A Multichannel Feature-Based Processing for Robust Speech Recognition

Mehrez Souden, Keisuke Kinoshita, Marc Delcroix, and Tomohiro Nakatani

NTT Communication Science Laboratories
Signal Processing Research Group
{mehrez.souden, kinoshita.k, marc.delcroix, nakatani.tomohiro}@lab.ntt.co.jp

Abstract

We propose a new approach for multichannel robust speech recognition. This approach extends the vector Taylor series (VTS)-based feature compensation from the single channel to the multichannel case. Precisely, we use the first order VTS to approximate each of the microphone feature vectors. Afterwards, these features are jointly processed to estimate the acoustic channel and noise statistics via expectation maximization (EM). Experimental results with TI-Digits and measured impulse responses show that the proposed method can achieve significant gains in terms of word recognition accuracy in different noise conditions.

Index Terms: Microphone array, vector Taylor series, robust speech recognition, environment compensation.

1. Introduction

In hands-free speech communication applications, the desired speech signals are generally corrupted by both reverberation and additive acoustic noise. Both distortions are known to have detrimental effects on speech recognition. Hence, reliable environment compensation techniques are essential to process the noisy and reverberant microphone observations before performing recognition.

Essentially, the robustness to acoustic environment distortions in distant-talking speech recognition can be achieved by either adjusting the model used for recognition also known as model adaptation to ideally match the propagation conditions [1, 2, 3] or enhancing the distorted observations of the desired features [4, 5, 6]. In either case, it is crucial to estimate the environment parameters. One possible way to do so consists in using the so-called stereo data where various combinations of propagation conditions and clean speech are required to train the acoustic model [2, 7]. However, such data may not be available in practice. In his pioneering work [4], Moreno proposed the VTS which is a simple yet very efficient approach to linearize the relationship between the features of clean speech, noise signals, and microphone observations, thereby leading to a tractable formulation. Significant robustness to additive noise was obtained even with zeroth and first order expansions. In [5, 6], Kim et al improved the latter approach by including an iterative EM estimation of the channel noise statistics jointly with a Bayesian adaptation.

It is not a secret that most robust speech recognition methods are designed to process only a single microphone output. On the other hand, microphone arrays are becoming commonplace in current speech communication devices. Consequently, it is worthwhile knowing whether the joint processing of the feature vectors of multiple microphone observations can lead to any improvement in environment compensation. Intuitively, the more observations (or additional useful information) of the target signal we have, the better is the expected optimal processing result. For example, in the linear (time/frequency) domain, it is known that joint sensor array processing outperforms its single channel counterpart since directional noise can be removed and higher signal-to-noise ratio (SNR) gains are achieved even in the extreme case of spatially uncorrelated noise. However, it is also known that speech processing in the linear domain does not generally yield high recognition performance as compared to its feature domain counterpart [2, 8].

In this paper, our contribution consists in exploiting the space dimension in the feature domain to further enhance the desired feature vector estimates. We take advantage of the VTS that has proven very successful in robust speech recognition by approximating each of the microphone feature vectors using first order Taylor series. This approximation order is convenient since the spatial information (e.g., spatial correlation of the noise) between sensors can be captured without compromising the tractability of our problem. In order to avoid the shortcoming of the stereo-data-based processing mentioned above, we estimate the acoustic channel and the noise statistics via expectation maximization. Finally, these environment parameters are fed to the optimal space-time MMSE filter in order to recover the desired features. We experimentally show that when using the proposed multichannel processing, the ensuing feature vector estimates are more reliable for recognition.

2. Problem Statement and Approximation

In the multidimensional spacetime domain, a spacetime feature vector of the noisy data is a non-linear transformation of the desired speech, noise, and acoustic channel. Without loss of generality, we consider the mel log-spectral features in this work. Let $S_t(p)$, $Y_{n,t}(p)$, $Q_n(p)$, and $V_{n,t}(p)$ denote the pth ($p = 1, ..., P$) feature components of the desired signal, the nth ($n = 1...N$) microphone observation, the nth acoustic channel, and the noise component at the nth microphone. t is the time frame index. These terms are related as [4, 5, 6]

$$Y_{n,t}(p) \approx S_t(p) + f\left(V_{n,t}(p), S_t(p), Q_{n,t}(p)\right) \quad (1)$$

