# SPEAKER ADAPTATION BASED ON SPECTRAL NORMALIZATION AND DYNAMIC HMM PARAMETER ADAPTATION

Ming-Whei Feng

GTE Laboratories Inc.
40 Sylvan Road
Waltham, MA 02254

## ABSTRACT

Speaker adaptation has received a considerable amount of attention in recent years. Most of the previous work focused on techniques which require a certain amount of speech to be collected from the target speaker. This paper presents two speaker adaptation methods, including a feature normalization and a HMM parameter adaptation, developed to improve a speaker-independent HMM-based speech recognition system. The proposed adaptation algorithms are text-independent and do not require target speech collection. By applying the feature normalization, the target speech is normalized to reduce the acoustic inter-speaker and environmental variability. By applying the HMM parameter adaptation, the recognition system parameters are dynamically modified to model the target speech. We carried out recognition experiments to assess the performance, using two different speaker-independent recognizers as the baseline systems: a continuous digit recognizer and a keyword recognition system. The results show that when both adaptation techniques are combined, the word error of the digit recognizer using the TI Connected Digit corpus is reduced by about 30% and the detection error of a keyword recognition system using the Road Rally corpora is reduced by about 40%.

## 1. INTRODUCTION

Statistical modeling approaches such as hidden Markov Models (HMMs) have been widely applied to develop speech recognition systems for practical applications. One of the major challenges in developing a speech recognition system with high accuracy is the vast acoustic variability among different speakers. To address this problem, in a speaker-dependent recognition system, speaker-specific statistical patterns are typically developed to directly model the acoustical characteristics of the target speaker. Since the target speaker's acoustic characteristics are taken into account, the resulting recognition performance is typically high. In a speaker-independent recognition system, statistical models are derived using the speech collected from a certain number of training speakers. The recognition performance may degrade significantly when the acoustic characteristics of the target speech is very different from that of the training speech.

Speaker adaptation methods [1-8] were also developed to deal with the inter-speaker acoustic variability. The work presented in [1,2] focused on transforming well-trained discrete HMMs derived from a prototype speaker, to model the speech from the target speaker, by using spectral mappings. Methods described in [5,6] applied a Bayesian technique to adapt the continuous density HMMs to model the target speech. Unsupervised speaker adaptation methods based on hierarchical spectral clustering [3], or linear spectral normalization [4,8] were also proposed to model the inter-speaker variability. However, most of the previous adaptation algorithms require the collection of a certain amount of target speech, which is sometimes undesirable or impractical for certain types of applications.

This paper presents two speaker adaptation algorithms which are text-independent and do not require previously collected target speech. Both methods were developed to improve the performance of a HMM-based speaker-independent recognition system. The first proposed method is a feature normalization procedure in which we model the inter-speaker variability as a linear transformation in the spectral domain. The linear transformation is mapped to a spectral shift vector in the logarithmic spectral (cepstral) domain. Several techniques were investigated to measure the spectral bias vector. The second adaptation method is a HMM parameter adaptation procedure which takes the decoded word label and the corresponding speech segments and derives a probabilistic score index which gives an indication about the correctness of the label and the segmentation. If the probabilistic score index is above a pre-estimated threshold, the decoded speech segment is used to re-estimate the corresponding HMMs to improve the probability of generating the segment given by the models.

To assess the performance of the proposed methods, we carried out recognition experiments using two speaker-independent recognition systems as the baseline systems: a continuous digit recognition system and a keyword recognition system [9,10]. For the continuous digit recognition, experiments were conducted using the TI Connected Digit Corpus. The results indicate that the average word error rate of the baseline system is reduced by about 30% when using a combination of both methods. For the keyword recognition, the Road Rally Corpora were used and the results show that the detection error rate of the baseline system is reduced by about 40% when combining both proposed methods.

The remainder of the paper is organized as follows. The feature normalization method and the techniques used to derive the normalization are described in Section 2. The HMM adaptation algorithm and the derivation of the probabilistic score index are illustrated in Section 3. Section 4 describes the recognition experiments and presents the results.

