FAST CONVOLUTIVE BLIND SPEECH SEPARATION VIA SUBBAND ADAPTATION

François Duplessis-Beaulieu and Benoît Champagne

Department of Electrical and Computer Engineering, McGill University
3480 University Street, Montréal, CANADA, H3A 2A7
fduplessis@tsp.ece.mcgill.ca, champagne@ece.mcgill.ca

ABSTRACT

In this paper, we consider the problem of *blind source separation* (BSS) applied to speech signals. Due to reverberation, BSS in the time domain is usually expensive in terms of computations. We propose in this paper a subband BSS system based on the use of adaptive feedback de-mixing networks in an oversampled uniform DFT filter bank structure. We show that the computational cost can be significantly decreased if BSS is carried out in subbands due to the possibility of reducing the sampling rate. Experiments with real speech signals, conducted with two-input two-output BSS systems using oversampled 32-subband and fullband adaptation, indicate that separation quality and distortion are similar for both systems. However, the proposed subband system is more than 10 times computationally faster than the fullband one.

1. INTRODUCTION

Blind source separation (BSS) attempts to recover a set of sources from a set of mixtures, without knowing the physical realization of the original sources, and how they were mixed in the first place. In this work, we are mainly interested in convolutive speech separation, where the aim is to isolate multiple speech sources captured by two or more microphones in a reverberant room. However, most of the ideas exposed here can easily be extended to other types of signals and convolutive situations as well.

Many BSS algorithms operating in the time domain have been proposed that can be applied to convolutive speech mixtures, e.g. see [1], [2]. Unfortunately, these algorithms are characterized by heavy computational requirements. A popular approach to reduce the computational burden consists in carrying out separation in the frequency domain [3]. Frequency domain BSS is attractive because each frequency bin corresponds to an instantaneous BSS problem. Moreover, combined with efficient block FFT processing, a computationally fast BSS algorithm can be derived.

Nevertheless, frequency domain BSS has its own drawbacks. Due to the difficulty in calculating the inverse of the room transfer function both for the desired and interfering signals [4], frequency domain BSS may fail to separate mixtures when the direct path is weak. Another problem is that for each frequency bin, the recovered sources may be permuted and must be reordered before the fullband signal is reconstructed, a difficult task [3]. Finally, block processing with FFT may introduce unacceptable delays in certain applications.

In this paper, we investigate a novel BSS system based on the subband realization of the adaptive feedback de-mixing network in [2]. The computational efficiency of this approach results from the possibility of reducing the sampling rate within each subband. To avoid decimation aliasing effects, the proposed system uses flexible oversampled uniform DFT filter banks that allow arbitrary oversampling, low complexity of implementation and low processing delay. The proposed subband BSS approach is evaluated with real speech signals recorded in a reverberant room in the case of a two-input two-output mixing situation. One particular realization of the subband BSS approach using 32 subbands is shown to run more than 10 times faster than its fullband counterpart without any noticeable loss in separation and distortion performance. Furthermore, the subband implementation is not affected by the above mentioned limitations of frequency domain BSS.

This work was supported in part by a grant form the Natural Sciences and Engineering Research Council of Canada.

2. BACKGROUND

Let us consider the situation where N persons are speaking in a room, with respective speech signals $s_i(n), 1 \leq i \leq N$, where n denotes discrete-time index. After propagation within the room, these signals are captured by N microphones, which pick up mixtures of those speakers, denoted $x_i(n), 1 \leq i \leq N$. The goal of BSS is to recover the source signals by processing all available mixtures. The resulting signals are denoted $y_i(n), 1 \leq i \leq N$.

In the time domain, it has been proposed to use a feedback de-mixing network for BSS of speech mixtures [2], as illustrated in Fig. 1 for a two-input two-output network. The output of the network can be expressed as follows:

$$y_i(n) = x_i(n) + \sum_{j \neq i} \boldsymbol{w}_{ij}^T \boldsymbol{y}_j(n), \tag{1}$$

where $\boldsymbol{y}_j(n) = [y_j(n), \ldots, y_j(n-L+1)]^T$ and $\boldsymbol{w}_{ij} \in \mathbb{R}^L$ represents an FIR filter of length L. The coefficients of $\boldsymbol{w}_{ij}$ can be updated using a steepest descent approach. As shown in [2], the update equation is

$$\boldsymbol{w}_{ij}(n+1) = \boldsymbol{w}_{ij}(n) - \mu \, \mathrm{sign}(y_i(n)) \, \boldsymbol{y}_j(n), \tag{2}$$

where $\mu > 0$ is a step size.

