LOGMAX OBSERVATION MODEL WITH MFCC-BASED SPECTRAL PRIOR FOR REDUCTION OF HIGHLY NONSTATIONARY AMBIENT NOISE

Tomohiro Nakatani Takuya Yoshioka Shoko Araki Marc Delcroix Masakiyo Fujimoto

NTT Communication Science Laboratories, NTT Corporation
2-4, Hikaridai, Seikacho, Sorakugun, Kyoto, 619-0237 Japan

ABSTRACT

This paper proposes a new single/multi-channel speech enhancement approach based on a LogMax observation model integrated with Gaussian mixture models of speech and noise mel-frequency cepstral coefficients (MFCC-GMM). It has been reported that the LogMax observation model has high potential for reducing highly nonstationary noise, for example, when it is combined with factorial hidden Markov models. In addition, it has recently been shown that a source location based speech enhancement approach can be easily incorporated into this model for more efficient and reliable estimation. However, the unique structure of the LogMax model has prevented us from using it with MFCC-GMMs, which is a fundamental limitation of this approach. Our proposal in this paper is aimed at overcoming this limitation. Experiments using the PASCAL CHiME separation and recognition challenge task show the superiority of the proposed approach as regards both speech quality and automatic speech recognition performance.

Index Terms: Speech enhancement, mel-frequency cepstral coefficients, automatic speech recognition, model-based approach

1. INTRODUCTION

When we capture our daily speech using distant microphones, various types of ambient noise, including time-varying noise, are mixed with the captured signals, and severely degrade the audible quality of the speech and the automatic speech recognition (ASR) performance.

To solve this problem, model-based noise reduction approaches have been extensively studied, where statistical models of speech log-spectra are utilized as prior knowledge to improve the noise reduction performance. A vector Taylor series (VTS) approximation approach [1] has been widely used for reducing stationary/slowly time-varying noise. Another important approach is based on the LogMax model [2], and has been proposed to cope well with highly nonstationary noise, such as extraneous speakers. This model assumes that the observed spectral value at each time frequency is equal to the maximum speech and noise spectral values, and the noise reduction is accomplished by finding a pair of log-spectra for speech and noise that best fits the observation. For more efficient and reliable estimation, a new technique has recently been proposed that extends the LogMax model approach to multi-microphone cases [3], where the location features of the speech and noise are utilized jointly with the spectral characteristics modeled by GMMs. It is referred to as DOminance based Locational and Power-spectral cHaracteristics INtegration (DOLPHIN), and has been shown to achieve accurate noise reduction even under highly nonstationary noise conditions, and to improve the ASR performance greatly [4].

With the LogMax model based noise reduction, we need to evaluate which of speech and noise has a larger spectral value at each time-frequency bin, and to do this in a computationally tractable way, the existing approaches assume that the spectral values over different frequencies are statistically independent of each other given the GMM indices [2], dealing with the indices as the parameters to be estimated. However, the characteristics of actual speech do not well meet this assumption, and thus the accuracy of the speech model is limited based on this assumption. For example, although MFCC-GMMs are currently thought to be one of the best statistical models for speech log-spectra, the existing approaches with the LogMax model cannot adopt them as the spectral models because the spectral values over different frequencies are essentially correlated when the corresponding MFCCs follow a Gaussian distribution.

This paper proposes a new estimation setting for DOLPHIN that allows us to use MFCC-GMM spectral priors with it. In this setting, the MFCCs are assumed to be generated by MFCC-GMMs, and their log-spectra are generated depending on the MFCCs, where the values of the log-spectra are assumed to be statistically independent over different frequencies given the MFCCs. Furthermore, the MFCCs of speech and noise are handled as parameters to be estimated based on Maximum a Posteriori (MAP) estimation. This setting allows DOLPHIN to evaluate efficiently which of speech and noise log-spectra has a larger value at each frequency, and thus to perform a computationally efficient noise reduction with MFCC-GMMs. In the rest of this paper, we first present a new formulation for DOLPHIN in Section 2, using a monaural speech enhancement scenario, namely a scenario without location features. We refer to this as DOLPHIN-MFCC-1ch, or DOLPHIN-MC1 for short. In Section 3, we extend DOLPHIN-MC1 to cope with locational features. The resultant algorithm is referred to as DOLPHIN-MC2. Section 4 provides experimental results showing the superiority of both proposed methods in terms of the audible quality and ASR of the enhanced speech compared with conventional approaches.

