AN EM ALGORITHM FOR TRAINING WIDEBAND ACOUSTIC MODELS FROM MIXED-BANDWIDTH TRAINING DATA

Michael L. Seltzer and Alex Acero

{mseltzer, alexac}@microsoft.com
Microsoft Research
1 Microsoft Way
Redmond, WA 98052, USA

ABSTRACT

One serious difficulty in the deployment of wideband speech recognition systems for new tasks is the expense in both time and cost of obtaining sufficient training data. A more economical approach is to collect telephone speech and then restrict the application to operate at the telephone bandwidth. However, this generally results in suboptimal performance compared to a wideband recognition system. In this paper, we propose a novel EM algorithm in which wideband acoustic models are trained using a small amount of wideband speech augmented by a larger amount of narrowband speech. Experiments performed using wideband speech and telephone speech demonstrate that the proposed mixed-bandwidth training algorithm results in significant improvements in recognition accuracy over conventional training strategies when the amount of wideband data is limited.

1. INTRODUCTION

The deployment of speech recognition systems for new tasks is often hindered by a lack of sufficient training. Collecting ample training data is especially problematic for applications that process wideband speech. For example, a large vocabulary desktop dictation system requires a large corpus of wideband training data. However, there are many resource-poor languages for which such a corpus does not exist. A similar lack of training data inevitably occurs as new tasks arise. For example, the amount of available wideband training data for automatic meeting transcription is currently very limited [1].

In such cases, collecting a sufficient amount of wideband training data may be prohibitively expensive and time-consuming. Alternatively, recording speech over the telephone is a relatively economical and efficient way to collect large amounts of data from a wide variety of geographic regions. However, this requires that the deployed speech recognizer process narrowband speech. This is suboptimal, as narrowband recognition systems generally perform worse than those that process wideband speech [2].

In [3], we proposed an alternative approach in which wideband acoustic models are trained using a small amount of wideband speech and a large amount of narrowband speech. In this work, a front-end processing stage called Feature Bandwidth Extension (FBE) was used to convert narrowband feature vectors into wideband features vectors. These estimated wideband features were then pooled with available wideband data to train the recognizer in the conventional manner. While moderate improvements over conventional training methods were obtained, this approach had some significant drawbacks. Because FBE only generates point estimates of the wideband features, the subsequent training algorithm implicitly assumed that these wideband feature estimates were as representative of the wideband speech as the actual wideband data. Because the estimation error was not accounted for, the resulting model parameters were suboptimal.

In this paper, we present a principled EM algorithm for training a wideband speech recognizer using mixed-bandwidth training data. In the proposed approach, the wideband model parameters are iteratively updated using both wideband and narrowband speech data. This is accomplished by treating the spectral components missing from the narrowband speech as additional hidden variables. By training the recognizer in this way, we overcome the drawbacks of the previous FBE method, and obtain wideband acoustic models that perform significantly better than fully trained narrowband models or wideband models trained from limited wideband data. Yet, because the proposed algorithm only requires a modest amount of wideband training data, we are still able to avoid the large costs associated with collecting large amounts of wideband speech.

The methods proposed in this paper are related to previous research in training mixture models from incomplete feature vectors [5]. However, this work is not directly applicable to speech recognition applications because of the idiosyncrasies of the feature extraction process, namely the computation of mel-frequency cepstral coefficients. Missing data techniques have also been used to improve the robustness of ASR systems to additive noise for decoding, e.g. [6, 7].

The remainder of the paper is organized as follows. In Section 2, we review the feature extraction process and introduce the missing data paradigm for mixed-bandwidth speech. We then show how to train a Gaussian mixture model from log mel spectral features using mixed-bandwidth training data in Section 3. We then describe the modifications required to generate models of cepstral features in Section 4 and show how the proposed algorithm can be used to train a large vocabulary HMM-based speech recognition system in Section 5. Section 6 describes a series of experiments that show the efficacy of the proposed method. Finally, we present some conclusions in Section 7.

2. FEATURE EXTRACTION FOR ASR

In this work, we assume that mel-frequency cepstral coefficients (MFCC) are the features used for recognition. For wideband data sampled at a 16 kHz, the log mel spectral coefficients represent the

0-7803-9479-8/05/$20.00 © 2005 IEEE

energy in a series of overlapping frequency regions which range from approximately 100 Hz to 8 kHz. This log mel spectral vector $\mathbf{x}$ is then converted to a cepstral vector $\mathbf{z}$ via a DCT as

$$\mathbf{z} = \mathbf{C}\mathbf{x}. \tag{1}$$

where $\mathbf{C}$ is the DCT matrix. Dimensionality reduction is also usually performed, so the DCT matrix $\mathbf{C}$ is $M \times L$ with $M \leq L$. Typically, 13-dimensional cepstra are computed from 20-40-dimensional log mel spectral vectors.

