Interference Reduction in Reverberant Speech Separation With Visual Voice Activity Detection

Qingju Liu, Andrew J. Aubrey, *Member, IEEE*, and Wenwu Wang, *Senior Member, IEEE*

***Abstract*—The visual modality, deemed to be complementary to the audio modality, has recently been exploited to improve the performance of blind source separation (BSS) of speech mixtures, especially in adverse environments where the performance of audio-domain methods deteriorates steadily. In this paper, we present an enhancement method to audio-domain BSS with the integration of voice activity information, obtained via a visual voice activity detection (VAD) algorithm. Mimicking aspects of human hearing, binaural speech mixtures are considered in our two-stage system. Firstly, in the off-line training stage, a speaker-independent voice activity detector is formed using the visual stimuli via the adaboosting algorithm. In the on-line separation stage, interaural phase difference (IPD) and interaural level difference (ILD) cues are statistically analyzed to assign probabilistically each time-frequency (TF) point of the audio mixtures to the source signals. Next, the detected voice activity cues (found via the visual VAD) are integrated to reduce the interference residual. Detection of the interference residual takes place gradually, with two layers of boundaries in the correlation and energy ratio map. We have tested our algorithm on speech mixtures generated using room impulse responses at different reverberation times and noise levels. Simulation results show performance improvement of the proposed method for target speech extraction in noisy and reverberant environments, in terms of signal-to-interference ratio (SIR) and perceptual evaluation of speech quality (PESQ).**

***Index Terms*—Adaboosting, binaural, blind source separation, interference removal, visual voice activity detection.**

I. Introduction

BLIND SOURCE separation (BSS), which aims to recover the unknown source signals from their mixtures, has been used to solve the 'cocktail party problem' [1], in the presence of reverberation and background noise. Different BSS techniques have been developed for this purpose, including independent component analysis (ICA) [2]–[5], beamforming [6], [7] and time-frequency (TF) masking [8]–[10]. However, in an adverse environment with long reverberation, strong background noise and multiple competing speakers, performance of the above audio-domain BSS algorithms degrades steadily. The interference distortion introduced by the residual from the interfering signals or background noise is a main cause for the performance degradation, and therefore needs to be attenuated to improve the quality and intelligibility of the separated speech.

Several interference reduction methods have been developed for this purpose [11]–[18]. Adaptive filtering [11]–[16] is perhaps one of the most popular methods, which aims to cancel the interference by applying a time-varying filter to the target BSS output, where the parameters of the filter are often estimated via a least mean squares based method. The BSS outputs not related to the target are used as reference signals, whose contributions to the target output are minimised after applying the adaptive filter. There are some other interference reduction algorithms available. For instance, in Choi *et al.* [17], probabilistic target absence/presence models are first built for the BSS outputs, and then used to remove the interference signal from the target in the power-spectral domain. In [18], the cepstral smoothing technique is applied to the binary spectral masks in order to preserve the speech structure of the target source while reducing the cross-talk by eliminating the random peaks from the interfering signals. However, the application of the methods mentioned above can be limited in a certain number of scenarios. For example, the adaptive filtering algorithm can be computationally expensive particularly for dealing with convolutive mixtures with long reverberation when the interference reduction algorithms are operated in multiple frequency channels [15], [16] or subbands [14]. The spectral cancellation approach in [17] assumes the probabilistic models obtained at the BSS outputs match approximately the original source absence/presence models, which is however not the case for the speech mixtures acquired in adverse conditions. The spectral smoothing method [18] does not explicitly use the information related to the interfering signals, and as a consequence, the spectral information related to the target source may also be conversely attenuated.

To improve the speech intelligibility in adverse conditions, additional information is needed. One such type of information comes from voice activity, which indicates whether the speaker is active or silent for a given time period. Voice activity cues play an important part in speech processing, with applications in speech coding and speech recognition [19]. During the detected silence periods of the target speech, the interference including the competing sounds and the background noise

Manuscript received August 04, 2013; revised December 08, 2013, and April 20, 2014; accepted May 04, 2014. Date of publication May 09, 2014; date of current version September 15, 2014. This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) of the U.K. under Grant EP/H012842/1. The associate editor responsible for coordinating the review of this paper and approving it for publication was Dr. Chong-Wah Ngo.

Q. Liu and W. Wang are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: q.liu@surrey.ac.uk; w.wang@surrey.ac.uk).

A. J. Aubrey was with School of Computer Science and Informatics, Cardiff University, Cardiff CF10 3AT, U.K. He is now with 3dMD, Atlanta, GA 30339 USA (e-mail: a.j.aubrey@cs.cardiff.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2014.2322824

1520-9210 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

can be estimated, which can be used to further enhance the corrupted speech source via, e.g., spectral subtraction. In this paper, we enhance the performance of speech separation in highly reverberant room environments, using a novel interference removal scheme that detects and removes the interference residual, where the voice activity cues are exploited to assist the interference detection process. To successfully implement the proposed interference removal scheme, a robust voice activity detection (VAD) algorithm with a high accuracy is essential.

There are many audio-domain VAD algorithms available. The ITU-T VAD [19] standard operates on a multi-boundary decision region with post-processing, in the space spanned by four statistical parameters, which has been widely used for fixed telephony and multimedia communications. The decision-directed parameter estimation method is employed in [20], with a first-order Markov process modelling of speech occurrences. A recent method in [21] is applied in non-stationary noise environments with low signal-to-noise ratios (SNRs), where the speech absence is adaptively estimated from a smoothed periodogram in two iterations.

However, the reliability of the above audio-domain VAD algorithms deteriorates significantly with the presence of highly non-stationary noise, e.g. the competing speech in a cocktail party scenario. Recent works show that the vision associated with the concurrent audio source contains complementary information about the sound [22]–[24], which is not affected by acoustic noise, and deemed to have the potential to improve audio-domain processing. Exploiting the bimodal coherence of speech, a family of visual VAD algorithms is developed, exhibiting advantages in adverse environments [25]–[27]. The algorithm in [25] uses a single Gaussian kernel to model the silence periods and Gaussian mixture models (GMM) for speech, with principal component analysis (PCA) for visual feature representation. A dynamic lip parameter is defined in [26] to indicate lip movements, which is low-pass filtered and thresholded to classify an audio frame. Hidden Markov models (HMMs) [28] with post-filtering are applied in [27], to model the dynamic changes of the motion field in the mouth region for silence detection. However, the algorithm in [25] does not consider dynamic features, which thereby suffers from information loss about the lip movements. Describing the visual stream only with the movement parameter, the algorithm in [26] does not consider static features. As a result its performance is not very promising. The HMM model in [27] is however trained only on the visual information from the silence periods, i.e. without using those from the active periods.

To address the above limitations, we have recently proposed a method in [29] using both static and dynamic geometric visual features with adaboost training [30]. Instead of statistically modelling the visual features as in [25], [27], we build a voice activity detector in the off-line training stage, by applying adaboost training to the labelled visual features obtained by lip-tracking. Rather than applying a hard threshold to fix a combination of the visual features as in [26], the optimal detection thresholds and their associated visual features are iteratively chosen for an enhanced detection.

In this paper, we propose a new method to mitigate the interference distortion in the BSS outputs. More specifically, the detected voice activity cues via the visual VAD [29] are integrated into a traditional audio-domain BSS method for the enhancement of the separated speech. Binaural mixtures in reverberant and noisy environments are considered, to mimic aspects of human hearing in an enclosed room environment. First, the audio-domain BSS method is applied to the mixtures in the time-frequency (TF) domain, which assigns probabilistically each TF point to the source signals exploiting the spatial cues of IPD and ILD [10], [31], to obtain approximately the source estimates. Afterwards, the interference removal scheme is applied to these source estimates, where contributions of the target speech are attenuated in the silence periods detected by the visual VAD. More specifically, the interference residual in the TF domain is first detected on a block-by-block basis based on the relation between the correlation and the energy ratio evaluated from the spectra of the BSS outputs, and then removed using a spectral subtraction technique.

