Blind Source Separation and Visual Voice Activity Detection for Target Speech Extraction

Qingju Liu and Wenwu Wang
Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, UK
{Q.Liu,W.Wang}@surrey.ac.uk

***Abstract*—Despite being studied extensively, the performance of blind source separation (BSS) is still limited especially for the sensor data collected in adverse environments. Recent studies show that such an issue can be mitigated by incorporating multimodal information into the BSS process. In this paper, we propose a method for the enhancement of the target speech separated by a BSS algorithm from sound mixtures, using visual voice activity detection (VAD) and spectral subtraction. First, a classifier for visual VAD is formed in the off-line training stage, using labelled features extracted from the visual stimuli. Then we use this visual VAD classifier to detect the voice activity of the target speech. Finally we apply a multi-band spectral subtraction algorithm to enhance the BSS-separated speech signal based on the detected voice activity. We have tested our algorithm on the mixtures generated artificially by the mixing filters with different reverberation times, and the results show that our algorithm improves the quality of the separated target signal.**

***Index Terms*—Blind source separation, multimodal enhancement, visual voice activity detection, multi-band spectral subtraction**

I. Introduction

Many blind source separation (BSS) algorithms have been proposed to recover the unknown source signals from their mixtures, using techniques such as independent component analysis (ICA) [1], [2]. However, in an adverse room environment, audio mixtures collected at the sensors are usually deteriorated by long reverberations and strong background noise. The performance of BSS will be degraded considerably in such an adverse situation. Denoising techniques for post-processing can mitigate this problem and spectral subtraction is an efficient and widely used method [3], [4]. To estimate the spectral variance of noise from the noise-embedded signals, inactive period (when the target speaker is silent), i.e., the voice activity needs to be detected. However, the audio-domain voice activity detection (VAD) is vulnerable to noise. For a BSS-separated target speech signal, the existence of residual interference from other speakers makes accurate VAD even more difficult due to the nonstationarity of the interference.

However, the video signal associated with a target speaker is not affected by acoustic noise, providing information complementary to the audio speech. Studies have shown both the production and perception of human speech are bimodal [5], [6]. For example, in a cocktail party scenario, looking at the speaker's face, or more precisely the movements of the lip region, helps one to comprehend the speech of interest. The bimodal coherence of audio and visual stimuli was shown to be useful for voice activity detection [7], [8], [9]. However, the above visual VAD algorithms use either only static or only dynamic features. Differences between speakers are not considered either. Also, head rotations which are inevitable in real video recordings, may greatly degrade the performance. Therefore, aiming to address these limitations, we propose a novel visual VAD method using adaboost [10] and then use it for the enhancement of the target speech separated by a BSS algorithm.

The main flow of the proposed system is as follows. First, in the off-line training stage, we apply the adaboost training algorithm to the labelled visual features, which are extracted from the video signal associated with a target speaker. we use the adaboost model for visual voice activity detection. Then we apply the visual VAD to detect the silent periods in the target speech, using the accompanied contemporary video. Finally we apply the spectral subtraction algorithm to the BSS-separated target speech signal, using the adaptive noise spectrum estimated in the silent periods detected via the visual VAD.

The remainder of the paper is organised as follows. Section 2 introduces the main flow of the proposed system. Section 3 presents the detailed visual VAD method using adaboost in the off-line training process. BSS enhancement using visual VAD and spectral subtraction is presented in Section 4. Experimental results are demonstrated in Section 5, followed by the conclusion.

II. Proposed system

Convolutive models are usually used to approximate the room mixing process. For simplicity, we consider a 2 × 2 system:

$$\mathbf{x}(n) = \mathbf{H}(n) * \mathbf{s}(n) + \boldsymbol{\xi}(n), \tag{1}$$

where $\mathbf{x}(n) = [x_1(n), x_2(n)]^T$ are two observations mixed by two sources $\mathbf{s}(n) = [s_1(n), s_2(n)]^T$ and $*$ denotes convolution; $\mathbf{H}(n)$ is the mixing matrix whose entry $h_{pk}(n)$ represents the impulse response from source k to sensor p; $\boldsymbol{\xi}(n)$ is the additive noise vector (omitted for convenience) and n is the discrete time index. To recover the source signals from their mixtures, beamforming techniques can be used if we have prior information about the position of sensor array and sources. An alternative is to solve this problem in the frequency domain (FD) via ICA or time-frequency masking.

