Compensating for Noise and Mismatch in Speaker Verification Systems Using Approximate Bayesian Inference

Ciira wa Maina and John MacLaren Walsh

Abstract— **This paper presents a feature domain approach to the problem of robust speaker verification in noisy acoustic environments. We derive a variational Bayesian algorithm that enhances the log spectra of noisy speech using speaker dependent priors. This algorithm extends prior work by Frey *et al.* where the Algonquin algorithm was introduced to enhance speech log spectra in order to improve speech recognition in noisy environments. Our work is built on the intuition that speaker dependent priors would work better than priors that attempt to capture global speech properties. Experimental results using the TIMIT data set and the MIT Mobile Device Speaker Verification Corpus (MDSVC) are presented to demonstrate the algorithm's performance.**

Index Terms— **Speaker verification, variational Bayesian inference.**

I. Introduction

The performance of speaker verification systems is affected by noise and mismatch between training and operating conditions. Therefore, systems that are robust to noise continue to be of interest (for example see [1]).

There are two main approaches to noise robust speaker recognition namely the model-domain approach and the feature-domain approach [2]. In the model-domain approach, speaker models are adapted to account for the various acoustic environments in which the system will be used [3]. Another model-domain approach involves training different models for different acoustic conditions. In [1] the authors present a system based on multicondition training where the speaker models are derived from speech distorted by different types of noise at various signal-to-noise ratios (SNRs).

In the feature domain approach, the speech or features derived from the speech such as log spectral parameters are enhanced to mitigate the effects of noise on the features. Speech enhancement is an important area of research and there are a number of techniques such as spectral subtraction and statistical model based speech enhancement algorithms [4]. Cepstral mean subtraction (CMS) and RASTA processing are frequently used to mitigate channel effects in the log spectral domain [5]. However, these techniques fail to exploit any prior information about the features. Recently, methods that rely on prior speech and interference models have been proposed [6], [7]. Using these priors, the clean speech features are estimated using Bayesian techniques. The Algonquin speech enhancement algorithm [8], [9] and some extensions [10], [11], [2] apply a variational inference technique to enhance noisy reverberant speech using a speaker independent mixture of Gaussians speech prior in the log spectral domain.

Another feature domain approach that has recently received significant attention is nuissance attribute projection (NAP) which was originally developed for use in support vector machines[12]. Recent work has extended NAP for use in feature compensation [13]. Here, the space in which the features live is assumed to contain a smaller subspace of nuissance attributes due to noise and channel distortion. A projection matrix applied to the observations can zero components in the direction of the nuissance space. This is similar to the approach introduced by Kenny *et al.* [14], [15] which is a model-domain technique. Here the means of a background Gaussian mixture model are adapted at enrollment time to determine the speaker dependent means. The technique is similar to the classical maximum *a posteriori* (MAP) adaptation technique used in state of the art speaker verification systems and is known as eigenvoice MAP. In eigenvoice MAP, the background model means are modified using a linear combination of the eigenvoice vectors which span the speaker space.

In this work we extend the Algonquin speech enhancement algorithm to use speaker dependent log spectrum priors and derive a variational Bayesian algorithm for inference. In addition to cleaning features in a manner that bears resemblance to Algonquin, our novel algorithm tackles the problem of performing speaker verification jointly with enhancing the speech. This allows speaker specific model information to be utilized to help clean the speech. In our earlier work, an acoustic domain variational Bayesian (VB) speech enhancement algorithm was derived which relied on speaker dependent speech priors in the autoregressive parameter domain [16]. Mel Frequency Cepstral Coefficients (MFCCs) obtained from the enhanced speech were shown to improve speaker identification in noisy acoustic environments. In this work, we enhance the log spectra using speaker dependent priors since the log spectra are 'closer' to the MFCC domain which is the most desirable domain for speaker recognition.

Variational inference methods have emerged as a powerful class of approximate inference techniques. In this approach inference is viewed as an optimization problem where an appropriate cost function is minimized [17]. Variational Bayesian inference [18], belief propagation (BP) and expectation propagation (EP)[19] fall in this category.

Variational Bayesian methods have been successfully applied to several signal processing problems such as source separation [20] and parameter estimation [21]. This provides motivation for the work presented here where variational Bayesian (VB) techniques are used to improve speaker verification performance in noisy environments.