where

$$f(a,b,c) = c + \log\left(1 + e^{a-b-c}\right) \quad (2)$$

for a given triple (a, b, c). Vector notations are used next and we define the overall observed feature vector $\mathbf{y}_t = \left[\mathbf{y}_{1,t}^T ... \mathbf{y}_{N,t}^T\right]^T$ where $\mathbf{y}_{n,t} = [Y_{n,t}(1) ... Y_{n,t}(P)]^T$, also we define $\mathbf{v}_t = \left[\mathbf{v}_{1,t}^T ... \mathbf{v}_{N,t}^T\right]^T$ where $\mathbf{v}_{n,t} = [V_{n,t}(1) ... V_{n,t}(P)]^T$, $\mathbf{q} = \left[\mathbf{q}_1^T ... \mathbf{q}_N^T\right]^T$ where $\mathbf{q}_n = [Q_n(1) ... Q_n(P)]^T$ and finally

$\mathbf{s}_t = [S_t(1) \; ... \; S_t(P)]^T$. Our objective is to obtain a reliable estimate of the desired signal $\mathbf{s}_t$ for recognition.

Before proceeding, it is important to recall that the VTS expansion is carried out per Gaussian component of the desired signal. In other words, we first assume that the distribution of the speech feature vector $\mathbf{s}_t$ is accurately represented by a Gaussian mixture model (GMM), i.e., $p(\mathbf{s}_t) = \sum_{k=1}^{K} c_k \mathcal{N}\left(\mathbf{s}_t, \boldsymbol{\mu}_{\mathbf{s},k}, \boldsymbol{\Sigma}_{\mathbf{ss},k}\right)$ where K is the number of Gaussians. We assume the knowledge of c_k , $\boldsymbol{\mu}_{\mathbf{s},k} = [\Upsilon_{s,k}(1) \; ... \; \Upsilon_{s,k}(P)]^T$, and $\boldsymbol{\Sigma}_{\mathbf{ss},k}$ that we can obtain through training. Furthermore, the speech GMM is trained for diagonal matrices $\boldsymbol{\Sigma}_{\mathbf{ss},k}$.

First order Taylor series expansion allows for feature enhancement without requiring a high complexity. At the same time, the spatial correlation of the noise (between sensors) can be captured. Therefore, we use this approximation order in the remainder. Similar to [5, 6], we assume that we have a single Gaussian model for the noise with mean $\boldsymbol{\mu}_{\mathbf{v}} = \left[\boldsymbol{\mu}_{\mathbf{v}_1}^T \; ... \; \boldsymbol{\mu}_{\mathbf{v}_N}^T\right]^T$, where $\boldsymbol{\mu}_{\mathbf{v}_n} = [\Upsilon_{v_n}(1) \; ... \; \Upsilon_{v_n}(P)]^T$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{vv}}$, which we found accurate enough in all the investigated scenarios. For the kth Gaussian component, $\mathbf{y}_t$ is approximated around the expansion point $\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\mu}_{\mathbf{s},k}, \mathbf{q}_0\right)$, where $\mathbf{q}_0$ chosen in ad-hoc way, by

$$\mathbf{y}_t \approx \left[\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)}\right]\mathbf{s}_t + \mathbf{D}_k^{(q)}\mathbf{q} + \mathbf{D}_k^{(v)}\mathbf{v}_t + \mathbf{g}_k \quad (3)$$

where $\mathbf{I}_{NP\times P}$ is an $NP \times P$–dimensional matrix with the nth $(n = 1, ..., N)$ $P \times P$-dimensional block being an identity matrix. $\mathbf{D}_k^{(s)}$ has the same structure but with the pth $(p = 1, ..., P)$ diagonal term of the nth $P \times P$ diagonal matrix, $\mathbf{D}_{n,k}^{(s)}$, being

$$\frac{\partial f}{\partial S_t(p)}\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\mu}_{\mathbf{s},k}, \mathbf{q}_0\right) = -\frac{e^{\Upsilon_{v_n}(p)}}{e^{\Upsilon_{s,k}(p)+Q_{0n}(p)} + e^{\Upsilon_{v_n}(p)}}.$$

