



SNR-Based Mask Compensation for Computational Auditory Scene Analysis Applied to Speech Recognition in a Car Environment

Ji Hun Park¹, Seon Man Kim¹, Jae Sam Yoon¹, Hong Kook Kim¹, Sung Joo Lee², and Yunkeun Lee²

¹School of Information and Communications

Gwangju Institute of Science and Technology, Gwangju 500-712, Korea

²Speech Processing Team, Speech and Language Information Research Division
Electronics and Telecommunications Research Institute, Daejeon 305-350, Korea

¹{jh_park, kobem30002, jsyoon, hongkook}@gist.ac.kr ²{lee1862, yklee}@etri.re.kr

Abstract

In this paper, we propose a computational auditory scene analysis (CASA)-based front-end for two-microphone speech recognition in a car environment. One of the important issues associated with CASA is the accurate estimation of mask information for target speech separation within multiple microphone noisy speech. For such a task, the time-frequency mask information is compensated through the signal-to-noise ratio resulted from a beamformer to adjust the noise quantity included in noisy speech. We evaluate the performance of an automatic speech recognition (ASR) system employing a CASA-based front-end with the proposed mask compensation method. In addition, we compare its performance with those employing a CASA-based front-end without mask compensation and the beamforming-based front-end. As a result, the CASA-based front-end achieves an average word error rate (WER) reduction of 8.57% when the proposed mask compensation method is applied. In addition, the CASA-based front-end with the proposed method provides a relative WER reduction of 26.52%, compared with the beamforming-based front-end.

Index Terms: Speech recognition, speech separation, computational auditory scene analysis, mask compensation, beamforming

1. Introduction

As supplementary operations associated with driving and vehicle maintenance continue to be recognized as powerful and valuable tools in assuring drivers' convenience and safety, telematics represents a key process within the automobile industry. In fact, speech recognition already has been put into operation with efficient and intelligent interfaces to operate automotive navigation systems more safely [1][2]. However, there are some problems with the degradation of speech recognition performance due to car noise, e.g., the din of traffic and music played from the automobile stereo system. Through separation of speech-of-interest from noisy speech by using two microphones, computational auditory scene analysis (CASA) [3] has become a widely used method into resolving such issues. In most CASA approaches, mask information indicates whether a particular frequency region for a given time frame includes dominantly target speech or dominantly noise, thus it is used to retrieve target speech. Accordingly, an accurate estimation of the mask information is required to obtain high-quality separation performance, thus it is critical to improve the performance of automatic speech recognition (ASR) system based on CASA.

A number of research works have been reported in estimating mask information in binaural environments [4][5]. Roman *et al.* have proposed a mask estimation method based

on a supervised learning algorithm in [4]. They trained Gaussian kernel-based mask model to obtain a mask pattern for each time-frequency (T-F) bin under conditions assumed to be independent over each analysis frame. On the other hands, we have proposed a hidden Markov model (HMM)-based mask estimation method to take the continuity of speech along the analysis frames into account [5]. The proposed mask estimation method tried to estimate the mask value for a given T-F bin by utilizing estimated likelihoods from the HMMs, representing statistical patterns of speech and noise under consecutive analysis frames. As a result, relatively highly accurate mask estimation was possible. However, all the above-mentioned methods utilize the interaural time differences (ITDs) and interaural level differences (ILDs) to estimate mask information. ITDs and ILDs represent time and level difference between left and right input channels, not being capable of reflecting actual signal-to-noise ratio (SNR). In other words, a mask pattern, as opposed to the spatial configuration, does not depend on the SNR. Thus, the mask compensation process is required to estimate suitable mask information according to SNR [6].

In this paper, we propose an SNR-based mask compensation method for a CASA-based front-end in a two-microphone environment. Towards the end, a beamforming technique is employed to estimate segmental SNR for each analysis frame. In particular, a beamformer is adapted to noise conditions according to the direction-of-arrival (DOA) information for each analysis frame obtained from the CASA analysis. By using the estimated segmental SNRs, the mask value is compensated and applied to noisy speech.