The rest of the paper is organized as follows. In section II we present the problem formulation and characterize the joint distribution of the parameters and observations in our model. In section III we give a brief introduction to variational Bayesian inference and present the variational approximation to the true posterior. Experimental results on the TIMIT data set and the MIT Mobile Device Speaker Verification Corpus are presented in section IV. Section V presents a discussion and concludes the paper.

II. Problem Formulation

We consider the enhancement of log-spectra of observed speech in order to improve the performance of speaker verification systems by using speaker specific speech priors in the log spectrum domain. In [22] an approximate relationship between the log spectra of observed speech and clean speech is derived. We assume that the clean speech is corrupted by a channel and additive noise. We have

$$y[t] = h[t] * s[t] + n[t], \tag{1}$$

where $y[t]$ is the observed speech, $h[t]$ is the impulse response of the channel, $s[t]$ is the clean speech $n[t]$ is the additive noise and $*$ denotes convolution.

Taking the DFT and assuming that the frame size is of sufficient length compared to the length of the channel impulse response we get

$$Y[k] = H[k]S[k] + N[k],$$

where k is the frequency bin index. Taking the logarithm of the power spectrum $\mathbf{y} = \log|Y[:]|^2$ it can be shown that [22]

$$\mathbf{y} \approx \mathbf{s} + \mathbf{h} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{h} - \mathbf{s})) \tag{2}$$

where $\mathbf{s} = \log|S[:]|^2$, $\mathbf{h} = \log|H[:]|^2$ and $\mathbf{n} = \log|N[:]|^2$. The approximate observation likelihood is given by

$$p(\mathbf{y}|\mathbf{s},\mathbf{h},\mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s}+\mathbf{h}+\log(\mathbf{1}+\exp(\mathbf{n}-\mathbf{h}-\mathbf{s})),\boldsymbol{\psi}) \tag{3}$$

where $\boldsymbol{\psi}$ is the covariance matrix of the modelling errors which are assumed to be Gaussian with zero mean.

In this work we assume that we can mitigate channel effects using methods such as mean subtraction and concentrate on mitigating the effects of additive distortion. In this case the observation likelihood becomes

$$p(\mathbf{y}|\mathbf{s},\mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s})), \boldsymbol{\psi}).$$

To complete the probabilistic formulation we introduce priors over $\mathbf{s}$ and $\mathbf{n}$. For a given speaker ℓ the prior over $\mathbf{s}$ is given by

$$p(\mathbf{s}|\ell) = \sum_{m=1}^{M_s} \pi^s_{\ell m} \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}^s_{\ell m}, \boldsymbol{\Sigma}^s_{\ell m}) \tag{4}$$

where $\ell \in \mathcal{L} = \{1, 2, \ldots, |\mathcal{L}|\}$ with $\mathcal{L}$ being the library of known speakers.

We find it analytically convenient to introduce an indicator variable $\mathbf{z}_s$ that is a $M_s|\mathcal{L}| \times 1$ random binary vector that captures both the identity of the speaker and the mixture coefficient 'active' over a given frame. We have

$$p(\mathbf{s}|\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} \left[\mathcal{N}(\mathbf{s}; \boldsymbol{\mu}^s_i, \boldsymbol{\Sigma}^s_i)\right]^{z_{s,i}}, \tag{5}$$

and

$$p(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} (\pi^s_i)^{z_{s,i}}. \tag{6}$$

We assume that the noise is well modelled by a single Gaussian. That is

$$p(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n). \tag{7}$$

This simplifies the derivation of the posterior and is sufficient for the noise types considered here. Extension to the Gaussian mixture model case is straightforward.

We can now write the joint distribution of this model as

$$p(\mathbf{y}, \mathbf{s}, \mathbf{z}_s, \mathbf{n}) = p(\mathbf{y}|\mathbf{s}, \mathbf{n})p(\mathbf{s}|\mathbf{z}_s)p(\mathbf{z}_s)p(\mathbf{n}). \tag{8}$$

Inference in this model is complicated due to the nonlinear likelihood term. To allow us to derive a tractable variational inference algorithm we linearize the likelihood as in [8], [9].

