TIME-VARYING RESIDUAL NOISE FEATURE MODEL ESTIMATION FOR MULTI-MICROPHONE SPEECH RECOGNITION

Takuya Yoshioka, Emmanuel Y. J. Ternon, Tomohiro Nakatani

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikari-dai, Seika-cho, Soraku-gun, Kyoto 619-0237, Japan

ABSTRACT

This paper proposes a method for compensating for the effect of noise remaining in a signal generated by a multi-microphone signal enhancer in the feature domain as a post-processing. The proposed method assumes that the multi-microphone signal enhancer generates estimates of both the target and original environmental noise signals. To obtain a time-varying residual noise feature model that responds to noise changes quickly and is consistent with a clean feature model, the proposed method leverages both the multiple signal estimates provided by the signal enhancer and the clean feature model. Specifically, the proposed method first roughly estimates residual noise features on a frame-by-frame basis by comparing the target and noise signal estimates. Then, these rough estimates are refined by using the clean feature model to yield a time-varying residual noise feature model. Experimental results show the effectiveness of the proposed method and its wide applicability.

Index Terms— Speech recognition, noise robustness, feature enhancement, maximum likelihood, multiple microphones

1. INTRODUCTION

Robustness against acoustic environmental noise has been one of the main topics in the automatic speech recognition (ASR) research. The need for dependable noise robustness techniques seems to be growing due to the rapid spread of speech recognition technology.

Signal enhancement and feature enhancement are popular approaches to noise robust ASR [1]. The signal enhancement approach attempts to estimate a clean signal from its noisy version to improve the signal-to-noise ratio. On the other hand, the feature enhancement approach attempts to estimate clean features underlying observed noisy features by exploiting a clean feature model. A Gaussian mixture model (GMM) or a hidden Markov model (HMM) is often used as the clean feature model. The use of a clean feature model allows us to compensate effectively for the mismatch between the noisy features and an acoustic model at a practical computational cost. The clean feature model can also be used for noise feature estimation.

On the other hand, noise robust ASR systems may also be classified as either monaural or multi-microphone systems. For monaural systems, many techniques tailored for noise robust ASR have been developed including ones based on feature enhancement. These techniques are very effective when the noise is stationary or changes slowly [2]. However, when the noise is significantly non-stationary (for example, when the noise consists of voices of interfering speakers), they become ineffective due to the difficulty of noise feature model estimation. In such severe environments, multi-microphone systems are much more advantageous.

The problem common to all multi-microphone systems is residual noise. Most multi-microphone techniques are based on the signal enhancement approach. Unfortunately, some of the noise inevitably remains in the enhanced signals. This residual noise causes a mismatch between the features obtained from the enhanced signals and an acoustic model, which results in a degraded word accuracy. To compensate for this mismatch, we consider using feature enhancement techniques to eliminate the residual noise from the features extracted from the enhanced signals. However, when the residual noise is significantly non-stationary (and this often occurs when the original environmental noise is also extremely non-stationary), existing feature enhancement methods cannot estimate an accurate residual noise feature model.

The goal of this paper is to accurately estimate a time varying model of residual noise features so that we can compensate for the effect of significantly non-stationary residual noise in the feature domain as a post-processing step for multi-microphone signal enhancement. We assume that a multi-microphone signal enhancer produces at least two signals. One is an estimate of the target signal and the other (or others) is an estimate of all or a part of the environmental noise. We refer to the target estimate and the noise estimates as the *main signal* and *side signals*, respectively.

To obtain a time-varying residual noise feature model that responds to noise changes quickly and at the same time is consistent with a clean feature model used for feature enhancement, we propose a two-step approach. The first step roughly estimates the residual noise features on a frame-by-frame basis by comparing the main and side signals. The temporal dynamics of these estimates is assumed to be close to that of the true residual noise features. However, there are inevitable errors between the true residual noise features and their rough estimates. Thus, the second step makes up for these errors by using the clean feature model and yields the time-varying residual noise feature model. Specifically, in this step, we estimate the static part of the sequence of those errors, which we call the bias. In addition, we model the error's dynamic part as being independent and identically distributed samples from a zero-mean normal distribution and estimate its variance alongside the bias estimation. These two types of parameters (i.e., the bias and variance) and additional convolutive distortion parameters are optimized jointly based on a maximum likelihood (ML) criterion in the feature domain by using the clean feature model. Thus, the second step makes the residual noise feature model consistent with the clean feature model. Note that the proposed method is different from the method of [3], which integrates a generalized sidelobe canceller (GSC) and feature enhancement, in that the proposed method takes the variance and convolutive distortion into account and that the application range of the proposed method is not limited to the GSC.

