NEW ARCHITECTURE COMBINING BLIND SIGNAL EXTRACTION AND MODIFIED SPECTRAL SUBTRACTION FOR SUPPRESSION OF BACKGROUND NOISE

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate school of information science
Nara Institute of science and technology
Ikoma, Nara, Japan

ABSTRACT

In this paper, we propose a new architecture for enhancing the speech in a hands-free human/machine communication scenario. The proposed architecture uses frequency domain blind source extraction (FD-BSE) for noise estimation and beamforming in the speech direction. Soft masks function of the SNR are computed from the FD-BSE outputs and use to enhance the performance of the spectral subtraction applied channel wise to suppress the noise. Simulation results show that the proposed architecture can achieve a comparable SNR as conventional spectral subtraction with less distortion of the speech.

Index Terms— Blind signal separation, diffuse noise suppression, spectral subtraction

1. INTRODUCTION

In order to improve the human/machine interface, implementing hands-free speech recognition is the most natural choice. But picking the user's voice at distance is not an easy task because of noise and reverberation. Microphone array techniques were used to improve the captured speech by reducing the effect of noise and reverberation ([1, 2]). In recent years, frequency domain blind signal separation (FD-BSS) has been used with success for recovering the speech by separating the observed signals in their different components (see review paper [3]).
FD-BSS is in particular efficient for speech/speech separation [4]. But in the human/machine communication where the user's voice has to be extracted from a diffuse background noise, FD-BSS gives a better estimate of the diffuse background noise than of the target speech. Consequently FD-BSS has to be combined with some post-filtering techniques in order to improve the quality of the captured speech [5, 6].

In this paper, we propose a new architecture that combines a frequency domain blind extraction (FD-BSE) module with a modified multichannel spectral subtraction in order to suppress the diffuse background noise in the human/machine communication scenario. FD-BSE extracts the speech and gives an estimate of the diffuse background noise at each of the microphone. These noise estimates are used to compute soft masks giving an approximation of the speech to noise ratio in each channel (these are different from binary masks used in [7, 8, 4]). Then a modified spectral subtraction, where the noise estimates and the subtraction parameters are modulated using the soft mask information, is applied channel wise. Finally beamforming is applied to the speech frame whereas for the noise frame median filtering is used to attenuate the residual musical noise after the spectral subtraction.

The proposed method is compared to FD-BSE alone and FD-BSE combined with conventional spectral subtraction in order to show that it achieves a good noise reduction in term of SNR without introducing as much distortion as the conventional spectral subtraction.

2. ESTIMATION OF SPEECH AND BACKGROUND NOISE AT MICROPHONE

In the hands-free interface for human/machine communication, the user is close to the machine whereas the other signals create a diffuse background noise. The propagation of sounds from their locations of emission to the microphone array is modeled by a convolutive mixture. After applying a F points short time Fourier transform (STFT) to the observed signals, the convolutive mixture is equivalent to F instantaneous mixtures in the frequency domain. At the fth frequency bin, the observed signals are

$$X(f,t) = A(f)S(f,t)$$

where the $n \times n$ complex valued matrix $A(f)$ represents the instantaneous mixture received by the n microphone array and $S(f,t) = [s_1(f,t), \ldots, s_n(f,t)]^T$ are the emitted signal components at the fth frequency bin. t denotes the frame index. Let us consider that $s_1(f,t)$ is the target speech signal and all the other components are the background noise. Then we can decompose the observed signals in target speech and background noise parts

$$\begin{aligned} X(f,t) &= A^{(1)}(f)s_1(f,t) + \sum_{i=2}^{n} A^{(i)}(f)s_n(f,t) \\ &= X_S(f,t) + X_N(f,t) \end{aligned}$$

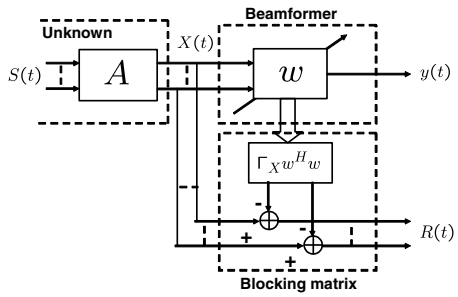

Fig. 1. BSE at frequency bin f.

where $A^{(i)}(f)$ denotes the ith column of $A(f)$.

