ACOUSTIC MODEL COMBINATION TO COMPENSATE FOR RESIDUAL NOISE IN MULTI-CHANNEL SOURCE SEPARATION

Jae Sam Yoon, Ji Hun Park, and Hong Kook Kim

Department of Information and Communications
Gwangju Institute of Science and Technology (GIST)
1 Oryong-dong, Buk-gu, Gwangju 500-712, Korea
{jsyoon, jh_park, hongkook}@gist.ac.kr

ABSTRACT

In this paper, we propose an acoustic model combination technique for reducing a mismatch in a multi-channel noisy environment. To this end, we first apply a mask-based multi-channel source separation method, typically computational auditory scene analysis (CASA), to separate the speech source from noise. However, a certain degree of noise remains in the separated speech source, especially under low signal-to-noise ratio (SNR) conditions since the estimated mask is not ideal. Thus, the performance of automatic speech recognition (ASR) is limited. To improve ASR performance, the remaining noise can be further compensated in the acoustic model domain under a framework of parallel model combination. In particular, a noise model for PMC is estimated from the noise remained after application of the mask-based source separation, and SNR for PMC is also estimated based on the average of relative magnitude of mask along the utterance. It is shown from the experiments that the proposed acoustic model combination method relatively reduces the word error rate by 52.14% compared to mask-based source separation alone.

Index Terms— Speech recognition, multi-channel source separation, parallel model combination, mask-based noise model estimation, mask-based SNR estimation, computational auditory scene analysis

1. INTRODUCTION

Selective hearing is a useful mechanism for extracting desired signals in noisy acoustic environments, often called the "cocktail party effect" [1]. In human auditory systems, a desired signal can be localized and separated based on the inter-aural time difference (ITD) and the inter-aural level difference (ILD), respectively. In contrast, a mask-based multi-channel source separation (MMSS) method such as computational auditory scene analysis (CASA) separates a target signal and noise using ITDs and ILDs based on a set-up of two or more microphones [2]. In MMSS, time-frequency (T-F) mask information is first estimated from ITDs and ILDs. After that, since the T-F mask information indicates the dominance of target speech at a particular T-F region, the target speech can be separated after applying the estimated T-F mask to noisy speech.

Among the research works investigating mask estimation in binaural (or two-microphone) environments, Roman *et al.* proposed a Gaussian kernel-based mask estimation method based on a supervised learning algorithm [3]. However, since residual noise signals can remain in target signals separated by the Gaussian kernel-based MMSS, the performance of automatic speech recognition (ASR) using the target signals can degrade especially under low signal-to-noise ratio (SNR) conditions. Thus, methods being capable of compensating for residual noise are required.

Acoustic model adaptation such as parallel model combination (PMC) can be a candidate approach to compensating for the remaining noise to further improve the ASR performance. PMC provides noise-corrupted models using the SNR-weighted combination of clean-trained models and a noise model obtained from a noisy input speech [4]. Thus, the acquisition of a well-estimated noise model and exact SNR is important in PMC to improve ASR performance in noisy environments.

In this paper, we propose a noise model estimation method and an SNR estimation method for PMC from the mask information obtained from MSS. Specifically, in the proposed mask-based noise model estimation (MNME) method, a noise mask, which is defined by a mask subtracted from 1, is applied to noisy speech, generating an estimated noise signal. After that, a noise model for PMC is obtained from the estimated noise signal. In the proposed mask-based SNR estimation (MSE) method, the ratio between the average values of a mask and a noise mask is calculated to estimate SNR for PMC.

The remainder of this paper is organized as follows. MMSS is briefly reviewed in Section 2, and an overview of the proposed acoustic model combination method including MNME and MSE is presented in Section 3. In Section 4, speech recognition experiments are performed. Finally, we conclude the paper in Section 5.

2. MASK-BASED MULTI-CHANNEL SOURCE SEPARATION

In this section, we summarize a mask-based multi-channel source separation (MMSS) method as a means of providing a further understanding of the motivation for the proposed approach. In the MMSS method utilized in this paper [2], separation of the target signal is performed on binaural input signals. To achieve this separation, the binaural signals are first decomposed into auditory spectral signals by employing a gammatone filterbank. The envelopes of the left and right auditory spectral signals are then calculated at each frame and frequency channel.