## 2 FEATURE NORMALIZATION

In a speaker-independent recognition system, it is hoped that the statistical models will assimilate sources of speaker variability and environmental variability so the remaining system is a robust characterization of the speech units. However, it is known that there are sources of variability

0-7803-2431-5/95 $4.00 © 1995 IEEE

including inter-speaker variability and environmental variability. One approach to deal with the sources of variability is to assume that the variability can be represented by a linear transformation in the spectral domain. In the logarithmic spectral domain, this variability is mapped to a fixed shift (spectral bias). Several techniques were investigated to model the cepstral bias vector.

### Cepstral-Mean Based Normalization

A cepstral-mean based normalization method which is similar to the method presented in [11,12] used for channel or environment normalization, was developed to compensate the bias vector in the cepstral feature domain for the target speaker. The is a simple and direct deconvolution method, where the bias vector is modeled as the difference between the average cepstrum of the training speech and the average cepstrum of the target speech. This spectral bias vector is separately estimated for each target sentence. The derived bias vector is then used to normalize the target speech before recognition decoding.

### Model-Based Normalization

The second normalization approach is a model-based approach, and is motivated by the work presented in [4,8]. In this approach, the speech from a standard speaker is modeled using a set of Gaussian density models $\mu(\bar{m}_i, \bar{\sigma}_i), 1 \leq i \leq M$. For simplicity, we assume all the covariance matrices are diagonal matrices. Given by a sequence of labeled cepstral feature vectors $\{\bar{x}_i, 1 \leq i \leq T\}$ from a given target speaker $q$, the cepstral shift vector $\bar{h}_q$ characterizing the speaker q can be estimated to maximize the likelihood function:

$$L = \prod_{1 \leq i \leq T} P(\bar{x}_i \mid \bar{m}_{c_i}, \bar{\sigma}_{c_i}, \bar{h}_q) \cdot P(\bar{m}_{c_i}, \bar{\sigma}_{c_i} \mid \bar{h}_q), \tag{1}$$

where $c_i$ is the label and $P(\bar{m}_i, \bar{\sigma}_i \mid \bar{h}_q)$ is assumed as uniform. By differentiating the log likelihood in (1) with respect to the unknown spectral bias vector and setting to zero, we obtain

$$\hat{\bar{h}}_q = \frac{1}{T} \sum_{i=1}^{T} (\bar{x}_{c_i} - \bar{m}_{c_i}). \tag{2}$$

In the current procedure, we use the training speech to derive the Gaussian models for the standard speaker. For each training sentence, we first label the corresponding observation sequence using the Gaussian models of the standard speaker. We then estimate the spectral bias vector using (2) and remove the spectral shift vector from that sentence. The normalized training speech is used to derive the HMM parameters for recognition. A similar normalization procedure is performed on the target speech before the recognition decoding.

From the recognition experiments, we observed that the performance improvement caused by the cepstral-mean-based approach is similar to that of the model-base normalization. However, the computational complexity of the model-based technique is much higher than the other. Therefore, we select the cepstral-mean based normalization as the normalization method.

## 3. HMM PARAMETER ADAPTATION

In a typical speaker-independent HMM-based recognition system, the HMM parameters are derived to model the training speech and the parameters remain unchanged when recognizing target speech. The recognition performance may drop significantly when the acoustic characteristics of the target speech does not match that of the training speech. To improve the recognition performance, we developed a dynamic HMM parameter adaptation approach, which attempts to adjust the HMM parameters to model the acoustic characteristics of the target speech.

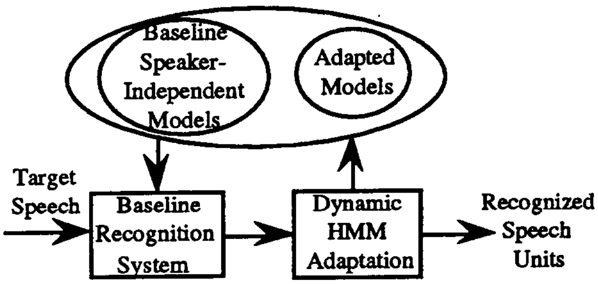

Figure 1. Dynamic HMM parameter adaptation paradigm.