2. FORMULATION FOR MONAURAL PROCESSING

Suppose $\mathbf{x} = [x_1, x_2, \ldots, x_K]$ is a log-mel filterbank output of an observed monaural signal at a time frame, where x_k is its k-th frequency element. Because the method proposed in this paper is applied to each time frame independently, the time frame indices of symbols are omitted hereafter. Based on the LogMax model, the observed signal is modeled as

$$x_k = \max\{s_k^{(1)}, s_k^{(2)}\}, \tag{1}$$

where $\mathbf{s}^{(l)} = [s_1^{(l)}, \ldots, s_K^{(l)}]$ for $l = 1$ and $l = 2$, respectively, are the unknown filterbank outputs of the speech and the noise. Hereafter, l is used as the index of the two sources, namely the speech ($l = 1$) and the noise ($l = 2$). Letting H be a discrete cosine transformation matrix, a pair of MFCCs for $\mathbf{s}^{(1)}$ and $\mathbf{s}^{(2)}$ are defined as

$$\mathbf{c} = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}\} \quad \text{where} \quad \mathbf{c}^{(l)} = H\mathbf{s}^{(l)}. \tag{2}$$

DOLPHIN-MC1 deals with the MFCCs as parameters to be estimated by the MAP estimation, which is defined as

$$\hat{\mathbf{c}} \quad = \quad \arg\max_{\mathbf{c}} p(\mathbf{x}, \mathbf{c}). \tag{3}$$

Using the estimated MFCCs, we then estimate clean speech spectra based on the minimum mean squared error (MMSE) estimation, which is defined as (14) in Section 2.2.

To solve the above problem, DOLPHIN-MC1 employs a generative model of the observed log-mel filterbank outputs as illustrated in Fig. 1 and the relationships defined in the following:

$$p(\mathbf{c}^{(l)}) = \sum_{i^{(l)}} p(i^{(l)}) p(\mathbf{c}^{(l)}|i^{(l)}) \tag{4}$$

$$p(\mathbf{s}^{(l)}|\mathbf{c}^{(l)}) = \mathcal{N}(\mathbf{s}^{(l)}; \boldsymbol{g}(\mathbf{c}^{(l)}), \Xi). \tag{5}$$

In (4), the MFCC $\mathbf{c}^{(l)}$ for each l is assumed to be generated by an MFCC-GMM. $p(i^{(l)})$ and $p(c^{(l)}|i^{(l)})$ are a mixture weight and a Gaussian probability density function (pdf) of the $i^{(l)}$-th component, respectively, and assumed to be trained in advance. In (5), we assumed that $\mathbf{s}^{(l)}$ can be predicted from $\mathbf{c}^{(l)}$ by a linear regression[1] $\boldsymbol{g}(\mathbf{c}^{(l)})$, and the prediction error follows a Gaussian pdf with a zero mean and a diagonal covariance matrix $\Xi = \mathrm{diag}\{\xi_k\}$. Then, similar to the conventional DOLPHIN approach, we introduce the following equations to deal with the LogMax model (1) in a probabilistic form.

$$p(x_k|d_k, s_k^{(1)}, s_k^{(2)}) \;=\; \delta(x_k - s_k^{(d_k)}) \tag{6}$$

$$p(d_k|s_k^{(1)}, s_k^{(2)}) \;=\; \begin{cases} 1 & \text{when } d_k = \arg\max_l s_k^{(l)} \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbf{d} = [d_1, \ldots, d_K]$ is a set of dominant source indices (DSI) that indicate which of speech ($d_k = 1$) and noise ($d_k = 2$) has the larger energy at frequency k. $\delta(\cdot)$ is the Dirac delta function. Then, the relationship between x_k, d_k, and $\mathbf{c}$ can be derived based on (5), (6), and (7), with marginalization over $s_k^{(1)}$ and $s_k^{(2)}$ as

$$\begin{aligned} p(x_k, d_k|\mathbf{c}) \;&=\; \iint p(x_k, d_k|s_k^{(1)}, s_k^{(2)}) \prod_l p(s_k^{(l)}|\mathbf{c}^{(l)}) ds_k^{(1)} ds_k^{(2)} \\ &=\; p(s_k^{(d_k)} = x_k|\mathbf{c}^{(d_k)}) \int_{-\infty}^{x_k} p(s_k^{(d'_k)}|\mathbf{c}^{(d'_k)}) ds_k^{(d'_k)}, \end{aligned} \tag{8}$$