The blind estimation of the speech and noise parts is possible using FD-BSS [9, 6]. Here we use the FD-BSE method proposed in [10]. Contrary to BSS, BSE estimates only one of the components of $S(f,t)$ in each frequency bin by taking

$$y(f,t) = w(f)X(f,t)$$

where $w(f)$ is a $1 \times n$ complex valued vector (see Fig.1). We call 'residuals' the contributions of all the signals other than $y(f,t)$ to the observations. The residuals are obtained by subtracting the orthogonal projection of the extracted signal from the observed signals

$$\begin{aligned} & R(f,t) = W_R(f)X(f,t), \\ \text{where} \quad & W_R(f) = I - \Gamma_X(f)w(f)^H w(f) \\ \text{with} \quad & \Gamma_X(f) = \mathcal{E}\{X(f,t)X^H(f,t)\}. \end{aligned}$$

The FD-BSE method can be seen as an adaptive beamformer and a blocking matrix as shown in Fig. 1.

In each frequency bin, the vector $w(f)$ extracting the speech component is iteratively determined using the update rule (dropping frame and frequency indices)

$$w_{k+1} \;=\; w_k - \mu_k \mathcal{E}\left\{\phi(y)R^H\right\} W_R \qquad (1)$$

where k is the iteration index, $\mu_k > 0$ is the adaptation step and $\phi(\cdot)$ is the score function associated with the extracted component. In the frequency domain, we can assume that all the components are circular (i.e. the joint density of their modulus and phase is separable) and use the approximation $\phi(y) = \tanh(|y|)\frac{y}{|y|}$ that is appropriate for speech extraction [11]. This update rule results in an extracted signal statistically independent of the residuals.

In the human/machine scenario, the speech extraction also uses the fact that the speech distribution is spikier than that of the diffuse background noise (To measure the spikiness of the distribution, we determine the parameter of the exponential distribution fitting the normalized modulus of $y(f,t)$ and $r_i(f,t)$ [10]). When $w(f)$ is such that the speech component is extracted, the residuals are estimates of the diffuse background noise at the microphone (equivalent to the projection back of FD-BSS). The extracted speech is also projected back

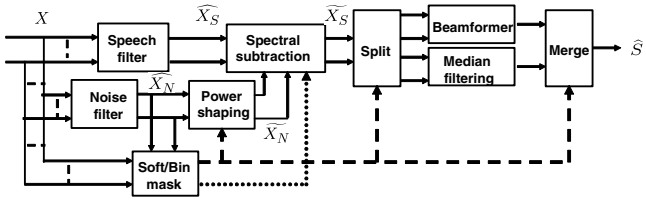

Fig. 2. Proposed architecture.

to the microphone array. Namely we have

$$\begin{aligned} \widehat{X_S}(f,t) &= \Gamma_X(f)w(f)^H w(f)X(f,t) \\ \widehat{X_N}(f,t) &= \left(I - \Gamma_X(f)w(f)^H w(f)\right) X(f,t). \end{aligned}$$

3. PROPOSED ARCHITECTURE

3.1. Overview

In the proposed architecture, all the processing is performed in the frequency domain by applying a short time Fourier transform to the observed signal received by the microphone array before processing and using overlap-add method to get the time domain signal after processing.

The block diagram in Fig 2 shows the processing in the frequency domain. First FD-BSE is used to obtain the estimate of $X_S(f,t)$ and $X_N(f,t)$ denoted by $\widehat{X_S}(f,t)$ and $\widehat{X_N}(f,t)$. The noise estimate and the observation are used to determine two type of masks: Soft masks (dotted line) and binary mask (dashed line). The spectrum of the noise estimate is modified using the binary mask (the observation is also used but the arrows to the power shaping block were omitted). Then the modified spectral subtraction is performed channel wise using the shaped noise spectrum and the soft masks.

After channel wise spectral subtraction, the signals are split in speech and noise frames (using the binary mask information). For the speech frames, the channels are beamformed using the vector $w(f)$ determined by the FD-BSE part whereas a median filter is applied on the noise frames. Finally the speech and noise frames are merged to give the speech estimate.

3.2. Soft masks creation

In the human/machine communication scenario, FD-BSS or FD-BSE give a good estimate $\widehat{X_N}(f,t)$ as it is possible to cancel the speech with a spatial null [12, 6]. Then considering a frame t_i where the speech is not active $X(f,t_i) = X_N(f,t_i)$ and $X(f,t_i) \approx \widehat{X_N}(f,t_i)$. On the contrary the more the speech is active in a given frame, the more $X(f,t_i)$ and $\widehat{X_N}(f,t_i)$ differ.