Furthermore, $\mathbf{D}_k^{(q)}$ and $\mathbf{D}_k^{(v)}$ are $NP \times NP$ diagonal matrices with their pth terms of their nth diagonal $P \times P$ matrices, $\mathbf{D}_{n,k}^{(q)}$ and $\mathbf{D}_{n,k}^{(v)}$, given by

$$\frac{\partial f}{\partial Q_n(p)}\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\mu}_{\mathbf{s},k}, \mathbf{q}_0\right) = 1 - \frac{e^{\Upsilon_{v_n}(p)}}{e^{\Upsilon_{s,k}(p)+Q_{0n}(p)} + e^{\Upsilon_{v_n}(p)}}$$

and

$$\frac{\partial f}{\partial V_n(p)}\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\mu}_{\mathbf{s},k}, \mathbf{q}_0\right) = \frac{e^{\Upsilon_{v_n}(p)}}{e^{\Upsilon_{s,k}(p)+Q_{0n}(p)} + e^{\Upsilon_{v_n}(p)}},$$

respectively. $f(\cdot)$ operates element-wise. Here, we assume that $\frac{\partial f(S(p),Q_n(p),V_n(p))}{\partial V_m(p')} = 0$, $\frac{\partial f(S(p),Q_n(p),V_n(p))}{\partial Q_m(p')} = 0$ and $\frac{\partial f(S(p),Q_n(p),V_n(p))}{\partial S(p')} = 0$ if $m \neq n$ or $p \neq p'$. Finally, $\mathbf{g}_k = \left[\mathbf{g}_{1,k}^T \; ... \; \mathbf{g}_{N,k}^T\right]^T = f\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\mu}_{\mathbf{s},k}, \mathbf{q}_0\right) - \mathbf{D}_k^{(s)}\boldsymbol{\mu}_{\mathbf{s},k} - \mathbf{D}_k^{(q)}\mathbf{q}_0 - \mathbf{D}_k^{(v)}\boldsymbol{\mu}_{\mathbf{v}}$. In our processing, we force the inter-frequency correlation to zero in the covariance matrices of the noise and noisy data and we compute their diagonal and subdiagonal terms that account for inter-microphone correlations.

3. Multichannel Filter

Given an observed space-time feature vector $\mathbf{y}_t$, the best estimate of the clean feature vector, $\mathbf{s}_t$, is known to correspond to the minimum mean square error (MMSE) solution, i.e.,

$$\hat{\mathbf{s}}_t = E\left\{\mathbf{s}_t|\mathbf{y}_t\right\}. \quad (4)$$

Since the distribution of the feature vector $\mathbf{s}_t$ is a Gaussian mixture, $\hat{\mathbf{s}}_t$ is given by

$$\begin{aligned}\hat{\mathbf{s}}_t &= \sum_{k=1}^{K} p(k|\mathbf{y}_t)\int_{\mathbf{s}_t} \mathbf{s}_t p(\mathbf{s}_t|\mathbf{y}_t, k)d\mathbf{s}_t \\ &= \sum_{k=1}^{K} p(k|\mathbf{y}_t)\left[\mathbf{H}_k\left(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k}\right) + \boldsymbol{\mu}_{\mathbf{s},k}\right] \quad (5)\end{aligned}$$

where $\boldsymbol{\mu}_{\mathbf{y},k} = E\left\{\mathbf{y}_t|k\right\}$, $\mathbf{H}_k = \boldsymbol{\Sigma}_{\mathbf{sy},k}\boldsymbol{\Sigma}_{\mathbf{yy},k}^{-1}$, $\boldsymbol{\Sigma}_{\mathbf{sy},k} = E\left\{(\mathbf{s}_t - \boldsymbol{\mu}_{\mathbf{s},k})(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k})^T|k\right\}$, $\boldsymbol{\Sigma}_{\mathbf{yy},k} = E\left\{(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k})(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k})^T|k\right\}$. These statistics depend on the Taylor series expansion order. When using the approximate linearization of (3),

$$\boldsymbol{\mu}_{\mathbf{y},k} = \left[\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)}\right]\boldsymbol{\mu}_{\mathbf{s},k} + \mathbf{D}_k^{(q)}\mathbf{q} + \mathbf{D}_k^{(v)}\boldsymbol{\mu}_{\mathbf{v}} + \mathbf{g}_k,$$

$$\boldsymbol{\Sigma}_{\mathbf{sy},k} = \boldsymbol{\Sigma}_{\mathbf{ss},k}(\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)})^T,$$

and

$$\boldsymbol{\Sigma}_{\mathbf{yy},k} = \tilde{\boldsymbol{\Sigma}}_{\mathbf{yy},k} + \mathbf{D}_k^{(v)}\boldsymbol{\Sigma}_{\mathbf{vv}}\mathbf{D}_k^{(v)T}$$

where $\tilde{\boldsymbol{\Sigma}}_{\mathbf{yy},k} = (\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)})\boldsymbol{\Sigma}_{\mathbf{ss},k}(\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)})^T$.