where d'_k is a non-dominant source index. The MAP function $p(\mathbf{x}, \mathbf{c})$ in (3) can finally be defined as

$$p(\mathbf{x}, \mathbf{c}) = \left(\prod_k \sum_{d_k} p(x_k, d_k|\mathbf{c}) \right) \left(\prod_l \sum_{i^{(l)}} p(\mathbf{c}^{(l)}, i^{(l)}) \right), \tag{9}$$

where d_k and $i^{(l)}$ are hidden variables. The first component on the right hand side reflects the LogMax model, while the second component reflects the MFCC-GMMs. With the above function, the MFCCs are estimated based on both models.

One important feature in the above setting is that the joint pdf of $\mathbf{x}$ and $\mathbf{d}$ given the parameters to be estimated, or (8) in the above case, is in a simple form, namely it is defined separately for individual frequencies. This simple form allows us to derive the computationally efficient estimation procedure described in the following paragraphs. Note that, to obtain this simple form, we handle the MFCCs as parameters to be estimated by the MAP estimation, and assume that the covariance matrix of $p(\mathbf{s}^{(l)}|\mathbf{c}^{(l)})$ in (5), namely Ξ, is diagonal. By contrast, if we deal with the Gaussian index pair $\mathbf{i} = \{i^{(1)}, i^{(2)}\}$ as parameters to be estimated similar to the conventional DOLPHIN approach and marginalize the generative model over the MFCC pair $\mathbf{c} = \{\mathbf{c}^{(1)}, \mathbf{c}^{(2)}\}$, we can no longer utilize a simple form such as that in (8) because the covariance matrix of $p(\mathbf{s}^{(l)}|i^{(l)})$ can never be diagonal with the MFCC-GMMs.

[1]For example, $\boldsymbol{g}(\mathbf{c}^{(l)})$ can be a pseudo-inversion of H. In our experiments, we parameterized $\boldsymbol{g}(\mathbf{c})$ as $\boldsymbol{g}(\mathbf{c}) = A\mathbf{c} + \mathbf{b}$, and set A and $\mathbf{b}$ as values that minimize $E\{|\mathbf{s} - \boldsymbol{g}(\mathbf{c})|^2\}$ in a training data set. We then set ξ_k as the average squared error at a frequency k.

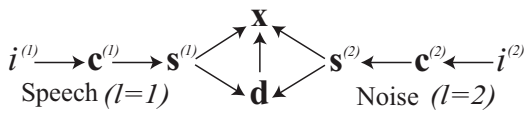

Fig. 1. Graphical model of DOLPHIN-MC1.

2.1. MAP estimation of MFCCs

Because the MAP function (9) includes hidden variables, DOLPHIN-MC1 uses the expectation maximization (EM) algorithm for the maximization, where the MFCCs are estimated by iterating the E- and M-steps derived with the EM algorithm. The auxiliary function, $Q(\mathbf{c}|\hat{\mathbf{c}}) = E\{\log p(\mathbf{x}, \mathbf{d}, \mathbf{c}, \mathbf{i})|\hat{\mathbf{c}}\} = \sum_k E\{\log p(x_k, d_k|\mathbf{c})|\hat{\mathbf{c}}\} + \sum_l E\{\log p(\mathbf{c}^{(l)}, i^{(l)})|\hat{\mathbf{c}}\}$, can be expanded as

$$Q(\mathbf{c}|\hat{\mathbf{c}}) \;=\; \sum_l Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}}) \tag{10}$$

$$Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}}) \;=\; \sum_k \Psi_k(\mathbf{c}^{(l)}; \hat{\mathbf{c}}) + \sum_i Z_i^{(l)} \log p(\mathbf{c}^{(l)}|i^{(l)} = i) \tag{11}$$

$$\begin{aligned} \Psi_k(\mathbf{c}^{(l)}; \hat{\mathbf{c}}) \;&=\; D_k^{(l)} \log p(s_k^{(l)} = x_k|\mathbf{c}^{(l)}) \\ &\quad + (1 - D_k^{(l)}) \log \int_{-\infty}^{x_k} p(s_k^{(l)}|\mathbf{c}^{(l)}) ds_k^{(l)} \end{aligned} \tag{12}$$

$$D_k^{(l)} \;=\; p(d_k = l|x_k, \hat{\mathbf{c}}) \quad \text{and} \quad Z_i^{(l)} = p(i^{(l)} = i|\hat{\mathbf{c}}^{(l)}). \tag{13}$$