In our algorithm, we apply the frequency domain BSS where the joint diagonalization [2] algorithm was used for each frequency channel. To solve its associated permutation and scaling problems, the correlation approach in [11] and the minimum distortion principle [12] are applied respectively. As mentioned earlier, the performance of conventional BSS deteriorates in adverse environments, and the separated signals may still contain a certain level of noise and interference from other speakers. Suppose $y_1(n)$ is the target speech that we are interested in, we can use multi-band spectral subtraction [4] to enhance $y_1(n)$ provided that we can accurately estimate the spectrum of the interference. To this end, we use the visual voice activity detector to find the silent periods in $s_1(n)$ and then use it for spectrum estimation of the interference. Fig. 1 shows the system diagram of the proposed method.

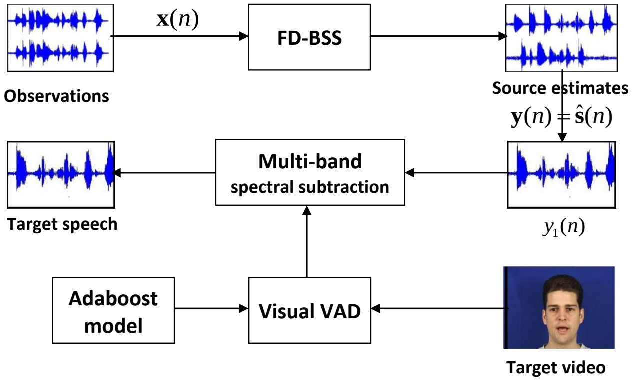


Fig. 1. The diagram of our proposed 2×2 system.

III. Visual VAD

Visual VAD is a classifier to determine whether a frame of the visual signal is silent or not, which is formed in the off-line training process based on the labelled visual features extracted from videos of a specific speaker. Since the lip region is more coherent with the audio speech, we use the following features:

$$\begin{aligned}\mathbf{v}(t) = [&W(t), H(t), A(t), d_W(T), d_H(T), d_A(T),\\ &d_W(T-1), ..., d_W(-T), d_H(-T), d_A(-T)]^T,\end{aligned} \tag{2}$$

where $W(t)$ is the outer lip width at the t-th frame, $H(t)$ is the outer lip height and $A(t)$ is the mean intensity of the rectangle in parallel to the lip corners. In equation (2), the difference feature $d_W(\tau)$ is defined as:

$$d_W(\tau) = W(t) - W(t - \tau), \tag{3}$$

where τ is the frame offset. Likewise, we define $d_H(\tau)$ and $d_A(\tau)$. Those features can be obtained from some key points of lip contours once the lip region is extracted. There are many lip tracking algorithms available, using e.g. complex statistical models [13] and active shape models [14]. We have proposed a 12-points lip tracking algorithm combining the snake algorithm with rotational template matching [15], which copes with head rotations and works well for videos with low-resolutions.

We then apply the adaboost algorithm over the labelled features (samples) to obtain parameters for the adaboost model (i.e., the strong classifier or the voice activity detector). In each iteration, it selects one optimum weak classifier under the current sample weights, and then updates the sample weights for the next iteration by giving more weighting to the misclassified samples. This weight update solves the redundancy problem in the feature space caused by the similarity between neighbouring frames. We use the same weak classifier $h_i(\mathbf{v}(t), m, p, \phi)$ as in Viola-Jones object detection algorithm [16]. In the i-th iteration, the m-th element of $\mathbf{v}(t)$, i.e. $v_m(t)$ is selected and compared with a threshold ϕ and a polarity of p:

$$h_i(\mathbf{v}(t), m, p, \phi) = \begin{cases} 1, & \text{if } p v_m(t) > p\phi \\ 0, & \text{otherwise} \end{cases}. \tag{4}$$

The final strong classifier for visual VAD is a weighted voting result over all the selected weak classifiers:

$$\mathcal{C}(\mathbf{v}(t)) = \begin{cases} 1, & \text{if } \sum_{i=1}^{I} w_i h_i(\mathbf{v}(t)) > \frac{1}{2}\sum_{i=1}^{I} w_i \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

where I is the total number of iterations, w_i is the weighting parameter, which is decided by the error rate of the i-th selected weak classifier with the updated sample weights in the i-th iteration. More details about the adaboost algorithm can be found from [10].

