REVERBERATION AND NOISE ROBUST FEATURE ENHANCEMENT USING MULTIPLE INPUTS

Shin Jae Kang, Tae Gyoon Kang, Kang Hyun Lee, Kiho Cho and Nam Soo Kim

Department of Electrical and Computer Engineering and INMC
Seoul National University, Seoul 151-742, Korea
E-mail: {sjkang, tgkang, khlee, khcho}@hi.snu.ac.kr, nkim@snu.ac.kr

ABSTRACT

We propose a novel approach to feature enhancement in multi-channel scenario. Our approach is based on the interacting multiple model (IMM), which was originally developed in single-channel scenario. We extend the single-channel IMM algorithm such that it can handle the multi-channel inputs under the Bayesian framework. The multi-channel IMM algorithm is capable of tracking time-varying room impulse responses and background noises by updating the relevant parameters in an on-line manner. In various environmental conditions, the performance gain of the proposed method has been confirmed.

Index Terms— Robust speech recognition, multi-channel, interacting multiple model (IMM), dereverberation

1. INTRODUCTION

The performance of an automatic speech recognition (ASR) system is usually degraded when the input speech is distorted by background noise or acoustic reverberation. In order to alleviate this performance degradation in adverse environments, a variety of techniques have been developed e.g., speech enhancement, feature compensation and model adaptation algorithms [1]-[8]. Though separate algorithms perform differently, their ultimate goal is to reduce the mismatch between the degraded input signal and the trained recognition model parameters. In this paper, we focus on the feature compensation techniques which directly enhance the distorted input feature vectors to match the characteristic of the training data before being decoded by the acoustic recognition model.

In most cases the target speech and noise or other interference sources reside in different spatial locations. Multiple microphone arrays are useful to extract the desired signal especially when each sound source is separated spatially. During the past several decades multi-channel based beamforming techniques such as the generalized sidelobe canceller [9] have been proposed to attenuate the coherent interfering sources and acoustic reverberation. Multi-channel based criteria are also directly applied in the feature domain for robust speech recognition [10].

In this paper, we propose a novel multi-channel feature enhancement technique applied in the log-spectral domain. In the proposed approach, we extend the interacting multiple model (IMM) algorithm [7] originally designed in the single-channel scenario so that it can fit to the multi-channel processing. The proposed method has mainly two advantages. First, no *a prior* knowledge of the room impulse response (RIR) is needed. Second, the parameters concerned with the acoustic reverberation and background noise are sequentially updated in a frame-by-frame manner instead of utterance-by-utterance or file-by-file basis for tracking their time-varying nature. This type of real-time update of the RIR parameters is very important in handling the possible movements of the talker or microphones.

2. OBSERVATION MODEL IN MULTI-CHANNEL REVERBERANT NOISY ENVIRONMENT

We consider a typical hands-free scenario for ASR in which multiple microphones are used. The target speaker is located in a certain distance from the microphones in an enclosed room, which results in acoustic reverberation. Let $y_i[l]$ be the signal obtained from the i-th microphone with $l \in \{0, 1, \cdots\}$ denoting the time index. If $x[l]$ is the target speech signal and $h_{i,l}[p]$ represents the RIR from the target speaker to the i-th microphone with corresponding tap index $p \in \{0, 1, \cdots\}$, then

$$y_i[l] = \sum_{p=0}^{\infty} h_{i,l}[p]x[l-p] + n_i[l] \quad (1)$$

where $n_i[l]$ is the background noise added to the i-th microphone input.

By using the formulation presented in [7], we can rewrite the relation of (1) in the logarithmic mel magnitude spectral

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012R1A2A2A01045874) and by the MSIP (Ministry of Science, ICT and Future Planning), Korea, under the ITRC (Information Technology Research Center) support program (NIPA-2013-H0301-13-4005) supervised by the NIPA (National IT Industry Promotion Agency).