Let $g([\mathbf{s}, \mathbf{n}]) = \log(\mathbf{1} + \exp(\mathbf{n} - \mathbf{s}))$. We linearize $g(.)$ using a first order Taylor series expansion about the point $[\mathbf{s}_0, \mathbf{n}_0]$. We have

$$g([\mathbf{s}, \mathbf{n}]) \approx g([\mathbf{s}_0, \mathbf{n}_0]) + \nabla g([\mathbf{s}_0, \mathbf{n}_0])([\mathbf{s}, \mathbf{n}] - [\mathbf{s}_0, \mathbf{n}_0]) \tag{9}$$

And the linearized likelihood is

$$\hat{p}(\mathbf{y}|\mathbf{s}, \mathbf{n}) = \mathcal{N}(\mathbf{y}|\mathbf{s} + g([\mathbf{s}_0, \mathbf{n}_0]) + \mathbf{G}([\mathbf{s}, \mathbf{n}] - [\mathbf{s}_0, \mathbf{n}_0]), \boldsymbol{\psi}) \tag{10}$$

Where $\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n] \overset{\text{def}}{=} \nabla g([\mathbf{s}_0, \mathbf{n}_0])$ with

$$\begin{aligned}
\mathbf{G}_s &= \operatorname{diag}\left[\frac{-\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{-\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right] \\
\mathbf{G}_n &= \operatorname{diag}\left[\frac{\exp(n_0^1 - s_0^1)}{1 + \exp(n_0^1 - s_0^1)}, \ldots, \frac{\exp(n_0^N - s_0^N)}{1 + \exp(n_0^N - s_0^N)}\right]
\end{aligned}$$

where N is the dimension of the Log-spectrum feature vector.

We can now derive a variational Bayesian inference algorithm to enhance the observed log spectrum.

III. Variational Bayesian Inference

In variational Bayesian inference, we seek an approximation $q(\Theta)$ to the intractable posterior $p(\Theta|\mathbf{y})$ over the model parameters Θ which minimizes the Kullback-Leibler (KL) divergence between $q(\Theta)$ and $p(\Theta|\mathbf{y})$ with $q(\Theta)$ constrained to lie within a tractable approximating family (in our case $\Theta = \{\mathbf{s}, \mathbf{z}_s, \mathbf{n}\}$). The KL divergence $D(q||p)$ is a measure of the distance between two distributions and is defined by [23]

$$D(q||p) = \int q(\Theta) \log \frac{q(\Theta)}{p(\Theta|\mathbf{y})} d\Theta.$$

To ensure tractability, the approximating family is selected such that the approximate posterior can be written as a product of factors depending on disjoint subsets of $\Theta = \{\theta_1, \dots, \theta_M\}$ [18], [24]. Assuming that each factor depends on a single element of Θ then

$$q(\Theta) = \prod_{i=1}^{M} q_i(\theta_i). \tag{11}$$

It can be shown that the optimal form of $q_j(\theta_j)$ denoted by $q_j^*(\theta_j)$ that minimizes $D(q||p)$ is given by [24]

$$\log q_j^*(\theta_j) = \mathbb{E}\{\log p(\mathbf{y}, \Theta)\}_{q(\Theta^{\setminus j})} + const. \tag{12}$$

We use the notation $q(\Theta^{\setminus j})$ to denote the approximate posterior of all the elements of Θ except θ_j. We obtain a set of coupled equations relating the optimal form of a given factor to the other factors. To solve these equations, we initialize all the factors and iteratively refine them one at a time using (12).

A. Approximate Posterior

Returning to the context of our model, we assume an approximate posterior $q(\Theta)$ that factorizes as follows

$$q(\Theta) = q(\mathbf{s})q(\mathbf{z}_s)q(\mathbf{n}). \tag{13}$$

The factorization used in this work differs from that in Frey *et al.* [8] by enforcing independence between the mixture coefficient indicator variable and the clean log spectra. Thus instead of a mixture of Gaussians posterior over the clean log spectra we have a single Gaussian. Additionally, the algorithm has been designed to jointly verify the speaker and enhance the speech using this information. In [8] the factiorization is

$$q(\Theta) = q(\mathbf{n}) \sum_{m=1}^{M} \rho_m q(\mathbf{s}|m) \tag{14}$$

where ρ_m is the posterior probability of the mth mixture component. The optimal forms of the approximate posterior when the factorization (14) is assumed are as follows

$$q(\Theta) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*) \sum_{m=1}^{M} \rho_m \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{\mathbf{s}}^{m,*}, \boldsymbol{\Sigma}_{\mathbf{s}}^{m,*}).$$

The update equations resulting from this factorization are presented in [9].