IV. Multi-band Spectral Subtraction

Spectral subtraction [3] is an effective method for noise reduction. In our proposed system, the multi-band spectral subtraction [4] algorithm is applied to suppress the non-uniform spectrum of the interfering speech. Suppose $Y_1(f, t)$ is the spectrum of the target source obtained from BSS at frequency bin f and time frame t via short-time Fourier transform, and $D(f, t)$ is the spectrum of the interference, then the enhanced power spectrum of $Y_{en}(f, t)$ is:

$$|Y_{en}(f, t)|^2 = |Y_1(f, t)|^2 - \alpha_i \sigma_i |D(f, t)|^2, \quad f \in \mathcal{F}_i \tag{6}$$

where α_i and σ_i control the subtraction level at the frequency band $\mathcal{F}_i$. α_i is decided by the signal-to-noise ratio in $\mathcal{F}_i$ and σ_i provides an additional degree of control, which is empirically determined to minimize speech distortion. We have used the same frequency band division and parameter setup as in [4].

The interference from the other speakers is non-stationary, therefore, we calculate and update $D(f, t)$ frame by frame from $Y_1(f, t)$ during the periods where the target speech is silent. To solve the non-stationarity problem, an alternative is to estimate the interference spectrum from the interfering speech estimated by BSS, and calculate the subtraction levels from periods where the target speaker is silent and the interference speaker is active. This needs further activity information about the interfering sources.

V. Experimental Results

A. Data Setup

1) Visual VAD: We downloaded a video clip of about 150 seconds available from http://www.youtube.com/watch?

v=zXBpW8GCDtY and cut it to 140 seconds by trimming the beginning two seconds and the end part of the signal. The first 120 seconds audio and video stimuli were used for the adaboost training, and the following 20 seconds were used as a target speech source and a target video. In the adaboost training process, we set $T = 10$, therefore the previous 10 frames and the next 10 frames also influence the activity detection of a current frame. Since the sampling rate of the video is 25 fps, about 800 ms ($2 \times T \times 25$) adjacent visual frames were considered for the activity detection of each frame. We set $I = 100$, that is, 100 weak classifiers were boosted.

2) Mixing process: The target source signal is the truncated speech of 20 seconds as mentioned above. The other source signal is the concatenated audio snippets from XM2VTS samples (each snippet lasts less than 20 seconds) available from http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/. Both signals were resampled to 16 kHz. The two source signals are shown in Fig. 2.

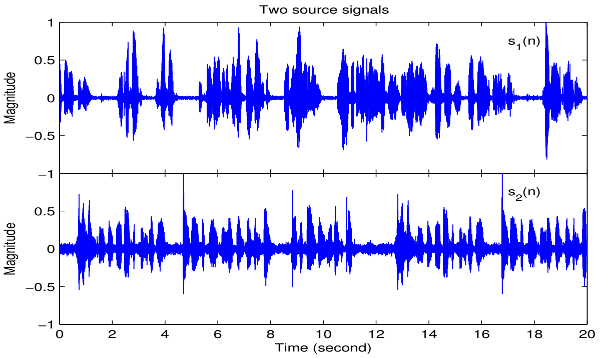


Fig. 2. The two source signals (the upper one is the target speech).

We used convolutive mixing process to model the room scenario, and two types of mixing filters were tested:

- Room impulse responses. Room impulse responses were down-sampled to a quarter of the original sampling rate, with a length of 8192 taps for each filter.
- Head related transfer functions (HRTFs). HRTFs were down-sampled to half of the sampling rate with a length of 64 taps for each filter.