coefficient (LMMSC) domain as follow:

$$\mathbf{y}_{i,m} = \ln\left(\sum_{\tau=0}^{L}\exp\left(\mathbf{x}_{m-\tau}+\mathbf{h}_{i,m,\tau}\right)+\exp\left(\mathbf{n}_{i,m}\right)\right)+\mathbf{v}_{i,m} \tag{2}$$

where $\mathbf{y}_{i,m}$, $\mathbf{x}_m$, $\mathbf{h}_{i,m,\tau}$, $\mathbf{n}_{i,m}$ and $\mathbf{v}_{i,m}$ respectively denote the Q-dimensional LMMSC vectors of the reverberant noisy speech, clean speech, time-variant log frequency response of the reverberant acoustic path for a tap index τ which is assumed to have finite length $(L+1)$, i.e., $\mathbf{h}_{i,m,\tau} = -\infty$ for all $\tau > L$, background noise and approximation error of the observation model at the m-th frame which are collected at the i-th microphone. The only difference of (2) from the formulation derived in [7] is that we now add subscript i to identify each microphone. The functions $\exp(\cdot)$ and $\ln(\cdot)$ in (2) are applied component-wisely and we assume that the error distribution at each microphone is given by

$$\mathbf{v}_{i,m} = \mathbf{v}_m \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{v}}, \boldsymbol{\Sigma}_{\mathbf{v}}\right) \tag{3}$$

in which $\mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ indicates a Gaussian PDF with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

3. MULTI-CHANNEL FEATURE ENHANCEMENT

3.1. A Bayesian framework

In our work, the clean speech component and the N log frequency responses are jointly handled as a state vector $\mathbf{z}_m$ at the m-th frame where N indicates the number of microphones. The core idea of our approach is to estimate the posterior probability $p(\mathbf{z}_m|\mathbb{y}_0^m)$ of the state vector

$$\mathbf{z}_m = \left[\mathbb{x}_m' \; \mathbb{h}_{1,m}' \; \mathbb{h}_{2,m}' \; \cdots \; \mathbb{h}_{N,m}'\right]' \tag{4}$$

$$\mathbb{x}_m = \left[\mathbf{x}_m' \; \mathbf{x}_{m-1}' \; \cdots \; \mathbf{x}_{m-L}'\right]' \tag{5}$$

$$\mathbb{h}_{i,m} = \left[\mathbf{h}_{i,m,0}' \; \mathbf{h}_{i,m,1}' \; \cdots \; \mathbf{h}_{i,m,L}'\right]' \tag{6}$$

conditioned on all the N observed reverberant noisy LMMSC vectors

$$\mathbb{y}_0^m = \left[\mathbf{y}_0' \; \mathbf{y}_1' \; \cdots \; \mathbf{y}_m'\right]' \tag{7}$$

$$\mathbf{y}_m = \left[\mathbf{y}_{1,m}' \; \mathbf{y}_{2,m}' \; \cdots \; \mathbf{y}_{N,m}'\right]' \tag{8}$$

where the prime denotes the transpose of a vector or matrix, $\mathbb{x}_m$ and $\mathbb{h}_{i,m}$ respectively mean a local clean speech and log frequency response LMMSC trajectories consisting of $(L+1)$ consecutive frames at the i-th microphone. In the above formulation, $\mathbf{x}_{m_1}^{m_2} = \left[\mathbf{x}_{m_1}' \; \mathbf{x}_{m_1+1}' \; \cdots \; \mathbf{x}_{m_2}'\right]'$ denotes a subsequence of vectors from frame index m_1 to m_2 and $\mathbf{y}_m$ means N observations at the m-th frame concatenated to a single vector.

A typical way to compute the posterior distribution of the state vector $\mathbf{z}_m$ based on a Bayesian inference is to recursively compute the predictive distribution $p(\mathbf{z}_m|\mathbb{y}_0^{m-1})$ and posterior distribution $p(\mathbf{z}_m|\mathbb{y}_0^m)$ given the previous reverberant noisy observations as follows:

$$p\left(\mathbf{z}_m|\mathbb{y}_0^{m-1}\right) = \int p\left(\mathbf{z}_m|\mathbf{z}_{m-1}, \mathbb{y}_0^{m-1}\right) \times p\left(\mathbf{z}_{m-1}|\mathbb{y}_0^{m-1}\right) d\mathbf{z}_{m-1} \tag{9}$$

$$p\left(\mathbf{z}_m|\mathbb{y}_0^m\right) = \frac{p\left(\mathbf{y}_m|\mathbf{z}_m, \mathbb{y}_0^{m-1}\right) p\left(\mathbf{z}_m|\mathbb{y}_0^{m-1}\right)}{\int p\left(\mathbf{y}_m|\mathbf{z}_m, \mathbb{y}_0^{m-1}\right) p\left(\mathbf{z}_m|\mathbb{y}_0^{m-1}\right) d\mathbf{z}_m}. \tag{10}$$

