COMPLEX NEWTON ALGORITHM FOR BLIND SIGNAL EXTRACTION OF SPEECH IN DIFFUSE NOISE

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

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

ABSTRACT

Several recent methods for speech enhancement in presence of diffuse background noise use frequency domain blind signal separation to estimate the diffuse noise and a nonlinear post filter to suppress this estimated noise. This paper presents a frequency domain blind signal extraction method for estimating the diffuse noise in place of the frequency domain blind signal separation. The method is based on the minimization by means of a complex Newton algorithm of a cost function depending of the modulus of the extracted component. The proposed complex Newton method is compared to the gradient descent on the same cost function and to the blind signal separation approach.

Index Terms— Blind signal extraction, speech enhancement, Newton method

1. INTRODUCTION

This papers deals with the enhancement of a target speech close to a microphone array in presence of diffuse background noise created by sources that are far from the array. In [1], the authors proposed an architecture that combines frequency domain blind signal separation (FD-BSS) to estimate the diffuse background noise with a nonlinear post filter for suppressing this noise.

FD-BSS is designed to separate an observed mixture in its different components, as a result when the number of microphones is greater than two, the FD-BSS based noise estimation also unnecessarily separates the noise in different components. An alternative is to use frequency domain blind signal extraction (FD-BSE) to extract the target speech and then linearly cancel this estimate to obtain the diffuse noise estimate used in the nonlinear post filter. In [2], we proposed an FD-BSE methods that minimizes a cost function based on the modulus of the extracted component. Unlike earlier work as [3], this method does not perform FD-BSS and then selects the source to be extracted but it directly extracts the desired source under some problem dependent assumptions. In this paper, we first give some additional insight concerning the local minima of this proposed cost function. Then we derive a complex Newton algorithm for minimizing this cost function. Finally we present some experimental results to show the performance of the proposed complex Newton method.

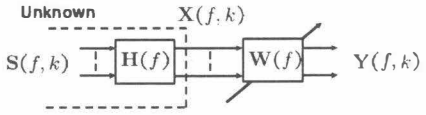

Fig. 1. BSS at the fth frequency bin

2. FREQUENCY DOMAIN BLIND SIGNAL SEPARATION

The frequency domain approach of BSS is often preferred to the time domain approach because using the short time Fourier transform (STFT) with F frequency bins, the time domain convolutive mixture is transformed in F instantaneous mixtures, one per frequency bin. Considering the specific problem of close speech enhancement in presence of diffuse background noise, the mixture model is

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

where $\mathcal{I}_n$ is the identity matrix of size n, $\mathbf{X}(f,k)$ is the STFT of the observed signals, $S_1(f,k)$ is the STFT of the speech component, $\mathbf{N}(f,k)$ is a vector containing the STFTs of 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).

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 non square instantaneous mixture is rewritten as

$$\mathbf{X}(f,k) = \mathbf{H}(f)\mathbf{S}(f,k). \tag{1}$$

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.

978-1-4244-4296-6/10/$25.00 ©2010 IEEE

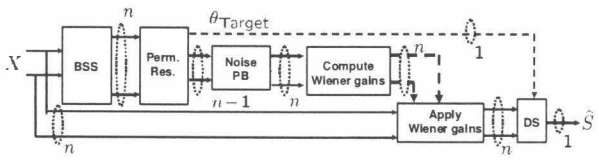

Fig. 2. BSS with channel-wise Wiener post filter at the fth frequency bin

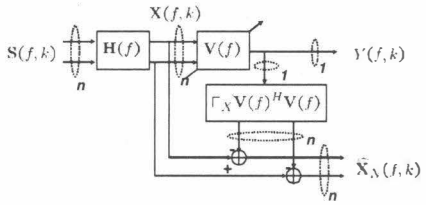

Fig. 3. BSE noise estimation at the fth frequency bin

FD-BSS can be achieved by estimating the frequency components $S(f,k)$ of the emitted signals in each of the frequency bin. At the fth frequency bin, the vector of estimated components is obtained by applying a complex valued separation matrix to the vector of observed signals (see Fig. 1)

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

For mixture of statistically independent point sources, the matrix $\mathbf{W}(f)$ is usually determined by an algorithm that minimizes a cost function measuring the statistical independence of the components of $\mathbf{Y}(f,k)$ [4] (the mutual information is a commonly used cost function, see review paper [5]).

