A ROBUST VOICE SEPARATION METHOD

Yijing Chu[1], Heping Ding[2], and Xiaojun Qiu[3]
[1,3]Key Laboratory of Modern Acoustics and Institute of Acoustics, Nanjing University, 210093, China
[2]Acoustics and Signal Processing, IMS, National Research Council, 1200 Montreal Rd., Ont., Canada K1A 0R6
[1]zhuyijing@nju.org.cn [2]heping.ding@nrc-cnrc.gc.ca [3]xjqiu@nju.edu.cn

1. Introduction

It is often needed to separate mixed voice signals, e.g., in a cocktail party environment. Blind source separation (BSS) techniques based on the independent component analysis [1] and so on work well in simulations where the source signals are independent and identically distributed (i.i.d.), but perform poorly in real life because actual sources are hardly i.i.d.. In addition, other practical issues such as complexity, convergence and tracking rate, and noise immunity also need to be resolved in the existing techniques.

In [3], a robust quasi-BSS method that addresses the above-mentioned practical issues was proposed. Assuming there are two voice sources and two microphones (the numbers can be generalized but are both two here for simplicity) in the room, Fig. 1 shows the signal flow diagram of the method, where the sources $s_1(n)$ and $s_2(n)$ acoustically go through linear mixing filters $\{H_{ij}, \forall i,j=1,2\}$ of the room before reaching the microphones. In the processing realm, the microphone signals $x_1(n)$ and $x_2(n)$ are fed to separation filters $\{W_{ij}, \forall i,j=1,2\}$, which are so chosen that $s_1(n)$ and $s_2(n)$ appear maximally separated in the outputs $u_1(n)$ and $u_2(n)$.

The key therefore is to find a solution to $\{W_{ij}, \forall i,j=1,2\}$. According to [3], such a solution exists if H_{1j} and H_{2j} $(j=1,2)$ share no common zeros and the noises in $x_1(n)$ and $x_2(n)$ are i.i.d. and independent of the sources. Next, [3] shows that a solution to $\{W_{j1}, W_{j2}\}$ $(j=1,2)$ can be estimated by minimizing

$$J_j(n) = \overline{u_j^2(n)} \ , \qquad (1)$$

a certain kind of time-average of the squares of $u_j(n)$ samples (or energy level of $u_j(n)$), during one source active (1SA) time periods where source $s_{3-j}(n)$ is active and $s_j(n)$ is silent. Furthermore, the solution can be found adaptively by using the affine projection algorithm (APA) [4] with a constraint that the first element of the impulse response in $\{W_{1j}, j=1,2\}$ be unity.

Simulations based on real audio recordings made in a typical listening room (RT30 at 1 kHz around 0.3 s), 8 kHz sampling rate, and each of $\{W_{ij}, \forall i,j=1,2\}$ having 800 coefficients [3] reveal that the proposed quasi-BSS method outperforms conventional methods, such as that in [2], in terms of convergence and tracking rate, achievable signal to interference ratio, and immunity to ambient noise.

Nevertheless, [3] did not give a complete solution to the BSS problem. In order to function properly, the proposed quasi-BSS method relies on

1. the condition that there are 1SA periods available, and
2. reliable detection of the 1SA periods.

In fact, "1" is not regarded as a problem since 1SA periods are almost always present in real human conversations, and "2" is a practical issue that [3] did not address.

In the following, we present a simple and effective scheme that detects the 1SA periods and controls the adaptation for $\{W_{ij}, \forall i,j=1,2\}$, so as to complete the work in [3] and form a complete BSS solution.

2. Detection and Control Scheme

Our proposed scheme is based on the proposition that, in general, a solution to $\{W_{j1}, W_{j2}\}$ $(j=1,2)$, which makes (1) satisfactorily low, only exists during a 1SA period. Given that, we introduce a pair of detection filters $\{D_1, D_2\}$, which are of the same structure as any $\{W_{j1}, W_{j2}\}$ $(j=1 \text{ or } 2)$ and are connected the same way as the latter, as shown in Fig. 2.

For fast convergence and tracking, $\{D_1, D_2\}$ always adapt using the APA at a large step size. The output energy level

$$J_D(n) = \overline{u_D^2(n)} \qquad (2)$$

is constantly monitored. As discussed above, (2) stays high as long as both sources $s_1(n)$ and $s_2(n)$ are active. The fact that (2) starts to drop is an indication of a 1SA period having been entered and $\{D_1, D_2\}$ converging on it. When $\{D_1, D_2\}$ have converged during the 1SA period, indicated by (2) having dropped below a certain threshold T, the coefficients of $\{D_1, D_2\}$ are copied to a pair of separation filters $\{W_{j1}, W_{j2}\}$ $(j=1 \text{ or } 2)$, which then start to produce the output $u_j(n)$ and adapt at a low rate (small step size) to track small variations (if any) in the mixing filters. When another

1SA (the other source active) period is entered, indicated by both (2) and $J_j(n)$ jumping, we stop adapting $\{W_{j1}, W_{j2}\}$ to prevent them from diverging. When (2) falls again – result of $\{D_1, D_2\}$ converging on this new 1SA period – we copy the $\{D_1, D_2\}$ to $\{W_{3-j,1}, W_{3-j,2}\}$, which then start to produce the output $u_{3-j}(n)$ and slowly adapt – as $\{W_{j1}, W_{j2}\}$ did earlier. From now on, $\{W_{j1}, W_{j2}\}$ $(j = 1 \text{ or } 2)$ adapt if $J_j(n)$ is small, and is frozen otherwise. Also at this point, $\{D_1, D_2\}$ can be decommissioned if large variations in mixing filters $\{H_{ij}, \forall i, j = 1,2\}$ are not expected in the future.