$\mathbf{H}_k$ is the multichannel feature-domain Wiener filter. In order to interpret (5), we can consider the most significant Gaussian component only as in [2]. In this case, the estimator in (5) becomes unbiased. More importantly, we can also demonstrate that the multichannel processing outperforms its single channel counterpart by comparing the ratio of the filtered speech variance to the variance of the residual noise at the filter output (known as output SNR in the linear-domain) [9]. Section 5 further corroborates the advantage of the multichannel processing experimentally in terms of word recognition accuracy.

4. Environment Parameters Estimation

The purpose here is to estimate the acoustic channel, $\mathbf{q}$, in addition to the noise mean $\boldsymbol{\mu}_{\mathbf{v}}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{vv}}$ iteratively using the EM algorithm. Our proposal extends the contribution in [5, 6] by estimating a more general form of the covariance matrix where the inter-microphone correlations are taken into account. Also, a simple alternative to the procedure suggested in the same references is given in order to mitigate a singularity that is inherent to the multichannel formulation as it will become clear below.

In the iterative EM algorithm, it is customary to start with some initial values of the hidden parameters $\lambda = \{\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Sigma}_{\mathbf{vv}}, \mathbf{q}\}$ and maximize the likelihood function for some unknown $\bar{\lambda} = \left\{\bar{\boldsymbol{\mu}}_{\mathbf{v}}, \bar{\boldsymbol{\Sigma}}_{\mathbf{vv}}, \bar{\mathbf{q}}\right\}$. This is achieved by maximizing the auxiliary function

$$\mathcal{Q}(\lambda, \bar{\lambda}) = E\left\{\log(p(\mathbf{S}, \mathbf{V}, \mathbf{K}|\bar{\lambda}))|\mathbf{Y}, \lambda\right\} \quad (6)$$

with respect to $\bar{\lambda}$, where $\mathbf{S}$ and $\mathbf{V}$ are formed by the feature vectors of the clean speech and noise at time instants $t = 1, ..., T$, T being the number of frames per utterance and $\mathbf{K}$ stands for the Gaussian components. $\mathcal{Q}(\lambda, \bar{\lambda})$ can be written as shown in (7) on the top of the next page. Next, we first determine $p(k|\mathbf{y}_t, \lambda)$, $p(\mathbf{v}_t|\bar{\boldsymbol{\Sigma}}_{\mathbf{vv}}, \bar{\boldsymbol{\mu}}_{\mathbf{v}})$, $p(\mathbf{v}_t|\mathbf{y}_t, k, \lambda)$. This will allow us to estimate the optimal noise parameters, denoted as $\bar{\boldsymbol{\mu}}_{\mathbf{v}}^{\mathrm{o}}$ and $\bar{\boldsymbol{\Sigma}}_{\mathbf{vv}}^{\mathrm{o}}$, using (7). In order to find the optimal channel distortion denoted as $\bar{\mathbf{q}}^{\mathrm{o}}$, a per-channel processing is needed as explained below.

First, we know that $p(\mathbf{y}_t|k, \lambda)$ is Gaussian with mean $\boldsymbol{\mu}_{\mathbf{y},k}$ and covariance $\boldsymbol{\Sigma}_{\mathbf{yy},k}$. This allows us to determine

$$\mathcal{Q}(\lambda,\bar{\lambda}) = \sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{y}_t,\lambda)\int_{\mathbf{v}_t} p(\mathbf{v}_t|\mathbf{y}_t,k,\lambda)\log\left[p(\mathbf{y}_t|\mathbf{v}_t,k,\bar{\mathbf{q}})p(\mathbf{v}_t|\bar{\mathbf{\Sigma}}_{\mathbf{vv}},\bar{\boldsymbol{\mu}}_{\mathbf{v}})c_k\right]d\mathbf{v}_t \quad (7)$$