Using (12) we obtain expressions for the optimal form of the factors for the factorization used in this work given by (13). We obtain

1)

$$q^*(\mathbf{s}) = \mathcal{N}(\mathbf{s}; \boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*) \tag{15}$$

with

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{s}}^* &= \Big[\boldsymbol{\psi}^{-1} + \mathbf{G}_s^T \boldsymbol{\psi}^{-1} \mathbf{G}_s + \boldsymbol{\psi}^{-1} \mathbf{G}_s \\ &+ \mathbf{G}_s \boldsymbol{\psi}^{-1} + \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i \boldsymbol{\Sigma}_i^{s-1}\Big]^{-1} \\ \boldsymbol{\mu}_{\mathbf{s}}^* &= \boldsymbol{\Sigma}_{\mathbf{s}}^* \Big[(\mathbf{I} + \mathbf{G}_s^T)\boldsymbol{\psi}^{-1}(\mathbf{y} - g([\mathbf{s}_0, \mathbf{n}_0]) \\ &- \mathbf{G}_n \boldsymbol{\mu}_{\mathbf{n}}^* + \mathbf{G}_s \mathbf{s}_0 + \mathbf{G}_n \mathbf{n}_0) \\ &+ \sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i \boldsymbol{\Sigma}_i^{s-1} \boldsymbol{\mu}_i^s\Big] \end{aligned}$$

2)

$$q^*(\mathbf{n}) = \mathcal{N}(\mathbf{n}; \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*) \tag{16}$$

with

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathbf{n}}^* &= \Big[\mathbf{G}_n^T \boldsymbol{\psi}^{-1} \mathbf{G}_n + \boldsymbol{\Sigma}_n^{-1}\Big]^{-1} \\ \boldsymbol{\mu}_{\mathbf{n}}^* &= \boldsymbol{\Sigma}_{\mathbf{n}}^* \Big[\mathbf{G}_n^T \boldsymbol{\psi}^{-1}(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{s}}^* - g([\mathbf{s}_0, \mathbf{n}_0]) - \mathbf{G}_s \boldsymbol{\mu}_{\mathbf{s}}^* \\ &+ \mathbf{G}_s \mathbf{s}_0 + \mathbf{G}_n \mathbf{n}_0) + \boldsymbol{\Sigma}_n^{-1} \boldsymbol{\mu}_n\Big] \end{aligned}$$

3)

$$q^*(\mathbf{z}_s) = \prod_{i=1}^{M_s|\mathcal{L}|} (\gamma_i)^{z_{s,i}} \tag{17}$$

where

$$\gamma_i = \frac{\rho_i}{\sum_{i=1}^{M_s|\mathcal{L}|} \rho_i}$$

and

$$\begin{aligned} \log \rho_i &= -\frac{1}{2}(\boldsymbol{\mu}_{\mathbf{s}}^* - \boldsymbol{\mu}_i^s)^T \boldsymbol{\Sigma}_i^{s-1} (\boldsymbol{\mu}_{\mathbf{s}}^* - \boldsymbol{\mu}_i^s) \\ &- \frac{1}{2}\log|\boldsymbol{\Sigma}_i^s| - \frac{1}{2}\mathrm{Tr}(\boldsymbol{\Sigma}_i^{s-1}\boldsymbol{\Sigma}_{\mathbf{s}}^*) + \log \pi_i^s. \end{aligned}$$

B. The VB Algorithm

To run the algorithm, the observed utterance is divided into K frames and each frame is enhanced. The linearization point is critical to the performance of the algorithm. As in [8], [9] we linearize the likelihood at the current estimate of the posterior mean $[\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*]$. The overall algorithm is summarized in algorithm 1.

```
for $k = 1, \dots, K$ do
    Initialize the posterior distribution parameters
    $\{\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*, \gamma_i\}$;
    for $n = 1$ to Number of Iterations do
        Set $[\mathbf{s}_0, \mathbf{n}_0] = [\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*]$;
        Compute $\mathbf{G} = [\mathbf{G}_s, \mathbf{G}_n]$ and $g([\mathbf{s}_0, \mathbf{n}_0])$;
        Update $\{\boldsymbol{\mu}_{\mathbf{s}}^*, \boldsymbol{\Sigma}_{\mathbf{s}}^*, \boldsymbol{\mu}_{\mathbf{n}}^*, \boldsymbol{\Sigma}_{\mathbf{n}}^*\}$ using (15)-(16);
        Update $\gamma_i$ using (17);
    end
end
```

