FREQUENCY DOMAIN SEMI-BLIND SIGNAL SEPARATION: APPLICATION TO THE REJECTION OF INTERNAL NOISES

Jani Even, Hiroshi Saruwatari, Kiyohiro Shikano

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

ABSTRACT

Recently, methods using blind signal separation were proposed to separate the signals received by a microphone array. In this paper, we propose a new frequency domain semi-blind source separation method for replacing the blind source separation method when it is possible to obtain additional information on some of the signals. This is of particular interest in situations like in hands-free speech recognition where the blind separation has to work on limited amount of data in a challenging environment. The proposed method incorporates references to some of the signals that are obtained by additional sensors. Some experimental results shows that the proposed method is able to incorporate the additional information efficiently and that the performances are improved in term of SNR and word accuracy in a speech recognition task.

Index Terms— Semi-Blind signal separation, speech processing

1. INTRODUCTION

Nowadays, communicating with machines is usually not natural and requires some adaptation or training. In order to improve the usability of these machines and reduce the burden for the users it is important to recreate the natural human communication interface: Speech. The most difficult task being to give machines the ability to listen. Speech recognition is working well if we use a microphone close to the user's mouth but this is not a natural interface and not a convenient one in many situations. For these reasons, the focus is now on hands-free speech recognition. In hands-free speech recognition, the user's voice is picked at distance by a microphone array making a more natural interface with the machine. However, the cost is that noise and reverberation deteriorate the received speech quality. Hence it is necessary to improve the quality of the received speech before speech recognition is performed.

In order to deal with the noise, blind signal separation (BSS) based techniques are strong candidates for processing the multidimensional observation given by microphone arrays (see review paper [1]). The goal of BSS is to separate the observed signal in its different components. Ideally, receiving the user's speech contaminated with noise, we would recover the speech and the noise separately. The frequency domain approach, referred to as FD-BSS, is especially of great interest since the convolutive mixture modeling the reverberant environment can be efficiently processed in the frequency domain. However, this is still a challenging task in a real environment where the number of interfering noise signals is large and the amount of data is limited.

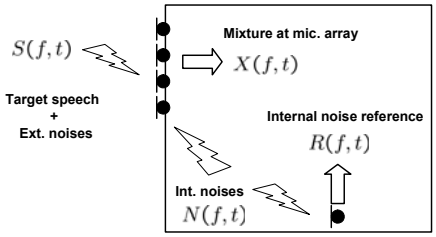

Fig. 1. General situation.

In this paper, we consider the case where some additional information is available. For example, consider a navigation system in a car with a hands-free speech recognition interface that uses FD-BSS to improve the received speech. If the driver listen to music then the system should use information from the music player to obtain better performance. This is a semi-blind approach because a reference to one of the signals is available. Note that the system knows what music was emitted but still have to determine the received music in the observed speech. Figure 1 illustrates the situation. The hands-free speech recognition system uses a microphone array that picks the user's speech and noises. The noises are composed of the exterior noises and the interior noises. The interior noises being the noises for which references are available. In a real environment, to improve the separation it seem necessary to exploit all information. For this purpose, we propose a semi-blind signal separation method that operates in the frequency domain in order to replace the FD-BSS approach. After presenting the new method, its performances are compared to the blind approach in a realistic environment.

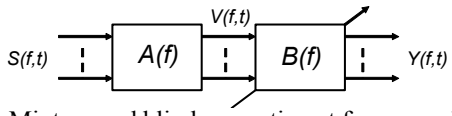

Fig. 2. Mixture and blind separation at frequency bin f.

2. FREQUENCY DOMAIN BLIND SIGNAL SEPARATION

In acoustic, the observed signals received by a microphone array in a reverberant environment are convolutive mixtures of some signals emitted from different locations. The goal of BSS is to recover the emitted signals knowing only the observed mixtures. In the frequency domain approach to BSS, a short time Fourier transform STFT is applied to the observed signals to get the frequency domain observations. Then the observed signals at the fth frequency bin are

$$V(f,t) = A(f)S(f,t) \tag{1}$$

where the $n \times n$ matrix $A(f)$ represents the mixture, $S(f,t)$ are the emitted signals at the fth frequency bin and t denotes the frame index. Consequently when using a F points analysis frame for the STFT the convolutive mixture is replaced by F instantaneous mixtures and the goal is to estimate the components of the emitted signals $S(f,t)$ in each frequency bin.