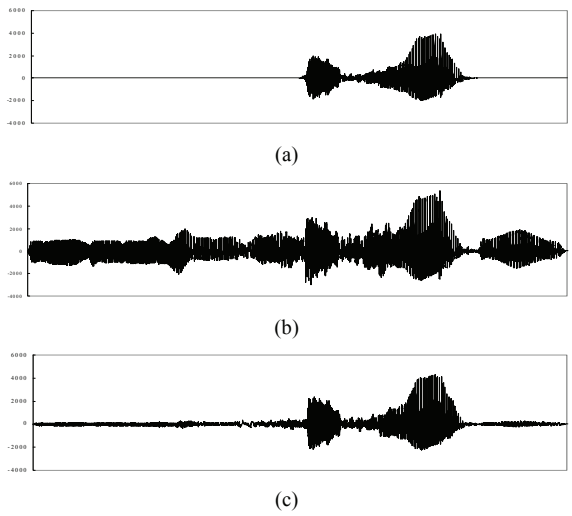

Figure 1: Comparison of waveforms of (a) original clean speech, (b) noisy speech at 0 dB SNR, and (c) speech signal separated by using a Gaussian kernel-based mask

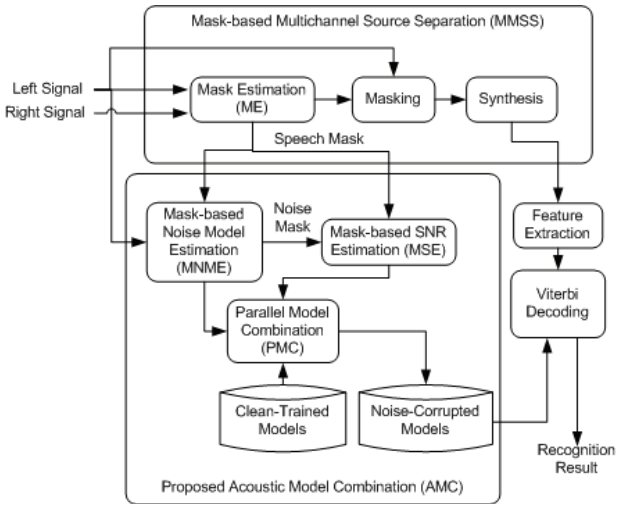

Figure 2: Procedure of the proposed acoustic model combination method in a multi-channel environment.

2.1. Binaural cue extraction

To estimate mask patterns, a pair of ITD and ILD for each frame and frequency channel must be extracted. For a given frame and frequency channel, the normalized cross-correlation between the auditory spectral signals from the left and right channels is first calculated. Then, ITD is estimated as the time lag where the normalized cross-correlation is maximized. Next, ILD is computed as the ratio of auditory envelopes obtained from the left and the right channel signals.

2.2. Masking for source separation

The mask information is extracted from the estimated ITDs and ILDs, and used to separate the desired speech from noisy speech. In this paper, we use the Gaussian kernel-based mask estimation method [3], where the mask models are trained by employing a Gaussian kernel density estimator for a given frame and frequency channel. Each trained model provides a mask value obtained from the ratio of speech probability and noise probability in the two-dimensional ITD-ILD plane.

As a reference, an ideal mask is also obtained as the ratio of the envelopes of the clean target signal and the noise signal, assuming that the background noise added to clean speech is completely known [5].

Finally, the auditory spectral signals of a target source are estimated by multiplying either a Gaussian kernel-based mask or an ideal mask to the input auditory spectral signals.

2.3. Generation of estimated clean speech

The estimated clean speech is obtained by inversely filtering the auditory spectral signals by the gammatone filterbank [6]. Fig. 1 illustrates the waveforms of original clean speech, noisy speech, and separated clean speech after applying the Gaussian kernel-based mask. Note here that when compared to the original clean speech signal, the estimated clean speech signal (Fig. 1(c)) is still noisy even its SNR is higher than the noisy speech signal shown in Fig. 1(b). If the estimated clean speech is directly used for ASR, the ASR performance might be degraded due to the remaining residual noise in the estimated clean speech signal. Thus, in the next section we propose a method that compensates for such residual noise in the acoustic model domain.

