INTEGRATED BIAS REMOVAL TECHNIQUES FOR ROBUST SPEECH RECOGNITION*

Craig Lawrence and Mazin Rahim†

University of Maryland, College Park, MD 20742
†AT&T Labs-Research, Murray Hill, NJ 07974

10.21437/Eurospeech.1997-398

ABSTRACT

In this paper, we present a family of maximum likelihood (ML) techniques that aim at reducing an acoustic mismatch between the training and testing conditions of hidden Markov model (HMM)-based automatic speech recognition (ASR) systems. We propose a codebook-based stochastic matching (CBSM) approach for bias removal both at the feature level and at the model level. CBSM associates each bias with an ensemble of HMM mixture components that share similar acoustic characteristics. It is integrated with hierarchical signal bias removal (HSBR) and further extended to accommodate for N-best candidates. Experimental results on connected digits, recorded over a cellular network, shows that the proposed system reduces both the word and string error rates by about 36% and 31%, respectively, over a baseline system not incorporating bias removal.

1. INTRODUCTION

In this paper, we will focus on an acoustic mismatch between the training and the testing conditions of hidden Markov model (HMM)-based telephone speech recognition systems. This mismatch, which may be caused by variations in the telephone channel and transducer equipments, will be modeled as a time-varying additive bias, b_t, in the cepstral, or log-spectral, domain. Specifically, we assume that the cepstrum of the received signal at time t is given by

$$y_t = x_t + b_t, \tag{1}$$

where x_t is the cepstrum of the undistorted signal. A family of compensation techniques for estimating the bias b_t is presented.

When an acoustic mismatch is known to exist, compensation may be carried out in one of two ways. One is to attempt to bring the testing data "closer" to the acoustic space of the training environment. Lines of research falling within this framework include spectral subtraction [9, 3], signal bias removal [11], cepstral mean subtraction [1], and stochastic matching [12]. Alternatively, we could transform the models to better characterize the test data. Typical techniques include Bayesian learning procedures [8], parallel model combination [5] and stochastic matching [12]. An excellent survey of techniques for robust speech recognition can be found in [7].

Both feature-based and model-based transformations are studied in this paper. A new method is proposed, referred to as codebook-based stochastic matching (CBSM) which applies the expectation-maximization (EM) algorithm [2] to compute a bias for each ensemble of HMM mixture components that share similar acoustic characteristics. CBSM enables integration of bias removal pre-processing methods and can accommodate for multiple search candidates. Our proposed system will be evaluated on a scenario in which a set of wireline-trained models are tested on data collected in a wireless environment. The time-varying nature of this problem provides a challenging testbed for the techniques considered in this paper.

2. ML STOCHASTIC MATCHING

Consider a sequence of feature vectors $Y = \{y_1, \ldots, y_T\}$, s.t. $y_t \in \Re^D$, and a set of trained HMMs Λ_X. Let the observation density for state n with M components be defined as:

$$p_X(y|n) = \sum_{m=1}^{M} w_{n,m} \mathcal{N}(y; \mu_{n,m}, C_{n,m}), \tag{2}$$

where $w_{n,m}$ is the mixture weight and $\mathcal{N}(y; \mu_{n,m}, C_{n,m})$ is a Gaussian distribution with corresponding mean vector $\mu_{n,m}$ and covariance matrix $C_{n,m}$. Given an estimated bias sequence $\hat{B} = \{\hat{b}_1, \ldots, \hat{b}_T\}$, we allow the action of *bias removal* to be either a feature-space transformation $\hat{x}_t = y_t - \hat{b}_t, \quad t = 1, \ldots, T$, or, a model-space transformation (model adaptation) $\hat{\Lambda}_Y = M_{\hat{B}}(\Lambda_X)$, where $M_{\hat{B}}(\cdot)$ is determined by the particular method being applied. In either case, the problem of robust speech recognition, in the framework we have chosen, is reduced to that of computing the bias estimate $\hat{B}$.