$D_k^{(l)}$ and $Z_i^{(l)}$ are posteriors of the hidden variables, d_k and $i^{(l)}$, updated in E-step. $Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})$ in (11) is the MFCC matching function used in M-step. Because (11) only contains $\mathbf{c}^{(l)}$ for a certain l, we can update the MFCCs of each source independently in M-step, which is a unique advantage of the DOLPHIN approach. The first and the second terms in (11) reflect the LogMax model and the MFCC-GMM prior, respectively. In addition, the first and the second terms in (12) ensure that the dominant and non-dominant sources take the same value as the observed value and any value smaller than the observed value, respectively.

One issue that makes the maximization somewhat complex is the treatment of the second term in (12). To maximize a function with such a non-linear term, we can generally use a gradient-descent approach. In particular, as discussed in [3], the partial derivative of this term on an MFCC dimension has a hinge function shape, so we can adopt the simplest gradient-descent approach, namely the Newton-Raphson method, for the maximization. Our preliminary experiments revealed the numerical stability and effectiveness of this method. In our experiments, we calculated the first and second order derivatives of the second term in (12) using the Matlab error function "erfcx".

2.1.1. Processing flow of MAP estimation

1. Initialize $\hat{\mathbf{c}}^{(l)}$ for all l as MFCCs of $\mathbf{x}$.

2. Iterate the following until convergence is achieved.

   (a) Update $D_k^{(l)}$ and $Z_i^{(l)}$ for all k, l, and i, as follows.

$$D_k^{(l)} = \frac{p(x_k, d_k = l|\hat{\mathbf{c}})}{\sum_{d_k} p(x_k, d_k|\hat{\mathbf{c}})} \text{ and } Z_i^{(l)} = \frac{p(i^{(l)} = i)p(\hat{\mathbf{c}}^{(l)}|i^{(l)} = i)}{\sum_{i^{(l)}} p(i^{(l)})p(\hat{\mathbf{c}}^{(l)}|i^{(l)})}.$$

   (b) Update $\hat{\mathbf{c}}^{(l)}$ for each l as

$$\hat{\mathbf{c}}^{(l)} = \hat{\mathbf{c}}^{(l)} - \left(\frac{\partial^2 Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})}{\partial \mathbf{c}^{(l)\,2}}\right)^{-1} \left(\frac{\partial Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})}{\partial \mathbf{c}^{(l)}}\right),$$

   where $\frac{\partial Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})}{\partial \mathbf{c}^{(l)}}$ and $\frac{\partial^2 Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})}{\partial \mathbf{c}^{(l)\,2}}$ are a gradient vector and a Hessian matrix of $Q^{(l)}(\mathbf{c}^{(l)}|\hat{\mathbf{c}})$, respectively.

2.1.2. Variation of spectral models

Because MFCCs are dealt with as parameters to be estimated in the MAP estimation, the update of each source model in E-step, or the update of $Z_i^{(l)}$ with the above procedure, can be separated from the other parts of the estimation. This allows us to introduce different types of spectral models in a rather flexible manner without greatly increasing the computational complexity. For example, the introduction of hidden Markov models for MFCCs is very straightforward. In addition, the adaptation of the model parameters given the observation can also be included in a straightforward way.

Here,let us define another useful spectral model for noise, for which all the model parameters are estimated from the observed signals with no prior training. This model is referred to as MFCC-GM, and is composed of a single Gaussian model defined as $p(\mathbf{c}^{(2)}) = \mathcal{N}(\mathbf{c}^{(2)}; \boldsymbol{\mu}, \Sigma)$. In the MAP estimation with this model, we update the model parameters, $\boldsymbol{\mu}$ and Σ, in M-step, instead of updating the posterior $Z_i^{(2)}$ in E-step. Because we have estimated values, $\hat{\mathbf{c}}^{(2)}$, $\boldsymbol{\mu}$ and Σ are calculated simply as a mean and a covariance matrix of $\hat{\mathbf{c}}^{(2)}$ over all time frames. In our preliminary experiments using the CHiME Challenge database [5], for the noise model, MFCC-GM with no prior training was superior to MFCC-GMM with no model adaptation. This is probably because the noise in the database is so diverse that an MFCC-GMM cannot represent the pdf precisely without any adaptation to the observation. So, we use an MFCC-GM for the noise model in the experiments described in this paper.