Following this introduction, Section 2 describes an overview of the CASA-based front-end combined with mask compensation. In Section 3, we propose a mask compensation method as a means of incorporating the SNRs estimated by beamforming. Next, the performance of an ASR system employing the proposed method is evaluated in Section 4. Finally, we conclude our findings in Section 5.

2. CASA-based front-end

In this section, a block diagram of a CASA-based front-end, combined with the proposed SNR-based mask compensation method, is described. As shown in Fig. 1, the front-end first extracts the auditory spectral signals from the binaural input noisy speech. After that, the mask information is estimated using ITDs and ILDs extracted from the binaural auditory spectral signals. Next, the estimated mask information is compensated based on the segmental SNRs by employing a generalized sidelobe canceller (GSC) beamformer. Finally, the target speech is separated from noisy speech by applying the compensated mask information.

10.21437/Interspeech.2010-270

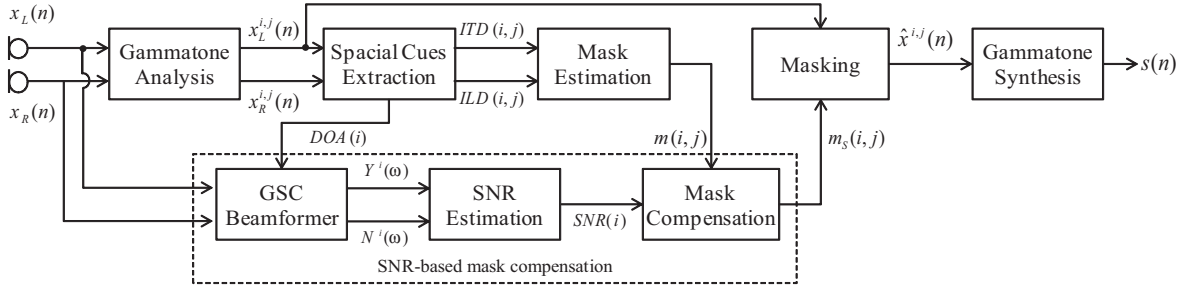


Figure 1: Block diagram of a CASA-based front-end combined with SNR-based mask compensation.

2.1. Auditory periphery

Binaural input signals with a sampling rate of 16 kHz are decomposed into auditory spectral signals by employing a gammatone filterbank [7] with 32 frequency bands, in which center frequencies are linearly spaced on an equivalent rectangular bandwidth (ERB)-scale [8] from 50 Hz to 8 kHz. Auditory spectral signals are then windowed using a rectangular window with a time resolution of 20 ms and a frame rate of 100 Hz, resulting in the left and right auditory spectral signals for the i -th frame and the j -th frequency band, $x_L^{i,j}(n)$ and $x_R^{i,j}(n)$, respectively.

2.2. Spatial cues extraction

In order to estimate mask patterns, an ITD and ILD are extracted for each T-F bin. First of all, normalized cross-correlation (CC) is initially computed between left and right auditory spectral signals, defined as

$$CC^{i,j}(\tau) = \frac{\sum_{n=0}^{N-1} x_L^{i,j}(n)x_R^{i,j}(n-\tau)}{\sqrt{\sum_{n=0}^{N-1} (x_L^{i,j}(n))^2} \sqrt{\sum_{n=0}^{N-1} (x_R^{i,j}(n))^2}} \quad (1)$$

where τ ranges from -16 to 16, corresponding to a range from -1 ms to 1 ms at a sampling rate of 16 kHz. In addition, N represents the number of speech samples per frame, being set at 320 in this paper. Next, an ITD for the (i,j) -th T-F bin is estimated as a time lag where the normalized CC is maximized. In other words, we have

$$ITD(i,j) = \arg \max_{\tau} CC^{i,j}(\tau). \quad (2)$$

In addition to the ITD extraction, the ILD for the (i,j) -th T-F bin is computed as the ratio of energies obtained from the left and the right auditory spectral signals using the following equation of

$$ILD(i,j) = 10 \log_{10} \left(\frac{\sum_{n=0}^{N-1} (x_L^{i,j}(n))^2}{\sum_{n=0}^{N-1} (x_R^{i,j}(n))^2} \right). \quad (3)$$