We assume that the narrowband speech has been upsampled to match the sampling rate of the wideband speech. If this speech is then transformed to a sequence of log mel spectral vectors, the components derived from mel filters that cover frequencies outside the original signal bandwidth will contain no information. We refer to these components as *missing*. In contrast, the components of the spectral vector that do contain reliable content are considered *observed*. Thus, a log mel spectral vector $\mathbf{x}$ can be partitioned as

$$\mathbf{x} = [\mathbf{x}^{\mathrm{o},T}\ \mathbf{x}^{\mathrm{m},T}]^T \tag{2}$$

where $\mathbf{x}^{\mathrm{o}}$ contains all components of $\mathbf{x}$ that are observed and $\mathbf{x}^{\mathrm{m}}$ contains all components that are missing. For wideband speech originally sampled at the target sampling rate, $\mathbf{x}^{\mathrm{o}} = \mathbf{x}$ and $\mathbf{x}^{\mathrm{m}} = [\,]$, i.e. there are no missing components.

In a similar manner, we can express a cepstral vector $\mathbf{z}$ as the sum of linear transformations of $\mathbf{x}^{\mathrm{o}}$ and $\mathbf{x}^{\mathrm{m}}$. Decomposing the DCT matrix into two sub-matrices, $\mathbf{C}^{\mathrm{o}}$, an $M \times L^o$ matrix, where L^o is the length of $\mathbf{x}^{\mathrm{o}}$ and $\mathbf{C}^{\mathrm{m}}$, an $M \times L^m$ matrix, where L^m is the length of $\mathbf{x}^{\mathrm{m}}$, we can write

$$\mathbf{z} = \mathbf{C}\mathbf{x} = [\mathbf{C}^{\mathrm{o}}\mathbf{C}^{\mathrm{m}}] \begin{bmatrix} \mathbf{x}^{\mathrm{o}} \\ \mathbf{x}^{\mathrm{m}} \end{bmatrix} = \mathbf{C}^{\mathrm{o}}\mathbf{x}^{\mathrm{o}} + \mathbf{C}^{\mathrm{m}}\mathbf{x}^{\mathrm{m}} = \mathbf{z}^{\mathrm{o}} + \mathbf{z}^{\mathrm{m}} \tag{3}$$

3. TRAINING A GAUSSIAN MIXTURE MODEL ON MIXED BANDWIDTH LOG SPECTRA

We are interested in training an HMM-based speech recognizer using cepstral features derived from mixed-bandwidth speech data. However, for simplicity, we begin by first discussing how to train a Gaussian mixture model (GMM) from mixed-bandwidth log mel spectral features. A GMM has the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|k)p(k) = \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(k) \tag{4}$$

where $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ and $p(k)$ are the mean vector, covariance matrix and prior probability of the kth Gaussian mixture component, respectively.

We seek to train this model using a combination of narrowband and wideband speech data using EM [4]. To do so, we use hidden variables to represent the unseen log mel spectral components $\mathbf{x}^{\mathrm{m}}$ in the narrowband training samples. Thus, we start with the following EM auxillary function

$$Q(\lambda, \hat{\lambda}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \int \log(p(\mathbf{x}_i^{\mathrm{o}}, \mathbf{x}^{\mathrm{m}}, k; \lambda))p(\mathbf{x}^{\mathrm{m}}, k|\mathbf{x}_i^{\mathrm{o}}; \hat{\lambda})d\mathbf{x}^{\mathrm{m}} \tag{5}$$

where i is the frame index, k is the hidden state variable indicating the Gaussian index, λ is the set of model parameters we seek to optimize and $\hat{\lambda}$ is the current estimate of these parameters.

In order to perform EM using (5), we need to factorize $p(\mathbf{x}|k)$ into its conditional and marginal densities as

$$p(\mathbf{x}|k) = p(\mathbf{x}^{\mathrm{o}}, \mathbf{x}^{\mathrm{m}}|k) = p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, k)p(\mathbf{x}^{\mathrm{o}}|k). \tag{6}$$

The marginal distribution can be expressed as

$$p(\mathbf{x}^{\mathrm{o}}|k) = \mathcal{N}(\mathbf{x}^{\mathrm{o}}; \boldsymbol{\mu}_k^{\mathrm{o}}, \boldsymbol{\Sigma}_k^{\mathrm{oo}}) \tag{7}$$

where $\boldsymbol{\mu}_k^{\mathrm{o}}$ and $\boldsymbol{\Sigma}_k^{\mathrm{oo}}$ are the mean and covariance of the observed components only. The conditional distribution can be expressed as

$$p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, k) = \mathcal{N}(\mathbf{x}^{\mathrm{m}}; \boldsymbol{\mu}_k^{\mathrm{m|o}}, \boldsymbol{\Sigma}_k^{\mathrm{m|o}}) \tag{8}$$

where $\boldsymbol{\mu}_k^{\mathrm{m|o}}$ and $\boldsymbol{\Sigma}_k^{\mathrm{m|o}}$ are the conditional mean and covariance, respectively, computed as

$$\boldsymbol{\mu}_k^{\mathrm{m|o}} = \boldsymbol{\mu}_k^{\mathrm{m}} + \boldsymbol{\Sigma}_k^{\mathrm{mo}}\boldsymbol{\Sigma}_k^{\mathrm{oo},-1}(\mathbf{x}^{\mathrm{o}} - \boldsymbol{\mu}_k^{\mathrm{o}}) \tag{9}$$

