DOMINANT SPEECH ENHANCEMENT BASED ON SNR-ADAPTIVE SOFT MASK FILTERING

So-Young Jeong, Jae-Hoon Jeong, Kwang-Cheol Oh

Corporate Technology Operations SAIT, Samsung Electronics Co., Ltd.
San #14-1, Nongseo-Dong, Giheung-gu, Yongin-Si, Gyeonggi-Do 446-712, Korea
{s.y.jeong, jaehoon8.jeong, okcheol}@samsung.com

ABSTRACT

In this paper, we present a SNR-adaptive soft mask filter for multi-channel noisy speech enhancement. Incorporating frame-by-frame spectral magnitude ratios into the time-frequency(T-F) mask filter framework, the adaptive filter can be designed robust to changing environments. Experimental results show that the proposed adaptive mask filter can effectively suppress non-stationary noise components even in a closely-spaced microphone pair. Moreover, the soft mask compressed with sigmoidal nonlinearity can reduce musical noises so that improved PESQ values are obtained.

Index Terms— time-frequency masking, adaptive soft mask, local SNR, speech enhancement

1. INTRODUCTION

Recently, the sparseness of sound mixture received much attention as *a priori* knowledge to separate sources where sufficient information about mixtures is not given[1, 3]. In particular, binary time-frequency(T-F) masking is a well-known technique to achieve the goal of computational auditory scene analysis(CASA) and has widespread applications such as speech separation, enhancement and automatic speech recognition[5, 6]. Moreover, in the missing data technique, also known as emerging issues in the fields of robust speech recognition, a mask filter gives reliability of each T-F elements so that only reliable region can contribute to the estimation of recognition score[4, 7]. The binary T-F masking can perfectly extract target sources from noisy under-determined mixtures if sources do not overlap in the time-frequency domain. This assumption works well in practical sound mixtures[1].

One of the key problems in T-F masking is to construct a reliable T-F mask from noisy mixtures. The binary mask typically suffers from musical noises due to its discontinuity among each T-F unit. A variety of techniques have been derived to mitigate this problem. For example, sigmoidal nonlinearity is introduced to have continuous values ranging from 0 to 1 in the soft decision-based fuzzy SNR mask[3, 4, 7]. Aarabi proposed phase-error based T-F mask filter and showed improvements of both SNR and recognition rates over noisy mixture signals[2]. Phase-errors calculated from each T-F block are compressed to take values in $[0, 1]$ and used for scaling factors in order to maintain spectral structure of speech source of interest and to suppress interfering noises coming from different directions.

In this paper, we present a SNR-adaptive soft mask filter to extract target speech from noisy mixtures. Mask filter parameters are not fixed but dependent on frame-by-frame spectral magnitude ratios, which are capable of coping with various mixture SNRs and noise types.

2. SNR-ADAPTIVE SOFT TIME-FREQUENCY MASK FILTER

To deal with the multi-channel speech enhancement, we start with the adaptive beamforming algorithm called generalized sidelobe canceller(GSC), developed by Griffiths and Jim[9]. The GSC consists of three components; fixed beamformer, blocking matrix, and adaptive noise canceller(ANC). The GSC showed good performance for enhancing signal of interest while suppressing interfering noises with low complexity. However, if noise-reference signal produced by blocking matrix contains target signal leakage, ANC may take the risk of suppressing target signal as well as noise signals. So the performance of GSC cannot be guaranteed. This problem often occurs where multi-channel microphones are closely spaced or direction of arrival(DOA) is mismatched, since input signals are highly correlated with each other[8, 9].

To overcome this difficulty, hybrid of beamformer and adaptive soft mask filter is proposed as Fig.1. In this paper, we used a adaptive mask filter as an alternative to ANC. We consider multi-channel input signals as $X_1(t), X_2(t), ..., X_N(t)$. Noisy inputs are processed with beamformer and followed by mask filter. The beamformer block generates primary and secondary signals which represent target-dominant signal and target-nulling signal, respectively.