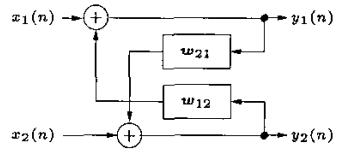

Fig. 1. Feedback convolutive de-mixing network for a two-input two-output system.

The time domain BSS algorithm described above is expensive in terms of computations. Indeed, for an $N \times N$ mixing network

(i.e. N sources and N microphones), it requires $2N(N-1)L$ real multiplications per time iteration. For the algorithm to operate adequately, the value of L must be comparable with the reverberation time of the acoustic enclosure (in samples). For example, in the application of BSS to speech mixtures recorded in a room with a small reverberation time of 50 ms at a basic sampling rate of 11 kHz, the required value of L is of the order of 550. For applications within enclosures with large reverberation times and higher sampling rate (e.g. high-quality audio), the required value of L may be significantly larger. Clearly, the computational complexity of $O(N^2L)$ required by the time-domain BSS algorithm (1)–(2) might be too high for practical real-time BSS processing.

In this work, we propose a subband-based BSS algorithm which lowers the computational complexity. As detailed in Sec. 3 and 4, the processing time can be significantly reduced by using subband instead of time-domain separation.

3. DFT FILTER BANKS FOR BSS

The proposed subband realization of BSS using oversampled uniform filter banks is illustrated in Fig. 2. Its operation is further described below.

In this approach, each microphone signal $x_i(n)$ $(i = 1, \ldots, N)$ is separated into K subbands by means of digital filters h_k, each of them corresponding to a certain frequency range of equal width. Since the bandwidth of the signal in each subband is now reduced, the sampling rate can be lowered by a factor $M \leq K$, resulting in the subband microphone signal $x_{ki}(m)$ $(k = 1, \ldots, K)$, where m denotes discrete-time index at the reduced sampling rate. The mixtures are then independently separated in each subband using a separate adaptive BSS algorithm operating at the reduced sampling rate. After the separation, the processed signals $y_{ki}(m)$ are digitally interpolated by a factor M and filtered by g_k so that their sampling rate is restored at its original value. The signals coming from the different subbands that belong to the same source are finally added together to reconstruct the fullband signal $y_i(n)$, which provides an approximation to the i-th speech source.

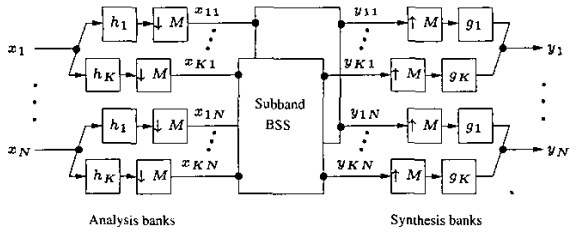

Fig. 2. A subband-based BSS system.

Filter bank requirements in the subband BSS application under study differ significantly from those commonly found in subband coding applications. Due to the presence of the separation network structure between the analysis and synthesis banks in Fig. 2, using critically sampled filter banks ($M = K$) would produce aliased audio components which are difficult to filter out [5]. For this reason, we propose to use oversampling (i.e. $M < K$) in subband to keep aliasing distortion in the fullband outputs of the synthesis banks below an acceptable level. The decimation factor M should be set to a value compatible with audio applications, while attempting to preserve computational efficiency of the subband approach. We also note that the perfect reconstruction property (PR) is not essential in subband BSS, as long as the reconstruction errors are inaudible. Accordingly, only a near-PR property is required for the analysis/synthesis banks.

In this work, we use modified uniform DFT filter banks, as described in [6], for the realization of the analysis/synthesis banks in Fig. 2. This approach has proved to be effective in the application of subband acoustic echo cancellation, where the filter bank requirements are similar to those described above. The operation of the DFT filter banks is illustrated in Fig. 3. In the analysis bank (top), the spectrum of the signal in subband k is shifted by $2\pi(k-1)/K$ to the left using a complex modulation. After modulation, the signal is filtered with $h(n)$, an FIR low-pass prototype filter of length D with cutoff frequency π/K, and then decimated by a factor M. In the synthesis bank (bottom), the signal in each subband is upsampled by M, and filtered by the synthesis prototype low-pass filter $g(n)$, also characterized by a length D and a cutoff frequency of π/K. After filtering, the subband signals are properly demodulated and added together so that a fullband signal is obtained.

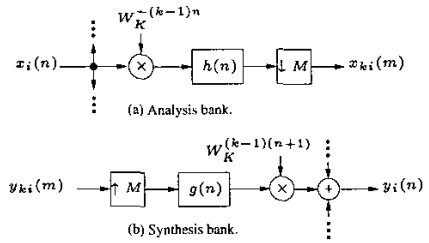

Fig. 3. DFT filter banks (for analysis and synthesis).