In the fth frequency bin, the estimates are obtained by applying an unmixing matrix $B(f)$ to the observed signals (see Fig.2)

$$Y(f,t) = B(f)V(f,t) = B(f)A(f)S(f,t). \tag{2}$$

A usual assumption in BSS is that the components of $S(f,t)$ are statistically independent in each frequency bin. Then the components of $Y(f,t)$ are statistically independent if and only if $B(f)$ is such that

$$Y(f,t) \quad = \quad P(f)\Lambda(f)S(f,t)$$

where $P(f)$ is a $n \times n$ permutation matrix and $\Lambda(f)$ is a diagonal $n \times n$ matrix [2].

As a consequence, in each frequency bin, it is possible to recover the components of $S(f,t)$ up to scale and permutation indeterminacy by finding the unmixing matrix $B(f)$ that gives statistically independent components. This problem is often referred to as independent component analysis (ICA). To complete the separation, it is necessary to match the components belonging to the same signal across all the frequency bins before applying the inverse STFT otherwise the time domain signals are still mixtures of the desired signals. Since our proposed semi-blind method is derived from the iterative INFOMAX method [3], we briefly present this method (see review [1] for reference to other methods). In the frequency bin f at the kth iteration, the separation equation is

$$Y^{(k)}(f,t) = B^{(k)}(f)V(f,t) \tag{3}$$

The mutual information of $Y^{(k)}(f,t)$ is minimized by updating the matrix $B(f)$ with the following rule (the frequency and frame indexes were dropped due to space limitation)

$$B^{(k+1)} = B^{(k)} + \mu(I - < \Phi(Y^{(k)})Y^{(k)H} >_t)B^{(k)} \tag{4}$$

where $< \cdot >_t$ denotes frame averaging and $\Phi(\cdot)$ denotes the vector of score functions. For $Y = [y_1, \ldots, y_p]^T$ this vector is defined by

$$\begin{aligned} \Phi(Y) \quad &= \quad [-\frac{\partial}{\partial y_1} \log P_{y_1}(y_1), \ldots, -\frac{\partial}{\partial y_p} \log P_{y_p}(y_p)]^T \\ &= \quad [\phi(y_1), \ldots, \phi(y_p)]^T \end{aligned}$$

where $P_{y_i}(y_i)$ is the probability density function of y_i. In practice the score functions are unknown and should be estimated from the data or prior knowledge on the signal densities is available.

3. PROPOSED METHOD

3.1. Block structure

The goal of the proposed semi-blind approach is also to recover some unknown signals when only some mixtures of these signals are available. However, contrary to the fully blind separation case, we are also given an additional information about the observed mixtures. We know that the mixing process has the following block structure

$$\begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix} = \begin{bmatrix} A(f) & B(f) \\ 0 & C(f) \end{bmatrix} \begin{bmatrix} S(f,t) \\ N(f,t) \end{bmatrix}. \tag{5}$$

The observed signals and the sources are both partitioned in two vectors. The first part of the observations $X(f,t)$, of size $(p \times T)$ with T the number of frame, is a mixtures of both $S(f,t)$ $(p \times T)$ and $N(f,t)$ $(q \times T)$ whereas the second part of the observations $R(f,t)$ $(q \times T)$ is only a function of $N(f,t)$. This structure corresponds to the situation described in fig.1, with p external signals and q internal noises. In the following we use the terms references for $R(f,t)$ and observations for $X(f,t)$. A diagram of the mixing is given in Fig.3.

The proposed demixer has a block structure of compatible dimensions with the matrices $A(f)$, $B(f)$ and $C(f)$.

$$\begin{bmatrix} Y(f,t) \\ Q(f,t) \end{bmatrix} = \begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

Compared to the blind problem of same dimension, the number of coefficients to update is reduced.

Using the results in [2] presented in Sect.2, the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent if and only if the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$ are such that

$$\begin{bmatrix} W_1(f) & W_2(f) \\ 0 & W_3(f) \end{bmatrix} \begin{bmatrix} A(f) & B(f) \\ 0 & C(f) \end{bmatrix} = \begin{bmatrix} P_1(f)\Lambda_1(f) & 0 \\ 0 & P_2(f)\Lambda_2(f) \end{bmatrix}$$