We have two main contributions in our proposed interference reduction scheme.

- **Visual VAD and its use for interference detection:** The reliable visual information obtained by visual VAD is exploited, to achieve better interference detection. This is of practical importance since audio-domain processing is often degraded by acoustic noise, while the associated visual stream is not affected. Furthermore, the visual stream contains complementary information to the audio stream. As a result, our algorithm has the potential to work in adverse conditions when audio signals are seriously corrupted while the quality of the visual signal in e.g. resolution, illumination, is assumed to be adequate for the extraction of the required visual information, such as the lip movement.
- **Correlation and the energy ratio based interference/target discrimination:** We propose a novel two-stage interference detection algorithm exploiting the joint relationship between the local mutual correlation and the energy ratio between the associated spectra of the BSS outputs. In the first stage, the distinctive regions on the mutual correlation and the energy ratio map that belong respectively to the interference and target source are detected and processed in order to guarantee accurate interference detection and avoids useful target source information from being removed. The second stage attempts to resolve the overlapping source/interference regions in order to refine the ambiguous source/interference detections in these regions, and to further reduce the interference residual remained within the source.

There are potentially several applications that may benefit from our research. For example, the proposed algorithm could be used for audio-visual surveillance in a noisy environment such as airport, supermarket, with cameras that could potentially zoom into the face of a particular speaker. Another real application that could benefit from our proposed algorithm is for voice enhancement in a noisy environment for smart phone users. When the handset is held towards the face of the user during conversation, the face of the user can be well captured by the camera with a reasonably good quality that may be sufficient for lipreading.

The remainder of the paper is organised as follows. Section II introduces the main flow of the proposed visual VAD-assisted

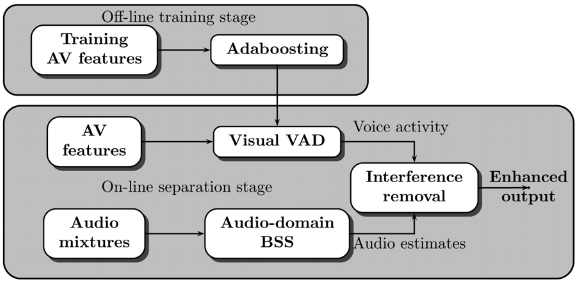


Fig. 1. The flow of our proposed two-stage AV-BSS system: VAD-incorporated BSS. In the training stage (upper shaded box), a visual VAD detector is trained via adaboost training. In the separation stage (lower shaded box), the detected voice activity cues are incorporated into the traditional audio-domain BSS algorithm through a novel interference removal scheme.

BSS system. Section III and Section IV present respectively the visual VAD technique and the audio-domain BSS used in the system. The proposed interference removal algorithm that enhances the BSS outputs is detailed in Section V. Experimental results are presented in Section VI, followed by the conclusion.

II. Proposed System

We present the block diagram of our proposed AV-BSS in Fig. 1, which contains the off-line training stage and the on-line separation stage, a similar strategy to those taken in our earlier work [29], [32], [33].

In the off-line training stage, we first extract the geometric visual features from the lip region, and then train a speaker-independent visual voice activity detector, i.e. a classifier, by applying the adaboost training [30] to these features.

In the on-line separation stage, the voice activities of the target speech are first detected by applying the trained visual VAD to the associated video signal. In parallel, the state-of-the-art audio-domain BSS algorithm proposed by Mandel *et al.* [10] is applied to the audio mixtures to obtain the source estimates. In this algorithm, the spatial cues of IPD and ILD evaluated from the binaural mixtures are used to estimate the TF masks for the separation of sources. However, in adverse environments especially with high reverberation and strong background noise, the interference distortion, similar to the cross-talk in a communication system, deteriorates greatly the intelligibility of the source estimates. The reason for the interference phenomenon is further explained in Section IV. To mitigate the degradation, we propose an interference removal scheme, integrating the VAD cues previously detected, which will be presented in detail in Section V.

For the completeness of the proposed system and readability of the subsequent sections, we introduce briefly the principles of the visual VAD as well as Mandel's BSS algorithms in the next two sections.

III. Visual VAD

The visual VAD is a classifier to determine whether a speaker is silent or not in a frame l using the associated video signal. The VAD is built on an off-line training process using the labelled visual features $\mathbf{v}(l)$ extracted from the associated video clips of the mouth region, i.e. region of interest (ROI). The appearance (intensity) based features are subject to variation of luminance and skin colour, therefore we restrict ourselves to using geometric features of the lip contours. In our proposed algorithm, the 38-point lip-tracking algorithm [34] is first applied to the raw video, for the extraction of the contours of both inner and outer lips. To accommodate individual differences in lip size, we normalise the mouth region such that, when the speaker is silent with naturally closed lips, the width of the outer lip should be of unit length.

From the lip contours, we can extract the $6(2*Q+1)$-dimensional geometric visual feature at each frame l where Q is the maximal frame offset

$$\mathbf{v}(l) = [\mathbf{g}^T(l), \mathbf{d}_l^T(-Q), \dots, \mathbf{d}_l^T(Q)]^T \tag{1}$$

where the superscript T denotes transpose and $\mathbf{g}(l)$ is the static visual feature containing six elements: the width, height and area of the outer and inner lip contours. In Equation (1), $\mathbf{d}_l(q)$ is the difference feature vector that models the dynamic visual changes, i.e. lip movements, defined as

$$\mathbf{d}_l(q) = \mathbf{g}(l) - \mathbf{g}(l-q) \tag{2}$$

where $q = [-Q, \dots, Q]$ is the frame offset. $\mathbf{v}(l)$ can also be obtained using other lip tracking algorithms, based on e.g. complex statistical models [35] and active shape models [29], [36].

We use our recently proposed visual VAD algorithm [29], which employs the adaboost training algorithm [30] to the manually labelled visual feature $\mathbf{v}(l)$. The same weak classifier used in the Viola-Jones object detection algorithm [37], denoted as $\mathcal{C}_{\mathrm{VJ}}(\cdot)$, is used in our algorithm as follows. In the i-th iteration, the κ_i-th element of $\mathbf{v}(l)$, i.e. $v_{\kappa_i}(l)$, is selected and compared with a threshold Δ_i and a polarity of $p_i \in \{1, -1\}$

$$\mathcal{C}_{\mathrm{VJ}}(\mathbf{v}(l), \kappa_i, p_i, \Delta_i) = \begin{cases} 1, & \text{if } p_i v_{\kappa_i}(l) > p_i \Delta_i, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

For simplicity, we denote $\mathcal{C}_{\mathrm{VJ}}(\cdot|i) = \mathcal{C}_{\mathrm{VJ}}(\mathbf{v}(l), \kappa_i, p_i, \Delta_i)$ as the weak classifier chosen at the i-th iteration, which contributes to the final strong classifier with a weighting parameter w_i. The parameter w_i is determined by the error rate ε_i in the i-th iteration in the training process, via

$$w_i = \ln \frac{1-\varepsilon_i}{\varepsilon_i} \tag{4}$$

Using the majority voting, the visual VAD detector is obtained

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

where I is the total number of iterations.

The trained visual VAD will be used to reduce interference residuals that may remain in the output of an audio-domain BSS. Next, we discuss a typical audio-domain BSS algorithm and analyse the reason why the interference distortion occurs.