The analysis prototype filter $h(n)$ is designed according to the procedure described in [6], which amounts to interpolate the impulse response of a tabulated quadrature mirror filter (QMF) by a factor of $K/2$. To eliminate phase distortion, the filter length D is constrained to be a multiple of the number of subband, K, and the synthesis prototype filter $g(n)$ is chosen such that $g(n) = h(D-n-1), n = 0, \ldots, D-1$.

If one implements the DFT filter bank straightforwardly as illustrated in Fig. 3, the overhead generated by subband analysis and synthesis can become expensive, considering that N pairs of analysis/synthesis banks are needed in the present subband BSS application. The weighted overlap-add (WOA) method [7] provides a more efficient realization. The WOA method interprets the DFT filter bank as a block transform, and uses the FFT to optimize the computations. The WOA method is fully described in [7], and its specialization to the above modified uniform DFT filter banks is considered in [6].

4. BSS USING COMPLEX SUBBAND ADAPTATION

Consider the subband realization of BSS using oversampled DFT filter banks, as illustrated in Fig. 2. In this work, we investigate the use of adaptive feedback de-mixing networks to separate the sources in each subband.

The basic feedback de-mixing network was described by (1)–(2) and illustrated in Fig. 1 for the case of for *real fullband signals*. Here, due to the use of complex modulation in the filter banks, the input signals in each subband, i.e. $x_{ki}(m)$, now becomes complex valued (except for subband 1 and $K/2+1$). Therefore, besides trivial changes in notation to accommodate subband index k and a

different discrete-time index m, further modifications are needed in the feedback de-mixing network equations (1)–(2) before they may be applied to complex subband signals.

To simplify the notation, let $\boldsymbol{y}_{ki} = [y_{ki}(m), \ldots, y_{ki}(m - L + 1)]^T$, $\boldsymbol{a}_{kij} = \boldsymbol{w}^R_{kij}$, and $\boldsymbol{b}_{kij} = \boldsymbol{w}^I_{kij}$, where superscript T denotes matrix transposition and superscripts R and I denote the real and imaginary parts, respectively. Hence, the output of the feedback de-mixing network can be written as follows:

$$y^R_{ki}(m) = x^R_{ki}(m) + \sum_{j \neq i} (\boldsymbol{a}^T_{kij} \boldsymbol{y}^R_{kj}(m) + \boldsymbol{b}^T_{kij} \boldsymbol{y}^I_{kj}(m)), \quad (3)$$

$$y^I_{ki}(m) = x^I_{ki}(m) + \sum_{j \neq i} (\boldsymbol{a}^T_{kij} \boldsymbol{y}^I_{kj}(m) - \boldsymbol{b}^T_{kij} \boldsymbol{y}^R_{kj}(m)). \quad (4)$$

According to [1], the cost function for BSS using a feedback de-mixing network, denoted ϕ, is given by

$$\phi = \sum_{i=1}^{N} \log p_i\big(y_i(n)\big), \quad (5)$$

where $p_i(\cdot)$ is the hypothesized pdf of source i, corresponding to real-valued baseband speech. In recent work, a Laplacian distribution is often used for the baseband speech samples [2]. To handle the case of a complex valued subband speech source, we must introduce an alternative pdf in the cost function (5) that is defined over the complex plane and properly represents the statistics of the source samples.

Through statistical analysis of experimentally collected subband speech samples at the output of the DFT analysis banks, we have been able to verify that the phase of subband speech is uniformly distributed. This can be justified from the central limit theorem on the basis of the complex modulation and FIR filtering involved in the subband decomposition (see Fig. 3(a)). Hence, the proposed pdf should only depend on the magnitude of the complex random variable.

The experimental magnitude distribution of subband speech is illustrated in Fig. 4. Points on this figure were generated by computing the histogram of the magnitude of subband speech samples using narrow bins. Also shown in Fig. 4 is the Gamma distribution $\Gamma(\alpha, 2)$ with parameter $\alpha = 125$ adjusted to match the distribution of the experimental data. It can be seen that the Gamma distribution provides a reasonable fit to the collected data, although a more accurate model could be found. However, the Gamma distribution leads to a simple form of the weight update equation when used in connection with the cost function (5), and the resulting algorithm exhibits a performance similar to that of fullband BSS (see Sec. 5).

