SPEECH ENHANCEMENT IN PRESENCE OF DIFFUSE BACKGROUND NOISE: WHY USING BLIND SIGNAL EXTRACTION?

†Jani Even, †Hiroshi Saruwatari, †Kiyorhiro Shikano, and ‡Tomoya Takatani

†Nara Institute of Science and Technology, Nara, Japan
‡ TOYOTA MOTOR CORPORATION, Aichi, Japan

ABSTRACT

This paper study the blind estimation of the diffuse background noise for the hands-free speech interface. Some recent papers showed that it is possible to use blind signal separation (BSS) to estimate the diffuse background noise by suppressing the speech component after all the components were separated. In particular, the scale indeterminacy of BSS is avoided by using the projection back method. In this paper, we study an alternative to the projection back for the noise estimation and justify the use of blind signal extraction BSE rather than BSS.

Index Terms— blind signal extraction, noise estimation, speech enhancement

1. INTRODUCTION

In the last decades, several efficient methods exploiting blind signal separation (BSS) where proposed for processing the multidimensional observation given by microphone arrays. A great number of these methods address the separation of speech signals, the so called *cocktail party problem*, using the frequency domain approach (FD-BSS) (see review paper [1]). Another promising application of FD-BSS in acoustic signal processing is the hands-free speech interface. In such interface, the user interacts with the system using his (or her) voice which is picked at a distance by a microphone array and processed by the system. This is not strictly speaking a separation problem but a speech enhancement problem as the goal is to improve the quality of the user's speech that is corrupted by the diffuse background noise. But, in such situation, FD-BSS is an efficient method for estimating the diffuse background noise [2]. After FD-BSS is performed, the noise estimate is obtain by discarding the speech component and projecting back [3] the noise components. Then noise suppression is conducted by means of a nonlinear filter using the noise estimate given by FD-BSS (for example using spectral subtraction or Wiener filtering) [2].

In this paper, we propose a study of the diffuse background noise estimation for the hands-free speech interface. In particular, we compare the conventional noise estimation method based on FD-BSS [2] to the noise estimation method proposed in [4] that relies on frequency domain blind signal extraction (FD-BSE). We first derive the diffuse background noise estimate obtained by the projection back method when considering the hands-free speech interface. Then we also derive the diffuse background noise estimate obtained by subtracting the orthogonal projection of the speech component from the observation [4]. These derivations do not appear in [2, 4] or other papers to the knowledge of the authors. By comparing these two noise estimation methods, we justify that extracting only the speech component is sufficient to get an accurate estimation of the diffuse background noise. This gives the grounds for replacing the FD-BSS based noise estimation used in hands-free speech interface by a FD-BSE based noise estimation. Finally a hands-free dictation task in presence of diffuse background noise is presented as an example.

2. HANDS-FREE SPEECH INTERFACE

Let us first present the model of the hands-free speech interface which is defined in the frequency domain. The frequency domain signals are obtained using a short time Fourier transform of size F. In the remainder f denotes the frequency bin and k denotes the frame index. Considering that the user is a point source, the mixing model in the fth frequency bin is

$$\mathbf{X}(f,k) = \mathbf{H}_\theta(f) S_1(f,k) + \mathbf{N}(f,k), \qquad (1)$$

where $S_1(f,k)$ is the speech component, $\mathbf{N}(f,k)$ is a vector containing the n components of the diffuse background noise and

$$\mathbf{H}_\theta(f) = \{\exp(j2\pi(f/F)f_s \frac{id}{c} \sin\theta(f))\}_{i\in[0,n-1]}$$

is a $n \times 1$ vector depending of the speech direction of arrival (DOA) $\theta(f)$ (also of the sampling frequency f_s, microphone inter spacing d, and sound velocity c). Note that the vector $\mathbf{H}_\theta(f)$ is function of the frequency. The reason is that the *apparent* DOA at a given frequency, that accounts for the effect of the reflection and the reverberation, differs from the *physical* DOA of the speech, which is the angle defined by the user's position relatively to the microphone array.