If both $p\left(\mathbf{z}_m|\mathbb{y}_0^{m-1}\right)$ and $p\left(\mathbf{z}_m|\mathbb{y}_0^m\right)$ are assumed to be Gaussian distributions, it is sufficient to revise the statistical moments up to the second-order which are defined as follows:

$$\begin{cases} \hat{\mathbf{z}}_{m|m-1} = E\left[\mathbf{z}_m|\mathbb{y}_0^{m-1}\right] \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m-1}} = E\left[\left(\mathbf{z}_m - \hat{\mathbf{z}}_{m|m-1}\right)\left(\mathbf{z}_m - \hat{\mathbf{z}}_{m|m-1}\right)'|\mathbb{y}_0^{m-1}\right] \end{cases} \tag{11}$$

$$\begin{cases} \hat{\mathbf{z}}_{m|m} = E\left[\mathbf{z}_m|\mathbb{y}_0^m\right] \\ \hat{\boldsymbol{\Sigma}}_{\mathbf{z}_{m|m}} = E\left[\left(\mathbf{z}_m - \hat{\mathbf{z}}_{m|m}\right)\left(\mathbf{z}_m - \hat{\mathbf{z}}_{m|m}\right)'|\mathbb{y}_0^m\right] \end{cases} \tag{12}$$

where $E[\cdot]$ indicates expectation. Interested readers are referred to [7] for further information.

3.2. Prior models

In this section, the prior models for clean speech, RIR, and background noise are described. As mentioned in [7], the *a priori* clean speech distribution is assumed as a mixture of K Gaussians to approximate a high degree of speech dynamics as follows:

$$p(\mathbb{x}_m) = \sum_{j=1}^{K} p(\gamma_m = j)\mathcal{N}\left(\mathbb{x}_m; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j\right) \tag{13}$$

where $\gamma_m \in \{1, 2, \cdots, K\}$ denotes the index of the mixture component at the m-th frame, and $p(\gamma_m = j)$, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ represent the weight, mean vector and covariance matrix of the j-th Gaussian distribution respectively.

We assume that the i-th log frequency response $\mathbb{h}_{i,m}$ at the m-th frame is statistically independent of the clean speech and background noise features. Since the RIR is non-stationary, the parameter estimate for $\mathbb{h}_{i,m}$ must be constantly updated to track its time evolution. In our work, we exploit a random walk process which is the simplest solution for predicting the next state as given by

$$\mathbb{h}_{i,m} = \mathbb{h}_{i,m-1} + \mathbf{w}_{\mathbb{h},i,m} \tag{14}$$

$$\mathbf{w}_{\mathbb{h},i,m} = \mathbf{w}_{\mathbb{h},m} \sim \mathcal{N}\left(\mathbf{0}_{Q(L+1)N}, \sigma_{\mathbb{h}}^2\mathbf{I}_{Q(L+1)N}\right) \tag{15}$$

where $\mathbf{0}_d$ represents the zero vector with dimension d and $\mathbf{I}_d$ denotes the identity matrix of size $d \times d$. When $\sigma_{\mathbb{h}}^2$ is small, this model is well suited to a slowly evolving RIR environment.

The characteristics of the background noise are very diverse. It is difficult to train all kinds of the background noise

in advance. In a short period before active speech activity occurs, however, we can assume that the background noise only exists and its characteristic is stationary. Furthermore, the complexity of the background noise model needs to be quite low to allow a fast and computationally efficient on-line tracking. For these reasons, we employ a single Gaussian background noise model as given by

$$\mathbf{n}_{i,m} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{n}_{i,m}}, \boldsymbol{\Sigma}_{\mathbf{n}_{i,m}}\right) \tag{16}$$

where the mean vector $\boldsymbol{\mu}_{\mathbf{n}_{i,m}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{n}_{i,m}}$ are unknown and should be estimated during the environment compensation procedure.