Therefore, a suitable model for BSS of complex subband speech is proposed as follows:

$$p_i\big(y^R_{ki}(m), y^I_{ki}(m)\big) = \frac{\alpha^2}{2\pi} e^{-\alpha \sqrt{\left(y^R_{ki}(m)\right)^2 + \left(y^I_{ki}(m)\right)^2}} \quad (6)$$

where α is a positive constant for normalization purposes. We shall refer to the above as a complex Laplacian pdf.

Substituting (6), the pdf of subband speech, in (5) yields the following cost function

$$\phi = -\alpha \sum_{i=1}^{N} \sqrt{\left(y^R_{ki}(m)\right)^2 + \left(y^I_{ki}(m)\right)^2} + c \quad (7)$$

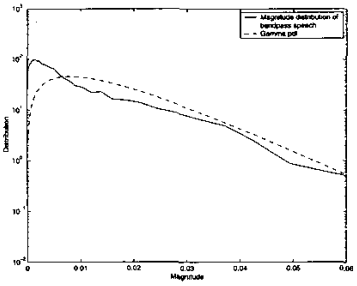

Fig. 4. Experimentally observed magnitude distribution of bandpass speech samples taken from subband 4 of a 16-subband filter bank versus a Gamma pdf.

where c is a constant. Note that the above cost function is real valued, but the filter coefficients to adapt are complex valued. Maximization of the cost function can be done using a complex gradient operator, commonly defined as

$$\nabla\phi = \frac{1}{2}\left(\frac{\partial\phi}{\partial \boldsymbol{a}_{kij}} + j\frac{\partial\phi}{\partial \boldsymbol{b}_{kij}}\right), \quad (8)$$

Applying definition (8) to (7) results in the following expression for its gradient:

$$\nabla\phi = -\frac{\alpha}{2}\frac{y^*_{ki}(m)}{\|y_{ki}(m)\|}\boldsymbol{y}_{kj}. \quad (9)$$

Hence, using a feedback network, the complex coefficients of the separating filters can be updated with the following stochastic gradient algorithm

$$\boldsymbol{w}_{kij}(m+1) = \boldsymbol{w}_{kij}(m) - \mu\frac{y^*_{ki}(m)}{\|y_{ki}(m)\|}\boldsymbol{y}_{kj}, \quad (10)$$

where $\mu = \mu'\alpha/2 > 0$ is the step size.

Note that since the sampling rate has been reduced by a factor of M in the subband, the number of taps we would use for fullband adaptation, L, can now be set to L/M for subband adaptation. This way, the same physical time span is covered by the subband and fullband adaptive BSS filters. Also note that due to the real nature of the microphone signals in this applications, subband signals $x_{ki}(m) = x_{k'i}(m)$, where $k' = K + 1 - k$. Therefore, BSS processing needs only be applied in subband 1 to $K/2 + 1$.

The computational complexity of the subband BSS system can be evaluated by counting the number of required multiplications per time iteration at the high sampling rate, assuming an $N \times N$ mixing network. If we take into account the savings resulting from the symmetry in the subband signals, as well as the overhead generated by the filter banks, the number of real multiplications per iteration for a system using a WOA realization can be found in Tab. 1. A computational gain γ is defined by considering the amount of computations needed for fullband adaptation versus subband adaptation:

$$\gamma = \frac{2}{3}\frac{M^2}{K}\left[1 + \frac{2}{3}\frac{M}{(N-1)L}\left(\frac{D}{K} + \log_2 K\right)\right]^{-1} \quad (11)$$

	Real multiplications
WOA	$2N(D + K\log_2 K)/M$
Output generation	$[2N(N-1)LK]/M^2$
Weight update	$[N(N-1)LK]/M^2$

Table 1. Number of real multiplications per iteration (at the high sampling rate) for subband BSS.

5. EXPERIMENTAL RESULTS

Schobben *et al.* proposed in [8] the following evaluation method for BSS systems. The idea is to record the sources in a real-world environment, while letting only one of them active at a time. The mixtures are then obtained by adding all contributions together, i.e.

$$x_i(n) = \sum_{j=1}^{N} \xi_{i,j}(n), \quad i = 1, \ldots, N, \tag{12}$$

where $\xi_{i,j}(n)$ denotes the contribution of speaker j to microphone i. Similarly, $\eta_{i,j}(n)$ denotes the recovered source i when speaker j is active. Separation quality is given by the following power ratio in decibels

$$S_i = 10\log\left(\frac{E\left[\eta_{i,i}^2(n)\right]}{E\left[\sum_{j\neq i}\eta_{i,j}^2(n)\right]}\right), \tag{13}$$

and distortion can be measured using

$$D_i = 10\log\left(\frac{E[(\xi_{i,i}(n) - \lambda_i y_i(n-d))^2]}{E[\xi_{i,i}^2(n)]}\right), \tag{14}$$

where $\lambda_i = E[\xi_{i,i}^2(n)]/E[y_i^2(n)]$, and d is a delay introduced by the BSS system.