Thus we propose to use the ratio of the power of $\widehat{X_N}(f,t)$ and $X(f,t)$ for a given frame as our belief in the fact that the

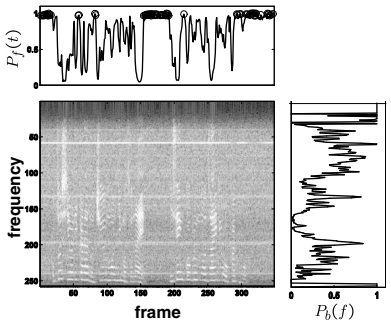

Fig. 3. Observed spectrum and corresponding frame and bin soft masks.

frame is composed of noise only. Thus we define

the frame soft mask as $P_f(t) = \frac{\sum_{f=1}^{F} |\widehat{X_N}(f,t)|^2}{\sum_{f=1}^{F} |X(f,t)|^2}$.

$P_f(t_i)$ measures our belief that during the frame t_i the speech is inactive. The frame soft mask can also be seen as

a measure of the frame SNR as $P_f(t) \approx (1 + \text{SNR}(t))^{-1}$

where $\text{SNR}(t)$ is ratio of speech and noise power in the frame.

In the remainder, we also define the $\gamma\%$ frame binary mask obtained by selecting the $\gamma\%$ most probable noise frames (binary mask set to one).

Similarly, by considering the frequency bins, we can

define a bin soft mask $P_b(f) = \frac{\sum_{t=1}^{T} |\widehat{X_N}(f,t)|^2}{\sum_{t=1}^{T} |X(f,t)|^2}$

that measures our belief that the speech is inactive in a given frequency bin.

The frame and bin soft masks are shown along with the observed signal in Fig. 3. The circle markers on the frame soft mask indicates frames selected for the 10% frame binary mask.

3.3. Power shaping

The role of the power shaping block is to match the estimated noise and the observed signal statistics for the frames we consider as noise only. This is done by setting the mean and variance of the spectrum of $\widehat{X_N}(f,t)$ to the same values as the mean and variance of the spectrum of $X(f,t)$ considering only the frames selected by the $\gamma\%$ frame binary mask.

3.4. Modified spectral subtraction

In each channel, the spectrum of the component of the power shaped noise estimate $\widetilde{X_N}(f,t)$ is subtracted from the spectrum of the component of the estimated speech $\widehat{X_S}(f,t)$

$$|\widetilde{X_S}(f,t)|^2 = \begin{cases} |\widehat{X_S}(f,t)|^2 - H(f,t)|\widetilde{X_N}(f,t)|^2 \\ \text{if } |\widehat{X_S}(f,t)|^2 - H(f,t)|\widetilde{X_N}(f,t)|^2 > 0 \\ \beta|\widehat{X_L}(f,t)|^2 \\ \text{else} \end{cases}$$

with β the flooring coefficient. Note in particular that the subtraction parameter (referred to as α in Sect.4) of conventional spectral subtraction [13] is replaced by a mask of the noise

spectrum defined by$H(f,t) = \delta_0 I + \delta_m P_b(f) P_f(t)$,

where δ_0 is the minimal subtraction and δ_m the additional subtraction modulated by the soft masks. Since $P_b(f)P_f(t)$ measures our belief in the absence of speech for a given time frequency value, the modified spectral subtraction only applies strong over subtraction where we believe there is no speech.

3.5. Median filtering

Spectral subtraction produces a very characteristic noise often referred to as musical noise. The musical noise is an impulsive noise in the time frequency domain. For example in the noise frames, some 'islands' of higher energy are left in the time frequency domain because of mismatch in the estimated noise and the spectral subtraction assumptions. For this reason, we apply a median filtering to the noise frames in order to attenuate the musical noise by discarding the higher values.

4. EXPERIMENTAL RESULTS

To demonstrate the effectiveness of the proposed approach we compared it to FD-BSE alone and FD-BSS with channel wise conventional spectral subtraction. A four microphone array (inter mic. spacing of 2.15cm) was used to record a diffuse noise (a vacuum cleaner at two meters from the array and -40^o) and several impulse responses (at one meter from the array with angles in $[-80^o, 80^o]$). The room reverberation time is T60 $= 200$ms. The recorded noise was mixed at different SNR with the convolution of the impulse responses and clean speech (20 signals from a database of Japanese utterances at 16kHz).

For the proposed method, three different $\gamma\%$ frame binary masks are considered 70%, 40% and 20% (respectively prop 1, 2 and 3 in Fig. 5). The modified subtraction parameters are $\beta = 0.003$, $\delta_0 I = 1$ and $\delta_m = 5$. The short time Fourier transform uses a 512 point hamming window with 50% overlap and pre-emphasis (a first order high pass filter $z_p = 0.97$). Speech extraction is performed by 600 iterations of the FD-BSE method with adaptation step of 0.3 divided by two every 200 iterations.