In order to estimate the DOA for the i -th frame, the pooled CC (PCC) is computed by summing CCs in Eq. (1) over all frequency bands and some adjacent time frames. After that, the DOA is estimated as a time lag where the PCC is maximized, such as

$$DOA(i) = \arg \max_{\tau} \sum_{i=0}^{N_i} \sum_{j=0}^{N_j} CC^{i,j}(\tau) \quad (4)$$

where N_i is the number of frames for DOA estimation and N_j represents the total number of gammatone filters. In this work, N_i and N_j are set to 5 and 32, respectively. The GSC

beamformer is adapted using the DOA information, as shown in Fig. 1.

2.3. Mask estimation and compensation

The mask information for each T-F bin is estimated using ITD and ILD. In this paper, a model-based approach is used for estimating mask, where the mask model is trained by employing the Gaussian kernel density estimator [9]. Initially, ITDs and ILDs are extracted from the training data. And then, a two-dimensional ITD-ILD plane is constructed for each frequency band, where axes of the ITD-ILD plane being linearly quantized via a minimum to maximum ITD and ILD value. Here, each axis is quantized with a step of 100. By using the Gaussian kernel density estimator, target and noise probabilities are then estimated for each region in the ITD-ILD plane. Finally, a mask model is composed of 10,000 ratios between speech and noise probabilities that are computed for each quantized ITD-ILD region.

The trained mask model provides a mask value for each T-F bin according to the estimated ITD and ILD value from Eqs. (2) and (3), respectively. In other words, the mask value, $m(i,j)$ for the (i,j) -th T-F bin could be estimated by searching ITD-ILD region of the mask model which corresponds to the ITD and ILD of the (i,j) -th T-F bin. The estimated mask value is then compensated using the segmental SNR estimated from the GSC beamformer, which is described further in Section 3.

2.4. Masking and synthesis

The auditory spectral signals of target speech are extracted by multiplying masks to auditory spectral signals obtained from the left input signal. The estimated target speech is then obtained by inversely filtering auditory spectral signals via the gammatone filterbank [10]. Finally, this synthesized target speech is utilized for ASR.

3. SNR-based mask compensation

In order to compensate the mask information according to the SNR, the segmental SNR is estimated for each time frame during the GSC beamforming process. For a given SNR, the mask information is compensated and then applied to noisy speech.

3.1. Beamforming-based SNR estimation

Fig. 2 shows how to estimate the segmental SNR by using the GSC beamformer. As shown in the figure, the binaural input signals, with a sampling rate of 16 kHz, are first windowed by a Hamming window whose length is 20 ms. Next, left and right spectral signals for the i -th analysis frame, $X_L^i(\omega)$ and $X_R^i(\omega)$, are obtained by applying a short-time Fourier transform (STFT) to two input noisy signals, $x_L(n)$ and $x_R(n)$,

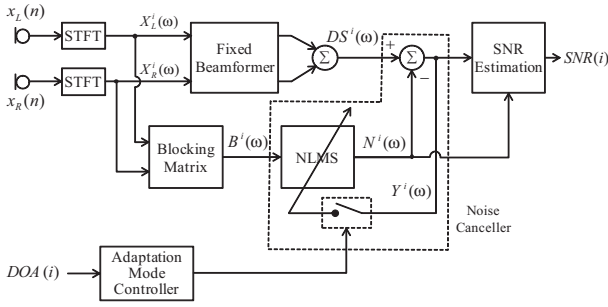


Figure 2: SNR estimation using the GSC beamformer.

respectively. After that, $X_L^i(\omega)$ and $X_R^i(\omega)$ are applied to both a fixed beamformer and a blocking matrix, representing the main components of the GSC beamformer. The target-directed signals are primarily secured by the fixed beamformer, while others are captured by the blocking matrix [11]. In this paper, a delay-and-sum beamformer is used for the fixed beamformer.

