TRAINING WIDEBAND ACOUSTIC MODELS USING MIXED-BANDWIDTH TRAINING DATA VIA FEATURE BANDWIDTH EXTENSION

Michael L. Seltzer and Alex Acero

Microsoft Research
1 Microsoft Way
Redmond, WA 98102
{mseltzer,alexac}@microsoft.com

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 sub-optimal performance. In this paper, we propose a new algorithm for training wideband acoustic models that requires only a small amount of wideband speech augmented by a larger amount of narrowband speech. The algorithm operates by first converting the narrowband features to wideband features through a process called Feature Bandwidth Extension. The bandwidth-extended features are then combined with available wideband data to train the acoustic models using a modified version of the conventional forward-backward algorithm. 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

One serious difficulty in the deployment of automatic speech recognition (ASR) systems for new tasks is the expense of obtaining sufficient training data. This is especially true for applications which process wideband speech, e.g. desktop applications. The cost and time required for data collection can be mitigated by collecting speech over the telephone. 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, collecting speech data in this manner has the drawback that the speech used to train the recognizer will be narrowband, typically sampled at 8 kHz with a bandwidth of 300-3400 Hz. This means that during decoding, the test speech must be restricted to the same bandwidth. However, all other things being equal, recognition systems that process narrowband speech perform worse than those that process wideband speech, i.e. speech sampled 16 kHz with a bandwidth of 0-8000 Hz [1]. Therefore, the performance obtained by restricting the bandwidth of the speech recognition system to that of telephone speech is sub-optimal.

Thus, when creating a new wideband speech recognition application, there are two options for collecting training data. The first is to collect enough wideband speech to adequately train the recognizer. This option is expensive in both time and cost, but yields the best performance. The second is to collect training data over the telephone and then restrict the bandwidth of the wideband test speech to match that of the telephone speech. This option is more cost-effective but results in sub-optimal recognition accuracy.

In this paper, we propose an alternative approach in which wideband acoustic models are trained using a small sample of wideband speech and a large sample of narrowband speech. In this approach, the narrowband speech features are first converted to wideband features using a process we call *Feature Bandwidth Extension* (FBE). These bandwidth-extended features are then pooled with features derived from available wideband speech and used to train the acoustic models using a modified version of the conventional forward-backward algorithm. We demonstrate that combining bandwidth-extended features together with wideband features in this manner produces acoustic models that outperform those trained on either the limited wideband speech or the abundant narrowband speech in isolation.

The method proposed in this paper is related to previous research in training mixture models from incomplete data [2]. However, this work is not directly applicable to speech recognition applications because of the idiosyncrasies of 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. [3].

The remainder of the paper is organized as follows. In Section 2, the feature extraction process for speech recognition is briefly reviewed and the missing data paradigm for mixed-bandwidth speech is introduced. In Section 3, we describe the proposed method for performing HMM training from mixed-bandwidth training data using FBE. Section 4 describes experiments that show the validity of the proposed method. Finally, we summarize this work and present some conclusions in Section 5.

2. FEATURE EXTRACTION FOR ASR

In this work, we assume that mel-frequency cepstral coefficients (MFCC) are the features used for recognition. We define $\mathbf{x}_i$ as the log mel spectrum of the ith frame of speech. For wideband speech, the log mel spectrum represents the energy in the mel filterbank, a series of overlapping frequency regions which range from approximately 100 Hz to 8 kHz. This log mel spectral vector is then converted to a cepstral vector $\mathbf{z}_i$ via a discrete-cosine transform (DCT) as

$$\mathbf{z}_i = \mathbf{C}\mathbf{x}_i, \tag{1}$$

where $\mathbf{C}$ is the DCT matrix. Dimensionality reduction is also usually performed, so the DCT matrix $\mathbf{C}$ is $M \times L$ with $M \leq L$.

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 narrowband speech, the observed and missing subvectors roughly correspond to the low and high frequency components, respectively. However, for telephone speech, the lowest mel components typically fall outside the telephone passband and are therefore considered missing as well. For wideband speech originally sampled at the target sampling rate, $\mathbf{x}^{\mathrm{o}} = \mathbf{x}$ and $\mathbf{x}^{\mathrm{m}} = [\,]$.