$p(k|\mathbf{y}_t,\lambda) = \frac{c_k p(\mathbf{y}_t|k,\lambda)}{\sum_{k=1}^{K} c_k p(\mathbf{y}_t|k,\lambda)}$. $p(\mathbf{v}_t|\bar{\mathbf{\Sigma}}_{\mathbf{vv}},\bar{\boldsymbol{\mu}}_{\mathbf{v}})$ is Gaussian. Also, $p(\mathbf{v}_t|\mathbf{y}_t,k,\bar{\lambda})$ is Gaussian and to determine its mean and covariance, we propose that instead of going through the identification procedure as suggested in [6] following Appendix V.2 of [7], it is simpler and more convenient to use the fact that the joint distribution of $\left[\mathbf{y}_t^T\ \mathbf{v}_t^T\right]^T$ is Gaussian with mean

$$\tilde{\boldsymbol{\mu}}_k(\lambda) = \left[\boldsymbol{\mu}_{\mathbf{y},k}^T\ \boldsymbol{\mu}_{\mathbf{v}}^T\right]^T \quad (8)$$

and covariance matrix

$$\tilde{\mathbf{\Sigma}}_k(\lambda) = \begin{bmatrix} \mathbf{\Sigma}_{\mathbf{yy},k} & \mathbf{\Sigma}_{\mathbf{yv},k} \\ \mathbf{\Sigma}_{\mathbf{vy},k} & \mathbf{\Sigma}_{\mathbf{vv}} \end{bmatrix}. \quad (9)$$

Then, we know that $p(\mathbf{v}_t|\mathbf{y}_t,k,\lambda)$ is Gaussian with mean

$$\begin{aligned}\tilde{\boldsymbol{\mu}}_{\mathbf{v},k}(\mathbf{y}_t,\lambda) &= \boldsymbol{\mu}_{\mathbf{v}} + \mathbf{\Sigma}_{\mathbf{vy},k}\mathbf{\Sigma}_{\mathbf{yy},k}^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k}\right) \quad (10)\\ &= \boldsymbol{\mu}_{\mathbf{v}} + \mathbf{\Sigma}_{\mathbf{vv}}\mathbf{D}_k^{(v)T}\mathbf{\Sigma}_{\mathbf{yy},k}^{-1}\left(\mathbf{y}_t - \boldsymbol{\mu}_{\mathbf{y},k}\right)\end{aligned}$$

and covariance matrix

$$\tilde{\mathbf{\Sigma}}_{\mathbf{vv},k}(\lambda) = \mathbf{\Sigma}_{\mathbf{vv}} - \mathbf{\Sigma}_{\mathbf{vv}}\mathbf{D}_k^{(v)T}\mathbf{\Sigma}_{\mathbf{yy},k}^{-1}\mathbf{D}_k^{(v)}\mathbf{\Sigma}_{\mathbf{vv}}. \quad (11)$$

In [5, 6], an inversion of $\tilde{\mathbf{\Sigma}}_{\mathbf{yy},k}$ is required. This matrix is of rank P, meaning that it is full-rank when $N = 1$. In this case, (10) and (11) coincide with the expressions found in [5, 6] (up to Woodburry identity). However, in our study (i.e, $N > 1$), this matrix is rank deficient and its inversion becomes problematic. In this sense, (10) and (11) are desirable because they do not involve the inversion of a singular matrix.

Since $p(\mathbf{y}_t|\mathbf{v}_t,k,\bar{\mathbf{q}})$ and c_k are independent of $\bar{\mathbf{\Sigma}}_{\mathbf{vv}}$ and $\bar{\boldsymbol{\mu}}_{\mathbf{v}}$, we are now able to find $\bar{\boldsymbol{\mu}}_{\mathbf{v}}^{\mathrm{o}}$ and $\bar{\mathbf{\Sigma}}_{\mathbf{vv}}^{\mathrm{o}}$ as shown in (12) and (13) on the top of the next page by setting the derivatives of $\mathcal{Q}(\lambda,\bar{\lambda})$ in (7) with respect to both variables to zero.