3. PROPOSED ACOUSTIC MODEL COMBINATION

Acoustic model adaptation such as parallel model combination (PMC) can be used to compensate for the remaining noise to further improve the ASR performance. In this section, we explain the proposed acoustic model combination method incorporated with multi-channel source separation, where a noise model and SNR are estimated on the basis of a mask obtained from MMSS described in Section 2.

3.1. Overview of the proposed method

Fig. 2 shows the procedure of the proposed acoustic model combination method. First, the process of MMSS described in Section 2 is performed. In other words, mask information is estimated from binaural signals, and then the estimate of clean speech is then obtained. From now on, we refer to such mask as speech mask because it contributes the estimation of clean speech. Next, a 39-dimensional feature vector is extracted from the target source signal, where 12 mel-frequency cepstral coefficients (MFCCs) and a log energy are concatenated with their deltas and delta-deltas. In the proposed method, a noise model is estimated using a noise mask which is defined by a speech mask subtracted from 1; the SNR value is estimated from the ratio between the average values of a speech mask and a noise mask over each utterance. After that, the noise-corrupted models are obtained by combining the clean-trained models and the estimated noise model with the estimated SNR value. Finally, the Viterbi decoding of the estimated target source signal is done using the noise-corrupted models.

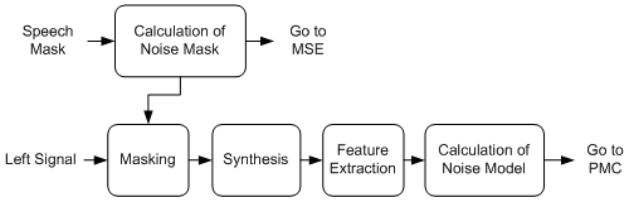

Figure 3: Procedure of the proposed noise model estimation method.

3.2. Parallel model combination

In this paper, we adopt a log-normal PMC [4]. To this end, the clean-trained models and the estimated noise model in the cepstral domain are first transformed into the linear spectral domain by taking the logarithm and an inverse discrete cosine transform (DCT). Next, the noise-corrupted models in the linear spectral domain are calculated by adding the clean-trained models and an SNR-weighted noise model.

$$\begin{aligned} \hat{\mu} &= \mu + g \cdot \tilde{\mu} \\ \hat{\Sigma} &= \Sigma + g^2 \cdot \tilde{\Sigma} \end{aligned} \tag{1}$$

where μ and Σ are the mean and variance of a clean-trained model, respectively; similarly, $\tilde{\mu}$ and $\tilde{\Sigma}$ are the mean and variance of the noise model, $\hat{\mu}$ and $\hat{\Sigma}$ are the mean and variance of a noise-corrupted model, and g is an estimated SNR value. Finally, the noise-corrupted models in the cepstral domain are obtained by taking the exponential followed by a DCT.

As we can see in Eq. (1), a well-estimated noise model and exact SNR calculation are required for PMC. Thus, we propose a noise estimation method and an SNR estimation method for PMC in the next subsections.

3.3. Mask-based noise model estimation

Fig. 3 shows the procedure of the proposed mask-based noise model estimation (MNME) method. Here, a noise mask, $m_N(i,j)$, is defined as

$$m_N(i,j) = 1 - m_S(i,j) \tag{2}$$

where i and j are the indices of the frequency channel and frame, respectively. Also, $m_S(i,j)$ is a speech mask obtained from the mask-estimation for estimating target speech. Similar to generating target speech, the noise signal is synthesized using the noise masks, and the MFCCs are then calculated from the synthesized noise signal. Finally, the noise model is estimated by calculating the means and variances of the MFCCs.

3.4. Mask-based SNR estimation for PMC

The SNR value, g, in Eq. (1) is estimated using noise masks and speech masks. Fig. 4 shows the procedure of the proposed SNR estimation method. As a first step for estimating the SNR value, we first find the non-target speech frames using speech masks. Then, the SNR value is calculated as the ratio of the average values of the noise masks and speech masks over each utterance.