$$\boldsymbol{\Sigma}_k^{\mathrm{m|o}} = \boldsymbol{\Sigma}_k^{\mathrm{mm}} - \boldsymbol{\Sigma}_k^{\mathrm{mo}}\boldsymbol{\Sigma}_k^{\mathrm{oo},-1}\boldsymbol{\Sigma}_k^{\mathrm{om}} \tag{10}$$

where $\boldsymbol{\mu}_k^{\mathrm{m}}$ and $\boldsymbol{\Sigma}_k^{\mathrm{mm}}$ are the mean and covariance of the unobserved components, respectively, and $\boldsymbol{\Sigma}_k^{\mathrm{mo}}$ and $\boldsymbol{\Sigma}_k^{\mathrm{om}}$ are the partitions of $\boldsymbol{\Sigma}_k$ that represent the covariance between the missing and observed components. For a derivation of these expressions, see [8].

Using these expressions, we can derive the update equations for the Gaussian mixture model parametes. The derivation is omitted due to space considerations. The updated prior probability of the kth Gaussian can be expressed as

$$p(k)^{new} = \frac{1}{N}\sum_{i=1}^{N} p(k|\mathbf{x}_i^{\mathrm{o}}) \tag{11}$$

where $p(k|\mathbf{x}_i^{\mathrm{o}})$ is the posterior probability of the kth Gaussian based only on the observed components of each feature vector. Recall that for wideband speech, because all log spectral components are observed, the posterior probabilities are computed from the full feature vector, i.e. $p(k|\mathbf{x}_i^{\mathrm{o}}) = p(k|\mathbf{x}_i)$.

To derive the update formulas for $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, we first define $\tilde{\mathbf{x}}_{ik}$ as

$$\tilde{\mathbf{x}}_{ik} = \begin{cases} \mathbf{x}_i & \text{if frame } i \text{ is wideband} \\ \begin{bmatrix} \mathbf{x}_i^{\mathrm{o}} \\ \hat{\boldsymbol{\mu}}_{ik}^{\mathrm{m|o}} \end{bmatrix} & \text{if frame } i \text{ is narrowband} \end{cases} \tag{12}$$

where $\hat{\boldsymbol{\mu}}_{ik}^{\mathrm{m|o}}$ is computed from (9). We can then express the mean update formula as

$$\boldsymbol{\mu}_k^{new} = \frac{\sum_{i=1}^{N} p(k|\mathbf{x}_i^{\mathrm{o}})\tilde{\mathbf{x}}_{ik}}{\sum_{i=1}^{N} p(k|\mathbf{x}_i^{\mathrm{o}})} \tag{13}$$

Thus, the mean update expression is similar to that of a conventional GMM, except that the missing vector components of each narrowband frame are replaced by the state-conditional posterior means.

The covariance update is also similar to that of a conventional GMM. We can express the covariance update formula as

$$\boldsymbol{\Sigma}_k^{new} = \frac{\sum_{i=1}^{N} p(k|\mathbf{x}_i^{\mathrm{o}})\left((\tilde{\mathbf{x}}_{ik} - \boldsymbol{\mu}_k)(\tilde{\mathbf{x}}_{ik} - \boldsymbol{\mu}_k)^T + \tilde{\boldsymbol{\Sigma}}_k^{\mathrm{m|o}}\right)}{\sum_{i=1}^{N} p(k|\mathbf{x}_i^{\mathrm{o}})} \tag{14}$$

where

$$\tilde{\boldsymbol{\Sigma}}_k^{\mathrm{m|o}} = \begin{cases} \mathbf{0} & \text{if frame } i \text{ is wideband} \\ \begin{bmatrix} \mathbf{0}^{\mathrm{oo}} & \mathbf{0}^{\mathrm{om}} \\ \mathbf{0}^{\mathrm{om}} & \hat{\boldsymbol{\Sigma}}_k^{\mathrm{m|o}} \end{bmatrix} & \text{if frame } i \text{ is narrowband} \end{cases} \quad (15)$$

The state-dependent conditional covariance $\hat{\boldsymbol{\Sigma}}_k^{\mathrm{m|o}}$ in (15) is computed from (10) and padded with appropriately-sized zero matrices to create $\tilde{\boldsymbol{\Sigma}}_k^{\mathrm{m|o}}$. This additional covariance term reflects the uncertainty associated with the absence of these components in the narrowband training vectors.

4. WORKING WITH CEPSTRAL PARAMETERS

In the previous section, it was assumed that the components of the feature vector could be partitioned into observed and missing subvectors. However, most speech recognition systems process *cepstral* features, obtained from the log mel spectra via a DCT. Because of the DCT operation, each cepstral coefficient is a linear combination of *all* log mel spectral features, and as a result, the cepstral vector cannot be partitioned into observed and missing components.

In order to perform the required marginalization, it is necessary to transform the cepstral model parameters back to the log mel spectral domain. If no truncation occurs, this transformation can be done trivially via an IDCT. However, most speech recognizers generate cepstra using a truncated DCT. As a result, the log mel spectral covariance matrices obtained from cepstral covariance matrices via an IDCT are rank-deficient. Specifically, if the DCT matrix is $M \times L$ with $M < L$, then the log mel spectral covariance matrix is an $L \times L$ matrix with at most rank M. This is problematic because the covariance matrix must be full rank in order for it to be invertible.