The output of the blocking matrix, $B^i(\omega)$, is considered to be a non-target signal. However, such a signal still contains some degree of the target signals, denoting diffused target speech components. Thus, adaptive filtering should be performed to reduce the target signal from $B^i(\omega)$. In other words, $B^i(\omega)$ is modified into $N^i(\omega)$ by using a frequency-domain adaptive filter which is generally implemented by a normalized least mean square (NLMS) algorithm [12]. In particular, in order to better reflect noise conditions and to prevent target-directed signals from being distorted, an adaptation mode controller (AMC) is employed. The AMC determines NLMS filter adaptation according to noise dominance of a given analysis frame, where noise dominance is determined by DOA information obtained from Eq. (4). In other words, if the DOA for the i -th frame, $DOA(i)$, is not equal to the target direction, then the frame is decided by a noise dominant frame. Then, the NLMS adaptive filter is updated to match given noise conditions. After that, the modified non-target signal, $N^i(\omega)$, is subtracted from the output of the fixed beamformer, $DS^i(\omega)$, resulting in the target signal, $Y^i(\omega)$. Finally, the segmental SNR for the i -th time frame is estimated accordingly,

$$SNR(i) = 10 \log_{10} \left(\frac{\sum_{\omega=0}^{\pi} (Y^i(\omega))^2}{\sum_{\omega=0}^{\pi} (N^i(\omega))^2} \right). \quad (5)$$

3.2. SNR-based mask compensation

By using the estimated SNRs in Eq. (5), the mask value for the (i, j) -th T-F bin, $m(i, j)$, is then compensated as

$$m_c(i, j) = m(i, j) + \delta(i) \quad (6)$$

where $m_c(i, j)$ represents the modified mask value for the (i, j) -th T-F bin. In addition, $\delta(i)$ denotes a mask compensation factor for the i -th frame and is obtained as follows:

$$\delta(i) = \frac{1}{1 + \exp(-a(SNR(i) - b))} - 0.5 \quad (7)$$

where a and b indicate the gradient and the displacement of the sigmoid function, respectively. In this paper, a and b are separately set to 0.1 and 10 such that the estimated mask pattern, $m(i, j)$, is optimized for 10 dB SNR.

The compensated mask, $m_c(i, j)$, is then smoothed by a moving average method to better refine the mask value. In other words,

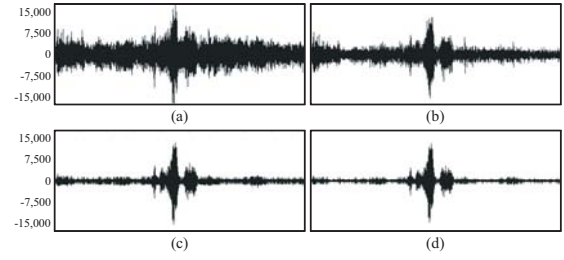


Figure 3: Comparison of waveforms of (a) input noisy speech obtained from the left channel at 0 dB SNR under a music noise condition and output speech signals obtained by (b) a GSC beamformer, (c) a conventional CASA-based front-end, and (d) a CASA-based front-end with the proposed mask compensation.

$$m_s(i, j) = \frac{1}{(2k+1)(2l+1)} \sum_{-k}^k \sum_{-l}^l m_c(i+k, j+l) \quad (8)$$

where $m_s(i, j)$ represents a smoothed mask value taken with respect to both the frequency and time axis. In addition, k and l are adjusting factors of the smoothing window in the time and frequency axes, respectively. All the factors are set as 1 in this paper.

Fig. 3 compares waveforms, applied to input signals (Fig. 3(a)), obtained by means of three different front-ends. As shown in the figure, the output speech signals processed by a conventional CASA-based front-end and the GSC beamformer reveal smaller noise components than input noisy speech signals. Fig. 3(d) also shows that noise components in output speech signals obtained by the CASA-based front-end coupled with the proposed mask compensation method are further reduced, resulting in more refined target speech.