Substituting (2) into (1), we can express the cepstral vector $\mathbf{z}$ as the sum of linear transformations of $\mathbf{x}^{\mathrm{o}}$ and $\mathbf{x}^{\mathrm{m}}$ as

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

where $\mathbf{C}$ has been partitioned into $\mathbf{C}^{\mathrm{o}}$, 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 $\mathbf{x}^{\mathrm{m}}$.

3. HMM TRAINING VIA FEATURE BANDWIDTH EXTENSION

The proposed method of HMM training using mixed-bandwidth speech data consists of two two stages. In the first stage, the bandwidth of the narrowband features is extended using *a priori* knowledge obtained from available wideband speech data. In the second stage, the wideband speech features and the bandwidth-extended features are used to train the HMMs. We now describe each of these two stages in more detail.

3.1. Bandwidth extension of narrowband speech features

In Feature Bandwidth Extension (FBE), we wish to infer the wideband cepstral vector $\mathbf{z}_i$ given the observed narrowband log mel spectral vector $\mathbf{x}_i^{\mathrm{o}}$. The minimum mean squared error (MMSE) estimate of $\mathbf{z}_i$ can be expressed as

$$\hat{\mathbf{z}}_i = E[\mathbf{z}_i^{\mathrm{o}} + \mathbf{z}^{\mathrm{m}} | \mathbf{x}_i^{\mathrm{o}}] = \mathbf{C}^{\mathrm{o}}\mathbf{x}_i^{\mathrm{o}} + \mathbf{C}^{\mathrm{m}} E[\mathbf{x}^{\mathrm{m}} | \mathbf{x}_i^{\mathrm{o}}] \tag{4}$$

where $E[\cdot]$ represents the expectation operatior.

In order to compute the expected value in (4), we utilize a prior model of wideband speech. We assume that wideband speech can be effectively modeled as a Gaussian Mixture Model (GMM). This GMM is trained in the log mel spectral domain from available wideband speech data using conventional EM. Thus, we have a distribution of wideband speech of the form

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}|k)p(k) = \sum_{k=1}^{K} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(k) \tag{5}$$

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

Using a GMM, the expected value in (4) can be rewritten as

$$E[\mathbf{x}^{\mathrm{m}} | \mathbf{x}_i^{\mathrm{o}}] = \sum_{k=1}^{K} \int \mathbf{x}^{\mathrm{m}} p(\mathbf{x}^{\mathrm{m}}, k | \mathbf{x}_i^{\mathrm{o}}) d\mathbf{x}^{\mathrm{m}} \tag{6}$$

$$= \sum_{k=1}^{K} p(k|\mathbf{x}_i^{\mathrm{o}}) \int \mathbf{x}^{\mathrm{m}} p(\mathbf{x}^{\mathrm{m}} | \mathbf{x}_i^{\mathrm{o}}, k) d\mathbf{x}^{\mathrm{m}} \tag{7}$$

where $p(k|\mathbf{x}_i^{\mathrm{o}})$ is the posterior probability of the kth Gaussian based only on the observed components of the feature vector. Computing the expected value in (7) requires the marginal and conditional probability density functions (PDFs) associated with $p(\mathbf{x}|k)$. Specifically, we need to factorize $p(\mathbf{x}|k)$ as

$$p(\mathbf{x}|k) = p(\mathbf{x}^{\mathrm{o}}, \mathbf{x}^{\mathrm{m}}|k) = p(\mathbf{x}^{\mathrm{o}}|k)p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, k). \tag{8}$$

To do so, we first sort the means and covariances into observed and missing partitions as

$$\boldsymbol{\mu}_k = \begin{bmatrix} \boldsymbol{\mu}_k^{\mathrm{o}} \\ \boldsymbol{\mu}_k^{\mathrm{m}} \end{bmatrix}, \quad \boldsymbol{\Sigma}_k = \begin{bmatrix} \boldsymbol{\Sigma}_k^{\mathrm{oo}} & \boldsymbol{\Sigma}_k^{\mathrm{mo}} \\ \boldsymbol{\Sigma}_k^{\mathrm{om}} & \boldsymbol{\Sigma}_k^{\mathrm{mm}} \end{bmatrix}. \tag{9}$$

Using (9), we can now express the marginal distribution as

$$p(\mathbf{x}^{\mathrm{o}}|k) = \mathcal{N}(\mathbf{x}^{\mathrm{o}}; \boldsymbol{\mu}_k^{\mathrm{o}}, \boldsymbol{\Sigma}_k^{\mathrm{oo}}) \tag{10}$$