IV. Audio-Domain BSS

To mimic aspects of human hearing, binaural mixtures are considered in our proposed system. We choose to use the state-of-the-art audio-domain method by Mandel *et al.* [10], where the TF masking technique is used to separate sources. The TF masks are determined by the evaluation of the spatial cues of IPD and ILD [10], [31], whose principles are as follows.

A source signal arrives at two ears with different time delays and attenuations, exhibiting

$$L(m,\omega)/R(m,\omega) = 10^{\frac{\alpha(m,\omega)}{20}} e^{\sqrt{-1}\beta(m,\omega)} \tag{6}$$

where $L(m,\omega)$ and $R(m,\omega)$ are the TF representations of the left-ear and right-ear signals respectively, indexed by time frame m and frequency bin ω, α and β are respectively the ILD and IPD. Two Gaussian distributions are used to model the ILD $\alpha(m,\omega)$ and IPD $\beta(m,\omega)$ for source k at the discretized time delay τ

$$\begin{cases} p_{\mathrm{IPD}}(m,\omega|k,\tau) \sim \mathcal{G}(\beta(m,\omega)|\xi_{k\tau}(\omega);\sigma_{k\tau}^2(\omega)) \\ p_{\mathrm{ILD}}(m,\omega|k) \sim \mathcal{G}(\alpha(m,\omega)|\mu_k(\omega);\eta_k^2(\omega)) \end{cases} \tag{7}$$

where $\mathcal{G}(\cdot)$ represents the Gaussian density function. $p_{\mathrm{IPD}}(m,\omega|k,\tau)$ and $p_{\mathrm{ILD}}(m,\omega|k)$ are respectively the likelihood of the IPD and ILD cues at the TF point (m,ω) originating from source k at time delay τ, parametrised by the means and variances $\Theta = \{\xi_{k\tau}(\omega), \sigma_{k\tau}^2(\omega), \mu_k(\omega), \eta_k^2(\omega)\}$. With the independence assumption of ILD and IPD, the likelihood of the point (m,ω) originating from source k at delay τ is

$$p_{\mathrm{IPD/ILD}}(m,\omega|k,\tau) = p_{\mathrm{IPD}}(m,\omega|k,\tau) p_{\mathrm{ILD}}(m,\omega|k), \tag{8}$$

and the posterior is calculated as

$$p(k,\tau|m,\omega) = \frac{p_{\mathrm{IPD/ILD}}(m,\omega|k,\tau)\varphi_{k\tau}}{\sum_{k,\tau} p_{\mathrm{IPD/ILD}}(m,\omega|k,\tau)\varphi_{k\tau}} \tag{9}$$

with $\varphi_{k\tau} = \sum_{m,\omega} p(k,\tau|m,\omega)$ being the prior probability of a TF point of the mixture coming from source k at delay τ. To recover the k-th source, an audio-domain separation mask is calculated via Equation (9) as $\mathcal{M}_k^a(m,\omega) = \sum_\tau p(k,\tau|m,\omega)$. This mask can be applied to $L(m,\omega)$, $R(m,\omega)$ or both[1] to estimate source k via e.g. $\hat{S}_k(m,\omega) = \mathcal{M}_k^a(m,\omega)L(m,\omega)$.

Residual Distortion: The success of Mandel's method relies on the sparsity assumption of the source signals, i.e., at most one source signal is dominant at each TF point. Therefore, in an ideal narrowband convolution process where the reverberation time is short, the mask $\mathcal{M}_k^a(m,\omega)$ is either 1 or 0 at the TF point (m,ω), depending on whether it is dominated or not by source k. However, with the presence of a high level of reverberation and noise, the mixing process becomes a wideband convolution process, introducing a large scale in ILD α and IPD β. As a result, variances in the parameter set Θ become large, which results in a relatively small disparity in the IPD/ILD evaluations for different sources in Equation (8). Consequently, a TF point may be assigned partially to an in-dominant source, determined by Equation (9), even though the in-dominant source contributes very little to that TF point. Due to the above reasons, an unwanted residual distortion is introduced, as demonstrated in Fig. 2, where the residual distortion is highlighted in ellipses.

To mitigate the interference distortion, we propose a novel interference removal scheme, which combines the voice activity cues detected by the visual VAD, with the magnitude spectrum of the source estimates $E_k(m,\omega) = |\hat{S}_k(m,\omega)|$. The detailed algorithm is presented as follows.

V. Proposed Interference Removal Algorithm

The main principle of the proposed interference removal method is to first detect the interference on a block-by-block basis and then to remove the contribution of the interference in those blocks where the interference is detected (see later in this section for the motivation of block-based processing). The most challenging part is the interference detection, for which we propose a two-stage scheme based on two quantities, namely the correlation coefficient and the energy ratio. As discussed more later in this section, on the scatter plot of the correlation coefficient versus the energy ratio calculated for all the blocks based on a dataset of real speech mixtures, the distributions of the interference and target source appear to be well apart from each other, with some overlap between them. Therefore, using this information, we could potentially distinguish the interference from the target speech within the separated speech signals (i.e. the BSS outputs).

Spectra of the BSS outputs are, however, corrupted by high-frequency noise, which adds difficulties to the distinction of the interference from the target. To mitigate the spectral noise, a 2D Gaussian filter is applied to smooth the spectrum $E_k(m,\omega)$ before interference reduction. Impact of the Gaussian filtering on the performance of the interference reduction algorithm is evaluated in Section VI-B. From the spectra of speech signals shown in Fig. 2(a) and 2(c), it is found that the lower the frequency, the more distinguished contours the spectra have, i.e., they are less affected by spectral noise. As a result, the smoothing filter should have less effect on the low frequency as compared to the high frequency, i.e., the standard deviation in the low frequency should be smaller than that in the high frequency. Consequently, a frequency-dependent smoothing filter $\mathcal{G}^\omega$ is required, and the smoothed spectrum is denoted as

$$\tilde{E}_k(m,\omega) = \mathcal{G}^\omega * E_k(m,\omega)$$

where $*$ denotes convolution. Mel-scale filterbanks can be exploited to determine $\mathcal{G}^\omega$ using the non-linear resolution of the human auditory system across an audio spectrum. Mel-scale frequency is related to frequency f in Hertz by