3.3. State transition formulation

As in [7], the transition process of the state vector $\mathbf{z}_m$ for the j-th Gaussian mixture component can be simply structured as follows:

$$\mathbf{z}_m = \mathcal{A}^{(j)}\mathbf{z}_{m-1} + \mathbf{b}_m^{(j)} \tag{17}$$

with

$$\mathcal{A}^{(j)} = \begin{bmatrix} \mathbf{A}\mathbf{B}^{-1} & \mathbf{O}_Q & & \\ \mathbf{I}_Q\ \mathbf{O}_Q \cdots \mathbf{O}_Q & & & \\ \mathbf{O}_Q\ \mathbf{I}_Q \cdots \mathbf{O}_Q & & \mathbf{O} & \\ \vdots\ \ddots\ \vdots\ \vdots & & & \\ \mathbf{O}_Q \cdots \mathbf{I}_Q\ \mathbf{O}_Q & & & \\ \hline & \mathbf{O}' & & \mathbf{I}_{Q(L+1)N} \end{bmatrix} \tag{18}$$

$$\mathbf{b}_m^{(j)} \sim \mathcal{N}\left(\boldsymbol{\mu}_{\mathbf{b}}^{(j)}, \boldsymbol{\Sigma}_{\mathbf{b}}^{(j)}\right) \tag{19}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_m,\mathbf{x}_{m-1})\ \mathrm{Cov}(\mathbf{x}_m,\mathbf{x}_{m-2}) \cdots \mathrm{Cov}(\mathbf{x}_m,\mathbf{x}_{m-L}) \end{bmatrix} \tag{20}$$

$$\mathbf{B} = \begin{bmatrix} \mathrm{Cov}(\mathbf{x}_{m-1},\mathbf{x}_{m-1}) & \mathrm{Cov}(\mathbf{x}_{m-1},\mathbf{x}_{m-2}) & \cdots & \mathrm{Cov}(\mathbf{x}_{m-1},\mathbf{x}_{m-L}) \\ \mathrm{Cov}(\mathbf{x}_{m-2},\mathbf{x}_{m-1}) & \mathrm{Cov}(\mathbf{x}_{m-2},\mathbf{x}_{m-2}) & \cdots & \mathrm{Cov}(\mathbf{x}_{m-2},\mathbf{x}_{m-L}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(\mathbf{x}_{m-L},\mathbf{x}_{m-1}) & \mathrm{Cov}(\mathbf{x}_{m-L},\mathbf{x}_{m-2}) & \cdots & \mathrm{Cov}(\mathbf{x}_{m-L},\mathbf{x}_{m-L}) \end{bmatrix} \tag{21}$$

and $\mathrm{Cov}(a, b)$ denotes the covariance between two vectors a and b. In addition,

$$\boldsymbol{\mu}_{\mathbf{b}}^{(j)} = \begin{bmatrix} \tilde{\boldsymbol{\mu}}_{\mathbf{b}} \\ \mathbf{0}_Q \\ \vdots \\ \mathbf{0}_Q \end{bmatrix}, \quad \boldsymbol{\Sigma}_{\mathbf{b}}^{(j)} = \begin{bmatrix} \tilde{\boldsymbol{\Sigma}}_{\mathbf{b}} & \mathbf{O}_Q \cdots \mathbf{O}_Q & \\ \mathbf{O}_Q\ \mathbf{O}_Q \cdots \mathbf{O}_Q & & \\ \vdots\ \vdots\ \ddots\ \vdots & & \mathbf{O} \\ \mathbf{O}_Q\ \mathbf{O}_Q \cdots \mathbf{O}_Q & & \\ \hline & \mathbf{O}' & \sigma_{\mathbb{h}}^2\mathbf{I}_{Q(L+1)N} \end{bmatrix} \tag{22}$$