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

$$p(\mathbf{x}^{\mathrm{m}}|\mathbf{x}^{\mathrm{o}}, k) = \mathcal{N}(\mathbf{x}^{\mathrm{m}}; \boldsymbol{\mu}_k^{\mathrm{m|o}}, \boldsymbol{\Sigma}_k^{\mathrm{m|o}}) \tag{11}$$

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

$$\boldsymbol{\mu}_k^{\mathrm{m|o}} = \boldsymbol{\mu}_k^{\mathrm{m}} + \boldsymbol{\Sigma}_k^{\mathrm{mo}} \boldsymbol{\Sigma}_k^{\mathrm{oo},-1} (\mathbf{x}^{\mathrm{o}} - \boldsymbol{\mu}_k^{\mathrm{o}}) \tag{12}$$

$$\boldsymbol{\Sigma}_k^{\mathrm{m|o}} = \boldsymbol{\Sigma}_k^{\mathrm{mm}} - \boldsymbol{\Sigma}_k^{\mathrm{mo}} \boldsymbol{\Sigma}_k^{\mathrm{oo},-1} \boldsymbol{\Sigma}_k^{\mathrm{om}} \tag{13}$$

Substituting (12) and (7) into (4) leads to the following solution for the MMSE estimate of $\mathbf{z}_i$ given the narrowband observation $\mathbf{x}_i^{\mathrm{o}}$

$$\hat{\mathbf{z}}_i = \mathbf{C}^{\mathrm{o}}\mathbf{x}_i^{\mathrm{o}} + \mathbf{C}^{\mathrm{m}} \left(\sum_{k=1}^{K} p(k|\mathbf{x}_i^{\mathrm{o}}) \boldsymbol{\mu}_k^{\mathrm{m|o}} \right). \tag{14}$$

where the posterior probability $p(k|\mathbf{x}_i^{\mathrm{o}})$ in (14) can be computed from (10) using Bayes rule.

3.2. HMM Training using bandwidth-extended features

Performing FBE on every frame of narrowband speech in the training set generates a sequence of bandwidth-extended feature vectors which we can now pool with the available wideband data and use to train the recognizer in the conventional manner. However, using the bandwidth-extended features without any changes to the training procedure implicitly makes the assumption that the MMSE estimates generated by FBE are error-free. Because the features have been inferred from narrowband data, they may in fact be erroneous. Therefore, intuitively, we should not trust the bandwidth-extended speech data as much as the actual wideband speech data. To reflect this "mistrust," we assign a weighting factor

to the posterior probability of each frame of bandwidth-extended speech when the Gaussian parameter updates are computed. More explicitly, for each frame i of bandwidth-extended speech, the posterior probability of being in Gaussian q of state k is modified as

$$\hat{\gamma}_{ikq} = \alpha\gamma_{ikq}, \qquad 0 \leq \alpha \leq 1 \tag{15}$$

where γ_{ikq} is the state-posterior probability computed from the E-step of the conventional forward-backward algorithm [4]. Thus, the updated mean vector for the qth Gaussian of state k is computed as

$$\boldsymbol{\nu}_{kq} = \frac{\sum_{i=1}^{N^w} \gamma_{ikq}\mathbf{z}_i + \sum_{j=1}^{N^b} \hat{\gamma}_{jkq}\hat{\mathbf{z}}_j}{\sum_{i=1}^{N^w} \gamma_{ikq} + \sum_{j=1}^{N^b} \hat{\gamma}_{jkq}} \tag{16}$$

where N^w is the total number of wideband frames, indexed by i, and N^b is the total number of bandwidth-extended frames, indexed by j. The Gaussian prior probabilities and covariance matrices are updated in a similar manner. Weighting the posterior probabilities in this manner is equivalent to MAP parameter estimation [5]. We discuss how to choose the value of α in the next section.

4. EXPERIMENTAL EVALUATION

In order to evaluate the proposed mixed-bandwidth training algorithm, we performed a series of experiments using the TIMIT corpus [6]. TIMIT is a phonetically labeled corpus of hand-designed utterances used to evaluate phonetic recognition accuracy. TIMIT was chosen because it was originally recorded as wideband speech with a sampling rate of 16 kHz and was later transmitted over the telephone network and released as the narrowband NTIMIT (Network TIMIT) corpus [7]. The useful bandwidth of the NTIMIT corpus is approximately 300-3400 Hz. By using these two parallel corpora, we can perform controlled experiments which explore the combination of wideband and narrowband speech, where the bandwidth of the data is the only variable.