The rest of this paper is organized as follows. Section 2 describes the requirements that the proposed method imposes on the multi-microphone signal enhancer. Section 3 describes the problem addressed in this paper and presents our solution. Section 4 reports some experimental results, and Section 5 concludes this paper.

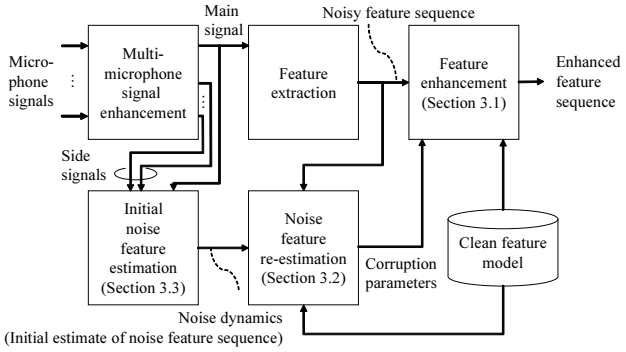

Fig. 1. Block diagram of proposed method.

2. PREREQUISITES

Let $s^{\mathrm{T}}(t)$ denote a clean speech signal of a target speaker. In addition, let $x_1^{\mathrm{T}}(t), \cdots, x_M^{\mathrm{T}}(t)$ denote M output signals from a multi-microphone signal enhancer. Superscript T indicates that these variables are defined in the time domain.

We assume that $x_1^{\mathrm{T}}(t)$ is an estimate of $s^{\mathrm{T}}(t)$ and each $x_m^{\mathrm{T}}(t)$ satisfying $m \geq 2$ is an estimate of all or a part of the environmental noise. We call this condition *the pre-separation condition*. As long as this condition is fulfilled, any algorithms can be used for multi-microphone signal enhancement performed as a pre-processing step for the proposed method. We call $x_1^{\mathrm{T}}(t)$ a main signal and $\{x_m^{\mathrm{T}}(t)\}_{m\geq 2}$ a set of side signals.

Now, we focus on two example applications to show that the pre-separation condition is not very restrictive and thus the proposed method has a wide range of applications. The first application is microphone array-based speech recognition in adverse acoustic environments. Signal enhancement techniques such as GSC and independent component analysis (ICA) can be used to reduce the noise contained in the microphone signals. Nevertheless, since the enhanced signal still contains non-negligible residual noise due to such factors as reverberation and speaker movement, compensation is needed for the residual noise. Fortunately, these techniques can generate estimates of both the target speaker's voice and the noise (or a set of noise components). Therefore, the proposed method can be used by considering the target voice estimate as a main signal and the noise estimates as side signals. The second application is meeting speech recognition using lapel microphones. This scenario considers a small group meeting where each participant wears a lapel microphone, and the goal is to recognize each person's voice separately. Although each lapel microphone picks up the associated speaker's voice with a mid-to-high signal-to-noise ratio (SNR), the noise, which consists of the other speakers' voices, has a detrimental effect on speech recognition. The relatively high SNR means that, when we recognize a specified speaker's voice, we can reasonably consider the signal of the target speaker's microphone to be the main signal and the signals of the other speakers' microphones to be side signals. Hence, the proposed method can be used to compensate for the interfering speakers' voices. The above two application scenarios are considered in the experiments reported in Section 5.

3. PROPOSED METHOD

Figure 1 shows a block diagram of a feature enhancement process based on the proposed method. A set of microphone signals is fed into the multiple microphone-based signal enhancement block to produce a main signal and a set of side signals. Then, the log mel frequency spectrum, or the feature, of the main signal is calculated in the feature extraction block for every short time frame. To highlight the fact that this main signal's feature is contaminated by some residual noise, we refer to this feature as a noisy feature. The main and side signals are also supplied to the initial noise feature estimation block, which compares the main and side signals to produce a rough estimate of the sequence of the residual noise features. This rough estimate is referred to as *noise dynamics* (the reason for this will be explained later). Then, the noise feature re-estimation block refines this rough estimate to yield a set of corruption parameters, which consists of a sequence of noise models and an estimated set of convolutive distortion parameters. Finally, the feature enhancement block removes the effect of the residual noise from the noisy feature sequence based on the vector Taylor series (VTS) method [1].