Algorithm 1: VB algorithm

C. Computational Complexity

The computational complexity of the algorithm is dominated by the cost to update the posterior distribution of the clean speech log spectra. From equation (15) we see that the computation of $\boldsymbol{\mu}_{\mathbf{s}}^*$ is dominated by the term $\sum_{i=1}^{M_s|\mathcal{L}|} \gamma_i \boldsymbol{\Sigma}_i^{s-1} \boldsymbol{\mu}_i^s$. Since the model covariance matrices are diagonal, evaluation of each term has a computational complexity of $O(N)$ where N is the dimension of the log spectral features. Thus each update of the mean parameters has a computational cost of $O(M_s|\mathcal{L}|N)$ which is linear in the number of mixture coefficients.

IV. Experimental Results

In this section we present experimental results that verify the performance of the algorithm. For the simulations we use the TIMIT database and the MIT Mobile Device Speaker Verification Corpus (MDSVC)[25]. The experiments investigate the equal error rate (EER) improvement obtained when the VB log spectral enhancemnent algorithm is used in speaker verification systems in noisy environments. To obtain noisy speech from TIMIT data, we add additive white Gaussian noise and realistic factory noise.

The TIMIT data set contains recordings of 630 speakers drawn from 8 dialect regions across the USA with each speaker recording 10 sentences [26]. The sampling frequency of the utterances is 16kHz with 16 bit resolution. In order to train the speaker models we used 8 sentences and used the other 2 for testing. The MIT Mobile Device Speaker Verification Corpus is a data set that is designed to test speaker verification systems with limited enrollment data in noisy acoustic conditions. The speech data consists of recordings of speakers saying ice cream flavor phrases and names. The recordings are done in an office, hallway and street intersection in order to provide realistic noisy speech.

A. System Descriptions

1) Baseline System: The basic task is to determine whether a given speaker is speaking in a particular speech segment. Thus given a speech segment X we test the following hypotheses

- H0: X is from speaker S
- H1: X is not from speaker S

Here the target speakers are modelled using speaker specific GMMs and a universal background model (UBM) is used to test the alternate hypothesis H1. The likelihood ratio is compared to a threshold in order to determine which hypothesis is correct. For each trial we compute the score

$$\mathsf{Score} = \log p(\mathbf{X}|\mathrm{TargetModel}) - \log p(\mathbf{X}|\mathrm{UBM}). \quad (18)$$

where $\mathbf{X}$ are the features computed from the test utterance. For the baseline system we use 13 dimensional MFCCs generated every 10ms using a 25ms window as features. Using the feature vectors extracted from training speech, we train speaker GMMs with 32 mixture coefficients.

2) Log Spectrum System: This system uses the log spectrum of the speech frames as features. Log spectra are generated every 10ms using a 25ms window which corresponds to 400 samples at 16kHz. The FFT length is 512 resulting in a feature vector of length 257. Using the feature vectors extracted from training speech, we train speaker GMMs with 8 mixture coefficients.

3) Variational Bayesian System: For this system, we form a library consisting of the target speaker and the UBM and run algorithm 1 to enhance the noisy log spectra. As with any iterative algorithm, initialization is very important and it affects the quality of the final solution. In our experiments, the following initialization scheme was found to work well: We initialize the posterior mean of the speech log spectrum, $\boldsymbol{\mu}_{\mathbf{s}}^*$, to the log spectrum of the noisy speech frame. The posterior covariance of the speech log spectrum, $\boldsymbol{\Sigma}_{\mathbf{s}}^*$, was initialized as the identity matrix. We initialize the posterior mean of the noise log spectrum, $\boldsymbol{\mu}_{\mathbf{n}}^*$, to the all zero vector. The posterior covariance of the noise log spectrum, $\boldsymbol{\Sigma}_{\mathbf{s}}^*$, was initialized as the identity matrix. Finally we initialize the parameters of $q(\mathbf{z_s})$ as $\gamma_i = \frac{1}{M_s|\mathcal{L}|}$.

Since we update the posterior parameters one at a time, we need to specify a parameter update schedule. The parameter update schedule is as follows:

1) Update the parameters of $q^*(\mathbf{n})$.
2) Update the parameters of $q^*(\mathbf{s})$.
3) Update the parameters of $q^*(\mathbf{z}_s)$.