The phonetic dictionary used for these experiments was identical to that used by Lee and Hon in their TIMIT baseline experiments [8]. The HTK speech recognition system was used to train 3-state context-dependent triphone models with 16 Gaussians per state. The feature vectors used for recognition were 13-dimensional cepstra derived from 40-dimensional log mel spectra, along with their delta and acceleration parameters. Frames were 25 ms with a 10 ms shift between successive frames. Cepstral mean normalization was performed prior to processing. A phonetic-bigram language model was used for decoding, with a language weight of 8.0. Performance was measured using the TIMIT core test set. A 50-speaker set of utterances not present in either the training or test sets was used as a development set for intermediate experiments.

In the first series of experiments, we evaluated the recognition performance when different amounts of wideband speech were used for training. The complete training set consists of approximately 3.1 hours of speech. Subsets of the training set, ranging from 1% up to 90% of the total training set were selected at random, and used to train the recognizer. Figure 1 shows the phonetic accuracy as a function of amount of data used to train the recognizer. Not surprisingly, the plot shows that the performance of the system degrades significantly with fewer training data.

We will now attempt to improve the performance when only limited wideband training data is available. In these experiments, we assume that only 10% of the wideband data is available, which corresponds to 0.3 hours of speech. The remaining 90% (2.8 hours)

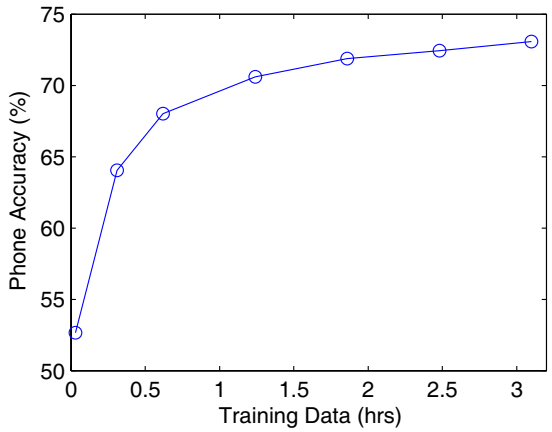

Fig. 1. Phone accuracy of the TIMIT core test set versus the amount of data used to train the recognizer. The leftmost data point represents 1% of the total training set (0.03 hrs) while the rightmost datapoint represents the full of the training set (3.1 hrs)

of the training data is taken from NTIMIT. In telephone speech that is upsampled to 16 kHz, the first 4 and last 13 components of the 40-dimensional log mel spectral vectors are outside of the telephone passband and therefore unobserved.

In order to perform FBE, a GMM of 256 densities was trained using the available 0.3 hours of wideband training data. Using this GMM, FBE was performed on the NTIMIT training vectors. In order to mitigate the spectral tilt induced by the telephone channel, mean normalization was performed on both the wideband log mel spectra used to train the GMM and the narrowband NTIMIT log mel spectra prior to FBE.

The 2.8 hours of bandwidth-extended cepstra from NTIMIT were then pooled with the 0.3 hours of wideband cepstra and used to train the recognizer. Multiple acoustic models were trained using a range of values between 0 and 1 for the weighting parameter α in (15). Note that setting $\alpha = 0$ is equivalent to using only the wideband data, while setting $\alpha = 1$ results in conventional HMM training. In order to determine the optimal value, the performance of each of the models was evaluated using the development set. Figure 2 shows the results of this experiment.

As the figure shows, optimum performance is obtained when $\alpha = 0.2$. We can also see that setting α greater than 0.8 results in performance that is worse than that obtained from the wideband data alone. Thus, the bandwidth-extended data is useful in training wideband acoustic models, but it clearly cannot be considered as useful as wideband data that is well matched to the test conditions.