In the following, we review the VTS method in Section 3.1. We then describe the noise feature re-estimation in Section 3.2 and the initial noise feature estimation in Section 3.3. Note that the role of the initial noise feature estimation is to calculate the noise dynamics. Although we present a binary mask-based initial noise feature estimation method in Section 3.3, other algorithms can be used for implementing this block. This is why we describe the noise feature re-estimation method in Section 3.2 before presenting our initial noise feature estimation method.

3.1. VTS feature enhancement

Let $s_{n,j}$ and $x_{n,j}$ denote the log mel frequency spectra (features) of the clean speech signal $s^{\mathrm{T}}(t)$ and the main signal $x_1^{\mathrm{T}}(t)$, respectively, where n and j are the short time frame and mel frequency band indices, respectively.

VTS feature enhancement calculates a minimum mean square error (MMSE) estimate of the clean feature $s_{n,j}$. This is done by using two models: a clean feature model and a corruption model.

The clean feature model describes prior knowledge about the clean feature distribution. A GMM is often used as the clean feature model, thus we have $p(s_{n,j}) = \sum_{k=1}^{K} \pi_k f_{\mathrm{N}}(s_{n,j}; \nu_{k,j}, \tau_{k,j}^2)$, where K is the number of mixture components and $f_{\mathrm{N}}(x; \mu, \sigma^2)$ denotes the probability density function (pdf) of the normal distribution with mean μ and variance σ^2. This GMM is trained in advance by using a corpus of clean speech.

The corruption model characterizes the process in which the clean feature $s_{n,j}$ is corrupted to become the noisy feature $x_{n,j}$. The corruption model takes the form of $x_{n,j} = f(s_{n,j}, r_{n,j}, h_j)$, where $r_{n,j}$ is the unknown noise remaining in $x_{n,j}$, h_j is the unknown convolutive distortion, and f is a nonlinear function of the form

$$f(s, r, h) = s + h + \log(1 + \exp(r - s - h)). \tag{1}$$

We assume that the convolutive distortion h_j is static during observation. On the other hand, we assume the residual noise $r_{n,j}$ to be distributed normally with unknown mean $\mu_{n,j}$ and unknown variance σ_j^2 as

$$p(r_{n,j}) = f_{\mathrm{N}}(r_{n,j}; \mu_{n,j}, \sigma_j^2). \tag{2}$$

Note that $\mu_{n,j}$ depends on time frame index n, allowing the model to capture the non-stationary characteristics of the residual noise. The parameter set, $\{\mu_{n,j}, \sigma_j^2, h_j\}_{n\in\mathbb{T}, j\in\mathbb{F}}$, of the corruption model is provided by the noise feature re-estimation block.

With the above clean feature and corruption models, the MMSE estimate can be calculated by linearizing the nonlinear function f

around the means of the clean feature and noise models. For more details, see Section 33.6.2 of [1] and the references therein.

3.2. Noise feature re-estimation

The role of noise feature re-estimation is to estimate the unknown corruption parameter set consisting of the convolutive distortion parameters $\{h_j\}_{j\in\mathbb{F}}$, the residual noise means $\{\mu_{n,j}\}_{n\in\mathbb{T},j\in\mathbb{F}}$ of all time frames, and the variances $\{\sigma_j^2\}_{j\in\mathbb{F}}$, where $\mathbb{T}$ and $\mathbb{F}$ denote the sets of all time frame indices and all mel frequency band indices, respectively. However, this is unfeasible without any additional information since the number of unknown parameters exceeds the size of the available data $\mathbb{X} = \{x_{n,j}\}_{n\in\mathbb{T},j\in\mathbb{F}}$. Previously proposed monaural feature enhancement methods assume $\mu_{n,j}$ not to change during observation or introduce strong dependency between $\mu_{n,j}$ and $\mu_{n-1,j}$ to avoid this problem. But, this prevents $\mu_{n,j}$ from following fast changes in noise characteristics.

Our approach is to decompose the mean sequence $(\mu_{n,j})_{n\in\mathbb{T}}$ into a dynamics part $(\hat{r}_{n,j})_{n\in\mathbb{T}}$ and a static bias b_j as

$$\mu_{n,j} = \hat{r}_{n,j} + b_j. \tag{3}$$

The key assumption is that the noise dynamics $(\hat{r}_{n,j})_{n\in\mathbb{T}}$ is obtained during initial noise feature estimation as a rough estimate of a sequence of residual noise features by exploiting the outputs of the multi-microphone signal enhancement block (i.e., the main and side signals). The bias b_j is needed because the noise dynamics $(\hat{r}_{n,j})_{n\in\mathbb{T}}$ is calculated without taking a clean feature model into consideration and thus is inconsistent with the clean feature model.