For the case of close speech enhancement in presence of diffuse background noise, FD-BSS estimates accurately the diffuse background noise by blindly steering a spatial null in the direction of the target speech to cancel it [1]. But with a limited number of microphones it is not possible to cancel the diffuse background noise and the target speech estimate is very poor. For this reason, the authors in [1] proposed to combine an FD-BSS based diffuse noise estimation with a nonlinear post filter for suppressing the diffuse noise.

An equivalent architecture is depicted in Fig. 2. After BSS, the estimate of the diffuse background noise is obtained by finding the speech component in each frequency bin and projecting back [6] the $n-1$ other estimated components to the microphone array. Then a channel-wise Wiener post filter is used to suppress the noise estimate and finally the channels are merged together with a delay and sum (DS) beamformer to get the final speech estimate (the beamformer direction is estimated from the row of the separation matrix corresponding to the speech estimate).

3. PROPOSED APPROACH

3.1. Blind signal extraction

In FD-BSE, at the fth frequency bin, the extracted component is obtained by applying a complex valued extraction vector to the vector of observed signals (see upper part in Fig. 3)

$$y(f,k) = \mathbf{V}(f)^H \mathbf{X}(f,k)$$

$$\text{with the constraint} \quad \mathcal{E}\left\{|y(f,k)|^2\right\} = 1. \tag{2}$$

Considering the target speech in diffuse noise problem, the extraction of the speech is expressed by $\mathbf{V}(f)^H\mathbf{H}(f) \approx \lambda \mathbf{e}_1$ where $\mathbf{e}_1$ is the first coordinate row vector and λ is an unknow complex scalar. The constraint forces $|\lambda|^2\mathcal{E}\left\{|s_1|^2\right\} \approx 1$. Then a n component noise estimate is obtained by taking

$$\widehat{\mathbf{X}}_N(f,k) = \left(\mathbf{I}_n - \mathbf{\Gamma}_{\mathbf{X}(f)}\mathbf{V}(f)\mathbf{V}(f)^H\right)\mathbf{X}(f,k). \tag{3}$$

where $\mathbf{\Gamma}_{\mathbf{X}(f)}$ is the covariance matrix of $\mathbf{X}(f)$. This corresponds to the lower part of Fig. 3.

3.2. Cost function and local minima

We use the cost function presented in [2] that is not based directly on statistical independence but on the sparseness of the extracted component. This cost function is defined by

$$J(\mathbf{V}(f)) = \frac{1}{2}\left(\mathcal{E}\left\{|y(f,k)|\right\}\right)^2 \tag{4}$$

where $\mathcal{E}\{\cdot\}$ is the expectation operator.

In this paper, we conduct the derivation of the local minima of this cost function for the two signal case to show that the cost function presents local minima when extraction is performed (This is a particular case as the cost function is defined for any number of signals in [2]).

Let us consider the mixture $y = v_1 s_1 + v_2 s_2$ of two independent complex valued random variables s_1 and s_2, the constraint (2) is enforced by taking

$$v_1 = \frac{|\cos\theta|}{\sigma_1} e^{\jmath\beta_1} \quad \text{and} \quad v_2 = \frac{|\sin\theta|}{\sigma_2} e^{\jmath\beta_2}$$

where $\sigma_i^2 = \mathcal{E}\left\{|s_i|^2\right\}$. Because of the symmetry of the problem we can consider the case $\theta \in [0, \frac{\pi}{2}]$. Let us moreover assume that the modulus and phase of s_1 and s_2 are independent and that their phases are uniformly distributed.