Therefore, we need to increase the rank of the cepstral model parameters, but must do so without affecting the discriminability of the models. To do so, we create L-dimensional cepstral models by augmenting the truncated M-dimensional cepstral models with an additional R dimensions, where $R = L - M$. The augmented cepstral models are of rank L, as are the resulting log mel spectral parameters. If we use the same R-dimensional parameters to augment all models, we ensure that the additional dimensions will have no effect on the computation of the posterior probabilities.

We define $\boldsymbol{\nu}_k$ and $\boldsymbol{\Phi}_k$ to be the the M-dimensional cepstral domain mean and variance, respectively, for the kth Gaussian in the mixture. The cepstral means $\boldsymbol{\nu}_k$ can be transformed into the log mel spectral domain as

$$\boldsymbol{\mu}_k = \mathbf{C}^{-1} \begin{bmatrix} \boldsymbol{\nu}_k \\ \boldsymbol{\nu}_R \end{bmatrix} = \begin{bmatrix} \mathbf{C}_M^{-1} \mathbf{C}_R^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\nu}_k \\ \boldsymbol{\nu}_R \end{bmatrix} = \mathbf{C}_M^{-1} \boldsymbol{\nu}_k + \mathbf{b} \quad (16)$$

where $\boldsymbol{\nu}_R$ represents the vector used to augment the cepstral means. $\mathbf{C}_M^{-1}$ represents the first M columns of the $L \times L$ IDCT matrix, $\mathbf{C}_R^{-1}$ represents the last R columns of this matrix, and $\mathbf{b} = \mathbf{C}_R^{-1} \boldsymbol{\nu}_R$.

Similarly, we can transform the cepstral variances to the log spectral domain as

$$\begin{aligned} \boldsymbol{\Sigma}_k &= \begin{bmatrix} \mathbf{C}_M^{-1} \mathbf{C}_R^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Phi}_k & \mathbf{0^T} \\ \mathbf{0} & \boldsymbol{\Phi}_R \end{bmatrix} \begin{bmatrix} \mathbf{C}_M^{-T} \\ \mathbf{C}_R^{-T} \end{bmatrix} \\ &= \mathbf{C}_M^{-1} \boldsymbol{\Phi}_k \mathbf{C}_M^{-T} + \mathbf{A} \end{aligned} \quad (17)$$

where $\boldsymbol{\Phi}_R$ is the $R \times R$ augmented covariance matrix, $\mathbf{0}$ is a $R \times M$ zero matrix, and $\mathbf{A} = \mathbf{C}_R^{-1} \boldsymbol{\Phi}_R \mathbf{C}_R^{-T}$. Both $\boldsymbol{\Phi}_k$ and $\boldsymbol{\Phi}_R$ are assumed to be diagonal, although (17) does not require it.

In this work, $\boldsymbol{\nu}_R$ and $\boldsymbol{\Phi}_R$ are computed by computing the global mean and covariance of L-dimensional cepstra from the wideband training data, and extracting the last R dimensions of these parameters.

Thus, in order to train a cepstral domain GMM from mixed bandwidth training data, the transformations shown in (16) and (17) are applied to all cepstral model parameters at the beginning of each EM iteration. Using the resulting log mel spectral parameters, the model parameters updates can be computed according to (11), (13), and (14). At the end of each iteration, a truncated DCT is used to transform the updated model parameters back to the cepstral domain.

5. HMM TRAINING WITH MIXED BANDWITH DATA

The proposed algorithm for training a GMM using mixed-bandwidth speech data can be readily extended to HMM training. In fact, the only change required is to replace the posterior probability $p(k|\mathbf{x}_i^{\mathrm{o}})$ with γ_{ikq}, the posterior probability of the qth Gaussian in HMM state k given the observation *sequence* $\mathcal{X}^{\mathrm{o}} = \{\mathbf{x}_1^{\mathrm{o}} \ldots \mathbf{x}_N^{\mathrm{o}}\}$. In our case, γ_{ikq} is defined as

$$\gamma_{ikq} = \frac{\alpha_{ik}\beta_{ik}}{\sum_{k'=1}^{K} \alpha_{ik'}\beta_{ik'}} \frac{p(\mathbf{x}_i^{\mathrm{o}}|k,p)c_{kq}}{\sum_{q'=1}^{Q} p(\mathbf{x}_i^{\mathrm{o}}|k,q')c_{kq'}} \quad (18)$$

where α_{ik} and β_{ik} are the conventional forward and backward variables, c_{kq} is the mixture weight of the qth Gaussian in state k, and $p(\mathbf{x}_i^{\mathrm{o}}|k,q) = \mathcal{N}(\mathbf{x}_i^{\mathrm{o}}; \boldsymbol{\mu}_{kq}^{\mathrm{o}}, \boldsymbol{\Sigma}_{kq}^{\mathrm{oo}})$, the likelihood of the given Gaussian measured using the observed components only.