2.2. MMSE estimation of log-power spectra

Because the goal of this paper is speech enhancement, we estimate high-resolution spectra of clean speech in the log-power spectral domain, denoted by $\tilde{\mathbf{s}}^{(1)}$, based on the MMSE estimation, which is defined as

$$\hat{\tilde{\mathbf{s}}}^{(1)} = \int \tilde{\mathbf{s}}^{(1)} p(\tilde{\mathbf{s}}^{(1)}|\mathbf{x}, \hat{\mathbf{c}}) d\tilde{\mathbf{s}}^{(1)}. \tag{14}$$

For this estimation, we introduce a linear regression[2] $\tilde{\boldsymbol{g}}(\mathbf{c}^{(l)})$ that predicts $\tilde{\mathbf{s}}^{(l)}$ from $\mathbf{c}^{(l)}$ as

$$p(\tilde{\mathbf{s}}^{(l)}|\mathbf{c}^{(l)}) = \mathcal{N}(\tilde{\mathbf{s}}^{(l)}; \tilde{\boldsymbol{g}}(\mathbf{c}^{(l)}), \tilde{\Xi}). \tag{15}$$

Then, the MMSE estimation of $\tilde{\mathbf{s}}^{(1)}$ becomes

$$\hat{\tilde{s}}_{\tilde{k}}^{(1)} = \tilde{D}_{\tilde{k}}^{(l)} \tilde{x}_{\tilde{k}} + (1 - \tilde{D}_{\tilde{k}}^{(l)}) \frac{\int_{-\infty}^{\tilde{x}_{\tilde{k}}} \tilde{s}_{\tilde{k}} p(\tilde{s}_{\tilde{k}}|\hat{\mathbf{c}}^{(l)}) d\tilde{s}_{\tilde{k}}}{\int_{-\infty}^{\tilde{x}_{\tilde{k}}} p(\tilde{s}_{\tilde{k}}|\hat{\mathbf{c}}^{(l)}) d\tilde{s}_{\tilde{k}}}, \tag{16}$$

[2]In our experiments, we also parameterized $\tilde{\boldsymbol{g}}(\mathbf{c}^{(l)})$ as $\tilde{\boldsymbol{g}}(c^{(l)}) = \tilde{A}\mathbf{c}^{(l)} + \tilde{\mathbf{b}}$, and determined the parameters using a training data set.

where $\tilde{k}$ is a frequency index in the log-power spectral domain, $\tilde{x}_{\tilde{k}}$ is the log-power spectral value of the observation, and $\tilde{D}_{\tilde{k}}^{(l)}$ is the posterior of a DSI defined in the log-power spectral domain and calculated using the estimated MFCC $\hat{\mathbf{c}}^{(l)}$ and $p(\tilde{\mathbf{s}}^{(l)}|\mathbf{c}^{(l)})$ in (15) in a way similar to that for $D_k^{(l)}$ in the filterbank domain.

The enhanced speech waveform is then calculated by using an inverse Fourier transform of $\exp(\hat{\tilde{\mathbf{s}}}^{(1)})$ with the phase of the observed signal followed by overlap-add synthesis.

3. INCORPORATION OF LOCATION MODEL

Similar to the conventional DOLPHIN approach, we can incorporate location based speech enhancement techniques [6] into DOLPHIN-MC1. We refer to this as DOLPHIN-MC2. In this paper, we introduce the same technique used in [4]. Because of the limited space, this paper only provides an outline of the method.

Let $\mathbf{a} = [a_1, \ldots, a_K]$ be additional observed features, referred to as location features, used for this incorporation, and $p(a_k^{(l)})$ be a pdf of the location feature of the l-th source. We assume that $p(a_k^{(l)})$ can be fixed in advance[3] based on prior training as in [4]. Then, we obtain $L_k^{(l)} = p(a_k^{(l)}) / \sum_l p(a_k^{(l)})$, which we refer to as normalized location posteriors, at all frequencies k. Finally, to incorporate the location model in the MAP estimation of DOLPHIN-MC1, we only need to modify the update of $D_k^{(l)}$ in step 2(a) as

$$D_k^{(l)} = \frac{L_k^{(l)} p(x_k, d_k = l|\hat{\mathbf{c}})}{\sum_{d_k} L_k^{(d_k)} p(x_k, d_k|\hat{\mathbf{c}})}. \tag{17}$$

For the MMSE estimation of log-power spectra, $\tilde{D}_{\tilde{k}}^{(l)}$ can also be estimated using $\tilde{L}_{\tilde{k}}^{(l)}$ defined in the log power spectral domain. $L_k^{(l)}$ and $\tilde{L}_{\tilde{k}}^{(l)}$ can be calculated using a technique proposed in [4].