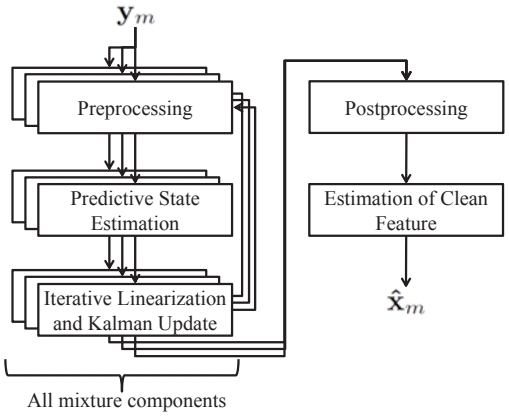

Fig. 1. A block diagram of feature enhancement algorithm

where

$$\tilde{\boldsymbol{\mu}}_{\mathbf{b}} = E[\mathbf{x}_m] - \mathbf{A}\mathbf{B}^{-1}\begin{bmatrix} E[\mathbf{x}_{m-1}] \\ E[\mathbf{x}_{m-2}] \\ \vdots \\ E[\mathbf{x}_{m-L}] \end{bmatrix} \tag{23}$$

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{b}} = \mathrm{Cov}(\mathbf{x}_m, \mathbf{x}_m) - \mathbf{A}\mathbf{B}^{-1}\mathbf{A}' \tag{24}$$

with $\mathbf{O}_Q$ and $\mathbf{O}$ respectively denoting a zero matrix with size $Q\times Q$ and $Q(L+1)\times Q(L+1)N$. The above formulation given by (17)-(24) is an extension of the state transition model derived in [7] to the case of multiple RIRs.

3.4. Function linearization

It is difficult to estimate directly the clean speech feature vector $\mathbf{x}_m$ and all the log frequency responses $\mathbf{h}_{i,m,\tau}$ of N-channel inputs from the speech distortion model (2) due to its nonlinearity. To alleviate its difficulty, we apply the piecewise linear approximation to the given nonlinear function by using Taylor series expansion. The first order form of Taylor series expansion at the i-th microphone input feature is given as in the following:

$$f_i\left(\mathbf{z}_m, \mathbf{n}_{i,m}\right) = \ln\left(\sum_{\tau=0}^{L}\exp\left(\mathbf{x}_{m-\tau} + \mathbf{h}_{i,m,\tau}\right) + \exp\left(\mathbf{n}_{i,m}\right)\right) \tag{25}$$

$$\approx \mathbf{G}_{i,m}\mathbf{z}_m + \mathbf{H}_{i,m}\mathbf{n}_{i,m} + \mathbf{q}_{i,m} \tag{26}$$

$$\mathbf{y}_m = \begin{bmatrix} \mathbf{y}_{1,m} \\ \mathbf{y}_{2,m} \\ \vdots \\ \mathbf{y}_{N,m} \end{bmatrix} = \begin{bmatrix} \mathbf{G}_{1,m} \\ \mathbf{G}_{2,m} \\ \vdots \\ \mathbf{G}_{N,m} \end{bmatrix}\mathbf{z}_m + \begin{bmatrix} \mathbf{H}_{1,m} & \mathbf{0}_Q & \cdots & \mathbf{0}_Q \\ \mathbf{0}_Q & \mathbf{H}_{2,m} & \cdots & \mathbf{0}_Q \\ \mathbf{0}_Q & \mathbf{0}_Q & \ddots & \mathbf{0}_Q \\ \mathbf{0}_Q & \mathbf{0}_Q & \cdots & \mathbf{H}_{N,m} \end{bmatrix}\begin{bmatrix} \mathbf{n}_{1,m} \\ \mathbf{n}_{2,m} \\ \vdots \\ \mathbf{n}_{N,m} \end{bmatrix} + \begin{bmatrix} \mathbf{q}_{1,m} \\ \mathbf{q}_{2,m} \\ \vdots \\ \mathbf{q}_{N,m} \end{bmatrix} + \begin{bmatrix} \mathbf{v}_{1,m} \\ \mathbf{v}_{2,m} \\ \vdots \\ \mathbf{v}_{N,m} \end{bmatrix} \tag{30}$$