We can reformulate (1) as a noiseless instantaneous mixture

$$\mathbf{X}(f,k) = \left[\ \mathbf{H}_\theta(f) \mid \mathcal{I}_n\ \right] \begin{bmatrix} S_1(f,k) \\ \mathbf{N}(f,k) \end{bmatrix},$$

where $\mathcal{I}_n$ is the identity matrix of size n.

For convenience we define

$$\mathbf{S}(f,k) = [S_1(f,k), S_2(f,k), \cdots, S_{n+1}(f,k)]^T$$
$$\text{with} \quad S_2(f,k), \cdots, S_{n+1}(f,k) = \mathbf{N}(f,k).$$

Then the noiseless instantaneous mixture is re-written as

$$\mathbf{X}(f,k) = \mathbf{A}(f)\mathbf{S}(f,k). \tag{2}$$

It is a realistic assumption that, in a given frequency bin, the target speech component is statistically independent of the diffuse background noise components. But the statistical independence of the diffuse background noise components is not assumed.

3. FD-BSS WITH NON LINEAR POST FILTER

In the fth frequency bin, the estimate $Y(f,t)$ is obtained by applying an unmixing matrices $W(f)$ to the observed signals

$$\mathbf{Y}(f,k) = \mathbf{W}(f)\mathbf{X}(f,k)$$

In [2], Takahashi et al. showed that in this situation the square matrix $\mathbf{W}(f)$ estimated by BSS is such that the row corresponding to the speech component estimate is a delay and sum (DS) beamformer in the direction of the speech's apparent DOA at that frequency. The other rows corresponding to the estimates of the noise components are null beamformers at the speech's apparent DOA at that frequency. After convergence of FD-BSS, we assume that the speech component is the first component of $\mathbf{Y}(f,k)$, the separation matrix has the form

$$\mathbf{W}(f) = \begin{bmatrix} \frac{1}{n}\mathbf{H}_\theta^H(f) \\ \mathbf{W}_\perp(f) \end{bmatrix} \tag{3}$$

where $\mathbf{W}_\perp(f)$ is a $n-1 \times n$ matrix of rank $n-1$ such that $\mathbf{W}_\perp(f)\mathbf{H}_\theta(f) = O_{n-1\times 1}$. We further assume that $\mathbf{W}(f)$ is invertible.

After separation, the noise estimate $\widehat{\mathbf{X}_N}(f,t)$ is obtained from the separated components (see next section). To suppress the diffuse background noise effect, a Wiener filter is applied on each component of the observed signals. The Wiener gain for the ith signal is

$$G_i(f,t) = \frac{|X_i(f,t)|^2}{|X_i(f,t)|^2 + \alpha|\widehat{X_{N_i}}(f,t)|^2}$$

where the subscript (i) denotes the ith component and α is a parameter controlling the noise reduction. The ith component of the filtered target speech is

$$\widehat{S}_i(f,t) = \sqrt{G_i(f,t)|X_i(f,t)|^2}\frac{X_i(f,t)}{|X_i(f,t)|}.$$

Finally the speech estimate $\widehat{S}(f,t)$ is obtained by applying a delay and sum (DS) beamformer in the direction θ of the target speech, see Fig. 1 (The angle θ is the average of the angles $\theta(f)$ estimated from $\mathbf{W}(f)$).

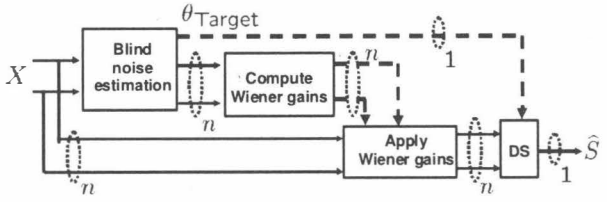


Fig. 1. Blind noise estimation with channel-wise Wiener post filter.