For simplicity, we assume that both target-enhancing beamformer and target-rejecting beamformer are composed of delay-and-sum beamformer in which their filter weights are

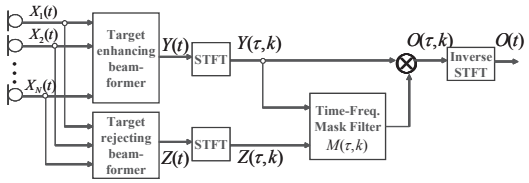


Fig. 1. Block diagram of multi-channel adaptive soft mask filter structure

$[1, 1, \ldots, 1]/N$,$[1, -1, \ldots, 1, -1]$, respectively. In this case, target speech is assumed to input as broadside direction to the microphone array. The target-enhancing beamformer generates primary output signal, $Y(t)$, which contains speech-dominant signal. Also, the target-rejecting beamformer outputs noise-dominant signal, $Z(t)$. However, instead of beamformers, other separation algorithms such as independent component analysis and target cancelling module may be used for generating primary and secondary signals[3, 6].

After converting time-domain signals into spectral domain spectrums via STFT(short-time Fourier transform), we define local SNR in each T-F unit as

$$SNR_{TF}(\tau, k) = \frac{|Y(\tau, k)|}{|Z(\tau, k)| + \varepsilon} \tag{1}$$

where τ and k is a frame and frequency index, respectively. $|Y(\tau, k)|$ and $|Z(\tau, k)|$ is a magnitude of complex spectrum of target-dominant signal and target-nulling signal, respectively. ε is a flooring term to overcome the divide-by-zero overflow.

The local SNR itself indicates reliability of each T-F unit, that is, a T-F unit with high SNR may come from target-dominant signal, while lower SNR unit come from interfering noises. Therefore, an appropriate thresholding parameter can determine whether corresponding T-F unit belongs to the target-dominant signal or not. Hence if we compress local SNR to take one of 0 or 1 value, a binary mask filter may be formed. The mask filter multiplied with dominant speech gives further enhanced target signal. However, the thresholding-based mask filter design discloses two problems. One is popular musical noises which originate from abrupt discontinuity among T-F units. The other is heuristically determined thresholding parameter, which does not cope with changing environments.

Therefore, we introduced a nonlinear sigmoid function to generate a soft T-F mask filter and adaptive mask parameters to accommodate various noise conditions.

$$\begin{aligned} M(\tau, k) &= g(SNR_{TF}(\tau, k)) \\ &= \frac{1}{1 + \exp(-\alpha(k) \cdot (SNR_{TF}(\tau, k) - \beta(\tau)))} \end{aligned} \tag{2}$$

where $\alpha(k)$ and $\beta(\tau)$ adjust the slope and bias of sigmoid function, respectively.

Fig.2 expresses sigmoidal output, $M(\tau, k)$, when $\alpha(k)$ and local SNR values are varied while $\beta(\tau)$ is fixed at 5.0. The sigmoidal slope, $\alpha(k)$, determines the compression ratio

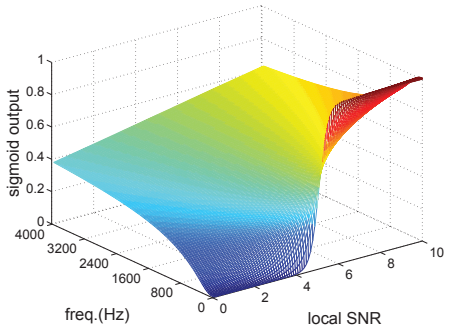


Fig. 2. Sigmoidal output is plotted when the slope($\alpha(k)$) has dynamic range $[0.5, 5.0]$ and number of FFT points is 512. Bias($\beta(\tau)$) is fixed at 5.0.

of difference between local T-F unit SNR, $SNR_{TF}(\tau, k)$, and frame SNR, $\beta(\tau)$. Since high frequency components(above $3KHz$) of speech have relatively low energies and are susceptible to noises, the local T-F SNR in the high frequency region is considered unreliable. Therefore, α is designed to be inversely proportional to the frequency index,k. It is desirable for the mask filter to have small dynamic ranges in the high frequency region than low frequency region.