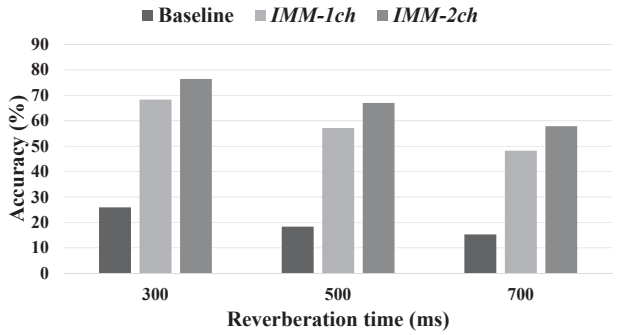

Fig. 2. Averaged word accuracies according to the reverberation time at SNR 5 dB

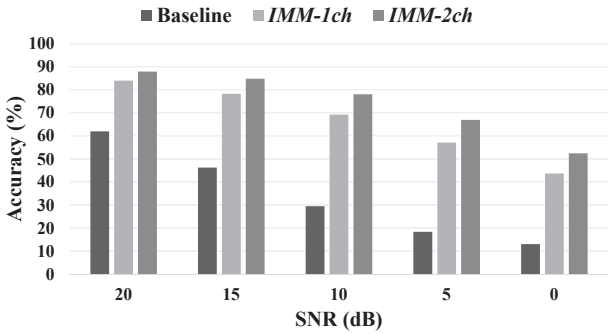

Fig. 3. Averaged word accuracies according to the SNR at reverberation 500 ms

where

$$\mathbf{G}_{i,m} = \left[\frac{\partial f_i}{\partial \mathbf{x}_m}' \cdots \mathbf{0}'_{Q(N+1)} \cdots \frac{\partial f_i}{\partial \mathbf{lh}_{i,m}}' \cdots \mathbf{0}'_{Q(N+1)} \cdots \right] \tag{27}$$

$$\mathbf{H}_{i,m} = \frac{\partial f_i}{\partial \mathbf{n}_{i,m}} \tag{28}$$

$$\mathbf{q}_{i,m} = f_i\left(\mathbf{z}_m^\circ, \mathbf{n}_{i,m}^\circ\right) - \mathbf{G}_{i,m}\mathbf{z}_m^\circ - \mathbf{H}_{i,m}\mathbf{n}_{i,m}^\circ \tag{29}$$

and $\mathbf{z}_m^\circ$ and $\mathbf{n}_{i,m}^\circ$ are constant vectors corresponding to the center of vector Taylor series expansion. In our work, we apply the statistical linear approximation [11] method for linear approximation. The matrix form of all the N observed reverberant noisy inputs can be shown as in (30).

3.5. Feature enhancement algorithm

For the clean speech estimate, a parallel extended Kalman filtering approach is applied in our algorithm, which is based on the IMM techniques [12]. The block diagram of the feature enhancement algorithm is given in Fig. 1, which is described in [7]. Unlike the single-channel algorithm, we form a super vector to accomodate all the multiple input features and utilize their correlation.

4. EXPERIMENTS

The proposed approach was applied to a connected digit recognition task using TI digits corpus. In our implementation, we employed the conventional front-end feature specified in the ETSI standard [13]. The baseline system of ASR was configured as proposed in [14], which was implemented by HTK software [15] for training and decoding. We assumed the clean training condition for the acoustic model of speech recognition in accordance with our purpose of estimating the clean feature vectors.

To simulate a reverberant noisy environment, a small rectangular room of dimensions 6 m × 4 m × 3 m (length × width × height) was configured. In our experiments, we used two omni-directional microphones ($N = 2$) placed at the height of 1.5 m. The distance between the two microphones was fixed at 10 cm. For the background noise, white noise was used as the diffuse noise source. In order to simulate various environmental conditions, we varied the distance between the target speaker and microphones from 2 m to 4 m. The target speaker was positioned at 10° deviated from the center of the microphones. The RIRs were simulated by Allen and Berkley's image method [16] using Habets's software [17]. The reverberation range was varied from 300 to 700 ms and all the conditions were tested in the SNR range 0 to 20 dB. For the results of single-channel based techniques, we took the average of the results obtained from the two separate channels. In our proposed method, there were a lot of parameters to be tuned. Due to the computational issue, we used $L = 2$ and $K = 16$ for all the experiments. For convenience, we denote the conventional single-channel algorithm by *IMM-1ch* and the proposed two-channel algorithm by *IMM-2ch*.