This schedule was observed in simulation to be numerically stable.

For our experiments, the algorithm was run for 5 iterations and the posterior mean of the speech log spectrum at the final iteration was used as the enhanced log spectrum of that frame. Using the enhanced log spectra for a given utterance, scores for each verification trial are computed using (18).

We also derive MFCCs from the enhanced log spectra and use these to compute scores for each verification trial. Thus for the VB system we have two results: one using the enhanced log spectra and the other using the MFCCs derived from these log spectra.

B. TIMIT Speaker Verification Results

We now turn to experiments aimed at determining the speaker verification performance of the systems in noisy conditions. We assume that the TIMIT data is clean and the SNR only accounts for the additive distortion we introduce. In this work the input SNR is defined as

$$\mathrm{SNR}_{in} \quad = \quad 10 \log \frac{\sum_t s^2[t]}{\sum_t (s[t] - y[t])^2}.$$

The UBMs were trained using the training data for a random 300 speaker subset of the 630 speaker TIMIT data set. The MFCC UBMs and speaker models had 32 mixtures while the log spectra UBMs and speaker models had 8 mixtures.

The verification experiments were performed with the test utterances corrupted by additive white Gaussian noise at various input SNRs. For each of the 630 speakers we have two test utterances yielding 1260 true trials. To generate impostor trials, a random set of ten speakers was selected from the remaining speakers and the corresponding test utterances used to generate 20 impostor trials per speaker. Thus there are a total of 12600 impostor trials.

Table I shows the equal error rates (EER) obtained in our verification experiments for the three systems at various input SNRs. Figure 1 shows the corresponding DET curves at 30dB. We see that the VB algorithm improves the performance of both the MFCC and log spectral systems. At 30dB the log spectral EER is reduced from 25.97% to 18.02% while the MFCC EER is reduced from 8.97% to 4.44%.

TABLE I

SPEAKER VERIFICATION EER (%) FOR THE ENTIRE TIMIT DATA SET

System	SNR (dB) 10	20	30
MFCCs (Baseline)	46.35	24.44	8.97
VB (MFCC)	31.11	13.97	4.44
Log Spectra	51.11	42.06	25.97
VB (Log Spectra)	42.94	28.73	18.02

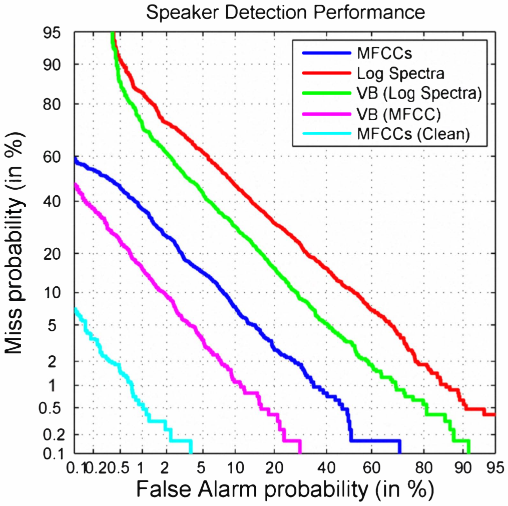

Fig. 1. Speaker verification performance for the entire TIMIT data set at 30dB.

1) TIMIT Speaker Verification Results in Realistic Noise: We now turn to experiments aimed at demonstrating the performance of the algorithm in realisitic noisy conditions. To this end we add noise from the NOISEX 92 data set [27] to the clean TIMIT data at various SNRs. This data set consists of recordings of various types of noise including factory noise and speech babble. The recordings are sampled at 19.98kHz and it is necessary to resample the recordings since TIMIT recordings are sampled at 16kHz.

The experiments using the entire TIMIT data set were repeated using factory noise. Table II shows the equal error rates (EER) obtained in our verification experiments for the three systems at various input SNRs using factory noise. These results are similar to those obtained using white noise. We see a significant improvement in EER at 10dB and 20dB with the MFCCs obtained from enhanced log spectra yielding the best performance.