For the conventional spectral subtraction the flooring is 0.003 and the subtraction parameter is $\alpha = 2$ (mild over-subtraction) and $\alpha = 5$ (strong over-subtraction).

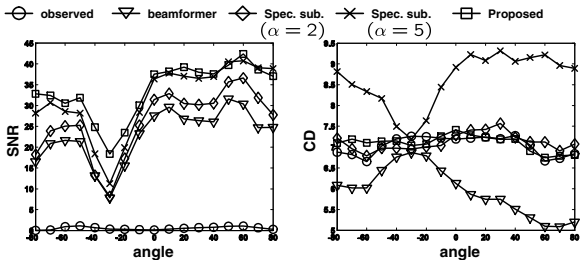

Fig. 4. SNR and cepstral distortion for all methods versus position of speaker for 10dB SNR

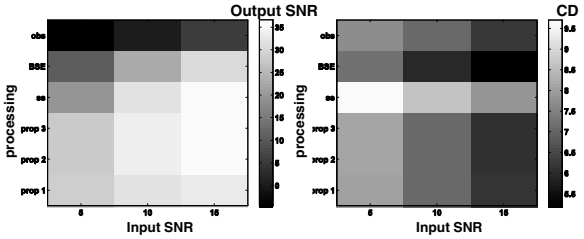

Fig. 5. Averaged performance at different input SNR.

The proposed method being highly non linear, the SNR estimation after processing is obtained by taking

$$\text{SNR} = \left(\frac{< yx_S >}{< yx_N >}\right)^2 \frac{< x_N x_N >}{< x_S x_S >},$$

where y is the output of the method and x_s and x_N are the true speech and noise at the microphone ($< \cdot >$ denotes time average).

Figure 4 shows the SNR and cepstral distortion (averaged on all speech signals) for all the positions of the speaker with the different methods ('observed' refers to the observation with pre-emphasis, 'beamformer' to FD-BSE, 'spec.sub.' to conventional spectral subtraction and 'proposed' to the proposed method with $\gamma = 40\%$). We can see that the proposed method achieves slightly better SNR as the spectral subtraction with strong over-subtraction but the distortion is equivalent to the mild over-subtraction case (Note that FD-BSE performs as a blind beamformer and that performance degrades in the direction of the noise).

In Fig. 5, the averaged values for all position shows that the best compromise between high SNR and low distortion is the proposed method with $\gamma = 40\%$ as FD-BSE alone introduce few distortion but does not improve significantly the SNR and conventional spectral subtraction results in higher distortion for comparable SNR.

5. CONCLUSION

In this paper, considering the suppression of the diffuse background noise in the human/machine communication scenario, we proposed an architecture that achieves high SNR but introduces few distortion to the speech estimate.

6. REFERENCES

[1] L.J. Griffiths and C.W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Trans. Antennas Prapagation*, vol. AP-30, pp. 27–34, 1982.

[2] S. Doclo et al., "Efficient frequency-domain implementation of speech distortion weighted multi-channel wiener filtering for noise reduction," *EUSIPCO'04*, pp. 2007–2010, 2004.

[3] M.S. Pedersen et al., "A survey of convolutive blind source separation methods," *Springer Handbook on Speech Comm.*, 2007.

[4] Y. Mori et al., "Blind source separation combining simo-ica and simo-model-based binary masking," *ICASSP'06*, pp. 81–84, 2006.

[5] J. Kocinski, "Speech intelligibility improvement using convolutive blind source separation assisted by denoising algorithms," *Speech Communication*, vol. 50, pp. 29–37, 2008.

[6] Y. Takahashi et al., "Blind spatial subtraction array with independent component analysys for hands-free speech recognition," *IWAENC'06*, 2006.

[7] R. Lyon, "A computational model of binaural localization and separation," *ICASSP 83*, pp. 1148–1151, 1983.

[8] N. Roman et al., "Speech segregation based on sound localization," *IJCNN 01*, pp. 2861–2866, 2001.

[9] P. Smaragdis, "Blind separation of convolved mixtures in the frequency domain," *Neurocomputing*, vol. 22, no. 1-3, pp. 21–34, 1998.

[10] J. Even et al., "Frequency domain blind signal extraction: Application to fast estimation of diffuse background noise," *HSCMA'08, Trent, Italy*, pp. 212–215, 2008.

[11] H. Sawada et al., "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fundamentals*, vol. E86-A, no. 3, pp. 590–596, 2003.

[12] H. Saruwatari et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP Jour. on Appl. Sig. Proc.*, vol. 2003, no. 11, pp. 1135–1146, 2003.

[13] S.F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. on ASSP*, vol. 27, no. 2, pp. 113–120, 1979.