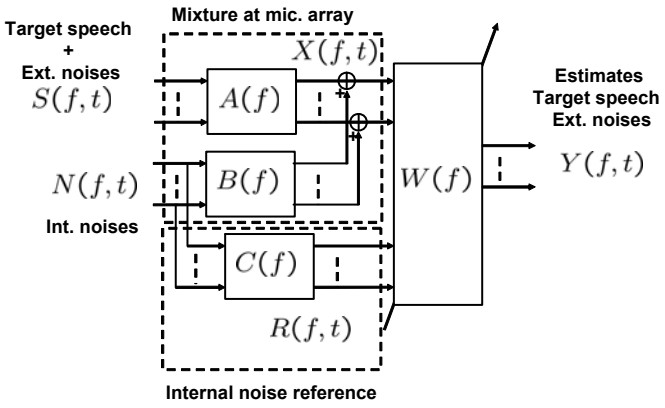

Fig. 3. Block structure of the mixture.

where $P_1(f)(p \times p)$ and $P_2(f)$ $(q \times q)$ are permutation matrices and $\Lambda_1(f)(p \times p)$ and $\Lambda_2(f)$ $(q \times q)$ are diagonal matrices. Consequently it is possible to estimate the components of $S(f,t)$ and $N(f,t)$ by updating $W_1(f)$, $W_2(f)$ and $W_3(f)$ until the components of $Y(f,t)$ and $Q(f,t)$ are all statistically independent (Note that an echo canceler [4] cancels the contribution of $N(f,t)$ in $X(f,t)$ but does not recover $S(f,t)$).

3.2. Proposed algorithm

The proposed semi-blind separation method uses the mutual information of $Y(f,t)$ and $Q(f,t)$ to measure the statistical independence of their components. The criterion is optimized by an iterative gradient descent on the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$. At iteration k, we have the following unmixing system

$$\begin{bmatrix} Y^{(k)}(f,t) \\ Q^{(k)}(f,t) \end{bmatrix} = \begin{bmatrix} W_1^{(k)}(f) & W_2^{(k)}(f) \\ 0 & W_3^{(k)}(f) \end{bmatrix} \begin{bmatrix} X(f,t) \\ R(f,t) \end{bmatrix}.$$

To obtain the update rules for these matrices we rewrite the update rule in the blind case eq.(4) with the proposed demixer structure

$$\begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix} = \begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix} - \mu \left(I_{p+q} - < \begin{bmatrix} \Phi(Y^{(k)}(f,t)) \\ \Phi(Q^{(k)}(f,t)) \end{bmatrix} \begin{bmatrix} Y^{(k)}(f,t) \\ Q^{(k)}(f,t) \end{bmatrix}^H >_t \right) \times \begin{bmatrix} W_1^{(k+1)}(f) & W_2^{(k+1)}(f) \\ 0 & W_3^{(k+1)}(f) \end{bmatrix}.$$

Then the update rules for the matrices $W_1(f)$, $W_2(f)$ and $W_3(f)$ are extracted (A semi-blind method for instantaneous mixtures in the time domain uses the same approach to get the update rules in [5]). The update rules for the matrices have the following form

$$W_j^{(k+1)}(f) = W_j^{(k)}(f) + \mu \Delta W_j^{(k)}(f)$$

where (dropping the frequency and frame indexes for $Y(f,t)$ and $Q(f,t)$)

$$\begin{aligned} \Delta W_1^{(k)}(f) &= \left(I - < \Phi(Y^{(k)}) Y^{(k)H} >_t \right) W_1^{(k)}(f) \\ \Delta W_2^{(k)}(f) &= \left(I - < \Phi(Y^{(k)}) Y^{(k)H} >_t \right) W_2^{(k)}(f) \\ &\quad - \left(< \Phi(Y^{(k)}) Q^{(k)H} >_t \right) W_3^{(k)}(f) \\ \Delta W_3^{(k)}(f) &= \left(I - < \Phi(Q^{(k)}) Q^{(k)H} >_t \right) W_3^{(k)}(f). \end{aligned}$$

The frequency domain signals are approximately circular because they were obtained by a STFT. For a circular random variable $y = |y| \mathbf{e}^{\mathbf{j} \arg y}$ we have

$$\phi(y) = \phi(|y|) \mathbf{e}^{\mathbf{j} \arg y}$$

Thus the unknown score functions can be estimated from the data using a kernel based estimate of the score function of their modulus.

After the semi-blind separation is performed in all the frequency bins, the permutation resolution is also simplified because of the block structure.

4. EXPERIMENTAL RESULTS

To demonstrate the importance of the internal noise reference we performed some experiments mixing the noise recorded in a train station as external noise and a synthetic non stationary noise as internal noise. The impulse response of the train station hall was also measured for a speaker at 50cm in front of a four microphone array (inter mic. spacing is 2.15cm). 200 Japanese sentences of different length were used as speech signals (2s to 14s at 16kHz from the JNAS database [6]). The observed signals are obtained in two steps. First a speech signal convoluted by the impulse response is mixed with the recorded noise. The SNR in this mixture is called SNR ext. Then the mixed speech and external noise is mixed with the internal noise that is filtered by a low pass filter. The SNR for this second mixture is SNR int. We also filter the internal noise to obtain the reference.