4. EXPERIMENTS

To evaluate DOLPHIN-MC1 (Prop-1ch) and DOLPHIN-MC2 (Prop-2ch), we used the PASCAL CHiME speech separation and recognition challenge database [5]. We adopted the same feature extraction procedure used in [4]. Spectral features, $\mathbf{x}$, were extracted after applying the delay-and-sum beamformer to the observed 2ch signal to enhance the front signal, and was used in both methods. Location features, $\mathbf{a}$, were extracted from the observed 2ch signal, and used by Prop-2ch. The frame size and shift were set at 100 ms and 25 ms, respectively. As prior training, speaker dependent MFCC-GMMs were trained on individual speakers in the training set. The dimensions of the MFCCs and filterbank outputs were set at 13 and 40, respectively, and the mixture component numbers were set at 256. As noise, we used an MFCC-GM with no prior training. A location model for speech and one for noise were trained on their respective training sets, which were the same as used in [4]. We compared Prop-1ch with the VTS approach [1] (Conv-1ch) as monaural speech enhancement techniques, and compared Prop-2ch with the conventional DOLPHIN (Conv-2ch) proposed in [4] as 2ch speech enhancement techniques. The same analysis conditions were adopted as those used in [4] for Conv-1ch and Conv-2ch. In particular, Conv-2ch used the GMMs of high-resolution log-power spectra for both speech and noise, and achieved the best performance in [4]. Note that we also used Conv-2ch in the initialization step of Prop-2ch because it achieved the best results in our preliminary evaluation using the development set.

[3]Or $p(a_k^{(l)})$ can also be learned from the observed signal as in [3].

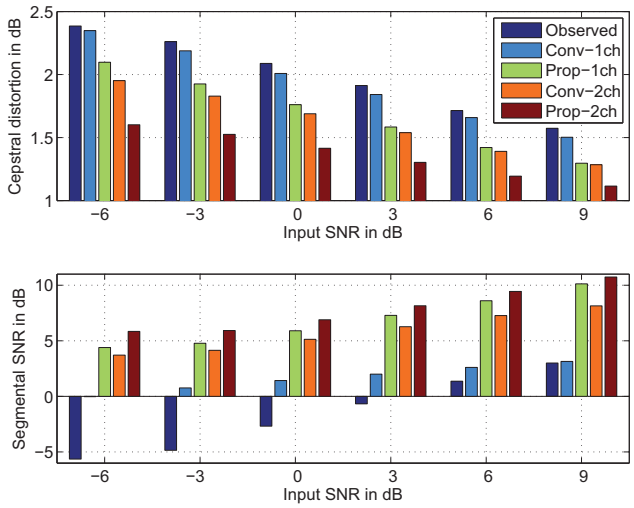

Fig. 2. Average CDs and average segmental SNRs of observed signals (development set), and those of speech signals enhanced by Conv-1ch, Prop-1ch, Conv-2ch, and Prop-2ch.

4.1. Quality of enhanced speech

Figure 2 shows the dependence of the quality of the enhanced speech on the signal-to-noise ratio (SNR) of the observation in terms of the average cepstral distortion (CD) calculated over the 1st to 12th order cepstral coefficients and the average segmental SNR. Segmental SNRs were calculated by extracting the noise remaining in the enhanced speech by subtracting the clean speech from the enhanced speech, and by calculating the power ratios of the clean speech to the extracted noise.

Under all SNR conditions, Prop-1ch and Prop-2ch substantially reduced the CDs and increased the segmental SNRs, and greatly outperformed Conv-1ch and Conv-2ch, respectively. In particular, Prop-1ch reduced the segmental SNRs much more than Conv-2ch although Prop-1ch was based only on monaural processing.

4.2. PASCAL CHiME keyword recognition task

We also evaluated the quality of the enhanced speech in terms of ASR performance. For the evaluation, we used the keyword recognition task for the evaluation set in the PASCAL CHiME challenge [5]. Acoustic models trained on clean speech (clean-condition training) and on enhanced speech (multi-condition training) were used for the recognition. Multi-condition training data were artificially created by adding noise from the training samples to the clean speech training data at different SNRs. We used speaker dependent acoustic models consisting of left-to-right HMMs trained with the SOLON recognizer [7]. For the clean acoustic model, the total number of HMM states was 254 and each state had 7 Gaussians. For the multi-condition model, each HMM state had 20 Gaussians.