In a nutshell, 1SA periods are detected and adaptation for $\{W_{ij}, \forall i, j = 1,2\}$ is controlled based on (1) $(j = 1,2)$ and (2).

3. Experimental Results

This proposed detection and control scheme has been applied to the method in [3] to process real audio recordings described in Section 1 (details in [3]).

The results are illustrated in Fig. 3. The microphone signals $x_1(n)$ and $x_2(n)$ consist of three consecutive periods: $s_1(n)$ active only, $s_2(n)$ active only, and both $s_1(n)$ and $s_2(n)$ active. At each of the two moments marked by the vertical lines, (2) becomes less than a threshold T. This triggers a copying action from $\{D_1, D_2\}$ to certain $\{W_{j1}, W_{j2}\}$, which then start to perform the mission task, as discussed in Section 2.

It can be seen that the proposed 1SA period detection and adaptation control scheme is quite effective and fast-acting. Audio demos showing the separation results under a couple of ambient noise levels will be played during the presentation of this paper.

4. Conclusion

This paper presents a scheme for controlling the adaptive filters associated with a previously proposed method for BSS. The scheme is proven to be effective and with a very small algorithmic overhead in implementation. The complete BSS solution can be further simplified by replacing the APA used with a fast affine projection (FAP), such as a one proposed in [5].

References

[1] S. Choi, A. Cichocki, H.M. Park and S.Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Information Processing - Letters and Reviews*, Vol. 6, No. 1, pp. 1–57, Jan. 2005

[2] L. Parra and C. Spence, "Convolutive Blind Separation of Non-Stationary Sources," *IEEE Trans. Speech and Audio Processing*, Vol. 8, No. 3, pp. 320–327, May 2000

[3] Y. Chu, H. Ding, and X. Qiu, "Quasi-Blind Source Separation Algorithm for Convolutive Mixture of Speech," *Proceedings of the 12th IEEE Digital Signal Processing Workshop*, pp. 233–238, Jackson Hole, WY, U.S.A., Sept. 2006

[4] H.C. Shin and A.H. Sayed, "Mean-square performance of a family of affine projection algorithms," *IEEE Trans. Signal Proc.*, Vol. 52, No. 1, pp. 90–102, Jan. 2004

[5] Heping Ding, "Fast Affine Projection Adaptation Algorithms with Stable and Robust Symmetric Linear System Solvers," *IEEE Transactions on Signal Processing*, Vol. 55, No. 5, pp. 1730–1740, May 2007

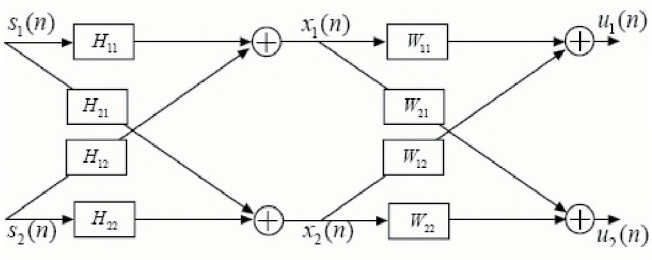

Fig. 1. Signal flow of proposed voice separation method [3]

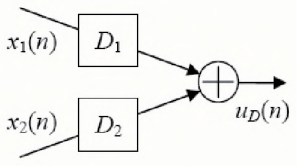

Fig. 2. Detection filters

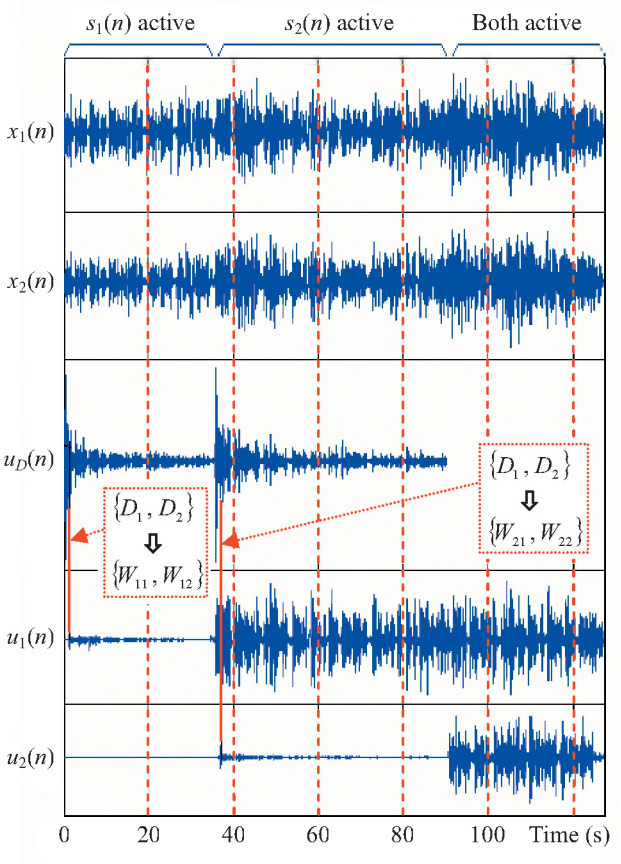

Fig. 3. The proposed detection and control scheme in action