4. NOISE ESTIMATION METHODS

4.1. Method using the noise components

After performing FD-BSS, the diffuse background noise is estimated by discarding the speech component and projecting back the noise components [2]. The projection back of the noise components is defined by

$$\widehat{\mathbf{X}_N}(f,k) = \mathbf{W}(f)^{-1}\mathbf{D}\mathbf{W}(f)\mathbf{X}(f,k)$$

where $\mathbf{D}$ is a diagonal matrix with entries $[0,1,\cdots,1]$ along the diagonal. To study the quality of the noise estimate given by the projection back, let us define the matrix

$$\mathbf{K}(f) = \mathbf{W}(f)^{-1}\mathbf{D}\mathbf{W}(f)\mathbf{A}(f) \tag{4}$$
$$\text{such that} \quad \widehat{\mathbf{X}_N}(f,k) = \mathbf{K}(f)\mathbf{S}(f,k).$$

Using Eq.(3), we get $\mathbf{W}(f)\mathbf{A}(f) = \begin{bmatrix} 1 & \frac{1}{n}\mathbf{H}_\theta^H(f) \\ O_{n-1\times 1} & \mathbf{W}_\perp(f) \end{bmatrix}$,

$$\text{then} \quad \mathbf{W}^{-1}(f) = \mathbf{A}(f)\begin{bmatrix} 1 & \frac{1}{n}\mathbf{H}_\theta^H(f) \\ O_{n-1\times 1} & \mathbf{W}_\perp(f) \end{bmatrix}^+$$

where $^+$ denotes the Moore-Penrose pseudo inverse. Using the expression of this Moore-Penrose pseudo inverse

$$\begin{bmatrix} 1 & \frac{1}{n}\mathbf{H}_\theta^H(f) \\ O_{n-1\times 1} & \mathbf{W}_\perp(f) \end{bmatrix}^+ = \begin{bmatrix} \frac{n}{n+1} & O_{1\times n-1} \\ \frac{1}{n+1}\mathbf{H}_\theta(f) & \mathbf{W}_\perp^+(f) \end{bmatrix},$$

Eq.(4) reduces to $\mathbf{K}(f) = \left[O_{n\times 1} \,\middle|\, \mathbf{W}_\perp^+(f)\mathbf{W}_\perp(f)\right]$

however $\mathbf{W}_\perp^+(f)$ is a right inverse of $\mathbf{W}_\perp(f)$ thus

$$\mathbf{W}_\perp^+(f)\mathbf{W}_\perp(f) \neq \mathcal{I}_n \quad \text{and} \quad \widehat{\mathbf{X}_N}(f,k) \neq \mathbf{N}(f,k).$$

The quality of the estimated noise obtained using the projection back depends of how close to $\mathcal{I}_n$ is $\mathbf{W}_\perp^+(f)\mathbf{W}_\perp(f)$.

For the right inverse we have the following equality

$$\|\mathbf{W}_\perp^+(f)\mathbf{W}_\perp(f) - \mathcal{I}_n\|_F^2 = 1 \tag{5}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Meaning that the average squared error on the entries of $\mathbf{W}_\perp^+(f)\mathbf{W}_\perp(f)$ is $\frac{1}{n^2}$.

4.2. Method using the speech component

Another way to estimate the diffuse background noise is to suppress the speech from the observation. After performing FD-BSS, this can be done by first projecting the estimated speech component on the observation and then subtracting

this projection from the observation. Note that the other $n-1$ components obtained by performing the FD-BSS are simply discarded.

The noise estimate obtained by subtracting the projection of the speech component from the observation is defined by

$$\widetilde{\mathbf{X_N}}(f,k) = \left(\mathcal{I}_n - \Gamma_{\mathbf{X}}(f)\lambda^H {\mathbf{W_1}}^H(f)\lambda\mathbf{W_1}(f)\right)\mathbf{X}(f,k)$$

where $\Gamma_{\mathbf{X}}(f)$ is the covariance of $\mathbf{X}(f,k)$, $\mathbf{W_1}(f)$ is the row of $\mathbf{W}(f)$ corresponding to the speech component, and λ is a scalar such that $z(f,k) = \lambda\mathbf{W_1}(f)\mathbf{X}(f,k)$ verifies $\mathcal{E}\{|z(f,k)|^2\} = 1$.