$\alpha(k)$ depends on a frequency index as

$$\alpha(k) = \frac{\sigma_2}{k^m} \tag{3}$$

where $\alpha(k) \in [\sigma_1, \sigma_2]$, σ_1 ,and σ_2 are the lower bound and upper bound of α, respectively. $m = \frac{\log(\sigma_2/\sigma_1)}{\log(NFFT/2)}$ indicates a smoothing parameter which is automatically set by the number of FFT points($NFFT$), lower and upper bounds.

$\beta(\tau)$ changes with frame index,τ, incorporating frame-by-frame energy variations as

$$\beta(\tau) = \lambda_1 \left(\frac{\sum_{\forall k} \|Z(\tau, k)\|}{\sum_{\forall k} \|Y(\tau, k)\| + \sum_{\forall k} \|Z(\tau, k)\|}\right) + \lambda_2 \tag{4}$$

where $\beta(\tau) \in [\lambda_2, \lambda_1 + \lambda_2]$ and both λ_1 and λ_2 are two variables for determining bounds of $\beta(\tau)$.

When the noise is dominant in the current frame, $\beta(\tau)$ becomes high and thus low values are assigned to the $M(\tau, k)$ for the wide range of local T-F SNR, *i.e.*, $SNR_{TF}(\tau, k)$. It can be inferred that the local T-F SNR competes with $\beta(\tau)$.

The estimated mask filter can be multiplied with the dominant-speech spectrum as given below

$$O(\tau, k) = M(\tau, k) \cdot Y(\tau, k) \tag{5}$$

where $O(\tau, k)$ is the resulting signal, from which inverse STFT be applied to reproduce the time-domain enhanced speech, $O(t)$.

3. EXPERIMENTAL RESULTS

To verify the performance of the proposed adaptive soft mask filter, we measured SNR and PESQ values over dual-channel noisy mixture signals.

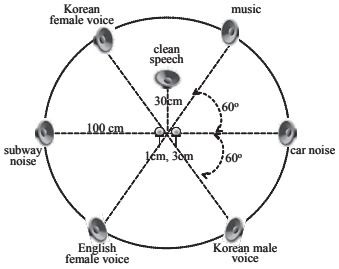


Fig. 3. Experimental setup. Data is collected in an anechoic room. Dual-channel recorded clean and noise signals are mixed together to produce noisy mixtures with varying SNRs.

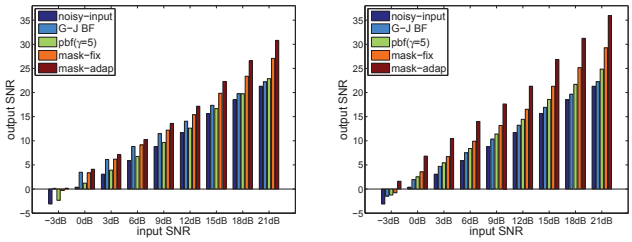
Fig.3 denotes the configuration to collect experimental data. In this configuration, we placed dual microphones at the center, a loudspeaker for playing target speech above the microphones with a distance of 30cm, and six loudspeakers for the noises around the microphones with 100cm apart. Each loudspeaker around the microphones simultaneously played with different noise sources of music, car noise, subway noise, Korean male voice, Korean female voice, and English male voice. We separately recorded clean speech and interfering noises. Then, we mixed them together in order to get noisy mixture signals with varying SNR from $-3dB$ to $21dB$.

Fig.4 shows the spectrograms of several dual-channel speech enhancement algorithms. The signal processed with Griffiths-Jim type beamformer(G-J BF) utilizes a VAD(voice activity detector) information which is precomputed from clean speech. In order to reduce speech distortion, the filter weights are only updated when the VAD signal indicates that the current sample belongs to noise parts[8, 9]. The length of FIR filter used in the adaptive noise cancelling is 64 and normalized LMS algorithm is used for the weight update. G-J BF can effectively suppress the noise components over 2KHz$\sim$3KHz seen in the noise-corrupted input signal, Fig.4 (b). However, non-stationary noise is not removed and thus its interfering harmonic components in about 2sec region, clearly remains in the resulting signal.