The derivative of the modulus of y with respect to θ is

$$\frac{d}{d\theta}|y| = \frac{d}{d\theta}\sqrt{yy^*} = \frac{1}{2|y|}\frac{d}{d\theta}\left(\frac{\cos^2\theta}{\sigma_1^2}|s_1|^2 + \frac{\sin^2\theta}{\sigma_2^2}|s_2|^2 + 2\frac{\cos\theta\sin\theta}{\sigma_1\sigma_2}|s_1||s_2|\cos\delta\right),$$

where $\delta = \gamma_1 - \beta_1 - \gamma_2 + \beta_2$ (with $s_i = |s_i|e^{\jmath\gamma_i}$). Then, after derivation and assuming that we can permute the operators $\frac{d}{d\theta}\cdot$ and $\mathcal{E}\{\cdot\}$, the derivative is

$$\frac{d}{d\theta}\mathcal{E}\{|y|\} = \sin\theta\cos\theta\,\mathcal{E}\left\{\frac{1}{|y|}B\right\} + \left(\cos^2\theta - \sin^2\theta\right)\mathcal{E}\left\{\frac{|s_1||s_2|}{|y|\sigma_1\sigma_2}\cos\delta\right\}$$

$$\text{where} \quad B = \frac{|s_2|^2}{\sigma_2^2} - \frac{|s_1|^2}{\sigma_1^2}.$$

Using the same approach with these assumptions we obtain for the second derivative

$$\frac{d^2}{d\theta^2}\mathcal{E}\{|y|\} = \left(\cos^2\theta - \sin^2\theta\right)\mathcal{E}\left\{\frac{1}{|y|}B\right\}$$
$$- \sin^2\theta\cos^2\theta\ \mathcal{E}\left\{\frac{1}{|y|^2}B^2\right\}$$
$$- \sin\theta\cos\theta\left(\cos^2\theta - \sin^2\theta\right)\mathcal{E}\left\{\frac{1}{|y|^3}B\frac{|s_1||s_2|}{\sigma_1\sigma_2}\cos\delta\right\}.$$

For $\theta = 0$, s_1 is extracted $|y| = \frac{|s_1|}{\sigma_1}$ and the last term of the derivative can be factorized as

$$\mathcal{E}\left\{\frac{|s_1||s_2|}{|y|\sigma_1\sigma_2}\cos\delta\right\} = \mathcal{E}\left\{\frac{|s_1||s_2|}{|y|\sigma_1\sigma_2}\right\}\mathcal{E}\{\cos\delta\}$$

because s_1 and s_2 have statistically independent modulus and phase. Since γ_1 is uniformly distributed in $[-\pi, \pi]$ we can prove using the characteristic function that $\mathcal{E}\{\cos\delta\} = 0$. Consequently the derivative is null for $\theta = 0$. Moreover, using the independence of s_1 and s_2, the second derivative is

$$\frac{\partial^2\mathcal{E}\{|y|\}}{\partial\theta^2} = \mathcal{E}\left\{\frac{B\sigma_1}{|s_1|}\right\} = \mathcal{E}\left\{\frac{\sigma_1}{|s_1|} - \frac{|s_1|}{\sigma_1}\right\}$$

using Jensen's inequality we get

$$\frac{\partial^2\mathcal{E}\{|y|\}}{\partial\theta^2} \geq \frac{\sigma_1}{\mathcal{E}\{|s_1|\}} - \frac{\mathcal{E}\{|s_1|\}}{\sigma_1}$$

which is always positive because $\mathcal{E}\{|s_1|^2\} - \mathcal{E}\{|s_1|\}^2 \geq 0$. So the modulus has local minima for $\theta = 0 + p\pi$ corresponding to $|y| = \frac{|s_1|}{\sigma_1}$. With the same reasoning, we can show that the modulus has local minima for $\theta = \frac{\pi}{2} + p\pi$ corresponding to $|y| = \frac{|s_2|}{\sigma_2}$ (the extension to the case of n signals is done by considering $n-1$ angles and splitting the n signals in two groups recursively).