Figure 1 shows the proposed recognition procedure which incorporates the HMM adaptation as part of the on-line recognition process following the regular recognition decoding procedure. A new target speaker starts with the baseline HMM system which was derived from the training speech. Following the regular recognition decoding process, the target speech sentence labeled by the recognition decoder is used to modify the baseline HMM parameters using the dynamic HMM parameter adaptation method. The modified HMMs replace the baseline HMMs and are used to recognize the next speech sentence from the same target speaker. This iterative recognition-adaptation process continues until the end of the target speaker turn. The HMM parameters of the recognition system are therefore dynamically updated during recognition.

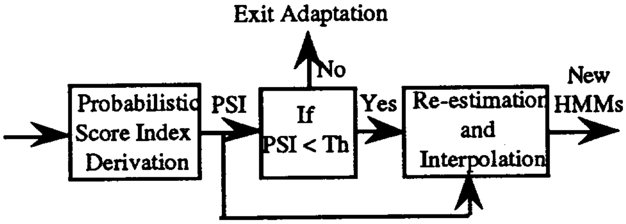
Exit Adaptation
No
New HMMs
Probabilistic Score Index Derivation
PSI
If PSI < Th
Yes
Re-estimation and Interpolation

Figure 2. HMM parameter adaptation block diagram.

Figure 2 shows the block diagram of the HMM parameter adaptation (HPA) algorithm. As shown in Figure 2, the HPA algorithm contains three steps. In the first step, it derives a probabilistic score index (PSI) for each word which is decoded in the preceding baseline recognition procedure. The derived PSI provides a confidence measure which indicates how well the segment matches the model. In the second step, each derived PSI is compared with an empirically derived threshold. If the derived PSI is above the threshold, we go to the third step to re-estimates the HMMs associated with the word label. When the calculated PSI is lower than the threshold, no model adaptation is performed because the label of the speech might not be correct. Below we describe the definition and the derivation of the probabilistic score index.

### Probabilistic Score Index

Given a hypothesized word label, the corresponding speech segment $O_1..O_T$ is mapped to every word model $\{HMM_i, 1 \le i \le N\}$, to calculate a maximum likelihood score :

$$ml = \max\{LS_i = \mathrm{Prob}(O_1 \ldots O_T \mid HMM_i),\, i = 1,..,N\}, \qquad (3)$$

where $LS_i$ is the likelihood score associated with the *i-th* model, and $N$ is the total number of word models used. In addition, we compute a differential maximum likelihood measure:

$$dml = ml - ml', \qquad (4)$$

where $ml'$ is the second largest likelihood score. The likelihood scores defined in (3) and (4) are similar to the *a-posteriori* measures used in the keyword recognition system described in [9,10].

Since the proposed HMM adaptation technique uses the decoded word label and the corresponding speech segment as the labeled target speech to re-estimate the associated HMM parameters, the correctness of the decoded word label and the speech segments are of great importance. To derive a reliable measure of the correctness of the label and the segment, the maximum likelihood score and the differential likelihood score for each word label $i$ are modeled as random variables $ML(i)$ and $DML(i)$ with discrete probability density functions (pdfs) denoted by $P_{ML(i)}, P_{DML(i)}$, for all $i$. In addition, the durations of the decoded speech segments for all word labels $i$ are also modeled as random variables with pdf $P_{D(i)}$.

Given a word label $k$ from the baseline recognition decoder, and the corresponding maximum likelihood score $ml$, we can calculate the probability of producing the score $ml$ for word label $k$ using the corresponding pdf:

$$PS_1 = P_{ML(k)}(ml), \qquad (5)$$

where $PS_1$ can be interpreted as a confidence factor which measures how well the decoded speech segment matches the corresponding HMM from the *a-posteriori* maximum likelihood score point of view. Similarly, we can calculate the confidence factor using the probability density function for the differential likelihood or the duration score. In the proposed method, we calculate the individual confidence score and derive a final probabilistic score index which is defined as a weighted linear combination of the three individual confidence measures:

$$PSI = w_1 \cdot P_{ML(k)}(ml) + w_2 \cdot P_{DML(k)}(dml) + w_3 \cdot P_{D(k)}(d), \qquad (6)$$

where the weighting factors $w_1, w_2, w_3$ are used to adjust the importance of the individual probabilistic score.