The above filters are available in the synthetic benchmarks [17]. We did not add any noise to the mixtures.

3) FD-BSS: The sliding Hamming window of 640 samples with 37.5% overlap was used for the short time Fourier transform, where the Fourier frame length of 640 samples was applied.

B. Experimental Results

We first tested the filters with long taps, and the two mixtures are shown in the right plot of Fig. 3(a). After the FD-BSS, we obtained two source estimates shown in the right plot of Fig. 3(b). Compared to the original sources, the BSS-separated sources still contain a considerable amount of noise and interference. Spectral subtraction was then applied to the BSS-separated target speech, based on the visual VAD obtained from the target video, the result is shown in the right part of Fig. 3(c).

We then tested the algorithm when the mixing filters were HRTFs. In this case, the reverberation time was shorter, and FD-BSS perfectly recovers the original sources when there is no additional noise (we can compare the original source signals in Fig. 2 and source estimates in the left plot of Fig. 3(b)). In this example, the performance improvement using the spectral subtraction, as shown in the left plot of Fig. 3(c), is less apparent.

VI. Conclusion

In this paper, we have presented a system for the enhancement of BSS-separated target speech, using the speaker-dependent visual voice activity detector obtained in the off-line training stage. We have applied the visual VAD for the noise and interference spectrum estimation, which is then suppressed by the multi-band spectral subtraction algorithm. The experimental results show that the system improves the quality of the target speech estimated from the reverberant mixtures.

Acknowledgment

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) (Grant number EP/H012842/1) and the MOD University Defence Research Centre on Signal Processing (UDRC).

References

[1] P. Comon, "Independent component analysis, a new concept?" *Signal Processing*, vol. 36, no. 3, pp. 287–314, Apr. 1994.

[2] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals," *IEE Proceedings-F*, vol. 140, pp. 362–370, 1993.

[3] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, pp. 113–120, 1979.

[4] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, May 2002.

[5] D. A. Bulkin and J. M. Groh, "Seeing sounds: visual and auditory interactions in the brain," *IEEE Transactions on Neural Networks*, vol. 16, no. 4, pp. 415–419, Aug. 2006.

[6] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, pp. B69–B78, 2004.

[7] P. Liu and Z. Wang, "Voice activity detection using visual information," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, May 2004.