While this is mathematically the only change required to apply the proposed mixed-bandwidth training algorithm to HMMs, some practical issues limit its direct application. In practice, when training a large vocabulary speech recognizer, there are many HMM states that have low occupancy counts, i.e. there are only a few observations which contribute to the sufficient statistics of that state. In such states, the covariance matrix $\boldsymbol{\Sigma}_{kq}^{\mathrm{oo}}$ obtained after marginalization is frequently rank-deficient, and thus, cannot be inverted. Because of this, the state posterior γ_{ikq} and the state-conditional posterior distribution $p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}},k,q)$, which both depend on $\boldsymbol{\Sigma}_{kq}^{\mathrm{oo},-1}$, cannot be computed.

5.1. Using Globally Shared Wideband Posterior Distributions

In cases of data sparseness such as this one, one method of improving the robustness of such calculations is to share data among different HMM states. In this work, this sharing is performed by assuming that the state-conditional posterior distribution of the wideband features can be approximated by a single global distribution shared by all states, i.e. we assume $p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}},k,q) \approx p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}})$. Thus, the posterior distribution is conditioned only on the observation but no longer on the state.

For each frame of narrowband speech $\mathbf{x}_i^{\mathrm{o}}$, we obtain this distribution using a front-end processing stage. Using a GMM that has been trained on the available wideband cepstra, a single E-step of the training algorithm described in Sections 3 and 4 is performed. This generates the state posterior probability $p(k|\mathbf{x}_i^{\mathrm{o}})$, and the posterior distribution $p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}},k) = \mathcal{N}(\boldsymbol{\mu}_{ik}^{\mathrm{m|o}}, \boldsymbol{\Sigma}_k^{\mathrm{m|o}})$ for each Gaussian k. The global distribution $p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}})$ is then obtained by

computing the first and second moments of $p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}}, k)$ and marginalizing over all Gaussians, as

$$E[\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}}] = \sum_k p(k|\mathbf{x}_i^{\mathrm{o}})\boldsymbol{\mu}_{ik}^{\mathrm{m|o}} \tag{19}$$

$$E[\mathbf{x}^{\mathrm{m}}\mathbf{x}^{\mathrm{m},T}|\mathbf{x}_i^{\mathrm{o}}] = \sum_k p(k|\mathbf{x}_i^{\mathrm{o}}) \left(\boldsymbol{\Sigma}_k^{\mathrm{m|o}} + \boldsymbol{\mu}_{ik}^{\mathrm{m|o}}\boldsymbol{\mu}_{ik}^{\mathrm{m|o},T}\right) \tag{20}$$

The mean and covariance of the global posterior distribution for frame i can then be easily computed from these parameters. Note that we can convert these parameters to the cepstral domain in order to obtain a globally shared *cepstral* posterior distribution $p(\mathbf{z}|\mathbf{x}_i^{\mathrm{o}}) = \mathcal{N}(\mathbf{z}; \hat{\mathbf{z}}_i, \boldsymbol{\Gamma}_i)$ where

$$\hat{\mathbf{z}}_i = \mathbf{C}^{\mathrm{o}}\mathbf{x}_i^{\mathrm{o}} + \mathbf{C}^{\mathrm{m}}\left(E[\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}}]\right) \tag{21}$$

$$\boldsymbol{\Gamma}_i = \mathbf{C}^{\mathrm{m}}\left(E[\mathbf{x}^{\mathrm{m}}\mathbf{x}^{\mathrm{m},T}|\mathbf{x}_i^{\mathrm{o}}] - E[\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}}]E[\mathbf{x}^{\mathrm{m}}|\mathbf{x}_i^{\mathrm{o}}]^T\right)\mathbf{C}^{\mathrm{m},T} \tag{22}$$

We assume that $\boldsymbol{\Gamma}_i$ is diagonal.

Notice that whereas the posterior mean for frame i was previously a function of both the state k and the observation $\mathbf{x}_i^{\mathrm{o}}$, it is now a function of the observation only. Additionally, the marginalization operation has resulted in a posterior variance which is now strictly a function of the observation, and not the state. Creating globally shared posterior distributions in this manner requires slight changes to the HMM update formulae. These will be detailed in Section 5.3.

5.2. Computing State Posteriors for the Narrowband Data

As (18) shows, the marginalized log spectral distributions $p(\mathbf{x}_i^{\mathrm{o}}|k, q)$ are required in order to compute γ_{ikq} for the narrowband data. However, as mentioned previously, there are many states in which $\boldsymbol{\Sigma}_{kq}^{\mathrm{oo}}$ is not invertible. Even in cases where it is invertible, performing Gaussian evaluation in the log spectral domain requires significantly more computation than in the cepstral domain where diagonal covariances can be used. For large training corpora, this increased computation may be prohibitively expensive.

In order to efficiently and robustly compute the state posteriors, we convert the marginalized log spectral models back to the cepstral domain using a $M \times L^{\mathrm{o}}$ DCT matrix $\mathbf{D}$, where L^{o} is the number of observed log spectral components. Thus, recalling (16) and (17), the narrowband cepstral model parameters are obtained from the wideband model parameters as

$$c_{qk}^{\mathrm{nb}} = c_{qk} \tag{23}$$

$$\boldsymbol{\nu}_{qk}^{\mathrm{nb}} = \mathbf{DP}\left(\mathbf{C}_M^{-1}\boldsymbol{\nu}_{qk} + \mathbf{b}\right) \tag{24}$$

$$\boldsymbol{\Phi}_{qk}^{\mathrm{nb}} = \mathbf{DP}\left(\mathbf{C}_M^{-1}\boldsymbol{\Phi}_k\mathbf{C}_M^{-T} + \mathbf{A}\right)\mathbf{P}^T\mathbf{D}^T \tag{25}$$

where $\mathbf{P}$ is an $L^{\mathrm{o}} \times L$ matrix which selects the observed components.