Table 1 shows the keyword recognition accuracy of each method for the evaluation set. Again, both Prop-1ch and Prop-2ch were comparable to, or greatly outperformed Conv-1ch and Conv-2ch, respectively, under all SNR conditions. In addition, as discussed in [8] in details, we also expect the ASR performance to be further improved combining this approach with other speech enhancement and robust ASR techniques.

Table 1. Keyword recognition accuracy (%) for observed signals (evaluation set), and for signals enhanced by Conv-1ch, Prop-1ch, Conv-2ch, and Prop-2ch.

(a) Clean-condition training

SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	Ave
Observed	46.8	53.6	64.5	76.0	83.3	91.8	69.4
Conv-1ch	56.4	63.6	74.1	83.0	86.8	**92.7**	76.1
Prop-1ch	**64.3**	**69.7**	**78.1**	**84.2**	**88.2**	92.2	**79.4**
Conv-2ch	69.8	76.1	83.3	88.1	91.2	93.6	83.7
Prop-2ch	**77.1**	**81.5**	**87.4**	**89.9**	**92.0**	**95.1**	**87.2**

(b) Multi-condition training

SNR	-6dB	-3dB	0dB	3dB	6dB	9dB	Ave
Observed	69.9	76.3	83.6	89.2	90.7	93.3	83.8
Conv-1ch	70.4	77.1	83.8	88.6	**90.8**	**93.7**	84.1
Prop-1ch	**74.6**	**79.6**	**84.6**	**89.4**	90.1	92.6	**85.1**
Conv-2ch	81.7	84.0	89.1	91.3	92.5	93.2	88.6
Prop-2ch	**84.4**	**88.3**	**90.3**	**92.1**	**93.2**	**94.1**	**90.4**

5. SUMMARY

This paper proposed a method for incorporating MFCC-GMM spectral priors into a single/multi-channel speech enhancement approach based on a LogMax observation model. The incorporation was made possible by separating the generative process of speech/noise spectra into the generation of MFCCs from GMMs and the generation of spectral shapes from the MFCCs, and by handling the MFCCs as parameters to be estimated using MAP estimation. In the experiments, thanks to the use of the MFCC-GMMs, the proposed methods greatly improved the quality of speech under various noise conditions in terms of cepstral distortions, segmental SNRs, and keyword recognition accuracies using the PASCAL CHiME challenge database.

6. REFERENCES

[1] P.J. Moreno, B. Raj, and R.M. Stern, "A vector Taylor series approach for environment-independent speech recognition," *ICASSP-96*, vol. II, pp. 733-736, 1996.

[2] S.J. Rennie, J.R. Hershey, and P.A. Olsen, "Single-channel multitalker speech recognition," IEEE SP Magazine, pp. 66–80, Nov. 2010.

[3] T. Nakatani, S. Araki, T. Yoshioka, and M. Fujimoto, "Joint unsupervized learning of hidden Markov source models and source location models for multichannel source separation," Proc. ICASSP-2011, 2011.

[4] T. Nakatani, S. Araki, M. Delcroix, T. Yoshioka, and M. Fujimoto, "Reduction of highly nonstationary ambient noise by integrating spectral and locational characteristics of speech and noise for robust ASR," Proc. Interspeech-2011, 2011.

[5] J. Barker, H. Christensen, N. Ma, P. Green, and E. Vincent, the PASCAL CHiME speech separation challenge website. `http://www.dcs.shef.ac.uk/spandh/chime/challenge.html`

[6] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdetermined convolutive mixtures," *Proc. WASPAA-2007*, pp. 139-142, 2007.

[7] T. Hori, C. Hori, Y. Minami, and A. Nakamura, "Efficient WFST-based one-pass decoding with on-the-fly hypothesis rescoring in extremely large vocabulary continuous speech recognition," IEEE Trans. SAP, vol. 15, no. 4, pp. 1352-1365, 2007.

[8] M. Delcroix et al., "Speech recognition in the presence of highly non-stationary noise based on spatial, spectral and temporal speech/noise modeling combined with dynamic variance adaptation," Proc. CHiME workshop, 2011.