4. Speech recognition experiments

In this section, ASR performance of a CASA-based front-end employing the proposed mask compensation method is evaluated by comparing performance with a conventional CASA-based front-end without the mask compensation method as well as the GSC beamforming-based front-end.

4.1. Speech database

As a training database for ASR, phonetically optimized words of 180,000 utterances recorded from about 1,500 speakers were used. Each utterance in the database was digitalized in 16-bit PCM at a sampling rate of 16 kHz and spoken in a quiet office environment. In addition, an ASR system was evaluated by using a test dataset composed of two sets of 452 Korean phonetically balanced words (PBWs), which corresponds to 904 words in total. Each PBW was spoken by one of 30 males and 30, being recorded in a clean environment.

In order to simulate a car noise environment, the configuration shown in Fig. 4 was constructed. As shown in the figure, each PBW was considered as a target speech signal and played from an acoustic speaker mounted below the headrest of the driver's seat. As background noise, the car audio system was turned on simultaneously. Thus, the target speech from a speaker as well as music noise from the automobile stereo system was recorded by a two-microphone array, with a space of 4 cm between microphones, located on the dashboard. To investigate the effect of the proposed method on different amount of background music, the car audio volume was increased such that SNRs were approximately measured from -5 dB to 20 dB at a step of 5 dB.

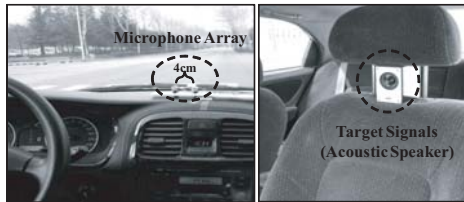


Figure 4: Environments for constructing the noisy speech database for speech recognition.

4.2. ASR system

As a recognition feature, 13 mel–frequency cepstral coefficients (MFCCs) were extracted for every 10 ms analysis frame. The 13 MFCCs were finally concatenated with their first and second derivatives, resulting in a 39–dimensional feature vector. During the training and testing, cepstral mean normalization was applied to feature vectors. In addition, an end–point detector, specialized for a car environment, was applied as shown in [2].

Acoustic models were 3,474 tied triphones represented by 3 states left–to–right HMM with a mixture of 16 Gaussians using diagonal covariance matrices. In order to match conditions between test and training, acoustic models were adapted. To this end, 8,000 utterances were collected from 80 speakers in moving cars (60 ~100 km/h) and the trained acoustic models were adjusted by the discriminative acoustic model adaptation method [13]. For the language model, the lexicon size was 452 words and a finite state network grammar was employed.

4.3. Experimental results

Table 1 compares average word error rate (WER) between the baseline system, the GSC beamforming–based front–end (BF), a conventional CASA–based front–end without mask compensation (CASA), and a CASA–based front–end employing the proposed SNR–based mask compensation (CASA–MC). Note that noisy speech from left channel was used for evaluating the performance of the baseline ASR system. On the other hand, the target speech $y^i(n)$, obtained by inverse STFT of $Y^i(\omega)$ shown in Fig. 2, was applied to the performance evaluation of an ASR system using BF. Note here that the output signals obtained by all the above–mentioned approaches were end–pointed prior to the Viterbi decoding.

As shown in the table, the baseline system gave the highest WER due to the absence of a speech separation process. Conversely, CASA–MC, representing the CASA front–end with the proposed mask compensation method, provided the smallest WER. In particular, the WERs of the CASA–MC were significantly reduced at low SNRs, compared with CASA or BF. Consequently, the CASA–MC provided relative WER reductions of 8.57% and 26.52% compared to the CASA and the BF, respectively.