The application of the EM algorithm for estimating the bias B has been described in the stochastic matching (SM) framework proposed by Sankar and Lee [12]. Given a word (or phone) sequence $\hat{W}$, the bias estimate

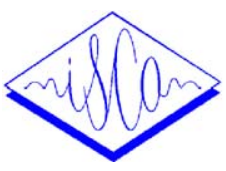
$$\hat{B} = \arg\max_{B} p(Y|B, \hat{W}, \Lambda_X) p(\hat{W}) \tag{3}$$

can be computed by maximizing the following auxiliary function:[1]

$$Q(\hat{B}|B) = E\{\log p(Y|\hat{B}, \hat{W}, \Lambda_X)|Y, \hat{W}, B, \Lambda_X\}. \tag{4}$$

*This work was performed at AT&T Labs- Research during the summer of 1996.

[1] Since $\hat{W}$ is fixed at this stage, the term $p(\hat{W})$ may be omitted.



EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

Upon conditioning the signal (or model), we can solve for the optimal W, i.e., $\hat{W}$, and reiterate this procedure to further refine the bias estimate.

At this point we make no assumptions regarding the domain in which the bias is applied (i.e., whether it is frame-based, state-based, etc.), other than it is additive as proposed in Eqn. (1). In general, let $\hat{b}(t,n,m)$ be the bias estimate associated with frame t, state n, and mixture component m. It can be shown (following the argument of [12]) that the auxiliary function reduces to

$$Q(\hat{B}|B) = \sum_{t=1}^{T}\sum_{n=1}^{N}\sum_{m=1}^{M}\gamma_t(n,m) \cdot \{-\frac{1}{2}[y_t - \hat{b}(t,n,m) - \mu_{n,m}]' C_{n,m}^{-1}[y_t - \hat{b}(t,n,m) - \mu_{n,m}]\}, \quad (5)$$

where $\gamma_t(n,m) = p(Y, s_t = n, c_t = m|B,\Lambda_X)$ and N is the number of states. As $Q(\hat{B}|B)$ is a concave function of $\hat{B}$, the optimal bias may be obtained by differentiating the above expression and solving for the zeros.

3. CODEBOOK-BASED STOCHASTIC MATCHING (CBSM)

In [6], we demonstrated that the recognition performance can be enhanced when computing multiple biases for each utterance. On the other hand, we have experienced no improvement when relaxing the stochastic constraints by increasing the bias resolution, such that a separate bias is assigned to each mixture component. To provide a trade-of between increasing the bias resolution and relaxing the stochastic constraints, we propose to associate each bias to a group of mixture components that share similar acoustic characteristics. The so called codebook-based SM (CBSM) approach uses the concept of "tying" among model parameters to determine the appropriate number of biases to be used.

Let $\{\mu_{n,m}\}$ be the set of mixture component mean vectors for Λ_X, where n and m represent the state and mixture component indices, respectively. Let $\Omega_1,\ldots,\Omega_K$ be K classes (or codewords) that span the entire space of the models with associated set of centroids $z_1,\ldots,z_K$. In the current study, the codebook Ω is constructed by clustering $\{\mu_{n,m}\}$ using the Lloyd algorithm with a Euclidean distance. For notational simplicity, we will assume that the elements of the sets Ω_k are the tuples which index the corresponding mixture component means (e.g., if $\mu_{n,m}$ is clustered to the codeword z_k, then $(n,m) \in \Omega_k$). Since we are interested in computing a bias for each codeword, then

$$\hat{b}(t,n,m) = \sum_{k=1}^{K}\hat{b}_k \mathbf{I}_{\Omega_k}(n,m), \quad (6)$$

where $\mathbf{I}_{\Omega_k}(\cdot)$ is the indicator function for the set Ω_k. Using this notation, we may rewrite the auxiliary function (5) as

$$Q(\hat{B}|B) = \sum_{t=1}^{T}\sum_{k=1}^{K}\sum_{(n,m)\in\Omega_k}\gamma_t(n,m) \cdot \left\{-\frac{1}{2}[y_t - \hat{b}_k - \mu_{n,m}]' C_{n,m}^{-1}[y_t - \hat{b}_k - \mu_{n,m}]\right\} \quad (7)$$