$$\mathrm{Mel}(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \tag{10}$$

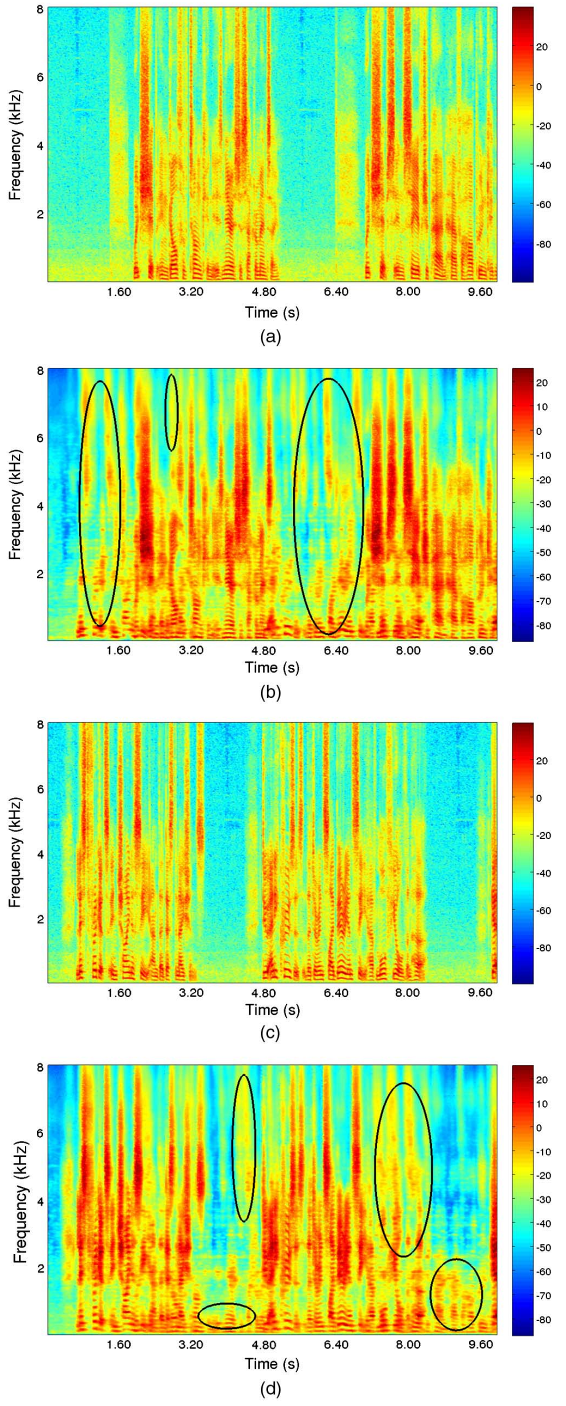
Specifically, a 2D filter $\mathcal{G}^\omega$ can be determined at each frequency bin ω, such that the $\mathcal{G}^\omega$ has the standard deviation vector composed of, e.g., 1/10 of the band of a Mel-scale filterbank centred at the ω-th frequency in the frequency dimension, as well as a fixed standard deviation δ_m in the time dimension. This process can be approximated with two Gaussian filters[2] $\mathcal{G}_1$ and $\mathcal{G}_2$

$$\tilde{E}_k(m,\omega) = \frac{\Omega - \omega}{\Omega}\mathcal{G}_1 * E_k(m,\omega) + \frac{\omega}{\Omega}\mathcal{G}_2 * E_k(m,\omega) \tag{11}$$

where Ω is the total number of the discretized frequency bins; $\mathcal{G}_1$ has a standard deviation vector of $\boldsymbol{\delta}_1 = [\delta_1, \delta_m]^T$, while

[1] We applied the audio mask to both of the binaural signals and calculated their average as the source estimate in our experiments.

[2] Previously we denote the frequency-dependent Gaussian function $\mathcal{G}^\omega$ with the superscript $\omega = [1, 2, \ldots, N_{\mathrm{FFT}}/2]$. Here we use two Gaussian functions with subscripts 1 and 2 to approximate $\mathcal{G}^\omega$.



(a)

(b)

(c)

(d)

Fig. 2. Spectrograms of the original source signals ((a) and (c)) and the source estimates ((b) and (d)) after the audio-domain BSS by [10]. The high reverberation (RT = 890 ms) severely degrades the separation performance, introducing interference distortion. As shown in the highlighted ellipses, there exists the residual from the competing speaker. For demonstration purposes, the above spectra are Gaussian smoothed and plotted on a decibel scale: (a) Magnitude spectrum of source 1; (b) Magnitude spectrum of source 1 estimate; (c) Magnitude spectrum of source 2; (d) Magnitude spectrum of source 2 estimate.

$\mathcal{G}_2$ has $\boldsymbol{\delta}_2 = [\delta_2, \delta_m]^T$, where $\delta_1 = \lceil\frac{\mathrm{Mel}(300)N_{\mathrm{FFT}}}{10F_s}\rceil$ and $\delta_2 = \lceil\frac{\mathrm{Mel}(F_s/2)N_{\mathrm{FFT}}}{10F_s}\rceil$ are the standard deviations in the spectral dimension. Here, N_{FFT} is the FFT size, F_s is the sampling rate and $\lceil\cdot\rceil$ rounds a number to its nearest integer. Note that the lowest frequency 0 Hz was not used for calculating δ_1, since in this case, $\delta_1 = 0$ and hence no smoothing would be applied. Therefore, we set $\delta_1 = \frac{\mathrm{Mel}(300)N_{\mathrm{FFT}}}{10F_s}$ instead, since 300 Hz is the beginning range of the voice band [38]. More specifically, we have $\mathcal{G}_1(\cdot) = \mathcal{G}(\cdot|0, \boldsymbol{\delta}_1)$ and $\mathcal{G}_2(\cdot) = \mathcal{G}(\cdot|\mathbf{0}, \boldsymbol{\delta}_2)$.

Then we detect the interference on a block-by-block basis, utilising the local mutual correlation and the energy ratio between $\tilde{E}_k$ and $\tilde{E}_j$ as follows, where k and j are the indices of two estimated sources. First, half-overlapping sliding blocks indexed by (b^m, b^ω) are applied to both spectra, to segment each spectrum into $B^m \times B^\omega$ blocks, with each block spanning the TF space of $L^m \times L^\omega$. The block spectrum associated with the (b^m, b^ω)-th block for the k-th source estimate is denoted as $\tilde{\mathbf{E}}_{kb^m b^\omega} = (\tilde{E}_k(m,\omega)) \in \mathbb{R}^{L^m \times L^\omega}$, where $m \in \mathbf{M}_{b^m} = \{(b^m - 1)\frac{B^m}{2} + [1 : L^m]\}$ and $\omega \in \mathbf{\Omega}_{b^\omega} = \{(b^\omega - 1)\frac{B^\omega}{2} + [1 : L^\omega]\}$. After that, in the (b^m, b^ω)-th block, the normalised correlation and energy ratio are calculated, respectively, as

$$\Gamma_{kj}(b^m, b^\omega) = \mathrm{Corr}(\tilde{\mathbf{E}}_{kb^m b^\omega}, \tilde{\mathbf{E}}_{jb^m b^\omega}), \tag{12}$$

$$\Upsilon_{kj}(b^m, b^\omega) = \|\tilde{\mathbf{E}}_{kb^m b^\omega}\|^2 / \|\tilde{\mathbf{E}}_{jb^m b^\omega}\|^2 \tag{13}$$

where $\|\cdot\|$ is the Euclidean norm, which is the square root of the summation of the squares of all the elements in the matrix. In Equation (12), $\mathrm{Corr}(\cdot,\cdot)$ first vectorises the two matrices in its arguments to obtain two vectors spanning the same length. After that, Pearson's correlation coefficient is calculated to obtain their similarity.

We want to attenuate the audio spectrum of the target speaker (suppose it is indexed by k) during the silence periods, so we integrate the voice activity cues into the energy ratio at the (b^m, b^ω)-th block as

$$\Upsilon_{kj}^{\mathrm{VAD}}(b^m, b^\omega) = \|\mathbf{E}_{kb^m b^\omega}^{\mathrm{VAD}}\|^2 / \|\mathbf{E}_{jb^m b^\omega}^{\mathrm{VAD}}\|^2 \tag{14}$$