Based on these results, we evaluated the performance of mixed-bandwidth training on the TIMIT core test set using acoustic models training with $\alpha = 0.2$. The results obtained are shown in Table 1. In this table, we compare the performance obtained using various training strategies possible when wideband training data is limited. The first row shows the accuracy obtained when telephone speech is used to train narrowband acoustic models, and the test data is downsampled and filtered to match the telephone speech characteristics. To improve the performance, we augmented the telephone speech with some telephonized wideband speech to match the test data. The second row shows the performance obtained when the telephone data is ignored and only

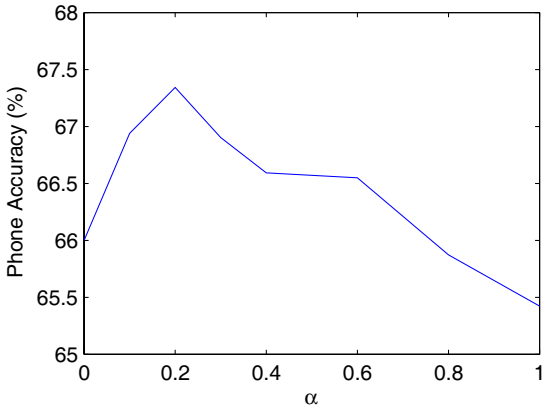

Fig. 2. Phone accuracy of the TIMIT development set versus the value of α used during mixed-bandwidth HMM training.

a limited amount of wideband data is used to train the system. The models are well matched to the test data but undertrained. The next two rows show the performance obtained when both the wideband and telephone speech are used for training using the proposed mixed-bandwidth training strategy. In the second case, after training was complete, supervised MLLR adaptation was performed using the limited wideband training data a second time [9]. Finally, the last row shows the performance obtained when the full wideband training set is used. This represents the upper bound on the phone recognition accuracy for the core test set.

The accuracy obtained by using limited wideband training data is 64.0%, while the accuracy of a fully trained wideband system is 73.1%. By augmenting the limited wideband speech data with bandwidth-extended telephone speech and using the proposed mixed-bandwidth training algorithm, we have narrowed the gap in performance between these two systems by 17.6% without using model adaptation and by 23.1% with adaptation. We note that no additional data was used for adaptation, just the same 0.3h of wideband data that was used in training. As these experiments show, the proposed mixed-bandwidth training algorithm results in a significant improvement over conventional training methods when the amount of wideband speech available for training is limited.

5. CONCLUSION

In this paper, we have proposed a two-stage method for training acoustic models for HMM-based speech recognition systems using mixed-bandwidth training data. In the first stage, wideband features are estimated from narrowband features using a process called Feature Bandwidth Extension (FBE). Then, the bandwidth-extended features are combined with available wideband speech data to train the acoustic models using a modified version of the forward-backward algorithm.

Through a series of experiments, we demonstrated that the proposed method is able to significantly improve speech recognition performance compared to systems trained solely on either narrowband data or a limited amount of wideband data. Using the proposed two-stage mixed-bandwidth training algorithm, we were able to reduce the difference in performance between a system trained from only limited wideband data and a fully trained wideband system by 23.1%.

Training Data	Phone Accuracy (%)
2.8h NTIMIT + 0.3h TIMIT-TB	62.3
0.3h TIMIT-WB	64.0
2.8h NTIMIT-BWE + 0.3h TIMIT-WB	65.6
2.8h NTIMIT-BWE + 0.3h TIMIT-WB, + MLLR using TIMIT-WB	66.1
3.1h TIMIT-WB	73.1

Table 1. The phone accuracy on the TIMIT core test under various training scenarios. TIMIT-WB is the original wideband TIMIT speech. TIMIT-TB is the TIMIT speech filtered to the telephone bandwidth, NTIMIT-FBE is NTIMIT with FBE applied.

We believe the performance of mixed-bandwidth training can be improved by explicitly incorporating a measure of the uncertainty of the bandwidth extension process into the training algorithm. For example, the variances associated with the MMSE estimates generated by FBE could be used to discount or deweight highly uncertain feature vectors in the parameter update formulas. In addition, we believe further improvement can be obtained by increasing the robustness of FBE to channel distortion and additive noise, both common in telephone speech.

6. REFERENCES

[1] 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.

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

[3] 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.

[4] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1990.

[5] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[6] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet, and N. Dahlgren, "The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," October 1990, NTIS order number PB91-505065.

[7] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," in *Proc. ICASSP*, Apr. 1990, vol. 1, pp. 109–112.

[8] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using Hidden Markov Models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[9] 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.