By differentiating this expression with respect to $\hat{b}_k$, and setting to zero, we may solve for the class biases, $\{\hat{b}_k^{(i)}\}_{i=1,D;k=1,K}$, yielding

$$\hat{b}_k^{(i)} = \frac{\sum_{t=1}^{T}\sum_{(n,m)\in\Omega_k}\gamma_t(n,m)\frac{y_{t,i} - \mu_{n,m,i}}{\sigma_{n,m,i}^2}}{\sum_{t=1}^{T}\sum_{(n,m)\in\Omega_k}\frac{\gamma_t(n,m)}{\sigma_{n,m,i}^2}}. \quad (8)$$

The means of mixture components are updated based on their codebook association, such that

$$\hat{\mu}_{n,m} = \mu_{n,m} - \sum_{k=1}^{K}\hat{b}_k \mathbf{I}_{\Omega_k}(n,m), \quad \forall(n,m). \quad (9)$$

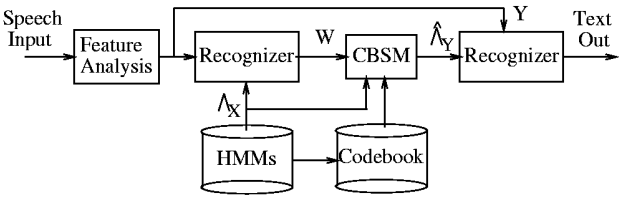


Figure 1: A block diagram of an ASR system incorporating CBSM.

A schematic characterization of CBSM is shown in Figure 1. Following feature analysis, the first-pass of recognition is performed using the original models Λ_X along with the features Y. The resultant transcription W and state segmentation are then used for model bias removal. CBSM is applied at this point aided by the codebook Ω_k to generate a new set of HMMs, $\hat{\Lambda}_Y$. A second-pass through the recognizer is then conducted using $\hat{\Lambda}_Y$ and Y to generate the recognized string.

3.1. INTEGRATED HSBR/CBSM

The basic strategy of hierarchical signal bias removal (HSBR) is described in two steps [11]. For a given codebook $\Omega = \{\mu_1,\ldots,\mu_K\}$ of size K, the first step includes estimating a bias for each codeword, i.e., $\{\hat{b}_k\}_{k=1}^K$. In the second step, a frame-dependent bias $\hat{b}_t$ is computed as a function of $\{\hat{b}_k\}_{k=1}^K$, and subtracted from the input speech signal at each time frame t, yielding the conditioned signal $\hat{x}_t$. The likelihood function in HSBR is based on a weighted acoustic distance with respect to the codebook Ω and neglects all the temporal constraints embedded in the

HMMs. The process may be iterated until some optimality criterion is satisfied. The hierarchical aspect of HSBR was inspired by [4] and operates by performing the signal conditioning as described above for multiple codebooks of increasing size in succession, beginning with $K = 1$ and ending with $K = K_{\max}$. The primary advantage of HSBR is that it is a one-pass scheme that generates a time-varying bias. Its primary disadvantage is that it is based on a memoryless system that does not directly minimize an acoustic mismatch between the HMM and the test signal.

CBSM, as illustrated in Figure 1, is a model-based transformation in which bias removal is applied to the model parameters leaving the original signal Y unchanged. There is clearly no reason why the recognition performance cannot be further enhanced by applying bias removal both at the feature level as well as the model level. In fact, applying signal conditioning, as it is the case in HSBR, can further help in enhancing the signal prior to recognition, thus improving the efficiency of the recognizer and minimizing search errors due to incomplete decoding paths. This is particularly important in situations where frequent beam failures are occuring due to a severe acoustic mismatch between the testing features and the training model. In these circumstances, SM-based approaches have little hope in improving performance since no segmentation information would be available.

In this study, we have integrated CBSM and HSBR for added robustness so that both feature and model space transformations are performed simultaneously. The procedure is similar to that shown in Figure 1 with the addition of the HSBR module prior to the first pass of recognition. Therefore, rather than applying y_t at each time frame in CBSM, we would simply use the conditioned signal $\hat{x}_t$.