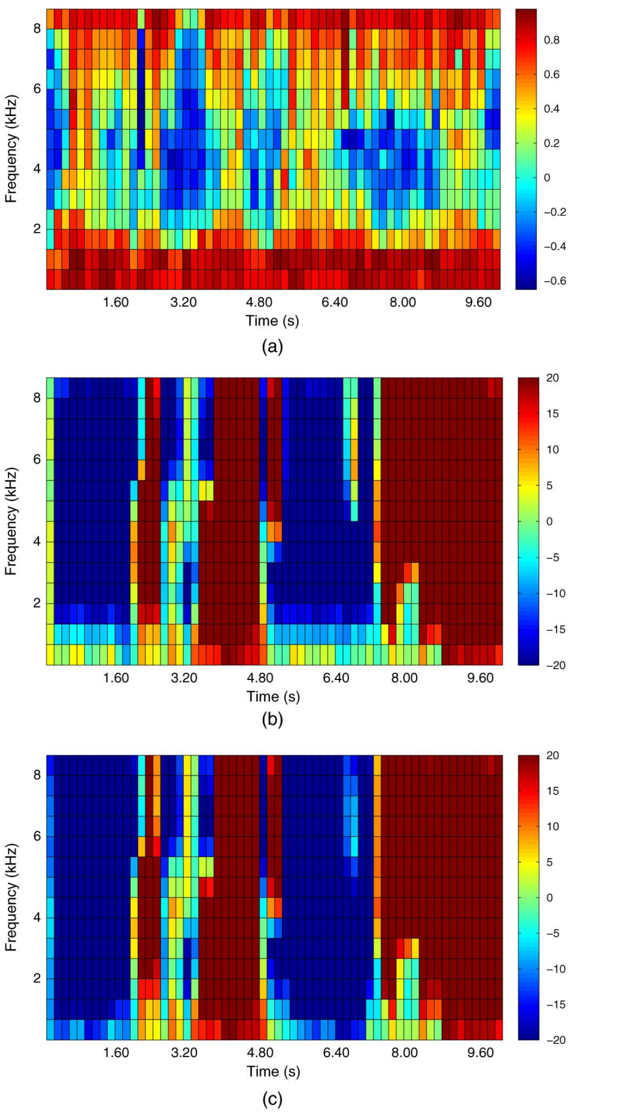
where $\mathbf{E}_{kb^m b^\omega}^{\mathrm{VAD}} = (E_k^{\mathrm{VAD}}(m,\omega)) \in \mathbb{R}^{L^m \times L^\omega}$, and

$$E_k^{\mathrm{VAD}}(m,\omega) = \begin{cases} \tilde{E}_k(m,\omega), & \text{if } \mathcal{C}(\mathbf{v}(\mathrm{m})) == 1 \\ \sqrt{\sigma}\tilde{E}_k(m,\omega), & \text{otherwise}, \end{cases} \tag{15}$$

where σ is a threshold in the range $(0, 1]$ rather than 0 to[3] accommodate the VAD false positive error (i.e. active being detected as silence) and the non-zero energy in silence periods. The detected VAD cues $\mathcal{C}(\mathbf{v}(l))$ are resampled to have the same temporal resolution as the spectrum, denoted as $\mathcal{C}(\mathbf{v}(m))$. The calculated block correlation and energy ratio (with or without integrating voice activity cues) are illustrated in Fig. 3.

As mentioned earlier, we can potentially distinguish interferences from the target source using the relationship between the correlation coefficient and the energy ratio defined in Equations (12) and (13). To see this, we first show an example of the scatter plot of the correlation coefficient versus the energy ratio, as in Fig. 4 which demonstrates the dependency of the interference/target distribution on the relation between the block correlation and the energy ratio. It can be observed that the

[3] It would have the same effect as the spectral subtraction [39] if we set $\sigma = 0$, which directly removes the information in detected silence periods. It may result in important information loss, considering that the visual VAD algorithm is not 100% accurate.



(a)

(b)

(c)

Fig. 3. The block correlation (upper) and the energy ratio (lower two) of the spectra of two source estimates. Energy ratio of the first source estimate over the second is shown. The estimated VAD cues are associated with source 1, and the third row shows the re-calculated energy ratio after combining the VAD cues. For demonstration purposes, the energy ratio is shown in decibel and thresholded with an upper bound of 20 dB and a lower bound of -20 dB. (a) Block correlation Γ (b) Block energy ratio Υ (c) Block energy ratio with VAD Υ^{VAD}.

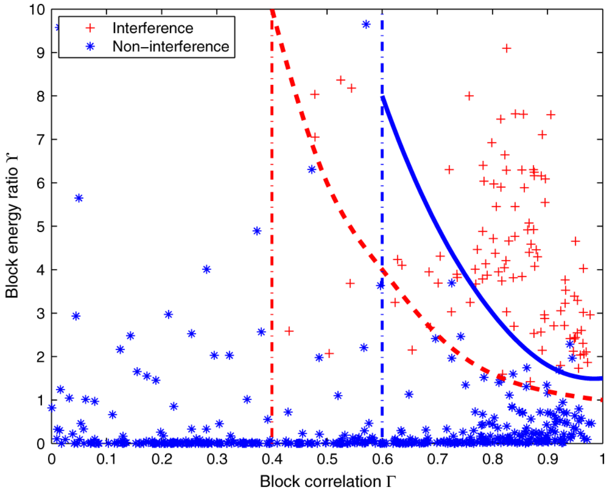


Fig. 4. The joint distribution of the block correlation Γ and the energy ratio Υ. The blocks are randomly chosen from the spectra of the BSS outputs from four different rooms. Each block is detected as interference (red cross) or non-interference (blue star). Details on how this figure is generated can be found in the texts below.

two regions that the interference and source are distributed, are, in general, distant from each other, with a certain amount of overlap between them. Based on this observation, we empirically split the scatter plot into three different regions: ‘strict’, ‘loose’, and ‘non’ interference regions, corresponding respectively to the region occupied mainly by the interference, the ambiguous region shared by both the interference and non-interference (i.e. target source), and the region occupied mainly by the non-interference.

Therefore, we propose to detect the interference in two stages, whose architecture is similar, in spirit, to the scheme taken in many hierarchical clustering algorithms [40]. The first stage aims to detect any signal that falls into the ‘strict’ region shown in Fig. 4, therefore it has a high detection accuracy. The second stage attempts to resolve the ambiguous ‘loose’ region, to further reduce the interference residual as much as possible.

The physical meaning of the proposed scheme is that: if the spectra of two different BSS outputs at the same TF position have a very similar structure (evaluated by the block correlation coefficient Γ defined in Equation (12)) and a very large energy difference (evaluated by the block energy ratio Υ, defined in Equation (13)), then the spectrum with low energy is likely to be the interference residual. This is, in practice, the case since two independent speech signals have different temporal and inter-frequency structures.

To determine the boundary between these regions, we first choose empirically several points (e.g. six in our work) according to the distribution plot to coarsely generate the boundaries between these regions. We then use a third-degree polynomial curve fitting technique to find the values of the other points on the coarse boundary, which results in the final boundary for splitting these three regions.

The scatter plot in Fig. 4 is generated as follows. We first applied BSS to four pairs of randomly-chosen mixtures in four different rooms, the room description is available in Section VI-B. For the ease of understanding, we denote the spectra of the two BSS outputs in a TF block as $\hat{\mathbf{E}}_1$ and $\hat{\mathbf{E}}_2$ (the block indices are dropped here for notational convenience), and the original source signals $\mathbf{E}_1$ and $\mathbf{E}_2$ in the associated block. Assuming that $\mathbf{E}_2$ and $\mathbf{E}_1$ are respectively the target source and interference signal, we attempt to find whether or not the source estimate $\hat{\mathbf{E}}_2$ is corrupted by the interference signal $\mathbf{E}_1$. In each TF block, we calculated the block correlation and energy ratio between $\hat{\mathbf{E}}_1$ and $\hat{\mathbf{E}}_2$, which correspond to one point (either cross or star) in the scatter plot of Fig. 4. Then, the following approach was used to determine if $\hat{\mathbf{E}}_2$ is corrupted by the interference $\mathbf{E}_1$ (cross) or not (star).