In all experiments we compared the iterative INFOMAX approach (blind) to the proposed approach (semi-blind). The STFT is performed with a 512 points hanning window with 256 points overlap. The matrices $B(f)$ are initialized to identity in all frequency bins then 200 iterations are performed with an adaptation step $\mu = 0.1$. The speech signal is selected out of the separated components in all the frequency bins using the same method for both approach. The INFOMAX method considers the reference signal as a fifth observation. Then both algorithms have the same amount of statistical information. The only difference is that the semi-blind approach knows that the mixture has the block structure showed in Fig.3.

The estimation quality is measured in term of noise reduction rate (NRR) defined as the difference of the SNR for

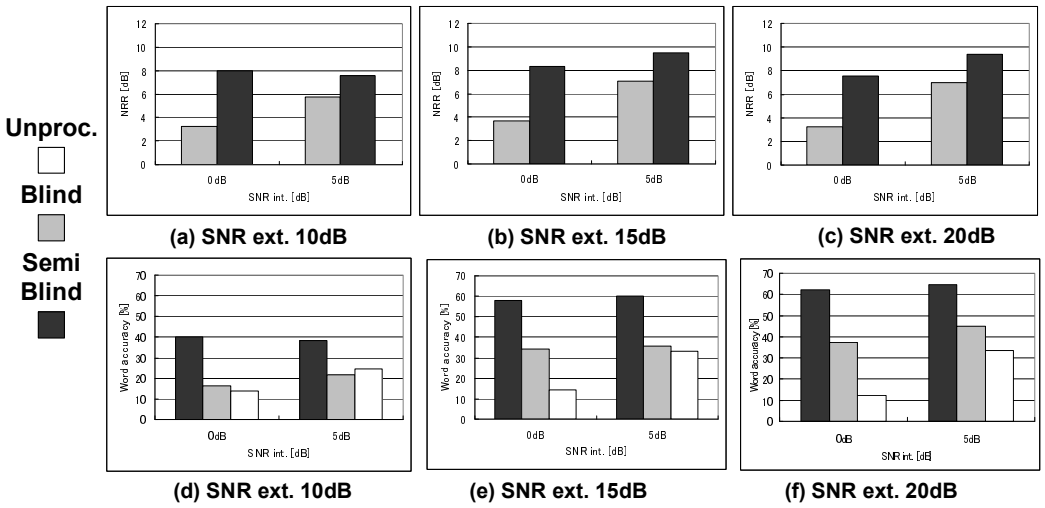

Fig. 4. Noise rate reduction (NRR) and word accuracy for different SNRs.

the speech estimates (after processing) and the SNR for the observations (before processing). Consequently, a positive NRR means that the speech estimate quality is improved. Figures 4(a), (b) and (c) show the NRR for mixtures at different SNRs (averaged on the 200 test signals). The second measure of performance is the word accuracy for a continuous speech recognition task. The speech recognition conditions are given in table 1 and the results in Figs. 4(d), (e) and (f).

The blind method is able to improve the speech signal but using the block structure gives the advantage to the semi-blind method when the number of iterations is limited. The performance of the blind method would increase if the number of iterations is larger but in a real situation computation time is limited. The performance difference is also larger for the shorter sentences.

Table 1. Conditions for speech recognition

Task	20k word newspaper dictation
Acoustic model	phonetic tied mixture, clean model [7]
Acoustic model training	260 speakers (150 sentences/speaker)
Decoder	JULIUS ver 3.2 [7]

5. CONCLUSION

In this paper we proposed a semi-blind separation approach that operates in the frequency domain. The method easily incorporates the information given by additional sensors to the BSS based approach. Experiments showed that this can be very beneficial in a hands-free speech recognition scenario.

6. REFERENCES

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

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

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

[4] J. Benesty et al., "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 156–165, 1998.

[5] M. Joho et al., "Combined blind/nonblind source separation based on the natural gradient," *IEEE Signal Processing Letters*, vol. 8, no. 8, pp. 236–238, 2001.

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

[7] A. Lee et al., "Julius - an open source real-time large vocabulary recognition engine," *EUROSPEECH,*, pp. 1691–1694, 2001.