TABLE II

SPEAKER VERIFICATION EER (%) FOR THE ENTIRE TIMIT DATA SET IN FACTORY NOISE

System	SNR (dB) 10	20	30
MFCCs (Baseline)	23.57	5.71	1.43
VB (MFCC)	10.63	3.17	1.43
Log Spectra	44.44	38.17	35.24
VB (Log Spectra)	39.92	36.03	35.48

C. MDSVC Speaker Verification Results

In the MDSVC data set, each speaker records 54 utterances in two sessions, one for training and the other for testing. The 54 utterances are recorded in three conditions: in an office, a hallway and a noisy street intersection. 18 utterances are recorded in each environment. The speaker models are trained using the 18 utterances recorded in an office since these are the closest to clean. Each utterance is approximately two seconds long. There are 48 target speakers in the data set with 22 female speakers and 26 male speakers. There are 40 impostors with 23 male and 17 female.

In our initial experiment we examine the performance of a baseline GMM-UBM speaker verification system. We investigate the EER performance of the system when the test utterances are recorded in the three different environments. Figure 2 shows the corresponding DET curves. We see that mismatch between training and testing data leads to performance degradation. The EER increases from 8.6% to 25.5% when the training data is recorded in an office but the test data is obtained in a noisy street intersection. These EERs are comparable to those obtained in [1, Fig. 7].

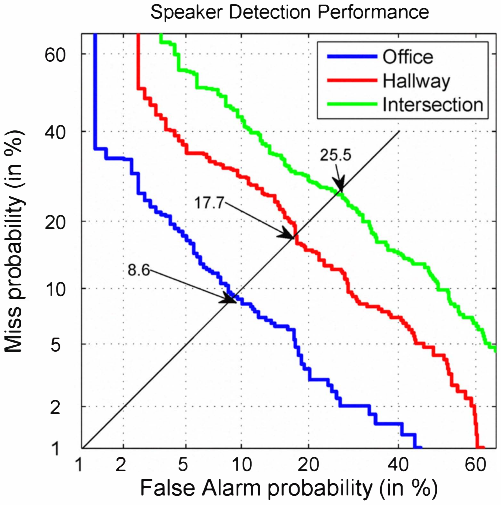

Fig. 2. Baseline GMM-UBM speaker verification system performance for test data drawn from different environments when training data was recorded in an office. These EERs are comparable to the baseline performance obtained in [1, Fig. 7].

In order to investigate the performance of the VB log spectral algorithm on this data set, experiments were performed to determine the EER improvement obtained when the test speech was recorded in various locations with both the MFCC and log spectral models trained using office speech. Table III shows the EERs obtained by the three systems described in section IV-A. Figure 3 shows the corresponding DET curves when the test data is recorded at a noisy street intersection. We see that the VB algorithm significantly improves the EER from 25.51% to 17.93%.

TABLE III
SPEAKER VERIFICATION EER (%) FOR THE MDSVC DATA SET

System	Location	
	Intersection	Hallway
MFCCs (Baseline)	25.51	18.94
VB (MFCC)	17.93	16.97
Log Spectra	37.88	32.32
VB (Log Spectra)	34.09	26.26

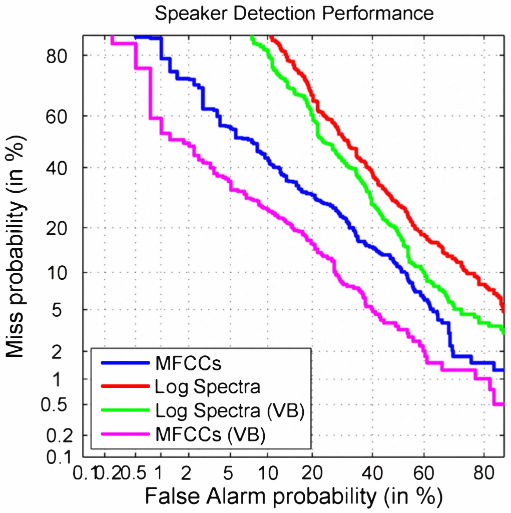

Fig. 3. Speaker verification system performance for test data drawn from a noisy street intersection for the VB log spectral enhancement algorithm.

V. DISCUSSION AND CONCLUSIONS

The experimental results reported in the previous section verify that the proposed log spectrum enhancement algorithm does indeed improve speaker verification in noisy environments. Significant improvements in EER of up to about 14% are obtained using MFCCs derived from enhanced log spectra when compared to MFCCs obtained directly from noisy speech. At 30dB the EER is reduced by about half from 8.97% to 4.44%. Also, the MFCCs obtained from the enhanced log spectra give the best performance at all SNRs reported.

