream prior evolution and posterior pooling* opens up many new research opportunities. We are currently studying other ways of prior evolution under different constraints and on different HMM parameters. We believe that the best setup will depend on the nature of the specific applications.

REFERENCES

[1] V. Digalakis, "On-line adaptation of hidden Markov models using incremental estimation algorithms," *Eurospeech-97* (Rhodes, Greece), 1997, pp.1859-1862.

[2] Q. Huo and C.-H. Lee, "On-line adaptive learning of the continuous density hidden Markov model based on approximate recursive Bayes estimate," *IEEE Trans. on Speech and Audio Processing*, Vol. 5, No. 2, pp.161-172, 1997.

[3] Q. Huo and C.-H. Lee, "On-line adaptive learning of the correlated continuous density hidden Markov models for speech recognition," *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 4, pp.386-397, 1998.

[4] Q. Huo and B. Ma, "A new CDHMM adaptation method: being incremental, adaptive and more efficient," *Proc. 1998 Int. Symp. on Chinese Spoken Language Processing*, Singapore, Dec. 1998, pp.71-74.

[5] C. J. Leggetter and P. C. Woodland, "Flexible speaker adaptation for large vocabulary speech recognition," *Eurospeech-95* (Madrid, Spain), 1995, pp.1155-1158.

[6] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book* (for HTK Version 2.1), Cambridge University, 1997.