5. Conclusion

In this paper, we proposed an SNR–based mask compensation method for a CASA–based front–end. With the help of a beamformer, an SNR value was able to be estimated for given two–microphone speech signals. Then, the mask information, originally estimated from CASA, was compensated on the basis of the estimated SNR. We performed ASR experiments under the simulated car noisy conditions, and compared the performance of an ASR system employing a CASA–based front–end with the proposed mask compensation method. As a

Table 1: Comparison of the average WERs (%) of the baseline system, beamforming–based front–end (BF), conventional CASA–based front–end (CASA), and the CASA–based front–end with the proposed mask compensation (CASA–MC).

SNR(dB)	Baseline	BF	CASA	CASA–MC
20	8.74	7.19	8.74	8.52
15	12.61	8.19	9.18	9.51
10	23.56	10.84	12.28	11.17
5	45.91	21.57	18.03	15.15
0	79.09	44.80	30.64	26.99
–5	94.58	80.97	60.62	56.19
Avg.	44.08	28.93	23.25	21.26

result, the CASA–based front–end with the proposed mask compensation method achieved relative WER reductions of 26.52% and 8.57%, compared that the GSC beamformer and a CASA–based front–end alone, respectively.

6. Acknowledgements

This work was supported in part by the basic research project through a grant provided by the GIST in 2010, and by the Industrial Strategic technology development program, 10035252, Development of dialog–based spontaneous speech interface technology on mobile platform funded by the Ministry of Knowledge Economy (MKE, Korea). We would also like to thank Professor Minsoo Hahn from Korea Advanced Institute of Science and Technology (KAIST) for providing the test database.

7. References

- [1] McCallum, M. C., Campbell, J. L., Richman, J. B., Brown, J. L., and Wiese, E., “Speech recognition and in–vehicle telematics devices: potential reductions in driver distraction,” *International Journal of Speech Technology*, vol. 7, no. 1, pp. 25–33, Jan. 2004.
- [2] Lee, S. J., Chung, H., Park, J. G., Jung, H.-Y., and Lee, Y., “A commercial car navigation system using Korean large vocabulary automatic speech recognizer,” in *Proc. Asia–Pacific Signal and Information Processing Association*, pp. 286–289, Oct. 2009.
- [3] Wang, D. L. and Brown, G. J., *Computational Auditory Scene Analysis: Principle, Algorithms and Applications*, IEEE Press, Wiley–Interscience, Sept. 2006.
- [4] Roman, N., Wang, D. L., and Brown, G. J., “Speech segregation based on sound localization,” *J. Acoust. Soc. Amer.*, vol. 114, no. 4, pp. 2236–2252, July 2003.
- [5] Park, J. H., Yoon, J. S., and Kim, H. K., “HMM–based mask estimation for a speech recognition front–end using computational auditory scene analysis,” *IEICE Trans. Inf. and Syst.*, vol. E91–D, no. 9, pp.2360–2364, Sept. 2008.
- [6] Jeong, S.-Y., Jeong, J.-H., and Oh, K.-C., “Dominant speech enhancement based on SNR–adaptive soft mask filtering,” in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1317–1320, Apr. 2009.
- [7] Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., and Rice, P., *An efficient auditory filterbank based on the gammatone function*, APU Report 2341, MRC, Applied Psychology Unit, Cambridge, U.K., 1988
- [8] Glasberg, B. R. and Moore, B. C. J., “Derivation of auditory filter shapes from notched–noise data,” *Hearing Research*, vol. 47, no. 1–2, pp. 103–138, Aug. 1990.
- [9] Parzen, E., “On Estimation of a Probability Density Function and Mode,” *The Annals of Mathematical Statistics*, vol. 33, no. 3. pp. 1065–1076. Aug. 1962.
- [10] Weintraub, M., *A theory and computational model of monaural auditory sound separation*, Ph. D. Thesis, Stanford University, 1985.
- [11] Brandstein, M. and Ward, D. *Microphone Arrays*, Springer, Heidelberg, New York, 2001.
- [12] Griffiths, L. J. and Jim, C. W., “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, Jan. 1982.
- [13] Kang, B. O., Jung, H.-Y., and Lee, Y., “Discriminative noise adaptive training approach for an environment migration,” in *Proc. Interspeech*, pp. 2085–2089, Aug. 2007.