We have tested the performance of the subband and fullband BSS systems for a two-input two-output mixing network. Speech signals for these tests were recorded with two omni-directional microphones in a small office room. The microphone outputs were sampled at 11 kHz using 16-bit precision. Results, obtained for three different sets of speaker positions labelled a, b and c, are summarized in Tab. 2. The strongest direct path is obtained in Position a, whereas the strongest cross-talk can be found in Position c. Position b offers a stronger cross-talk than a, but weaker than c. Two numbers, separated by a '/', are given for each entry. These numbers correspond to the first and second source. The performance rates reported in Tab. 2 were computed after 8 iterations (i.e. each set of files was processed 8 times). The step size μ was set to 1×10^{-5} and to 5×10^{-2} for the fullband and subband system, respectively. The prototype filter for the $K = 32$ subbands system was obtained by interpolating a QMF 12A (tabulated in [7]) by a factor of 16 (thus $D = 192$), and the downsampling factor M was set to 24. The choice of $M = 24$ was justified experimentally by varying M and looking at the various distortion measures D_i. $M = 24$ was the highest M that produces a low distortion output.

We may note from Tab. 2 that the position of the speakers relative to the microphones has a direct influence on the performance rates. Better separation and lower distortion are obtained when the direct path is strong, and cross-talk is weak. Furthermore, both BSS systems exhibit about the same performance. Permutations of the recovered sources, a noticeable problem with FFT based frequency domain BSS, were not observed in our experiments with subband BSS.

	Pos.	Fullband BSS	32-subband BSS
Separation (dB)	a	8.37/6.33	9.51/7.90
	b	6.73/3.82	8.41/4.73
	c	6.08/4.28	6.34/6.17
Distortion (dB)	a	-8.02/-6.35	-8.64/-7.00
	b	-6.92/-4.44	-7.95/-4.92
	c	-6.31/-4.74	-5.60/-5.13

Table 2. Performance of BSS systems ($L = 1152$).

Computational complexity is given in Tab. 3. These results represent the average processing time (one iteration) of 10-second speech files, and were obtained with the same parameters used to generate the results in Tab. 2. A *Pentium 3* processor clocked at 933 MHz was used to carry out the computations. Gains in terms of processing times with respect to the fullband system are also given. The theoretical gain γ is obtained according to (11).

	Time (s)	Real gain	γ
Fullband BSS	60.1	–	–
32-subband BSS	5.4	11.1	10.4

Table 3. Computational speed and time gains.

The 32-subband BSS system with WOA realization is about 11 times faster than the corresponding fullband system. Hence, the goal of reducing the number of computations is reached, without a significant impact on the separation and distortion rates.

6. REFERENCES

[1] K. Torkkola, "Blind separation of convolved sources based on information maximization," in *Proc. IEEE Workshop on Neural Networks for Signal Processing*, Kyoto, Japan, Sept. 1996, pp. 423 – 432.

[2] A. Koutras, E. Dermatas, and G. Kokkinakis, "Blind speech separation of moving speakers in real reverberant environments," in *Proc. ICASSP*, Istanbul, Turkey, June 2000, vol. 2, pp. 1133 – 1136.

[3] K. Torkkola, "Blind separation for audio signals – are we there yet?," in *Proc. of the ICA '99 Workshop*, Aussois, France, Jan. 1999, pp. 239 – 244.

[4] S. Araki, S. Makino, T. Nishikawa, and H. Saruwatari, "Fundamental limitation of frequency domain blind source separation for convolutive mixture of speech," in *Proc. ICASSP*, Salt Lake City, USA, May 2001, vol. 5, pp. 2737 – 2740.

[5] F. Duplessis-Beaulieu, "Fast convolutive blind speech separation via subband adaptation," M.S. thesis, McGill University, Sept. 2002.

[6] Q.-G. Liu, B. Champagne, and D. K.C. Ho, "Simple design of oversampled uniform DFT filter banks with applications to subband acoustic echo cancellation," *Signal Processing*, vol. 80, pp. 831–847, June 2000.

[7] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing*, Prentice-Hall, 1983.

[8] D. Schobben, K. Torkkola, and P. Smaragdis, "Evaluation of blind signal separation methods," in *Proc. of the ICA '99 Workshop*, Aussois, France, Jan. 1999, pp. 261–266.