Replacing $\mathbf{W_1}(f)$ by its value given in Eq.(3) we have

$$z(f,k) = \lambda\left[\ 1\ \middle|\ \tfrac{1}{n}\mathbf{H}_\theta(f)\ \right]\mathbf{S}(f,k)$$

then the constraint on $z(f,k)$ gives

$$|\lambda|^2 = \frac{1}{\sigma_1^2 + \sum_{i=2}^{n+1}\frac{1}{n^2}\sigma_i^2} \quad \text{where} \quad \sigma_i^2 = \mathcal{E}\{|S_i(f,k)|^2\}.$$

Let us denote by $\mathbf{Q}(f)$ the matrix such that

$$\widetilde{\mathbf{X_N}}(f,k) = \mathbf{Q}(f)\mathbf{S}(f,k).$$

Re-writing (we drop the frequency index for convenience) the covariance of $\mathbf{X}(k)$ as $\Gamma_{\mathbf{X}} = \mathbf{A}\Gamma_{\mathbf{S}}\mathbf{A}^H$ where $\Gamma_{\mathbf{S}} = \mathrm{diag}\{\sigma_1^2,\cdots,\sigma_{n+1}^2\}$ is the covariance of $\mathbf{S}(k)$ we get

$$\mathbf{Q} = \mathbf{A}(\mathbf{I}_{n+1} - \Gamma_{\mathbf{S}}\mathbf{A}^H|\lambda|^2{\mathbf{W_1}}^H\mathbf{W_1}\mathbf{A}).$$

Using Eq.(3), we can express the last term as

$$\mathbf{I}_{n+1} - \Gamma_{\mathbf{S}}\mathbf{A}^H|\lambda|^2{\mathbf{W_1}}^H\mathbf{W_1}\mathbf{A} =$$
$$\mathbf{I}_{n+1} - |\lambda|^2\Gamma_{\mathbf{S}}\begin{bmatrix}1\\ \frac{1}{n}\mathbf{H}_\theta\end{bmatrix}\begin{bmatrix}1\\ \frac{1}{n}\mathbf{H}_\theta\end{bmatrix}^H.$$

Then by matrix manipulation we obtain

$$\mathbf{Q} = \left[\mathbf{H}_\theta\,\middle|\,\mathcal{I}_n\right] - |\lambda|^2\left(\sigma_1^2\mathcal{I}_n + \frac{1}{n}\Lambda\right)\left[\mathbf{H}_\theta\,\middle|\,\tfrac{1}{n}\mathbf{H}_\theta\mathbf{H}_\theta^H\right],$$

where $\Lambda = \mathrm{diag}\{\sigma_2^2,\cdots,\sigma_{n+1}^2\}$.

Let us define $\mathbf{R} = |\lambda|^2\left(\sigma_1^2\mathcal{I}_n + \frac{1}{n}\Lambda\right)$

the diagonal matrix of general term $R_i = \dfrac{n\sigma_1^2 + \sigma_i^2}{n\sigma_1^2 + \sum_{k=2}^{n+1}\frac{\sigma_k^2}{n}}$

and the scalars σ_L and σ_H such that $\sigma_L^2 \le \sigma_k^2 \le \sigma_H^2$ for $k \in [2, n+1]$ (σ_L^2 and σ_H^2 are the minimal and maximal diffuse noise power across the microphones). Then we can write

$$\mathbf{Q} = \left[\ (\mathcal{I}_n - \mathbf{R})\,\mathbf{H}_\theta\ \middle|\ \mathcal{I}_n - \tfrac{1}{n}\mathbf{R}\mathbf{H}_\theta\mathbf{H}_\theta^H\ \right] \text{ with}$$