The signal processed with the phase-error based filter(PBF) proposed by Aarabi[2] is shown in Fig.4 . In this case, γ=5 is used throughout the experiments as in the [2]. The pbf processing can suppress interfering noises more than G-J BF processing with VAD as shown in Fig.4 (d).

The spectrograms processed by the proposed soft mask filters are given in the Fig.4 (e) and (f). As for the fixed parameters in the Fig.4 (e), we used α=0.5, β=5.0. It can be noticed that interfering harmonic components existing about 2sec region, are surely disappeared so that target speech is relatively emphasized, which results in better enhanced signal. As illustrated in the Fig.4(f), adaptive processing of the soft mask filter further suppresses noise as well as retains clean speech at high frequency region so that speech distortion is diminished.

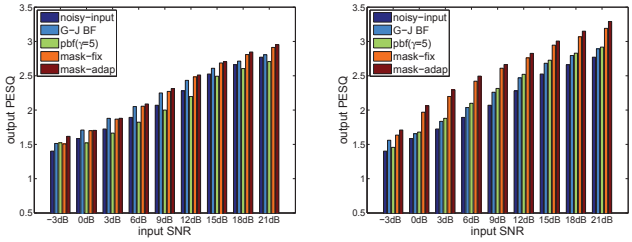
Fig.5 and 6 illustrate improvements in SNR and PESQ measure. Fig.5 denotes input SNR vs. output SNR, where the inter-microphone distance is 1cm (a), 3cm (b), respectively. Fig.6 gives input SNR vs. output PESQ values, where



(a) SNR with 1cm inter-mic. (b) SNR with 3cm inter-mic.

Fig. 5. SNR improvements. Figures denote input SNR vs. output SNR where the inter-microphone distance is 1cm (a), 3cm (b), respectively.

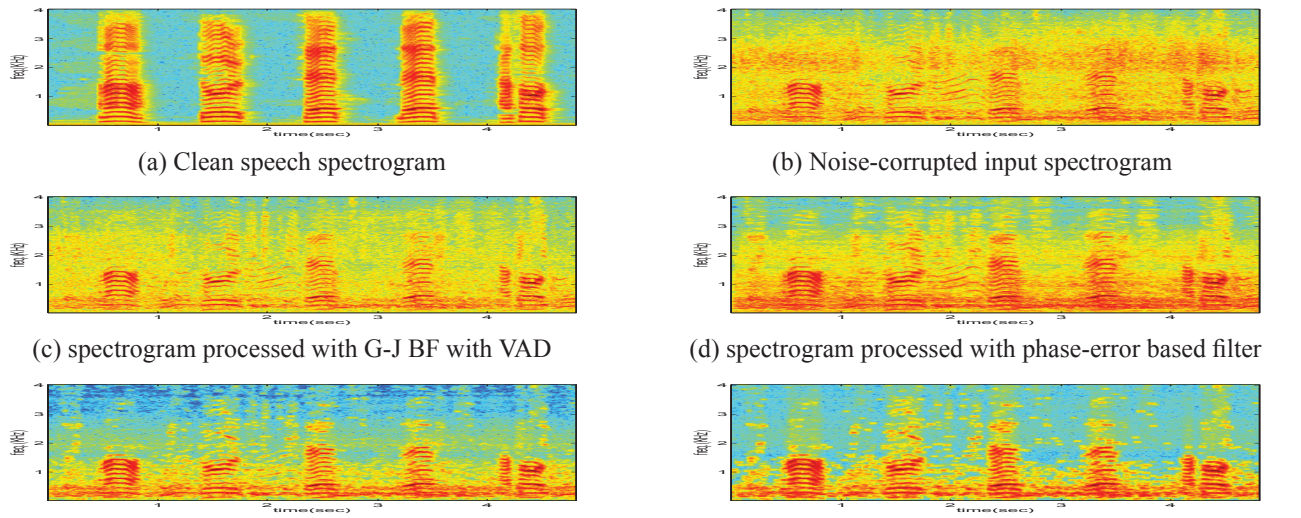
the inter-microphone distance is 1cm (a), 3cm (b), respectively. For the 1cm inter-microphone distance case, G-J BF algorithm with VAD gives consistently high values with the PBF, however, low results when the inter-mic distance is 3cm. The low performance in the PBF-processing with 1cm inter-microphone distance indicates that the phase information may not be sufficient and even give negative effect when noisy mixtures collected from a closely-spaced microphone pair are highly correlated each other. On the other hand, our local T-F SNR based adaptive soft mask filter gives higher SNR than other methods even where inter-microphone distance is only 1cm.