### Reestimation of HMMs

In the HMM parameter adaptation procedure, we focus on modifying the pdfs associated with HMMs. Given the decoded word label, the segmented observation sequence $\overline{O} = O_1..O_T$, and the current model parameters $\lambda = (\overline{\pi}, \overline{\overline{A}}, \overline{\overline{B}})$, our goal is to obtain a new set of pdfs associated with the decoded word label, to maximize the likelihood $P(\overline{O}|\lambda)$.

We use the Baum-Welch re-estimation procedure [13] to derive the maximum likelihood estimate of the new pdfs $\overline{b_s'}(l)$ for all symbols $V_l$ and all associated states $s$ in word model $k$. We then interpolate the newly estimated $\overline{b_s'}(l)$ with the current pdfs $\overline{b}_s(l)$ to form the final pdfs, for all associated $s$ and all $l$:

$$\overline{b}_s''(l) = \overline{b}_s(l) + PSI \cdot G \cdot \overline{b_s'}(l) \qquad (7)$$

where factor $G$ is used to adjust the contribution of the new estimate.

## 4. RECOGNITION EXPERIMENTS

In the recognition experiments, we applied the proposed adaptation methods to two separate speaker-independent recognition systems: a continuous digit recognition system and a keyword recognition system. In both experiments, we used 17 features (8 cepstral, 8 differential cepstral, and 1 differential power) computed every 16 msec., using a 32 msec. analysis window.

### Continuous Digit Recognition

We carried out experiments using the standard speaker-independent TI Connected Digit corpus [14]. The corpus consists of speech sentences with 1,2,3,4,5 and 7 digit strings, spoken by the speakers representing 22 dialectical regions. We used 2640 sentences from 110 speakers for training, and 2712 sentences from an independent set of 113 speakers for testing. All speech sentences were downsampled from 20Khz to 8kHz. Word-based HMMs with discrete pdfs were used to model each digit individually.

We used the word error rate defined as the sum of substitution error, deletion error, and insertion error, as the performance measure. Table 1 shows the results of applying the baseline (no adaptation), baseline with feature normalization algorithm (FNA), baseline with HMM parameter adaptation (HPA), and baseline with a combination of both schemes. As shown in table 1, the FNA method reduces the baseline word error by about 7% and the HPA method reduces baseline word error by about 26%. The combination of the two proposed schemes decreases the baseline error by about 31% .

| | word error rate (%) |
|---|---|
| Baseline | 4.5 |
| FNA + Baseline | 4.2 |
| HPA + Baseline | 3.3 |
| FNA+HPA+Baseline | 3.1 |

Table 1. Performance comparison of a continuous digit recognizer with and without adaptation techniques.

Keyword recognition

We conducted experiments using the Road Rally corpora with 20 pre-defined keywords. The Road Rally Corpora contain two sub-corpora, "Stonehenge" and "Waterloo". The Stonehenge speech was collected using high-quality microphones and a telephone-simulator filter, while the Waterloo speech was collected using actual telephone lines. We manually segmented all conversations and read passages to form natural sentences. The training speech contains sentences from 56 Waterloo read passages and 18 Stonehenge conversations. The test speech includes sentences from an independent set of 18 Stonehenge conversations. We applied the proposed adaptation techniques to a baseline keyword recognition which is similar to the one described [9,10], to assess the performance. Word-based HMMs with discrete pdfs were used to model both keyword and non-keyword elements.