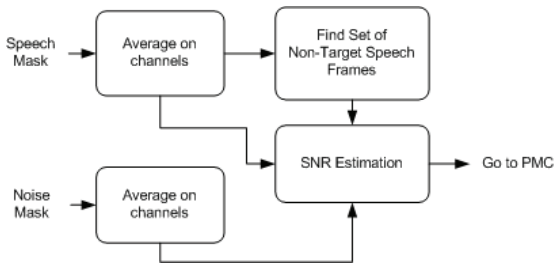

Figure 4: Procedure of the proposed SNR estimation method.

3.4.1. Detection of non-target speech frames

In order to detect non-target speech frame, we first calculate the speech mask averaged over all the frequency channels, $\overline{m}_S(j)$, as

$$\overline{m}_S(j) = \frac{1}{I}\sum_{i=1}^{I} m_S(i,j) \tag{3}$$

where I is the number of filterbank channels. Next, in order to estimate a threshold η for the detection of non-target speech frames, we estimate the mean, μ_m, and variance, σ_m^2, from the initial M frames that are assumed to be non-target speech. In other words,

$$\mu_m = \frac{1}{M}\sum_{n=1}^{M}\overline{m}_S(j) \quad \text{and} \quad \sigma_m^2 = \frac{1}{M}\sum_{n=1}^{M}\left(\overline{m}_S(j) - \mu_m\right)^2 \tag{4}$$

where the first M frames are assumed to be non-target speech. In this paper, M is set to 20. Finally, a set of the non-target speech frames, $\mathbf{S}$, can be obtained as

$$\mathbf{S} = \left\{ j \mid \overline{m}_S(j) \le \eta \right\} \tag{5}$$

where $\eta = \mu_m - \kappa\sigma_m$. Here κ is set to a value such that around 90% of the initial M frames are included in $\mathbf{S}$.

3.4.2. SNR estimation for PMC

To estimate the SNR parameter for PMC, we calculate the average values of the noise masks and speech masks on the set of non-target speech frames, denoted by g_S and g_N, respectively, in the following equation.

$$\begin{aligned} g_S &= \frac{1}{\#|\mathbf{S}|}\sum_{j\in\mathbf{S}}\overline{m}_S(j) \\ g_N &= \frac{1}{\#|\mathbf{S}|}\sum_{j\in\mathbf{S}}\overline{m}_N(j) \end{aligned} \tag{6}$$

where $\#|\mathbf{S}|$ is the number of frames belonging to the set $\mathbf{S}$, and $\overline{m}_N(j) = \frac{1}{I}\sum_{i=1}^{I} m_N(i,j)$ is the noise mask averaged over all the frequency channels. After that, the SNR parameter is estimated as

$$g = \frac{g_S}{g_N} \quad . \tag{7}$$

Finally, we apply g to Eq. (1) to obtain the means and variances of noise-corrupted acoustic models.

Table 1. Word error rates (%) of the baseline system, the mask-based multi-channel source separation (MMSS) for an ideal mask, a Gaussian kernel-based mask, and the proposed model combination method (MMSS+PMC).

	Baseline			MMSS						MMSS+PMC		
Mask Type	-			Ideal mask			Gaussian kernel-based mask			Gaussian kernel-based mask		
Angle \ SNR(dB)	0	10	20	0	10	20	0	10	20	0	10	20
10°	98.77	62.28	13.86	4.56	3.68	4.21	38.60	9.47	5.44	14.56	6.84	5.26
20°	97.88	56.14	11.75	4.39	3.86	4.39	42.11	9.12	5.09	14.21	6.67	5.09
40°	94.82	43.16	10.70	4.39	4.04	4.91	35.61	9.47	5.26	11.40	7.19	5.44
Average	97.16	53.86	12.10	4.45	3.86	4.50	38.77	9.35	5.26	13.39	6.90	5.26
	54.37			4.27			17.80			8.52		

4. SPEECH RECOGNITION EXPERIMENTS

For the speech recognition experiments, a binaural database was artificially constructed using an HRTF function [7] in conjunction with a Korean speech corpus [8]; 18,240 utterances of the corpus were used to train the acoustic model, and 570 utterances were used as the target speech data. Each target speech utterance was transformed into a binaural signal and mixed with female speech localized at 10°, 20°, and 40° from the target signal.