Now to determine $\bar{\mathbf{q}}^{\mathrm{o}}$, one might use the fact that $p(\mathbf{y}_t|\mathbf{v}_t,k,\bar{\mathbf{q}})$ is Gaussian with mean $\tilde{\boldsymbol{\mu}}_{\mathbf{y},k}(\mathbf{v}_t,\bar{\lambda}) = \left[\mathbf{I}_{NP\times P} + \mathbf{D}_k^{(s)}\right]\boldsymbol{\mu}_{\mathbf{s},k} + \mathbf{D}_k^{(q)}\bar{\mathbf{q}} + \mathbf{D}_k^{(v)}\mathbf{v}_t + \mathbf{g}_k$ and covariance $\tilde{\mathbf{\Sigma}}_{\mathbf{yy},k}$. Since $\tilde{\mathbf{\Sigma}}_{\mathbf{yy},k}$ is singular, $\bar{\mathbf{q}}$ cannot be uniquely identified by attempting to maximize (7) with respect to $\bar{\mathbf{q}}$. Nevertheless, we can mitigate this problem by applying the same reasoning above to each of the N channels. Indeed, for the nth microphone, we express its auxiliary function in a similar way to (7) and demonstrate that $\bar{\mathbf{q}}_n^{\mathrm{o}}$ is given by (14)–(15) where

$$\tilde{\mathbf{\Sigma}}_{\mathbf{y}_n\mathbf{y}_n,k} = \left(\mathbf{I}_{P\times P} + \mathbf{D}_{n,k}^{(s)}\right)\mathbf{\Sigma}_{\mathbf{ss},k}\left(\mathbf{I}_{P\times P} + \mathbf{D}_{n,k}^{(s)}\right)^T$$

which is invertible and

$$\tilde{\boldsymbol{\mu}}_{\mathbf{v}_n,k}(\mathbf{y}_{n,t},\lambda) = \boldsymbol{\mu}_{\mathbf{v}_n} + \mathbf{\Sigma}_{\mathbf{v}_n\mathbf{v}_n}\mathbf{D}_{n,k}^{(v)T}\mathbf{\Sigma}_{\mathbf{y}_n\mathbf{y}_n,k}^{-1}\left(\mathbf{y}_{n,t} - \boldsymbol{\mu}_{\mathbf{y}_n,k}\right).$$

In our experiments, at most 10 iterations were allowed for the EM algorithm to obtain the final estimates of λ which is fed to the MMSE estimator in (5).

5. Experimental Results

We evaluate the performance of our algorithm using several speech samples from the TI-Digit database which were sampled at 8 kHz. The feature compensation algorithm is implemented in the log-spectral domain as specified above. Afterwards,

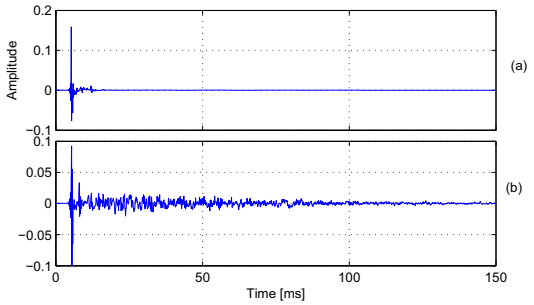

Figure 1: *Impulse response at the first microphone: (a) first reverberation condition, (b) second reverberation condition.*

the processed features are transformed to the cepstral domain via discrete cosine transform to obtain 13 Mel-frequency cepstral coefficients (MFCCs). The corresponding delta and delta-delta features were also appended to obtain our 39-dimensional MFCC feature vector. Viterbi decoding for recognition is performed using HTK tool [10]. The training data consists of 8440 clean speech utterances from the TI-Digit database. We train our GMM for $K = 256$ Gaussians of the log-spectra of the clean training data. Also, we estimate the acoustic model of the speech recognizer from the MFCCs using the same clean training data set. The recognizer acoustic model consists of speaker independent word based hidden Markov models with 18 states per word and three Gaussians per state.

To generate our multichannel testing data from the TI-Digit database and approximate some real propagation conditions, we considered 1001 other clean continuous digit utterances. Each utterance is convolved with real measured impulse responses [11]. Also, we added babble (segments of the noise file from the AURORA-2 database [12]) and computer generated white Gaussian noise signals at different SNR levels (computed per utterance) as specified in Tables 1 to 4. We present our results in the case of an array of 2 microphones (the first two impulse response measurements corresponding to the circular array with the source located at 2 m from the array center at an angle of 90 degrees, as described in [11]) with two reverberation conditions $T_{60} \approx 0$ ms and $T_{60} = 310$ ms. Figures 1 (a) and (b) show the impulse responses seen by the first microphone in both reverberation conditions (also denoted as reverberation (a) and reverberation (b) in the following), respectively. Our experimental setting of recognition almost corresponds to AURORA 2 noisy digits recognition tasks [12] but we consider the multichannel scenario with different reverberation and noise conditions.