First we calculate the ideal binary mask (IBM) [41] by comparing the elements of $\mathbf{E}_1$ with $\mathbf{E}_2$. If one TF point of $\mathbf{E}_1$ has a greater value than the corresponding TF point of $\mathbf{E}_2$, then the IBM is assigned with 1 at the associated TF point, otherwise 0. If more than 80% of the IBM values in this block are 1, it means $\mathbf{E}_1$ is the dominant source. As a result, $\mathbf{E}_1$ may corrupt the source estimate $\hat{\mathbf{E}}_2$. Then we calculate the correlation coefficients between the BSS outputs and the original source signals. If $\mathrm{Corr}(\hat{\mathbf{E}}_1, \mathbf{E}_1) > 0, \mathrm{Corr}(\hat{\mathbf{E}}_2, \mathbf{E}_1) > 0$, and $\mathrm{Corr}(\hat{\mathbf{E}}_1, \mathbf{E}_2) < 0, \mathrm{Corr}(\hat{\mathbf{E}}_2, \mathbf{E}_2) < 0$, then within this block, $\hat{\mathbf{E}}_2$ is dominated by the interference $\mathbf{E}_1$. Therefore this block is labelled as interference.

We should note that these blocks are randomly chosen from different room mixtures, and the distribution remains similar even if we change the candidate blocks or swap the order of the BSS outputs. Therefore, the relation between the correlation and energy ratio can be applied generically to different datasets, and hence the proposed algorithm is not over-fitted to the dataset used in this work.

We have also applied SVM [42] to the same data to automatically find the detection boundary, with 10-fold cross validation. However, the SVM method produces a relatively high negative error rate, i.e., non-interference being detected as interference. This may introduce greater information loss as compared with the proposed method. For this reason, the SVM results are not included.

With a hard threshold of $\Gamma_{kj} > 0.6$ in the first stage[4], we consider the audio block (b^m, b^ω) as interference for the j-th source, if it is above a third-degree polynomial curve parametrised by $[q_3, q_2, q_1, q_0]$

$$\Upsilon_{kj}^{\mathrm{VAD}} > q_3\Gamma_{kj}^3 + q_2\Gamma_{kj}^2 + q_1\Gamma_{kj} + q_0. \tag{16}$$

The motivation behind the use of the third-degree polynomial curve fitting technique is mainly to consider the fitting accuracy of the interpolation (curve fitting) functions and its associated computational cost. There are other curve fitting techniques that can serve for the same purpose. For instance, we tested three other curve fitting methods, respectively, linear interpolation (one-degree polynomial), quadratic interpolation (two-degree polynomial) and the exponential curve fitting (which has two parameters to tune). For example, in Fig. 4, the fitting errors by these methods obtained from the six manually selected points are [1.48 20.4 0.39], respectively, while the third-degree polynomial curve fitting has an error smaller than 0.01. Moreover, the computational complexity of the third-degree polynomial curve is lower than that of the exponential functions, and with only a few parameters involved in such a fitting function, the proposed algorithm is less likely to have the over-fitting problem.

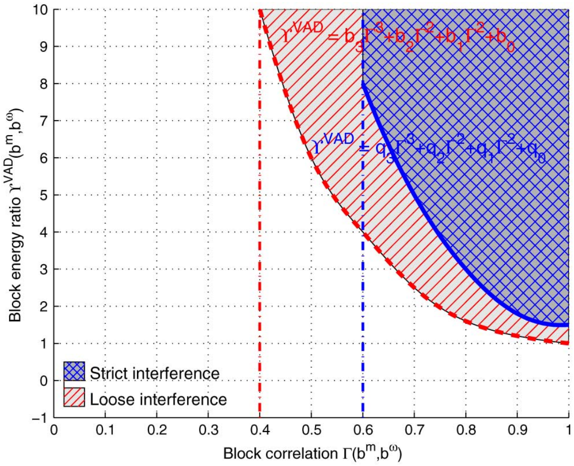


Fig. 5. The two-stage boundaries for interference detection. In the first stage, if $\Gamma_{kj}(b^m, b^\omega)$ and $\Upsilon_{kj}^{\mathrm{VAD}}(b^m, b^\omega)$ fall in the cross hatched region, we consider the block (b^m, b^ω) as an interference for source j. The cross-hatched region is thresholded by a strict boundary, shown by the solid curve. Then in the second stage, we detect the affected neighboring blocks determined by $\mathcal{A}(b^\omega)$. If Γ_{kj} and $\Upsilon_{kj}^{\mathrm{VAD}}$ associated with one affected block fall in the single-hatched region, this block is considered an interference for source j. The single-hatched region is margined by a loose boundary, shown by the dashed curve.

If the block (b^m, b^ω) is labelled as interference, exploiting the speech resolution characteristics, the neighbouring blocks of the same Mel filterbank, denoted as $(b^m, b^\omega + [-\mathcal{A}(b^\omega), \mathcal{A}(b^\omega)])$, are likely to be interference, where $\mathcal{A}(b^\omega)$ measures the affected neighbouring block number. Using the similar approximation for the spectrum smoothing, we calculate the ‘affected area’ using the Mel-scale filterbanks

$$\mathcal{A}(b^\omega) = \left\lceil \frac{B^\omega - b^\omega}{B^\omega} \frac{\mathrm{Mel}(300)B^\omega}{F_s} + \frac{b^\omega}{B^\omega} \frac{\mathrm{Mel}(F_s/2)B^\omega}{F_s} \right\rceil. \tag{17}$$

Then, in the second stage, for a labelled block (b^m, b^ω), we further detect the interference in its affected neighbouring blocks using a loose boundary parametrised by $[b_3, b_2, b_1, b_0]$ with a hard threshold of $\Gamma_{kj} > 0.4$

$$\Upsilon_{kj}^{\mathrm{VAD}} > b_3\Gamma_{kj}^3 + b_2\Gamma_{kj}^2 + b_1\Gamma_{kj} + b_0. \tag{18}$$

The same third-degree polynomial curve fitting technique as used in the first stage is employed to find the parameters in Equation (18). With the second stage, the overlap regions of the interference and source in the correlation and energy ratio plot can be better resolved. As a result, the interference residual can be further reduced. The two-stage boundaries in Equations (16) and (18) are shown in Fig. 5.

The block-wise principle used in our interference detection scheme aims to facilitate the evaluation of the structure similarity and energy ratio between the spectra of the BSS outputs, and thus to detect reliably the TF regions of the separated sources that still contain interference residuals. With block-wise processing, such regions can be identified more efficiently as compared to the evaluation of the above cues at each TF point of the spectra of the separated speech sources. However, the block-wise processing may introduce processing artefacts, which can be further reduced using some post-processing techniques such as spectral smoothing [18]. The influences of the

[4] We picked this value 0.6 empirically from Fig. 4 when the maximal non-interference correlation in the overlapped region is smaller than 0.6. In the same way, we obtained another hard threshold of 0.4 in the second stage associated with Equation (18) when the minimal interference correlation in the overlapped region is bigger than 0.4.

artefacts can be observed later in the performance evaluation in Section VI-B2, especially when the input angle is small e.g. 15°. Spectral smoothing is a popular technique to mitigate artefacts introduced in the processing technique. However, we did not combine it for further improving the results since this is out of the scope of this paper. This block-wise processing is a trade-off between the performance and computational complexity, which will be demonstrated later in the computational load comparison between our proposed algorithm and a point-wise processing method in Section VI-B2.

Finally, the interference at the block (b^m, b^ω) is removed from the j-th source estimates if it is labelled as interference:

$$\mathbf{E}_{jb^m b^\omega} \leftarrow \tilde{\mathbf{E}}_{jb^m b^\omega} - \frac{1}{\sqrt{\Upsilon_{kj}(b^m, b^\omega)}} \tilde{\mathbf{E}}_{kb^m b^\omega}. \tag{19}$$

The spectra after the interference reduction are transformed back to the time domain for source estimates reconstruction.

The interference removal scheme is summarised in Algorithm 1.