Thus, we can now compute the HMM state posterior probabilities in the cepstral domain for narrowband data. Of course, the narrowband log spectra must be converted to cepstra as $\mathbf{z}_i^{\mathrm{nb}} = \mathbf{D}\mathbf{x}_i^{\mathrm{o}}$ in order to do so.

5.3. Implementation Details

The training data for the proposed mixed-bandwidth training algorithm now consists of a sequence of wideband cepstra computed from the available wideband speech, a sequence of narrowband cepstra computed from the narrowband speech, and a wideband cepstral posterior distribution $p(\mathbf{z}|\mathbf{x}_i^{\mathrm{o}}) = \mathcal{N}(\mathbf{z}; \hat{\mathbf{z}}_i, \boldsymbol{\Gamma}_i)$ for each frame of narrowband speech. Incorporating this data into the proposed mixed-bandwidth EM algorithm, we can rewrite the HMM update formulae as

$$c_{kq} = \frac{\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq} + \sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq}^{\mathrm{nb}}}{\sum_{q'=1}^{Q}\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq'} + \sum_{q'=1}^{Q}\sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq'}^{\mathrm{nb}}} \tag{26}$$

$$\boldsymbol{\nu}_{kq} = \frac{\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq}\mathbf{z}_i + \sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq}^{\mathrm{nb}}\hat{\mathbf{z}}_j}{\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq} + \sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq}^{\mathrm{nb}}} \tag{27}$$

$$\boldsymbol{\Phi}_{kq} = \frac{\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq}(\mathbf{z}_i - \boldsymbol{\nu}_{kq}) + \sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq}^{\mathrm{nb}}((\hat{\mathbf{z}}_j - \boldsymbol{\nu}_{kq})^2 + \boldsymbol{\Gamma}_j)}{\sum_{i=1}^{N^{\mathrm{wb}}}\gamma_{ikq} + \sum_{j=1}^{N^{\mathrm{nb}}}\gamma_{jkq}^{\mathrm{nb}}} \tag{28}$$

where i indexes the wideband data, j indexes the narrowband data, γ_{ikq} is the posterior probability of a wideband cepstral vector $\mathbf{z}_i$ computed using the wideband models and $\gamma_{jkq}^{\mathrm{nb}}$ is the posterior probability of a narrowband cepstral vector $\mathbf{z}_j^{\mathrm{nb}}$, computed using the narrowband models obtained using (23)–(25). N^{wb} is the total number of wideband observations and N^{nb} is the total number of narrowband observations.

Training is performed as follows. Using the wideband cepstra, an initial wideband HMM, typically a monophone model with a single Gaussian per state, is trained using the conventional Baum-Welch algorithm. At this point, mixed-bandwidth training proceeds by splitting the accumulation of the sufficient statistics into two parts, as implied by (26)–(28). In the first part, the sufficient statistics are accumulated using the wideband models and the wideband cepstra in the usual manner. In the second part, the state posterior probabilities $\gamma_{jkq}^{\mathrm{nb}}$ are computed using the narrowband cepstra and the narrowband models generated from the current wideband models. Using these state posteriors, the sufficient statistics are accumulated using the wideband posterior means $\hat{\mathbf{z}}_j$ and variances $\boldsymbol{\Gamma}_j$. Once all the wideband and narrowband data have been processed, the sufficient statistics computed by each part are aggregated to compute the updated wideband model parameters. From this model, a new updated narrowband model is produced and the process is repeated until the convergence.

6. EXPERIMENTAL EVALUATION

In order to evaluate the proposed mixed-bandwidth training algorithm, we performed a series of experiments using the Wall Street Journal (WSJ0) corpus [9]. In order to perform controlled experiments in which the proportion of wideband to narrowband data is the only variable, we created a parallel telephony training corpus by passing the WSJ0 training set through a telephony filter designed to the G.712 specifications. The useful bandwidth of the telephony speech was assumed to be 300-3400 Hz.

The HTK speech recognition system was used to train 3-state context-dependent triphone models with 24 Gaussians per state. The feature vectors used for recognition were 13-dimensional cepstral vectors derived from 40-dimensional log mel spectra, along with their delta and acceleration parameters. Frames were 25 ms in duration with a 10 ms shift between successive frames. Cepstral mean normalization was performed prior to processing. A trigram language model was used for decoding. The speech recognizer was trained using the SI84 training set. Performance was measured

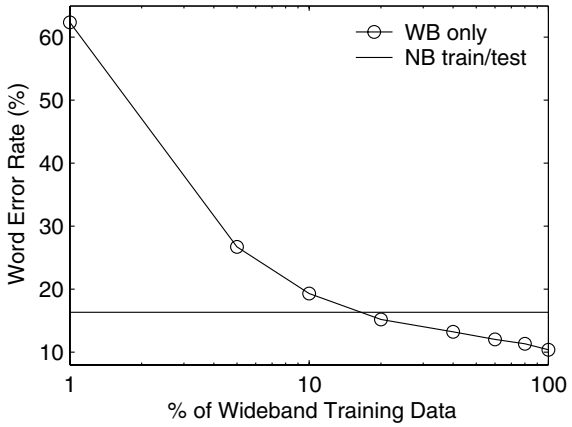

Fig. 1. Word Error Rate (WER) of the WSJ0 20k test set set versus the amount of data used to train the recognizer. The leftmost data point represents 1% of the total training set (0.12 hrs) while the rightmost datapoint represents the full training set (12.0 hrs). The figure also shows the WER obtained by a fully trained narrowband recognition system.