An interesting property of the cost function $J(\theta) = \frac{1}{2}\left(\mathcal{E}\{|y|\}\right)^2$ is the dependency of the local maximum position to $\Delta J = J(\pi/2) - J(0)$. In Fig. 4(a) we plotted $J(\theta)$ versus θ for $\Delta J \approx 0.15$, whereas in Fig. 4(b) it is plotted for for $\Delta J \approx 0.7$. Increasing ΔJ results in imperceptible minima in $\theta = \pi/2 + \pi$ as the two adjoining local maxima get very close (in Fig. 4(b) the local minima can be seen if we magnify the graph). To underline this interesting property, we calculated the cost function variation versus θ when s_1 is a frequency domain speech components and s_2 is a diffuse noise components (the data are the same as in the simulation part). The results for all the frequency bins are plotted in Fig. 4(c). Note that the local minima in $\theta = \frac{\pi}{2} + p\pi$ are imperceptible contrary to the ones in $\theta = 0 + p\pi$ because ΔJ is large for all frequency bins. As a consequence, the basins of attraction corresponding to the extraction of the diffuse noise are very small compared to the ones for the speech extraction. This feature of the proposed cost function is very interesting as it guarantees the extraction of the close speech component from the diffuse background noise without the need of any additional processing (unlike FD-BSS that requires to find the speech component out of the separated components).

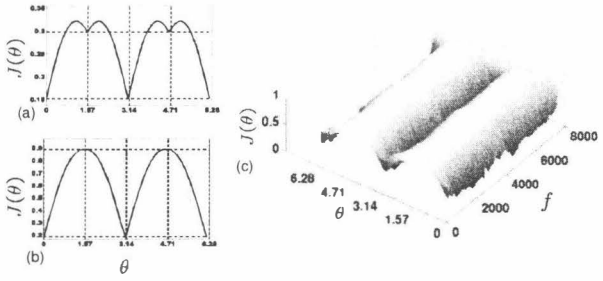

Fig. 4. (a)(b) Cost function for mixture of signals with different values of ΔJ versus angle, (c) Cost functions for mixture of speech and noise components versus frequency and angle (real data)

3.3. Complex Newton algorithm

Let us first define the complex derivative operator

$$\frac{\partial(\cdot)}{\partial\boldsymbol{\zeta}} = \begin{pmatrix} \frac{\partial(\cdot)}{\partial\mathbf{V}} & \frac{\partial(\cdot)}{\partial\mathbf{V}^*} \end{pmatrix}$$

where $*$ denotes the complex conjugate and the derivatives relatively to $\mathbf{V}$ and $\mathbf{V}^*$ are row operators (Wirtinger calculus was also used in [7] to present complex Newton ICA methods based on the kurtosis cost function).

The derivative of the cost function is (dropping frequency and frame indices)

$$\frac{\partial J(\mathbf{V})}{\partial\boldsymbol{\zeta}} = \frac{1}{2}\mathcal{E}\{|y|\}\begin{pmatrix} \mathcal{E}\left\{\frac{y\mathbf{X}^H}{|y|}\right\} & \mathcal{E}\left\{\frac{y^*\mathbf{X}^T}{|y|}\right\} \end{pmatrix}$$

and the Hessian is

$$\mathcal{H}_{\mathbf{V}}J(\mathbf{V}) = \frac{\partial}{\partial\boldsymbol{\zeta}}\left(\frac{\partial J(\mathbf{V})}{\partial\boldsymbol{\zeta}}\right)^H = \begin{bmatrix} \mathcal{H}_{\mathbf{VV}} & \mathcal{H}_{\mathbf{V}\cdot\mathbf{V}} \\ \mathcal{H}_{\mathbf{VV}\cdot} & \mathcal{H}_{\mathbf{V}\cdot\mathbf{V}\cdot} \end{bmatrix}$$

where $\mathcal{H}_{\mathbf{VV}} = \frac{1}{4}\left(\mathcal{C}\mathcal{C}^H - \mathcal{E}\{|y|\}\mathcal{E}\left\{\frac{\mathbf{X}\mathbf{X}^H}{|y|}\right\}\right)$,

and $\mathcal{H}_{\mathbf{V}\cdot\mathbf{V}} = \frac{1}{4}\left(\mathcal{C}\mathcal{C}^T - \mathcal{E}\{|y|\}\mathcal{E}\left\{\frac{(y^*)^2\mathbf{X}\mathbf{X}^T}{|y|^3}\right\}\right)$

with $\mathcal{C} = \mathcal{E}\left\{\frac{y^*\mathbf{X}}{|y|}\right\}$.