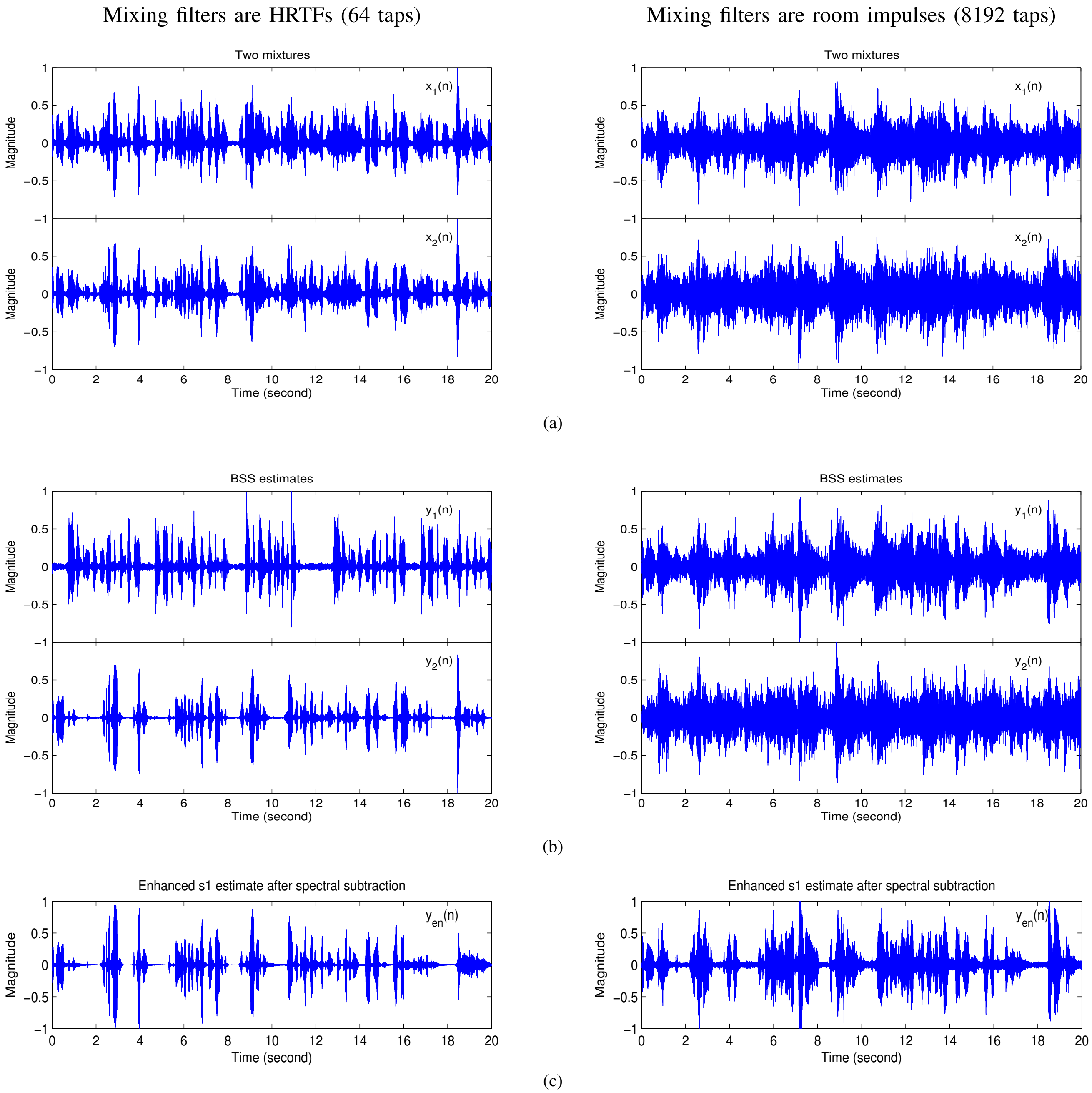
[8] D. Sodoyer, B. Rivet, L. Girin, C. Savariaux, J.-L. Schwartz, and C. Jutten, "A study of lip movements during spontaneous dialog and its application to voice activity detection," *The Journal of the Acoustical Society of America*, vol. 125, no. 2, pp. 1184–1196, Feb. 2009.

[9] A. Aubrey, Y. Hicks, and J. Chambers, "Visual voice activity detection with optical flow," *Image Processing, IET*, vol. 4, no. 6, pp. 463–472, Dec. 2010.

[10] Y. Freund and R. E. Schapire, "A short introduction to boosting," 1999.

[11] D. T. Pham, C. Serviere, and H. Boumaraf, "Blind separation of convolutive audio mixtures using nonstationarity," in *Proc. ICA*, 2003, pp. 975–980.

[12] K. Matsuoka, "Minimal distortion principle for blind source separation," in *the 41st SICE Annual Conference*, vol. 4, Aug. 2002, pp. 2138–2143.



(a)

(b)

(c)

Fig. 3. (a) The convolutive mixtures generated artificially. (b) The sources estimated by FD-BSS. (c) The target speech enhanced by multi-band spectral subtraction.

[13] A. Liew, S. Leung, and W. Lau, "Lip contour extraction using a deformable model," in *International Conference on Image Processing (ICIP)*, vol. 2, Sept. 2000, pp. 255–258.

[14] K. S. Jang, "Lip contour extraction based on active shape model and snakes," in *International Journal of Computer Science and Network Security (IJCSNS)*, vol. 7, Oct. 2007, pp. 148–153.

[15] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," in *submitted to Sensor Signal Processing for Defence (SSPD 2011)*.

[16] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2001, pp. 511–518.

[17] *ICA '99 SYNTHETIC BENCHMARKS*, 1999. [Online]. Available: http://sound.media.mit.edu/ica-bench/