Similarly, improvements in EER are obtained when training data is obtained in an office but test data is recorded at a noisy street intersection. This demonstrates mismatch compensation.

REFERENCES

[1] Ji Ming, T.J. Hazen, J.R. Glass, and D.A. Reynolds. Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(5):1711–1723, July 2007.

[2] Li Deng, J. Droppo, and A. Acero. Estimating cepstrum of speech under the presence of noise using a joint prior of static and dynamic features. *IEEE Transactions on Speech and Audio Processing*, 12(3):218–233, May 2004.

[3] R. Vogt and S. Sridharan. Experiments in session variability modelling for speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, page I, May 2006.

[4] P. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[5] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, (4):578–589, 1994.

[6] Hagai Attias, John C. Platt, Alex Acero, and Li Deng. Speech denoising and dereverberation using probabilistic models. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[7] Jiucang Hao, H. Attias, S. Nagarajan, Te-Won Lee, and T.J. Sejnowski. Speech Enhancement, Gain, and Noise Spectrum Adaptation Using Approximate Bayesian Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):24–37, Jan. 2009.

[8] B. J. Frey, T. T. Kristjansson, L. Deng, and A. Acero. ALGONQUIN Learning dynamic noise models from noisy speech for robust speech recognition. In *Advances in Neural Information Processing Systems 14*, pages 1165–1172, January 2002.

[9] Kristjansson, T. *Speech Recognition in Adverse Environments: a Probabilistic Approach.* PhD thesis, 2002.

[10] Li Deng, J. Droppo, and A. Acero. Recursive estimation of nonstationary noise using iterative stochastic approximation for robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 11(6):568–580, Nov. 2003.

[11] Li Deng, J. Droppo, and A. Acero. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. *IEEE Transactions on Speech and Audio Processing*, 12(2):133–143, March 2004.

[12] A. Solomonoff, C. Quillen, W. M. Campbell. Channel Compensation for SVM Speaker Recognition. In *In Proc. Odyssey: The Speaker and Language Recognition Workshop*, pages 41–44, Toledo, Spain, June 2004.

[13] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface, C. Vair. Compensation of Nuisance Factors for Speaker and Language Recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):1969 – 1978, September 2007.

[14] P. Kenny, G. Boulianne, P. Dumouchel. Eigenvoice Modeling with Sparse Training Data. *IEEE Transactions on Speech and Audio Processing*, 13:345–359, May 2005.

[15] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel. Speaker and session variability in GMM-based speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1448–1460, May 2007.

[16] Ciira wa Maina and John MacLaren Walsh. Joint Speech Enhancement and Speaker Identification Using Approximate Bayesian Inference. In *Conference on Information Sciences and Systems (CISS)*, Mar. 2010.

[17] M. J. Wainwright and M. I. Jordan. A Variational Principle for Graphical Models. In S. Haykin, J. Príncipe, T. J. Sejnowski, and J. McWhirter, editor, *New Directions in Statistical Signal Processing From Systems to Brains*, pages 155–202. MIT press, 2005.

[18] Hagai Attias. A Variational Bayesian Framework for Graphical Models. In *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.

[19] Thomas P. Minka. Expectation Propagation for approximate Bayesian inference. In *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[20] A. Taylan Cemgil, Cédric Févotte, and Simon J. Godsill. Variational and stochastic inference for Bayesian source separation. *Digital Signal Processing*, 17(5):891–913, 2007.

[21] S.J. Roberts and W.D. Penny. Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, 50(9):2245–2257, Sep 2002.

[22] B. Frey, L. Deng, A. Acero, and T. Kristjansson. Algonquin: iterating Laplace's method to remove multiple types of acoustic distortion for robust speech recognition. In *Eurospeech*, pages 901–904, January 2001.

[23] T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley and Sons, 2006.

[24] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[25] R. Woo, A. Park, T. J. Hazen. The MIT Mobile Device Speaker Verification Corpus: Data collection and preliminary experiments. In *Proc. Odyssey: The Speaker and Language Recognition Workshop*, pages 1–6, San Juan, Puerto Rico, 2006.

[26] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett and N.L. Dahlgren. The DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM , 1993. http://www.ldc.upenn.edu/Catalog.

[27] Andrew Varga and Herman J.M. Steeneken. Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, 12(3):247 – 251, 1993.