Tables 1 to 4 contain our recognition results in the four combinations of noise and reverberation conditions. To demonstrate the advantage of the proposed multichannel feature processing, we compare its performance to the single channel processing when only the first or second microphone is used. Baseline results are also given (with the first microphone only, for conciseness). Our speech recognizer achieves 99.36% word accuracy when using the clean testing data. It is clear that the joint processing of the noisy microphone observations allows for increased word recognition accuracy in all scenarios. The word accuracy gains over single channel processing can be as substantial as $\sim 9\%$ in the case of babble noise with reverberation (b) at low input SNR values (e.g., 0 and 5 dB).

In Sections 1, we argued that the multichannel feature com-

$$\bar{\boldsymbol{\mu}}_{\mathbf{v}}^{\circ} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{y}_t,\lambda)\tilde{\boldsymbol{\mu}}_{\mathbf{v},k}(\mathbf{y}_t,\lambda)}{T} \tag{12}$$

$$\bar{\boldsymbol{\Sigma}}_{\mathbf{vv}}^{\circ} = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{y}_t,\lambda)\left[\tilde{\boldsymbol{\Sigma}}_{\mathbf{vv},k}(\lambda)+\tilde{\boldsymbol{\mu}}_{\mathbf{v},k}(\mathbf{y}_t,\lambda)\tilde{\boldsymbol{\mu}}_{\mathbf{v},k}^{T}(\mathbf{y}_t,\lambda)\right]}{T} - \bar{\boldsymbol{\mu}}_{\mathbf{v}}^{\circ}\bar{\boldsymbol{\mu}}_{\mathbf{v}}^{\circ T}. \tag{13}$$

$$\mathbf{q}_n^{\circ} = \left[\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{y}_{n,t},\lambda)\mathbf{D}_{n,k}^{(q)T}\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}_n\mathbf{y}_n,k}^{-1}\mathbf{D}_{n,k}^{(q)}\right]^{-1} \times \left[\sum_{t=1}^{T}\sum_{k=1}^{K} p(k|\mathbf{y}_{n,t},\lambda)\mathbf{D}_{n,k}^{(q)T}\tilde{\boldsymbol{\Sigma}}_{\mathbf{y}_n\mathbf{y}_n,k}^{-1}\left(\mathbf{y}_{n,t}-\mathbf{m}_{n,k}\right)\right] \tag{14}$$

$$\mathbf{m}_{n,k} = \left(\mathbf{I}_{P\times P}+\mathbf{D}_{n,k}^{(s)}\right)\boldsymbol{\mu}_{\mathbf{s},k}+\mathbf{g}_{n,k}+\mathbf{D}_{n,k}^{(v)}\tilde{\boldsymbol{\mu}}_{\mathbf{v}_n,k}(\mathbf{y}_{n,t},\lambda) \tag{15}$$

pensation is more beneficial than the single channel processing. Intuitively, the more data observations of the desired signal we have, the better is the expected optimal processing result. Following the analogy to the linear-domain processing in Section 3, it is possible to confirm that when more microphone observations are available, increased ratio of the desired signal variance to the residual noise at the output of the feature compensation filter (analogous to output SNR [9] in the feature-domain) can be obtained. Both arguments suggest that using microphone arrays for feature compensation is more advantageous than the single channel processing. Our experimental findings lend credence to our intuition and analogy.