We used the detection error rate, rejection rate, and the false alarm per keyword per hour as performance measures. In Table 2, we present the performance resulting from applying the baseline system, baseline with the FNA method, baseline with the HPA method, and baseline with both adaptation techniques. Table 2 shows that the application of the FNA method reduces not only 21% of the baseline detection error, but also reduces the rejection rate by about 59%. By using the HPA algorithm only, the baseline detection error is reduced by 27% and the rejection error is reduced by 60%. By combining the two methods, the baseline detection error is reduced by 40% and the rejection error is reduced by 62%.

| | detection error rate (%) | false alarm per kw per hour | rejection rate (%) |
|---|---|---|---|
| Baseline | 27.0 | 16.4 | 7.5 |
| FNA + Baseline | 21.3 | 17.1 | 3.1 |
| HPA + Baseline | 19.8 | 16.5 | 3.0 |
| FNA+HPA+Baseline | 16.0 | 16.8 | 2.8 |

Table 2. Performance comparison of a keyword recognition system with and without adaptation techniques.

The above results show that the HPA method offers similar performance improvement in both cases. However, with the FNA method, we observe higher performance improvement for the keyword recognition experiment than for the continuous digit recognition experiment. This performance difference may be related to the fact that the Road Rally test speech and the majority of the training speech were recorded in very different acoustic environments. The FNA method appears to be particularly valuable in cases where such environmental differences exist.

## 5. CONCLUSION

We presented two speaker adaptation algorithms developed to improve the recognition performance of a speaker-independent recognition system. The proposed speaker adaptation algorithms, including a feature normalization and a HMM parameter adaptation, are text-independent and do not require previously collected target speech. By using the feature normalization, the target speech is normalized to reduce the acoustic inter-speaker and environmental variability. By applying the HMM parameter adaptation, the recognition system parameters are dynamically modified to model the target speech. Experiments on two speech corpora show that the average word error rate and detection error rate are reduced by 30-40% when both techniques are combined.

## References

[1]M. W. Feng, F. Kubala, R. Schwartz, and J. Makhoul, "Improved speaker adaptation using text-dependent spectral mappings," *Proc. ICASSP'88*, April 1988, pp. 131-134.

[2]M. W. Feng, R. Schwartz, F. Kubala, and J. Makhoul, "Iterative normalization for speaker-adaptive training in continuous speech recognition," *Proc. ICASSP'89*, May 1989, pp. 612-615.

[3]S. Furui, "Unsupervised speaker adaptation based on hierarchical spectral clustering," *IEEE Trans. on Acoust., Speech and Signal Proc.*, Vol.37, No.12, pp.1923-1930, Dec. 1989.

[4]S. Cox and J. Bridle, "Unsupervised speaker adaptation by probabilistic spectrum fitting", *Proc. ICASSP'89*, May 1989, pp. 294-297.

[5]C. H. Lee and C. H. Lin, and B. H. Juang, "A study on speaker adaptation of the parameters of the continuous density hidden Markov models," *IEEE Trans. on Signal Processing*, Vol.39, No.4, pp.806-814, April 1991.

[6]B. Necioglu, M. Ostendorf, J. Rohlicek, "A Bayesian approach to speaker adaptation for the stochastic segment model," *Proc. ICASSP'92*, March 1992, pp. 437-440.

[7]X. Huang and K. Lee, "On speaker-independent, speaker-dependent, and speaker-adaptive speech recognition," *Proc. ICASSP'91*, May 1991, pp. 877-880.

[8]Y. Zhao, "A new speaker adaptation technique using very short calibration speech," *Proc. ICASSP'93*, April 1993, pp.562-565.

[9]M. W. Feng and B. Mazor, "Continuous word spotting for applications in telecommunications", *Proc. of International Conference on Spoken Language Processing (ICSLP) 1992*, Oct. 1992, pp. 21-24.

[10]B. Mazor and M. W. Feng, "Improved a-posteriori processing for keyword spotting," *Proc. Eurospeech'93*, Nov. 1993, pp. 2231-2234.

[11]A. Acero and R. M. Stern, "Environmental robustness in automatic speech recognition," *Proc. ICASSP'90*, April 1990, pp. 849-852.

[12]S. Lerner and B. Mazor, "Telephone channel normalization for automatic speech recognition," *Proc. ICASSP'92*, March 1992, pp. 261-264.

[13]L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, pp. 1554-1563, 1966.

[14]R. Leonard, "A database for speaker-independent digit recognition," *Proc. ICASSP'84*, March 1984, pp. 42.11.1-4.