3.2. N-BEST CBSM

Ideally, to improve recognition performance through SM, one would like to reduce the mismatch between the *correct* hypothesis and the model parameters. In practice, however, since SM deals with the *recognized* hypothesis, this level of performance can probably never be achieved. One possibility for reducing this problem is to provide bias removal with multiple candidates (i.e., alternative transcriptions) with the hope that one of the candidates is truly the correct one. In this study, we have incorporated multiple candidates during bias removal using an N-best search. From our experience with digit recognition, we have found that for those strings that are incorrectly recognized 35-60% of the time the correct string appears in the top four candidates.

Our approach is to incorporate the statistics of the N-best candidates into the process of estimating the bias. During the first-pass through the recognizer, N-best candidates are generated with corresponding state segmentations $\{\hat{s}_t^l\}_{t=1}^T$, $l = 1, \ldots, N$. The computed bias for the k^{th} codeword is

$$\hat{b}_k^{(i)} = \frac{\sum_{l=1}^{N}\sum_{t=1}^{T}\sum_{(n,m)\in\Omega_k}\gamma_t^l(n,m)\frac{y_{t,i}-\mu_{n,m,i}}{\sigma^2_{n,m,i}}}{\sum_{l=1}^{N}\sum_{t=1}^{T}\sum_{(n,m)\in\Omega_k}\frac{\gamma_t^l(n,m)}{\sigma^2_{n,m,i}}}, \quad (10)$$

$i = 1, \ldots, D$, $k = 1, \ldots, K$, where $\gamma_t^l(n,m)$ is computed for the l^{th} candidate. The computed biases are then used to update all model parameters according to Eqn. 9. This approach of using N-best candidates with CBSM will be referred to as NB-CBSM.

4. EXPERIMENTAL RESULTS

A speaker-independent telephone-based connected digits database was used in this study. Digit strings ranging from one to sixteen digits in length were extracted from different field-trial collections with varied environmental conditions and transducer equipment. The training set consisted of 16089 digit strings and were used for designing the recognition models. The testing set consisted of 402 nine-digit strings collected in a wireless telephone environment from speakers using mostly analog handsets. The recording conditions varied from call to call, ranging from 5dB to 35dB SNR.

During feature extraction, each 30 msec frame was represented by 12 linear predictive coding (LPC) liftered cepstral coefficients, along with a normalized logarithmic energy. This feature vector was augmented with its first and second order time derivatives, the so called delta cepstrum/delta energy and delta-delta cepstrum/delta-delta energy, resulting in a vector of 39 features per frame. For recognition, each digit was modeled by a set of left-to-right continuous density HMMs which captured all possible inter-digit coarticulation [10].[2] A total of 274 context-dependent subword models were used, trained by ML estimation. Subword models consisted of 3 to 4 states, each having a mixture of 8 Gaussian distributions. The background noise model (i.e., silence) included a single state with 32 Gaussian distributions. All experiments were performed using the same training model without any type of conditioning or bias removal and with a known length grammar.

Figure 2 illustrates the effect of varying the CBSM codebook size (i.e., number of biases utilized) on the recognition performance. The experiment which was conducted with 0, 1, 2, 4, 8, 16 and 32 biases seem to indicate that a CBSM codebook size of 8 and beyond produces the lowest error rate. Therefore, the CBSM codebook was fixed at 8 in all remaining experiments.

Results for integrating HSBR and CBSM are shown in Table 1 (column 3). These results were obtained when first processing the signal with HSBR and

[2] Each digit is divided into three segments, a head, a body and a tail. A digit has one body and multiple heads and tails depending on the preceding and following context.