Algorithm 1: Framework of the proposed interference removal scheme.

```
Input: Magnitude spectra of the source estimates
       after the audio-domain TF masking
       $E_k(m,\omega), k = 1,2,\ldots,K$, parameters
       $[q_3, q_2, q_1, q_0]$ and $[b_3, b_2, b_1, b_0]$, block size
       $L^m \times L^\omega$, and visual VAD cues $\mathcal{C}(\mathbf{v}(m))$.
Output: Interference-reduced audio spectra $E_k(m,\omega)$.
% Gaussian smoothing
Obtain $\tilde{E}_k(m,\omega)$ using Equation (11).
% Segmentation
Obtain half-overlapping blocks $\tilde{\mathbf{E}}_{kb^m b^\omega}$.
% Mutual correlations and energy ratios
Calculate $\Gamma_{kj}(b^m, b^\omega)$ and $\Upsilon_{kj}^{\mathrm{VAD}}(b^m, b^\omega)$ with
 Equations (12) to (15) for each block.
% Interference detection
Detect interference with Equation (16).
Further detect interference with Equations (17) and (18).
% Interference reduction
Remove the detected interference with Equation (19).
```

VI. Experimental Results

To demonstrate our proposed method on real speech signals, we applied our algorithm on the LILiR main dataset [43], which was recorded in a controlled environment with each subject uttering continuous speech, and the frontal face of the subject being captured with a high-resolution camera, with a certain degree of head movements. Sequences were obtained from several recordings, sampled at $F_s = 16$ kHz. To obtain a speaker-independent visual VAD, recordings of two speakers were used. The first 4 sequences for subject 1 and the first 3 sequences for subject 2 were used for training, which last 41267 frames in total (approximately 28 minutes). The remaining 5 sequences lasting about 12 minutes were used for testing.

A. Visual VAD

Data and Parameter Setup: A 38-point lip contour extraction algorithm [34] was applied for both inner and outer lips to extract the lip contours, and we set $Q = 10$ to obtain the visual features $\mathbf{V} = (\mathbf{v}(l))$.

We manually labelled the clean audio activity for training at each frame l, denoted by $a(l)$. For the VAD training, we set $I = 100$ in the adaboost training, which means 100 weak classifiers will be combined. The error rate was used as a criterion to evaluate the performance over the total L frames

$$\epsilon = \frac{1}{L} \sum_l \# \left(\mathcal{C}(\mathbf{v}(l)) \neq a(l) \right) \tag{20}$$

where $\#(\cdot)$ counts the number.

In the same way, we defined the false positive rate ϵ_p and the false negative rate ϵ_n, which evaluate respectively the ratio of voice being detected as silence and its converse. We are more tolerant to ϵ_n as compared to ϵ_p.

For comparison purposes, we also implemented two baseline visual VAD methods.

The first baseline algorithm uses support vector machine (SVM) training to the same visual features as used in our algorithm. The linear SVM [44] was used to accommodate the high-dimension and the large scale.

The second one is the method by Aubrey *et al.* [27] using the optical flow algorithm [45], where one HMM model was built on the 18156 silence frames. However, there are instances of significant rigid head motion which greatly affects the performance, therefore, we centralised the cropped lip region based on the lip tracking results. Different recordings were scaled such that in a natural pose with closed lips, the width of the lips is 100 pixels. The cropped, raw images have a dimension of 96 × 128. We then applied the optical flow [45] to produce a 24×32 motion field, which was later projected onto a 10-dimensional space via principal component analysis (PCA). Finally, a 20-state HMM was trained on the visual feature space, where the conditional probability distribution is a Gaussian function for each state. When we applied the trained HMM on the testing data, the HMM evaluation results was normalised into [0 1], and we set the threshold[5] of 0.7 for VAD detection, which was filtered using a low-pass filter with 5 taps. The same low-pass filter was also applied to our proposed algorithm and the SVM based method.

Results Comparison and Analysis: First, the trained visual VAD detector was applied to the testing data, and our algorithm successfully detected most of the frames with a high accuracy. We noticed that our proposed algorithm suffers from a relatively high false positive error rate. However, the false negative error rate is much lower than the two competing methods, as shown in the dot-dashed curve in Fig. 6.

[5] This threshold was chosen from a series of candidates with an increment of 0.1 from 0.6 to 0.9, which gave the best results.

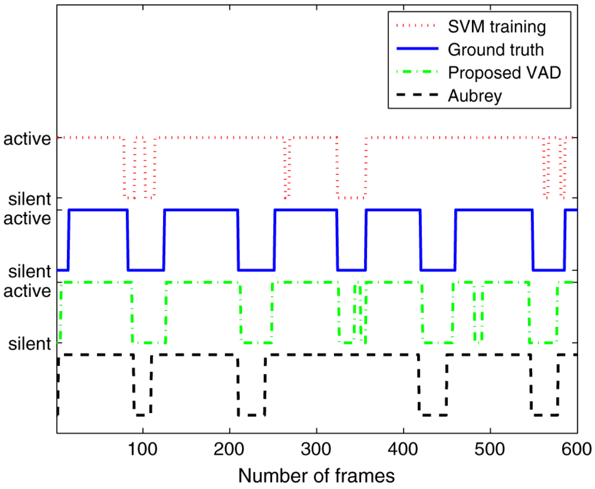


Fig. 6. Comparison of our proposed visual VAD algorithm with the ground-truth, Aubrey *et al.* [27], and the SVM [44] method.

We then quantified the detection performance using the testing data. Using our method, we obtained $\epsilon_p = 0.03$, and $\epsilon_n = 0.25$, while the total error rate $\epsilon = 0.11$. With Aubrey's method, we obtained $\epsilon_p = 0.01$, $\epsilon_n = 0.32$ and $\epsilon = 0.12$. Using SVM, we got $\epsilon_p = 0.01$, $\epsilon_n = 0.72$ and $\epsilon = 0.27$. Aubrey *et al.* [27] and the SVM [44] method achieve the best false positive rate, however at the expense of resulting in a high false negative rate, especially for the SVM training. The method of Aubrey *et al.* is likely to be affected by the following issues. First, a much large-scale data is used, rather than 1500 frames used in [27]. Second, shifts of head positions affect the performance, even though they are already alleviated by scaling and centralisations in the pre-processing. Third, some irregular lip movements in the silence periods (e.g. those caused by laughing and head rotations) are matched with movements in voice, which results in a high ϵ_n.

Then we compared the algorithm complexity, ignoring the training stage and post-processing in the detection stage. For our proposed algorithm and SVM training, the same visual features were used. The complexity is mainly caused by the lip-tracking algorithm. For the detection part, our method has $2I$ comparisons and I summations, which are neglect-able as compared to feature extraction. For the SVM detection, $6(2Q+1)$ multiplications are involved, which are also neglect-able as compared to feature extraction. For Aubrey *et al.*'s method, lip tracking is also required to centralise the lip region, which has the same complexity as our proposed algorithm. Then we need PCA projection to reduce the redundancy. A forward-type calculation for the likelihood of the testing data is applied to the 20-state HMM, whose complexity is mainly taken up by 400 multiplications and 20 exponential calculations for 10-dimensional data, which is much higher than our proposed algorithm.