The acoustic models were based on left-to-right triphone HMMs, and trained using the HTK version 3.2 Toolkit [9]. The number of Gaussian mixtures was 4 per state and all the triphone models were expanded from 42 monophones that included a silence and a short pause model. The triphone model states were tied by employing a decision tree. As a result, we had 7,577 triphones and 2,487 states. For a language model, the lexicon size was 2,250 words and a finite state network grammar was employed.

Table 1 shows the word error rates (WERs) of the baseline system, MMSS based on the ideal mask, the Gaussian kernel-based mask, and the proposed model combination method (MMSS+PMC). In the baseline, noisy speech from the left channel was directly used for ASR. As shown in the table, the baseline system gave the highest WER because no noise compensation method was used. In contrast, MMSS with the ideal mask had the smallest WER since the ideal mask could result in the best performance for signal separation. However, the WER of MMSS with the Gaussian kernel-based mask was considerably increased at low SNRs compared with MMSS with the ideal mask. On the other hand, the WER of the proposed model combination method was significantly reduced at low SNRs, compared with MMSS with the Gaussian kernel-based mask. As a result, it can be seen that the proposed method relatively decreased the WER by 52.14% compared with the MMSS with the Gaussian kernel-based mask.

5. CONCLUSION

Although multi-channel source separation approaches are generally acceptable in noisy environments, the performance improvement is restrictive at low SNRs due to residual noise signals contaminated in the estimated target signals. To reduce the effects of residual noise in the acoustic model domain, a parallel model combination technique was proposed here. In order to realize PMC, we also proposed a mask-based noise model estimation and a mask-based SNR estimation method. It was shown from speech recognition experiments that the proposed method achieved the relative WER reduction of 52.14% compared to a mask-based source separation using a Gaussian kernel-based mask.

6. ACKNOWLEDGEMENT

This work was supported in part by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2007-314-D00245), in part by the Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program supervised by the Institute for Information Technology Advancement (IITA) (IITA-2008-C1090-0804-0007), and in part by the basic research project through a grant provided by the Gwangju Institute of Science and Technology in 2009.

7. REFERENCES

[1] B. Arons, *A Review of the Cocktail Party Effect*, MIT Media Lab., Dec. 2006.

[2] D. L. Wang and G. J. Brown, *Computational Auditory Scene Analysis: Principle, Algorithms and Applications*, IEEE Press, Wiley-Interscience, 2006.

[3] N. Roman, D. L. Wang, and G. J. Brown, "Speech segregation based on sound localization," *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, July 2003.

[4] M. Gales and S. J. Young, "Robust continuous speech recognition using parallel model combination," *IEEE Trans. Speech and Audio Proc.*, vol. 4, no. 5, pp. 352–359, Sept. 1996.

[5] K. J. Palomaki, G. J. Brown, and D. L. Wang, "A binaural processor for missing data speech recognition in the presence of noise and small-room reverberation," *Speech Communication*, vol. 43, no. 8, pp. 361–378, Mar. 2004.

[6] M. Weintraub, *A theory and computational model of monaural auditory sound separation*, Ph.D. Thesis, Stanford University, 1985.

[7] W. G. Gardner and K. D. Martin, "HRTF measurements of a KEMAR," *J. Acoust. Soc. Amer.*, vol. 97, no. 6, pp. 3907–3908, June 1995.

[8] S. Kim, S. Oh, H.-Y. Jung, H.-B. Jeong, and J.-S. Kim, "Common speech database collection," in *Proc. Acoust. Soc. Korea*, vol. 21, no. 1, pp. 21–24, July 2002.

[9] S. Young, *et al*., *The HTK Book (for HTK Version 3.2)*, Microsoft Corporation, Cambridge University Engineering Department, 2002.