By Hermitian symmetry we have

$$\mathcal{H}_{\mathbf{V}\cdot\mathbf{V}\cdot} = \mathcal{H}_{\mathbf{VV}}^* \quad \text{and} \quad \mathcal{H}_{\mathbf{VV}\cdot} = \mathcal{H}_{\mathbf{V}\cdot\mathbf{V}}^*$$

Then the complex Newton method update is

$$\Delta\mathbf{V} = \frac{1}{2}\mathcal{E}\{|y|\}\left(\mathcal{H}_{\mathbf{VV}} - \mathcal{H}_{\mathbf{V}\cdot\mathbf{V}}\mathcal{H}_{\mathbf{V}\cdot\mathbf{V}}^{-1}\mathcal{H}_{\mathbf{VV}\cdot}\right)^{-1}$$
$$\left(\mathcal{H}_{\mathbf{V}\cdot\mathbf{V}}\mathcal{H}_{\mathbf{V}\cdot\mathbf{V}}^{-1}\mathcal{E}\left\{\frac{y}{|y|}\mathbf{X}^*\right\} - \mathcal{E}\left\{\frac{y^*}{|y|}\mathbf{X}\right\}\right)$$

and the update rule is $\mathbf{V}_{j+1} = \mathbf{V}_j + \mu_j\Delta\mathbf{V}_j$ where j denotes the iteration number and μ_j a positive adaptation step.

In practice, an additional cost function

$$J_N(\mathbf{V}) = \frac{1}{2}(\mathbf{V}^H\Gamma_{\mathbf{X}}\mathbf{V} - 1)^2$$

is added to $J(\mathbf{V})$ to implement the constraint (2).

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)

Table 2. *Word accuracy and computation time.*

OBS	BSS	BSE(grad.)	BSE(Newton)
67.92%	79.15%	78.39%	78.96%
	91.78 s	16.03 s	18.9 s

4. SIMULATIONS

In a train station hall (900ms of reverberation), a eight linear microphone array with spacing of 2.15 cm was used to record the diffuse noise and obtain the estimate of the impulse response for a user standing at 50cm in front of the array. Then we generated one hundred test utterances with 10dB of SNR for performing the 20K-word Japanese dictation task from JNAS [8].

The data are processed with the architecture presented in Fig. 2 where the noise estimation is performed by FD-BSS (Infomax like see [5]) or using FD-BSE with the proposed complex Newton method or the gradient based method [2]. For FD-BSE we take $\widehat{X}_N(f,k)$ as noise estimate (see Eq.(3) and Fig.3). For the Newton method, a fixed step $\mu = 0.1$ is used during 100 iterations whereas for the gradient method an initial step of $\mu = 0.1$ is divided by two every 50 iterations until 200 iterations are performed. FD-BSS has an initial step of $\mu = 0.5$ which is divided by two every 50 iterations until 200 iterations are performed. The gain of the channel-wise Wiener post filter is set to similar values for the three methods.

The recognizer is JULIUS [9] and 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 word accuracy and averaged computation times for the three methods and for the unprocessed signal (OBS) are given in Table 2. The proposed Newton method achieves similar word accuracy as the other two methods using half the number of iterations but, in its current implementation, computation time is higher than that of the gradient based BSE method because of the higher complexity of each iteration. An advantage of the Newton method is also its robustness to the selection of the adaptation step compared to the gradient based method.

5. CONCLUSION

In this paper, we derived the local minima of the BSE cost function we presented in [2] and proposed a complex Newton algorithm for its minimization. The simple implementation of this complex Newton method gives promising results and we are now working on a more efficient implementation to replace the gradient based optimization.

6. REFERENCES

[1] 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.

[2] 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.

[3] H. Sawada, S. Araki, R. Mukai, and S. Makino, "Blind extraction of dominant target sources using ica and time-frequency masking," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2165–2173, 2006.

[4] P. Comon, "Independent component analysis, a new concept ?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

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

[6] 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.

[7] L. Hualiang and T. Adali, "A class of complex ica algorithms based on the kurtosis cost function," *IEEE Transaction on neural networks*, vol. 19, no. 3, pp. 408–420, 2009.

[8] 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.

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