Owing to this assumption, the problem boils down to a joint estimation of the convolutive distortion parameters $\{h_j\}_{j\in\mathbb{F}}$, the biases $\{b_j\}_{j\in\mathbb{F}}$, and the variances $\{\sigma_j^2\}_{j\in\mathbb{F}}$. Obviously, the number of unknown parameters is now smaller than the data size.

We solve this parameter estimation problem by maximizing the likelihood function, $p(\mathbb{X};\Theta)$, where Θ is the parameter set $\{h_j, b_j, \sigma_j^2\}_{j\in\mathbb{F}}$. According to the VTS method, the likelihood function can be factorized as $p(\mathbb{X};\Theta) = \prod_{n\in\mathbb{T}}\prod_{j\in\mathbb{F}}\sum_{k=1}^{K}\pi_k p(x_{n,j}|k;\Theta)$, where each constituent pdf is approximated by a Gaussian as

$$p(x_{n,j}|k,;\Theta) = f_{\mathrm{N}}(x_{n,j};\psi_{n,k,j},\upsilon_{n,k,j}^2). \tag{4}$$

The mean $\psi_{n,k,j}$ and the variance $\upsilon_{n,k,j}^2$ are calculated as

$$\psi_{n,k,j} = f(\nu_{k,j}, \hat{r}_{n,j} + b_j, h_j) \tag{5}$$

$$\upsilon_{n,k,j}^2 = g(\nu_{k,j}, \hat{r}_{n,j} + b_j, h_j)^2\tau_{k,j}^2 + (1 - g(\nu_{k,j}, \hat{r}_{n,j} + b_j, h_j))^2\sigma_j^2, \tag{6}$$

where $g(s,r,h)$ is the partial derivative of the nonlinear function $f(s,r,h)$ with respect to s, i.e., $g(s,r,h) = 1/(1+\exp(r-s-h))$.

To solve this maximum likelihood problem, we use a variant of the twofold expectation maximization (EM) algorithm [4]. In the proposed method, the estimation process for the biases and variances and that for the convolutive distortion parameters are interleaved and repeated. Here, due to space limitations, we describe only the algorithm for estimating the biases and variances. The algorithm for estimating the convolutive distortion parameters can be derived similarly.

In the proposed bias and variance estimation algorithm, the GMM component index and the residual noise feature of each time frame are regarded as latent variables. Let $\hat{\Theta}$ denote a tentative estimate of Θ. Then, the auxiliary function to be maximized at each EM iteration is given by

$$Q(\Theta;\hat{\Theta}) = \sum_{n\in\mathbb{T}}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})\int q_{n,k,j}(r;\hat{\Theta})\log f_{\mathrm{N}}(r;\hat{r}_{n,j}+b_j,\sigma_j^2)dr. \tag{7}$$

Here, $\gamma_{n,k}(\hat{\Theta})$ is the conditional posterior probability of the kth GMM component being active at frame n given $\hat{\Theta}$. $q_{n,k,j}(r;\hat{\Theta})$ is the conditional posterior pdf over the residual noise feature at frame n and band j given GMM component index k and $\hat{\Theta}$. $\gamma_{n,k}(\hat{\Theta})$ and $q_{n,k,j}(r;\hat{\Theta})$ are calculated in the E-step as

$$\gamma_{n,k}(\hat{\Theta}) = \frac{\prod_{j\in\mathbb{F}} p(x_{n,j}|k;\hat{\Theta})}{\sum_{k=1}^{K}\prod_{j\in\mathbb{F}} p(x_{n,j}|k;\hat{\Theta})} \tag{8}$$

$$q_{n,k,j}(r;\hat{\Theta}) = f_{\mathrm{N}}(r;\kappa_{n,k,j}(\hat{\Theta}),\lambda_{n,k,j}^2(\hat{\Theta})) \tag{9}$$

$$\kappa_{n,k,j}(\hat{\Theta}) = \hat{r}_{n,j} + \hat{b}_j + \hat{g}_{n,k,j}\hat{\sigma}_j^2(x_{n,j} - \hat{\psi}_{n,k,j})/\hat{\upsilon}_{n,k,j}^2 \tag{10}$$

$$\lambda_{n,k,j}^2(\hat{\Theta}) = \hat{\sigma}_j^2\hat{g}_{n,k,j}^2\tau_{k,j}^2/\hat{\upsilon}_{n,k,j} \tag{11}$$

$$\hat{g}_{n,k,j} = g(\nu_{k,j}, \hat{r}_{n,j} + \hat{b}_j, \hat{h}_j) \tag{12}$$