Table 1: *Word accuracy* (%). *Babble noise & reverberation (a).*

SNR[dB]	0	5	10	15	20
Baseline	41.97	62.02	78.23	89.83	94.34
Microphone 1	63.37	82.62	92.39	95.86	97.54
Microphone 2	59.47	80.96	92.63	96.41	97.76
Multichannel	**72.00**	**88.36**	**95.33**	**97.27**	**98.10**

Table 2: *Word accuracy* (%). *White noise & reverberation (a).*

SNR[dB]	0	5	10	15	20
Baseline	11.08	32.82	66.07	85.42	92.75
Microphone 1	60.91	79.80	88.82	94.57	96.87
Microphone 2	61.93	80.78	89.28	94.26	96.75
Multichannel	**70.28**	**83.67**	**91.53**	**95.55**	**97.57**

Table 3: *Word accuracy* (%). *Babble noise & reverberation (b).*

SNR[dB]	0	5	10	15	20
Baseline	24.50	35.00	44.58	53.05	58.49
Microphone 1	46.95	63.46	72.86	76.33	77.03
Microphone 2	45.44	63.06	72.55	76.24	77.34
Multichannel	**56.56**	**72.34**	**78.60**	**79.15**	**79.00**

Table 4: *Word accuracy* (%). *White noise & reverberation (b).*

SNR[dB]	0	5	10	15	20
Baseline	10.52	20.35	42.32	60.33	61.68
Microphone 1	42.46	54.47	62.76	67.67	70.68
Microphone 2	43.08	52.96	62.20	67.24	71.20
Multichannel	**47.47**	**55.08**	**62.88**	**68.47**	**72.28**

Finally, by comparing the results in Tables 1 and 2 to those in Tables 3 and 4, respectively, we observe a remarkable deterioration of the recognition performance because of the reverberation even after feature compensation. Actually, this deterioration is essentially caused by the inaccuracy in modeling the acoustic channel effect (a multiplicative scalar per band). The frames used in our processing, in particular, and speech recognition, in general, are very short as compared to the channel impulse response. In this situation, modeling the channel effect by a frequency depend scaling factor is not accurate and the convolutive effect [3] has to be considered to achieve more reliable feature estimates.

6. Conclusions

In this paper, we presented a multichannel approach for feature vector enhancement. First order VTS approximation was used to linearize the space-time feature vector at the microphone array output and the iterative EM was developed to estimate the acoustic channel and the noise mean vector in addition to its auto and cross-covariances across the microphones. Experiments were carried out on speech signals from the TI-Digit database that were convolved with measured impulse responses and corrupted by additive noise at different SNR levels. Our results suggest that the proposed method outperforms the single channel VTS-based feature enhancement. Finally, it is worth noting that in contrast to the traditional beamforming methods, the proposed processing disregards the phase information. This is inherent to the feature transformation itself. It is known that in microphone arrays, the phase can contain some relevant information (e.g., speaker location) and could potentially be an additional cue for improved speech recognition. However, including it in feature-based processing is not clear yet. This topic will be investigated in the future.

7. References

[1] D. K., Kim and M.J.F. Gales, "Noisy Constrained Maximum-Likelihood Linear Regression for Noise-Robust Speech Recognition," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 19, pp. 315–325, Feb. 2011.

[2] L. Deng, A. Acero, M. Plumpe, and X. Huang, "Large-vocabulary speech recognition under adverse acoustic environments," in *Proc. ICSLP,* 2000, pp. 806–809.

[3] A. M. Sehr, "Reverberation modeling for robust distant-talking speech recognition," *PhD thesis*, Verlag Dr. Hut, Munchen, 2009.

[4] P. J. Moreno, "Speech Recognition in Noisy Environments," *PhD thesis*, Carnegie Mellon University, 1996.

[5] N. S. Kim, "Application of VTS to environment compensation with noise statistics," *ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels,* 1997, pp. 99–102.

[6] D. Y. Kim, C. K. Un, and N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," *Speech communication,* pp. 39–49, 1998.

[7] A. Acero, *Acoustical and environmental robustness in automatic speech recognition,* Kluwer academic publishers, Norwell, 1993.

[8] B. Milner, J. Darch, and S. Vaseghi, "Applying noise compensation methods to robustly predict speech features from MFCC vectors in noise," in *Proc. IEEE ICASSP,* 2008, pp. 3945–3949.

[9] M. Souden, J. Benesty, and S. Affes, "On optimal frequency-domain multichannel linear filtering for noise reduction," *IEEE Trans. Audio, Speech, Lang. Process.,* vol. 18, pp. 260–276, Feb. 2010.

[10] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, Version 3.4.*, Cambridge University, Eng. Dept., 2006.

[11] *http://tosa.mri.co.jp/sounddb/indexe.htm*

[12] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condition," in *Proc. ISCA,* 2000, pp. 18–20.