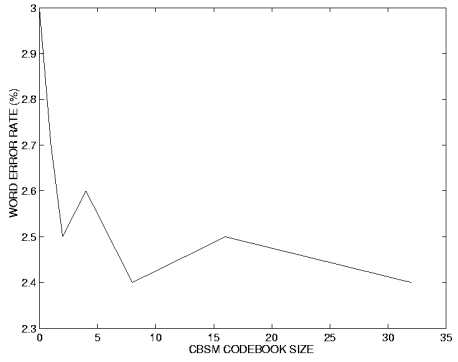


Figure 2: Variations of the word error rate as a function of the CBSM codebook size

then performing CBSM as described in Section 3.1. They translate to a 27% and 23% reduction in the word and string error rates, respectively, when compared to those of the baseline (see column 2). Also note that the rejection rate when applying integrated HSBR/CBSM was down to zero. This is attributed to applying HSBR. Column 4 presents the performance when integrating HSBR with NB-CBSM as described in Section 3.2. Four best candidates were chosen for this experiment in computing the biases in Eqn. 10. When compared to the baseline system, integrated HSBR/NB-CBSM has resulted in 36% and 31% reduction in the word and string error rates, respectively, with a zero rejection rate. These results are encouraging considering that no retraining was performed at any stage in our experiments.

5. CONCLUSIONS

A codebook-based stochastic matching method was proposed in this paper. CBSM associated a separate bias with an ensemble of mixture components that shared similar acoustic characteristics. This method was found to produce improvement in recognition performance when using 8 or more codewords (or biases). Further, we integrated hierarchical signal bias removal with an extension of CBSM that accommodated for four best candidates. Without having to retrain the original HMM system, we experienced a reduction in the word and string error rates by 36% and 31%, respectively, when testing on data collected from a cellular environment. The major disadvantage of stochastic matching based techniques is that they require multiple passes through the recognizer. In Table 1, we found HSBR/CBSM and HSBR/NB-CBSM to increase the processing time by a factor of two to four folds. This is a major problem when trying to operate a real-time service. A bias removal framework which can accommodate for sequential processing is clearly needed and remains a challenging problem in speech recognition research.

Table 1: Baseline results for digit recognition before and after processing with HSBR/CBSM and HSBR/NB-CBSM.

	Baseline	HSBR/ CBSM	HSBR/ NB-CBSM
Word Error (%)	3.3	2.4	2.1
String Error (%)	14.1	10.8	9.7
Rejection (%)	1.0	0.0	0.0
CPU (sec)	1831	3419	7063

6. REFERENCES

1. B. Atal. Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *J. Acoust. Soc. Am.*, 55:1304–1312, 1974.
2. A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statisc. Soc.*, 39:1–38, 1977.
3. Y. Ephraim, D. Malah, and B.-H. Juang. On the application of hidden markov models for enhancing noisy speech. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37:1846–1856, 1989.
4. S. Furui. Unsupervised speaker adaptation based on hierarchical spectral clustering. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 37:1923–1930, 1989.
5. M. J. F. Gales and S. Young. An improved approach to hidden markov model decomposition of speech and noise. In *Proc. Int. Conf. Acoust., Speech, Signal Processing*, pages 223–226, 1992.
6. C. Lawrence and M. Rahim. Integrated bias removal techniques for robust speech recognition. *submitted to Comput. Speech Language*, 1997.
7. C.-H. Lee. On feature and model compensation approach to robust speech recognition. In *Robust Speech Recognition for Unknown Communication Channels*, pages 45–54, 1997.
8. C.-H. Lee, C.-H. Lin, and B.-H. Juang. A study on speaker adaptation of the parameters of continuous density hidden markov models. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 39(4):806–814, 1991.
9. D. Mansour and B.-H. Juang. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 29:113–120, 1979.
10. R. Pieraccini and A. E. Rosenberg. Coarticulation models for continuous digit recognition. In *Proc. Acoust. Soc. Am.*, page 106, May 1990.
11. M. G. Rahim, B.-H. Juang, W. Chou, and Buhrke E. Signal conditioning techniques for robust speech recognition. *IEEE Signal Processing Letters*, 3(2):107–109, April 1996.
12. A. Sankar and C.-H. Lee. A maximum-likelihood approach to stochastic matching for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 4(3):190–202, May 1996.