Then, in the M-step, we update the estimates of the bias and variance for each $j \in \mathbb{F}$ as

$$\hat{b}_j = \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{T}}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})(\kappa_{n,k,j}(\hat{\Theta}) - \hat{r}_{n,j}) \tag{13}$$

$$\hat{\sigma}_j^2 = \frac{1}{|\mathbb{T}|}\sum_{n\in\mathbb{T}}\sum_{k=1}^{K}\gamma_{n,k}(\hat{\Theta})(\kappa_{n,k,j}(\hat{\Theta})^2 + \lambda_{n,j,k}(\hat{\Theta})^2) - \hat{b}_j^2. \tag{14}$$

3.3. Initial noise feature estimation

The initial noise feature estimation block calculates the noise dynamics. The noise dynamics $(\hat{r}_{n,j})_{n\in\mathbb{T}}$ needs to be a good estimate of a sequence of true residual noise features $(r_{n,j})_{n\in\mathbb{T}}$ up to a static bias.

At the heart of the initial noise feature estimation method presented here lies the idea that changes in the noise spectrum can be detected much more easily in the high-dimensional power spectrum domain than in the low-dimensional feature domain. In light of this, the method consists of two steps. The first step obtains an estimate of the noise log power spectrum of each short time frame, which we represent as $\hat{R}_{n,i}$, where i is the high-dimensional frequency bin index. $\hat{R}_{n,i}$ is transformed into a feature vector in the second step, and the result is used as the noise dynamics component $\hat{r}_{n,j}$.

The residual noise log power spectrum estimation in the first step is performed based on the binary mask concept [5], which is widely employed for single- and multi-channel separation of speech from highly non-stationary environmental noise such as interfering human voices. A binary mask $A_{n,i}$ is a binary variable that indicates the presence (0) or absence (1) of target speech at its associated time frequency point (n,i). Leveraging the pre-separation condition, $A_{n,i}$ is set at 1 if there exists $m \neq 1$ such that $X_{1,n,i} < X_{m,n,i}$ and 0 otherwise. Here, $X_{m,n,i}$ is the log power spectrum of the mth output signal of a multi-microphone signal enhancement block ($m = 1$ corresponds to the main signal). By using these binary masks, $\hat{R}_{n,i}$ is calculated as a locally weighted average of the log power spectra of the main signal. Specifically,

$$\hat{R}_{n,i} = \frac{\sum_{\tau=-\Delta_{\mathrm{T}}}^{\Delta_{\mathrm{T}}}\sum_{\phi=-\Delta_{\mathrm{F}}}^{\Delta_{\mathrm{F}}} A_{n+\tau,i+\phi}X_{1,n+\tau,i+\phi}}{\sum_{\tau=-\Delta_{\mathrm{T}}}^{\Delta_{\mathrm{T}}}\sum_{\phi=-\Delta_{\mathrm{F}}}^{\Delta_{\mathrm{F}}} A_{n+\tau,i+\phi}}, \tag{15}$$

where Δ_{T} and Δ_{F} specify the window for the local averaging. Our current implementation uses $\Delta_{\mathrm{T}} = 3$ and $\Delta_{\mathrm{F}} = 2$.

4. EXPERIMENTAL RESULTS

We conducted two experiments to confirm the effectiveness of the proposed method. As noted in Section 2, the first experiment considered microphone array-based recognition of digit strings corrupted

by interfering voices while the second experiment performed large vocabulary meeting speech recognition. The results of the respective experiments are reported in Sections 4.1 and 4.2.

4.1. Microphone array-based speech recognition

The purpose of this experiment was to recognize a spoken digit string corrupted by a different speaker's voice when a set of two different mixtures of the target and interfering speech signals was given. The vocabulary for this experiment was limited to a set of digits to evaluate the proposed method with a focus on acoustic factors.

To create each pair of mixed signals, we convolved a pair of target and interfering speech signals with a two-input two-output (TITO) room impulse response measured in a room with a reverberation time of 0.13 s. The room was 4.45 m wide and 3.35 m long with a 2.5 m high ceiling. To measure the TITO room impulse response in this room, we used a two-element microphone array placed at the center of the room and two loudspeakers, one placed to the left of the microphone array and one to the right, both at an angle of 30 degrees. The loudspeakers were 1 m from the microphone array. The clean subset of the Aurora2 test set was used as the target speech signals while the interfering speech signals were taken from the TIMIT corpus. Specifically, we made 4004 pairs of target and interfering speech signals and convolved each pair with the above TITO room impulse response to create a pair of mixed signals.