(a) PESQ with 1cm inter-mic. (b) PESQ with 3cm inter-mic.

Fig. 6. PESQ improvements. Figures denote input SNR vs. output PESQ where the inter-microphone distance is 1cm (a) and 3cm (b), respectively.

The adaptive soft mask has improvements in PESQ values



(a) Clean speech spectrogram

(b) Noise-corrupted input spectrogram

(c) spectrogram processed with G-J BF with VAD

(d) spectrogram processed with phase-error based filter

(e) spectrogram processed with fixed parameter-based filter

(f) spectrogram processed with adaptive parameter-based filter

Fig. 4. signal spectrograms. Noisy input has 9dB SNR and inter-microphone distance is 3cm. (a) clean speech signal, (b) noise-corrupted input signal, (c) signal processed with G-J BF with VAD, (d) signal processed with phase-error based filter(pbf) , (e) signal with fixed soft mask filter and (f) signal with adaptive soft mask filter are shown.

over noisy mixtures about 0.3∼0.6. Signal processed with the fixed mask filter gives lower PESQ improvements than the the adaptive mask filter about 0.2. It can be said that adaptive mask filter reduces speech distortion caused by both musical noises and over-subtracted high-frequency speech components, even without explicit VAD information.

4. CONCLUSIONS

In this paper, we introduced adaptive slope and bias for the sigmoidal soft mask filter. The adaptive slope depends on a frequency index and the bias is dependent on a frame index associated with frame-by-frame magnitude ratios. We tested the proposed mask filter with enhancement problems for recovering clean speech from noisy mixtures. For signals processed with the proposed algorithms, estimated SNR and PESQ values outperformed the conventional G-J beamformer with VAD and the phase-error based mask filter method even in a closely spaced microphone pair.

5. REFERENCES

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

[2] P. Aarabi and G. Shi,"Phase-based dual-microphone robust speech enhancement," *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, vol. 34, no. 4, pp. 1763–1773, Aug. 2004.

[3] S. Araki, H. Sawada, R. Mukai, and S. Makino, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proc. IWAENC*, (Paris, France), Sep. 2006.

[4] I. A. McCowan, A. Morris, and H. Bourlard, "Improving speech recognition performance of small microphone arrays using missing data techniques," in *Proc. ICSLP*, (Denver, USA), pp. 2181–2184, Sep. 2002.

[5] S. Srinivasan, N. Roman and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition ," in *Speech Communication*, vol. 48, pp. 1486–1501, 2006.

[6] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proc. ICASSP*, (Montreal, Canada), pp. 373–376, May 2004.

[7] J. Barker, L. Josifovski, M. Cooke and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP*, (Beijing, China), pp. 373–376, 2000.

[8] S. Gannot, D. Burshtein, E. Weinstein,, "Signal Enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. Signal Processing* , vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[9] L.J. Griffiths and C.W. Jim,, "An alternative approach for linearly constrained adaptive beamforming," *IEEE Trans. Antennas Propagat.* , vol. 30, pp. 27–34, Jan. 1982.