$$\frac{\frac{\sigma_L^2}{\sigma_1^2} - \frac{\sigma_H^2}{\sigma_1^2}}{n + \frac{\sigma_L^2}{\sigma_1^2}}\mathcal{I}_n \le \mathcal{I}_n - \mathbf{R} \le \frac{\frac{\sigma_H^2}{\sigma_1^2} - \frac{\sigma_L^2}{\sigma_1^2}}{n + \frac{\sigma_H^2}{\sigma_1^2}}\mathcal{I}_n$$
$$\frac{n + \frac{\sigma_L^2}{\sigma_1^2}}{n^2 + \frac{\sigma_H^2}{\sigma_1^2}}\mathcal{I}_n \le \frac{1}{n}\mathbf{R} \le \frac{n + \frac{\sigma_H^2}{\sigma_1^2}}{n^2 + \frac{\sigma_L^2}{\sigma_1^2}}\mathcal{I}_n \qquad (6)$$

The noise estimate obtained by this method is of comparable quality to the one obtained by projecting back the noise component. It depends explicitly of the power of the speech component relatively to the diffuse noise components. In particular, if $\sigma_L^2 = \sigma_H^2$ we have

$$\mathbf{Q} = \left[\ O_{n\times 1}\ \middle|\ \mathcal{I}_n - \tfrac{1}{n}\mathbf{H}_\theta\mathbf{H}_\theta^H\ \right].$$

Note that we also have the equality $\| - \frac{1}{n}\mathbf{H}_\theta\mathbf{H}_\theta^H\|_F^2 = 1$

meaning that the average mean square error on the entries of $\mathcal{I}_n - \frac{1}{n}\mathbf{H}_\theta\mathbf{H}_\theta^H$ is also $\frac{1}{n^2}$ as in Eq. (5).

5. RATIONALE FOR USING FD-BSE

From the above analysis, we can see that, for the hands-free speech interface case, the quality of the diffuse background noise estimate is not degraded by considering only the speech component and discarding the noise components obtained by the FD-BSS algorithm. Consequently, using FD-BSS to estimate a separation matrix is not necessary. It is wiser to use an FD-BSE method that estimates only a row vector for extracting the speech component and obtain the diffuse background noise estimate necessary for the post filter using the approach presented in Sect. 4.2.

In FD-BSE, at the fth frequency bin, the estimate $y(f,t)$ is obtained by applying an extracting vector $\mathbf{M}(f)$ to the observed signals

$$y(f,k) = \mathbf{M}(f)\mathbf{X}(f,k)$$

The vector $\mathbf{M}(f)$ that extract the speech component can be obtained by the method presented in [4] that minimize the cost function

$J(\mathbf{M}(f)) = \frac{1}{2}\mathcal{E}\left\{|y(f,k)|\right\}^2$ under the constraint

$\mathcal{E}\left\{|y(f,k)|^2\right\} = 1$ with an iterative gradient descent.

Then the diffuse background noise is estimated with the method in Sect. 4.2, replacing $\mathbf{W}_1(f)$ by $\mathbf{M}(f)$.

6. SIMULATION RESULTS

Experiments were conducted using measurements and recordings from a train station hall (using a four microphone array). Since our goal is to perform speech recognition, a 20K-word Japanese dictation task from JNAS [5] is used as performance measure. The recognizer is JULIUS [6], the conditions used in recognition are given in Table 1. The acoustic model is a clean model with super-imposed noise (office noise 25dB SNR). The test sentences are convoluted with the measured impulse response (in front of array at 50cm) and mixed with the recorded diffuse background noise at different SNR levels.