The acoustic model for this experiment was trained on the Aurora2 clean training set according to the complex back-end recipe. Thus, the acoustic model consisted of speaker-independent word HMMs with 16 states and three Gaussians per state. The acoustic features consisted of 13 MFCCs augmented with their velocity and acceleration parameters.

The results were as follows: the word error rate (WER) was 178.25 % when we fed the mixed signals directly into the speech recognizer. By separating individual digit strings from interfering voices with frequency-domain ICA [6] before performing speech recognition, the WER decreased to 25.78 %. This WER was further improved down to 3.21 % when we used the proposed method after the ICA-based source separation. However, when we applied a single-channel VTS feature enhancement algorithm combined with an expectation maximization-based noise estimator to the separated signals, the WER was 20.41 %. The limited improvement provided by the signle-channel algorithm is attributed to the stationarity assumption made by the noise estimator, which in turn shows that our proposed method could compensate effectively for the effect of non-stationary residual interfering speech.

4.2. Meeting speech recognition using lapel microphones

The goal of this experiment was to recognize the voice of each meeting participant wearing a lapel microphone. This experiment was aimed at evaluating the proposed method in a large vocabulary multi-speaker speech recognition task.

For this experiment, we recorded 16 sessions of Japanese meeting in two different rooms. 8 sessions were used as a development data set, and the remaining 8 sessions were used for the test. Each session involved 4 speakers and lasted approximately 15 minutes. We manually segmented each session into separate utterances and each of these utterances were used for evaluation. Our acoustic model was trained by using the CSJ corpus, which consists of academic and simulated presentations. For language model training, we used a set of meeting data excluding the test set, the CSJ corpus, and sentences extracted from the Web. Our speech recognizer used for this experiment consists of a discriminatively trained acoustic model, a Kneser-Ney smoothed word trigram language model, and a weighted finite state transducer-based decoder. The configurations of our speech recognizer is almost the same as those used for our previous meeting speech recognizer, which is described in [7].

The results were as follows: the WER obtained with the raw data captured by the lapel microphones was 59.6 %. The WER was reduced to 47.2 % by performing feature enhancement with the proposed method. This result shows that the proposed method can also be employed for improving meeting speech recognition performance based on the use of lapel microphones. However, when we used headset microphones to capture each person's voice accurately, the WER was 35.7 %, which indicates the need for further study to close the gap between the performance when using lapel microphones and that when using headset microphones. Thorough investigation of the results of this experiment will be described in a separate paper.

5. CONCLUSION

This paper described a method for estimating a time-varying residual noise feature model that is needed to compensate for the effect of significantly non-stationary residual noise in the feature domain. The noise model consists of a sequence of means and a static variance for each log mel frequency band. The key feature of the method is to regard the mean sequence as being composed of noise dynamics and a static bias. The noise dynamics is obtained by exploiting multiple microphones while the bias is estimated jointly with the variance and additional convolutive distortion parameters by using a clean feature model. The proposed method was shown to be effective in both a microphone array-based speech recognition task and a lapel microphone-based meeting speech recognition task, which demonstrates its wide applicability.

6. REFERENCES

[1] J. Droppo and A. Acero, "Environmental robustness," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds., pp. 653–679. Springer, 2008.

[2] S. Rennie, et al., "Dynamic noise adaptation," in *Proc. Int'l Conf. Acoust., Speech, Signal Process.*, 2006, pp. 1197–1200.

[3] X. Zhao and Z. Ou, "Closely coupled array processing and model-based compensation for microphone array speech recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1114–1122, 2007.

[4] Y. Zhao and B.-H. Juang, "A comparative study of noise estimation algirhtms for VTS-based robust speech recognition," in *Proc. Interspeech*, 2010, pp. 2090–2093.

[5] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Trans. Signal Process.*, vol. 52, no. 7, pp. 1830–1847, 2004.

[6] H. Sawada, et al., "Measuring dependence of bin-wise separated signals for permutation alignment in frequency-domain BSS," in *Int'l Symp. Circ., Syst.*, 2007, pp. 3247–3250.

[7] T. Hori, et al., "Low-latency real-time meeting recognition and understanding using distant microphones and omni-directional camera," *IEEE Trans. Audio, Speech, Language Process.*, 2011, to appear.