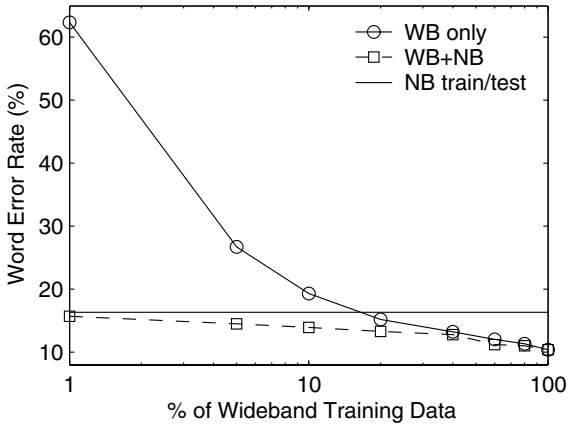

Fig. 2. WER of the WSJ0 20k test set using the proposed mixed-bandwidth EM training algorithm as a function of the amount of wideband data available. For comparison, the WERs obtained by a recognizer trained from limited wideband data only and by a fully trained narrowband recognizer are also shown.

using the WSJ0 20k test set, which consists of 333 utterances (approximately 42 for each of 8 speakers), and covers a 20,000 word vocabulary.

In the first series of experiments, we evaluated the wideband recognition performance when various amounts of wideband speech were used for training. The complete training set consists of approximately 12 hours of speech. Subsets of the training set, from 1% up to 80% of the total training set were selected at random, and used to train the recognizer. Figure 1 shows the resulting Word Error Rate (WER) as a function of the amount of data used for training. Note that the x-axis of the figure is displayed on a logarithmic scale. The rightmost point of 10.4% represents the WER when the full training set is used. This is the upper bound on performance in this experimental framework. For comparison, the figure also shows the WER obtained by a narrowband recognition system trained using the full training set. Not surprisingly, the figure shows that the performance of the wideband system degrades significantly with fewer training data. As the amount of wideband training data falls below 20% (approximately 2.4 hours of speech), better performance is obtained from a fully trained narrowband system.

6.1. Experiments with Telephony Speech

We will now attempt to improve the performance of wideband speech recognition systems when the wideband data are limited. We assume that only a limited percentage of the wideband training data is available and that the remainder of the training corpus is available as telephone-bandwidth speech. Telephone speech that is upsampled to 16 kHz and converted to 40-dimensional log mel spectral features has 17 out of 40 components that fall outside of the telephone passband. Specifically, the first 4 and last 13 components of the log mel spectral vectors are unobserved.

In order to generate the wideband posterior distribution for each frame of telephone speech, a GMM was trained from 39-dimensional cepstral vectors using the available wideband training data. Using this GMM, a wideband cepstral posterior distribution was estimated for each narrowband training vector. Mean normalization was performed on both the wideband cepstra used to train the GMM and the narrowband log mel spectra prior to processing in order to mitigate the spectral tilt induced by the telephone channel.

For a given wideband/narrowband partition of the training utterances, a wideband HMM was trained according to the procedure described in Section 5.3, and then used to decode the WSJ0 20k test set. This experiment was performed for partitions of the training set in which the wideband data accounted for between 1% and 80% of the training corpus, with the narrowband data accounting for the rest. In all experiments, the front-end GMM was trained using the available wideband data only.

The performance of the proposed mixed-bandwidth training algorithm is shown in Figure 2 as a function of the amount of wideband training data used. For comparison, the WERs obtained by a system trained with limited wideband data only and by a fully trained narrowband system are also shown. As the figure shows, at all percentages, a significant improvement in the WER is obtained over the use of the wideband data alone. Perhaps more importantly, the figure also shows that the proposed mixed-bandwidth training method results in better performance than a fully trained narrowband recognizer in all cases. Of course, we expect that as the amount of wideband training data approaches zero, the narrowband system will outperform the proposed method, as there simply will not be enough wideband data to train a reliable GMM.