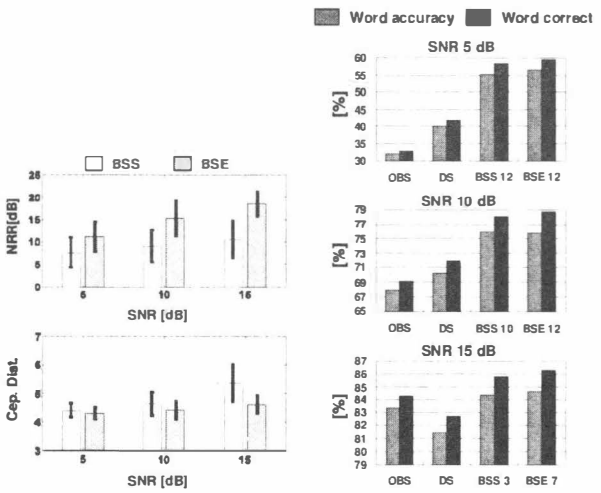
The FD-BSE method [4] is compared with a FD-BSS method (modified INFOMAX algorithm [7]). The quality of the noise estimate is measured in term of noise reduction rate

Table 1. *System specifications.*

Sampling frequency	16 kHz
Frame length	25 ms
Frame period	10 ms
Pre-emphasis	$1 - 0.97z^{-1}$
Feature vectors	12-order MFCC, 12-order ΔMFCCs 1-order ΔE
HMM	PTM , 2000 states
Training data	Adult and Senior (JNAS)
Test data	Adult and Senior female (JNAS)

(NRR) defined as the difference of the SNR before and after processing (taking the noise as signal). We also compute the cepstral distance (CD) between the estimated noise and the true noise to measure the distortion.

In Fig. 2(a), we can see that the NRRs are higher and the CDs are lower for the noise estimate obtained with FD-BSE (the values are averaged on the 100 signals and error bars are plotted). The FD-BSE noise estimate is especially better at higher SNRs (as we can expect from Eq.(6)). The FD-BSE based noise estimation is also on average 2.9 time faster than the FD-BSS based one. The error bars in Fig. 2(a) and the computation time standard deviations are relatively large because the signals from the JNAS database have variable lengths (2.4 s to 13.8 s). The word accuracies for the speech recognition task obtained using the unprocessed signal (OBS), the delay and sum beamformer in the target speech direction (DS), FD-BSS with multichannel Wiener filter (BSS+W α) and FD-BSS with channel-wise Wiener filter (BSE+W α) are given in Fig. 2(b).



(a) Noise estimate quality (b) Word Accuracy

Fig. 2. Simulation results.

7. CONCLUSION

In this paper, we justify the use of FD-BSE for estimating the diffuse background noise in the hands-free speech interface case. The reason is that the quality of the diffuse background noise estimate obtained by using only the estimated speech component does not differ from the one obtained using all the estimated noise components. Consequently, it is unnecessary to estimate a matrix with an FD-BSS method. Estimating a vector with an FD-BSE method is a better option as the computation cost is reduced while the noise estimation quality is maintained. Note that, in the real data simulation, the quality of the noise estimate obtained with FD-BSE is higher as the FD-BSE algorithm is trying to extract a speech like signal (see [4]) whereas the FD-BSS method is trying to recover the statistical independence of the speech and the noise which is a more challenging task.

8. REFERENCES

[1] M.S. Pedersen, J. Larsen, U. Kjems, and L.C. Parra, *A Survey of Convolutive Blind Source Separation Methods*, Springer, 2007.

[2] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Transaction on Audio, Speech and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009.

[3] N. Murata, S. Ikeda, and A. Zieh, "An approach to blind source separation based on temporal structure of speech signals," *Neurocomputing*, vol. 41, no. 1-4, pp. 1–24, 2001.

[4] J. Even, H. Saruwatari, and K. Shikano, "Blind signal extraction based speech enhancement in presence of diffuse background noise," *2009 IEEE Workshop on Statistical Signal Processing (SSP2009), Cardiff, Wales, UK*, pp. 513–516, 2009.

[5] K. Ito et al., "Jnas: Japanese speech corpus for large vocabulary continuous speech recognition research," *The Journal of Acoust. Soc. of Japan*, vol. 20, pp. 196–206, 1999.

[6] "Julius, an open-source large vocabulary csr engine - http://julius.sourceforge.jp," .

[7] A. J. Bell and T. J. Sejnowski, "An information maximization approach to blind separation and blind deconvolution," *Neural Computation*, vol. 7, no. 6, pp. 1129–1159, 1995.