We present the averaged word accuracy results according to the reverberation time and SNR in Figs. 2 and 3, respectively. We calculated the average of the results obtained from the distances. From the results of Fig. 2, we can see that the performance degradation was severe as the reverberation time became longer due to the dispersive effect of reverberation. Although our proposed algorithm did not use any explicit information on the reverberation time, better performance was achieved. From a number of experiments, we can see that the proposed algorithm outperformed the single-channel algorithm.

5. CONCLUSION

In this paper, we have proposed a novel approach to estimate the clean feature vectors in multi-channel environment, which was obtained by extending the single-channel IMM algorithm to a multi-channel version. The proposed method is based on a sequential Bayesian inference framework. From various experiments in reverberant noisy environments, it has been confirmed that the proposed algorithm outperformed the traditional single-channel algorithm.

6. REFERENCES

[1] M. Miyoshi and Y. Kaneda, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 36, no. 2, pp. 145-152, Feb. 1988.

[2] T. Nakatani, B. -H. Juang, T. Yoshioka, K. Kinoshita, M. Delcroix, and M. Miyoshi, "Speech dereverberation based on maximum-likelihood estimation with time-varying Gaussian source model," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 16, no. 8, pp. 1512-1527, Nov. 2008.

[3] M. Miyoshi, M. Delcroix, K. Kinoshita, T. Yoshioka, T. Nakatani, and T. Hikichi, "Inverse filtering for speech dereverberation without the use of room acoustics information," in *Speech Dereverberation*, P. A. Naylor and N. D. Gaubitch, Eds. Springer, 2010.

[4] K. Lebart, J. M. Boucher, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359-366, 2001.

[5] K. Kinoshita, M. Delcroix, and T. Nakatani, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 17, no. 4, pp. 534-545, May 2009.

[6] E. Habets, S. Gannot, and I. Cohen "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Process. Lett.*, vol. 16, pp. 770-773, Sep. 2009.

[7] C. W. Han, S. J. Kang, and N. S. Kim, "Reverberation and noise robust feature compensation based on IMM," *IEEE Trans. Audio, Speech and Lang. Process.*, vol. 21, no. 8, pp. 1598-1611, Aug. 2013.

[8] H. -G. Hirsch and H. Finster "A new approach for the adaptation of HMMs to reverberation and background noise," *Speech Commun.*, vol. 50, no. 3, pp. 244-263, 2008.

[9] S. Markovich, S. Gannot, and I. Cohen "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 17, no. 6, pp. 1071-1086, Aug. 2009.

[10] R. Rotili, E. Principi, S. Cifani, S. Squartini, and F. Piazza, "Multichannel feature enhancement for robust speech recognition," in *Speech Technologies*, I. Ipsic, Eds. InTech, 2011.

[11] N. S. Kim, "Statistical linear approximation for environment compensation," *IEEE Signal Process. Lett.*, vol. 5, no. 1, pp. 8-10, Jan. 1998.

[12] N. S. Kim, "IMM-based estimation for slowly evolving environments," *IEEE Signal Process. Lett.*, vol. 5, no. 6, pp. 146-149, Jun. 1998.

[13] ETSI Std. Document, "Speech processing, transmission and quality aspects (STQ); distributed speech recognition; front-end feature extraction algorithm; compression algorithm," ETSI ES 201108 V1.1.3, Sep. 2003.

[14] D. Pearce and H. -G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," *Proc. Int. Conf. Spoken Lang. Process.*, Oct. 2000.

[15] S. Young et al., *The HTK book*, Cambridge University Engineering Dept., 2006.

[16] J. B. Allen and D. A. Berkley "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943-950, Apr. 1979.

[17] E. habets, "Room impulse response (RIR) generator," Sep. 2010 [On-line]. Available: http://home.tiscali.nl/ehabets/rir_generator.html