6.2. Model Adaptation Reusing the Wideband Training Data

Because the training algorithm maximizes the likelihood of the *total* pool of training data, it may not necessarily be ideally matched to the wideband data. For example, since the wideband posterior distributions were generated using a front-end GMM, rather than from the HMMs themselves, there may be a bias in the model parameter estimates. As a result, we attempted to improve the model performance further by reusing the wideband training data to perform supervised model adaptation on the final wideband acoustic models. This is different from typical model adaptation in that

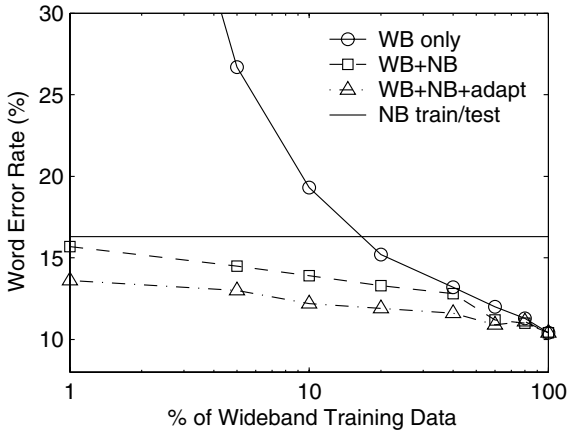

Fig. 3. WER of the WSJ0 20k test set after reusing the wideband training data to adapt the models generated by the proposed mixed-bandwidth training. For comparison, the WERs obtained by three other training methods are also shown: the proposed algorithm prior to model adaptation, a recognizer trained from limited wideband data only, and a fully trained narrowband recognizer.

Table 1. A comparison of the WER obtained using Feature Bandwidth Extension (FBE) and the proposed mixed-bandwidth EM algorithm (MIXBW-EM) on the WSJ0 20k test set for different proportions of wideband and narrowband training data.

Training Data	FBE	MIXBW-EM
20% WB + 80% NB	13.5	13.3
10% WB + 90% NB	18.3	13.9

we are not introducing new adaption data, but simply reusing the available wideband training data. Mean and variance adaptation was performed using MLLR [10] with two regression classes. The results obtained after model adaptation are shown in Figure 3. As the figure shows, significant improvements are seen at all wideband-narrowband combinations.

6.3. Comparison with Feature Bandwidth Extension

Finally, we compared the proposed mixed-bandwidth EM training algorithm to the Feature Bandwidth Extension (FBE) algorithm proposed in [3], The results are shown in Table 1. As described in Section 1, FBE generates point estimates of the wideband features from the narrowband observations, which are then pooled with actual wideband features to train wideband acoustic models. Because the error in FBE is unaccounted for during training, the resulting model parameters are suboptimal. In contrast, the proposed algorithm uses marginalization to compute the state posteriors for the narrowband data, and includes the uncertainty associated with the wideband feature estimates in the model parameter updates. We note that as more and more wideband data is available, the wideband feature estimation can be expected to improve, and thus, the performance of FBE will approach that of the proposed mixed-bandwidth training algorithm.

7. CONCLUSION

In this paper, we have proposed a method for training wideband acoustic models for HMM-based speech recognition systems using mixed-bandwidth training data. In this method, a limited amount of wideband training data is augmented with narrowband training data in order to train a speech recognizer for the recognition of wideband speech. The wideband acoustic models are trained using an EM algorithm in which hidden variables are assigned to the missing spectral components of the narrowband observations.

Through a series of experiments using parallel corpora of wideband and telephone speech, we demonstrated that the proposed method generates acoustic models that significantly outperform both a wideband recognizer trained from limited data and a fully trained narrowband recognizer. Thus, it is a viable method for training a wideband speech recognition system when collecting large amounts of wideband training data is not feasible. Moving forward, we believe the performance of the proposed algorithm can be further improved by exploring alternative methods for sharing data among the HMM states, rather than the global front-end-based approach used in this work.

8. REFERENCES

[1] J. S. Garofolo, C. D. Laprun, and J. G. Fiscus, "The rich transcription 2004 spring meeting recognition evaluation," in *Proc. NIST RT04 Meeting Recognition Workshop*, Montreal, Canada, 2004.

[2] P. Moreno and R. M. Stern, "Sources of degradation of speech recognition in the telephone network," in *Proc. ICASSP*, Adelaide, Australia, Apr. 1994, vol. I, pp. 109–112.

[3] M. L. Seltzer and A. Acero, "Training wideband acoustic models using mixed-bandwidth training data via feature bandwidth extension," in *Proc. ICASSP*, Philadelphia, PA, Mar. 2005, vol. 1, pp. 921–924.

[4] A. P. Dempster, N. M. Laird, D. B., and Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of Royal Statistical Society*, vol. 39, no. 1, pp. 1–38, 1977.

[5] Z. Ghahramani and M. I. Jordan, "Supervised learning from incomplete data via an EM approach," in *Advances in Neural Information Processing Systems*, 1994.

[6] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of damaged spectrographic features for robust speech recognition," in *Proc. ICSLP*, Beijing, China, Oct. 2000.

[7] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, June 2001.

[8] J. M. Mendel, *Lessons in Estimation Theory for Signal Processing, Communications, and Control*, Prentice Hall, New Jersey, 1995.

[9] D. B. Paul and J. M. Baker, "The design of the Wall Street Journal-based CSR corpus," in *Proc. ARPA Speech and Nat. Lang. Workshop*, Harriman, NY, Feb. 1992, pp. 357–362.

[10] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of HMMs using linear regression," Tech. Rep. CUED/F-INFENG/ TR. 181, Cambridge University, Cambridge, UK, June 1994.