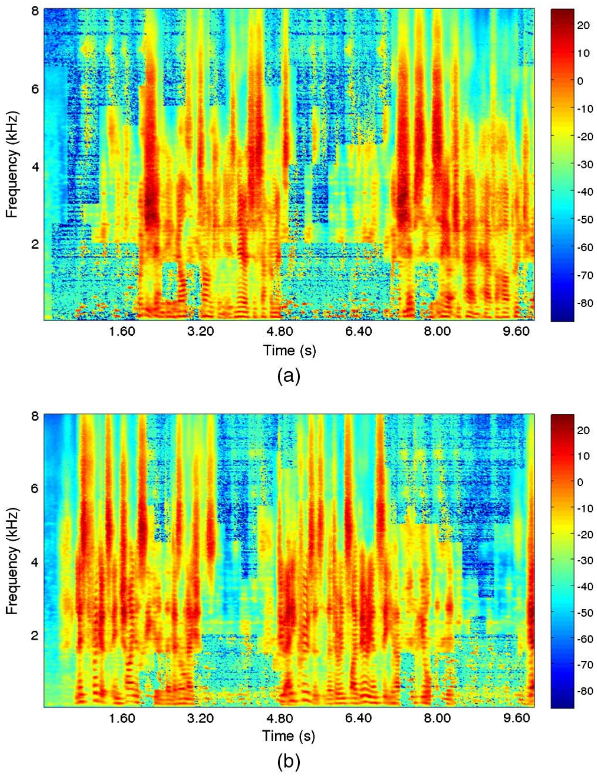
B. VAD-Incorporated BSS

Data, Parameter Setup: Considering the real-room time-invariant mixing process, the binaural room impulse responses (BRIRs) [46] were used to generate the audio mixtures, which are recorded with a dummy head in four reverberant rooms indexed by A, B, C and D, with reverberation times of [320, 470, 680, 890] ms respectively. These four rooms have different acoustic properties: a typical medium-sized office that seats 8 people, a medium small class room with the small shoebox shape, a large cinema style lecture theatre with soft seating and a low ceiling, and a typical medium/large sized seminar room with presentation space and a very high ceiling. To simulate the room mixtures, we set the target speaker in front of the dummy head, and we changed the azimuth of the competing speaker on the right hand side, varying from 15° to 90° with an increment of 15°. The source signals from the LiLIR dataset [43], each lasting 10 s, were passed through the BRIRs to generate the mixtures. In each of the six interference angles, 15 pairs of source signals were randomly chosen from the testing sequences associated with the target speaker and the competing speaker. This essentially facilitates the quantitative evaluations of the average performance of the proposed method under different room environments and for different speakers. When applying our proposed interference removal algorithm to a set of audio samples from another multi-modal database XM2VTS [47], we observed similar performance improvements as shown later, which further confirms the robustness of our proposed interference detection scheme. Due to space limitations, these results are omitted here. To test the robustness of the proposed algorithm to acoustic noise, Gaussian white noise was added to the mixtures at a SNR of 10 dB.

To implement the interference removal, we set $\sigma = 0.2$ which is found as follows. We varied its values among $\{0.01, 0.1, 0.2, 0.3, 0.4\}$ in Room A without additional noise, and obtain the average PESQ results from 90 independent tests, which are listed as $[2.57, 2.58, 2.58, 2.58, 2.57]$. Comparing the above results, the value of 0.2 is therefore chosen. We set $\delta_m = 3$ for the 2D smoothing in the time dimension. The value of δ_m equals δ_1 such that the 2D filter balances the temporal smoothing with the spectral smoothing. Each of the half-overlapping blocks spans the TF space of $L^m \times L^\omega = 64 \times 20$, which is equivalent to 1 kHz $\times$ 320 ms when N_{FFT} is set to 1024. L^m is set to span 320 ms (64 samples) so that it can cover the mean duration of English phonemes, even for the short vowels such as /oh/ (310 ms) [48], in which period the temporal structure of a phoneme can be captured and compared with the block correlation. To choose L^ω, we varied its values among $\{10, 20, 40\}$ in Room A in noise-free conditions, and evaluated the average PESQ results from 90 independent tests, which are listed as $[2.57, 2.58, 2.55]$ respectively. As a result, the value 20 is chosen.

The coefficients for fitting the boundary curves were obtained as $[q_3, q_2, q_1, q_0] = [-27.30, 109.65, -138.29, 57.41]$ and $[b_3, b_2, b_1, b_0] = [-22.15, 84.29, -98.08, 37.23]$.

We compared our proposed AV-BSS method, i.e. 'Proposed', with several other competing algorithms. The first one is Mandel's state-of-the-art audio-domain method, as introduced in Section IV, which we denoted as 'Mandel'. The second one uses IBM [41], denoted as 'Ideal', assuming contributions of each source signal to the audio mixtures are known in advance. The third one is proposed in our previous paper [49], which exploits the AV coherence to address the permutation problem associated with ICA, denoted as 'AV-LIU'. The forth one is the combination of the proposed interfering removal scheme with the ground truth VAD, which is manually

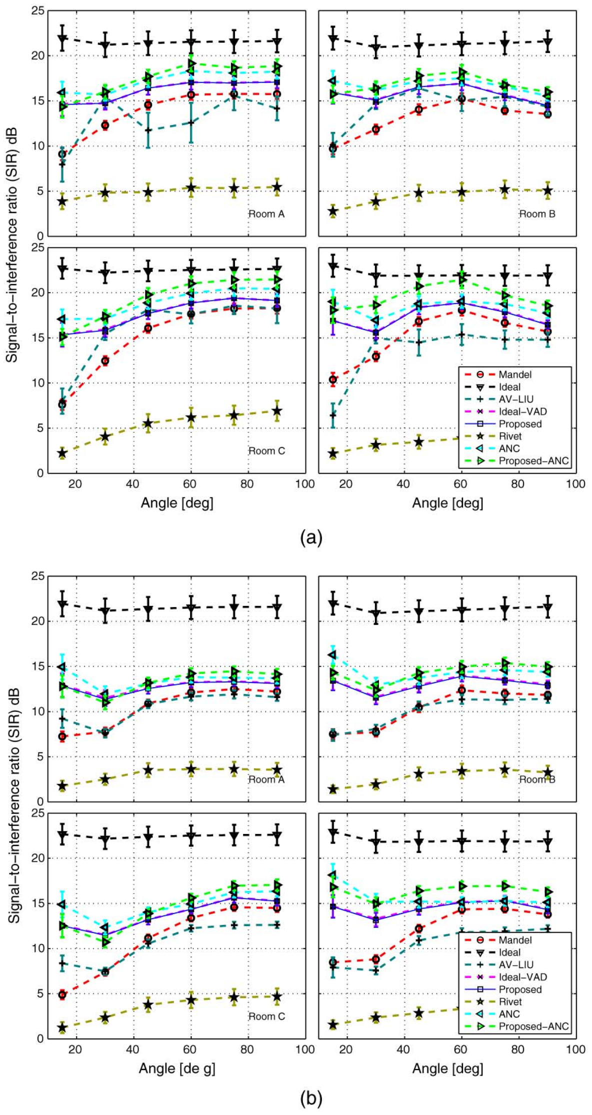


(a)

(b)

Fig. 7. Spectrograms of the source estimates after applying the interference removal scheme to enhance audio-domain BSS. The magnitude spectra are plotted in a decibel scale. Spectrograms of the associated original source signals are plotted in Fig. 2(a) and (c), and the spectra obtained by the audio-domain algorithm proposed by Mandel *et al.* are shown in Fig. 2(b) and (d). With further interference reduction, enhanced spectra with much less residual distortion are obtained, as demonstrated in the above figure. (a) Magnitude spectrum of source 1 estimate after the VAD-BSS. (b) Magnitude spectrum of source 2 estimate after the VAD-BSS.

labelled from the testing data, denoted as 'Ideal-VAD'. The fifth one is proposed by Rivet *et al.* [50] denoted as 'Rivet', where the visual VAD cues detected by our proposed visual VAD algorithm are used to regularise the permutations, which is also an ICA-based algorithm as [49]. The sixth one uses adaptive noise cancellation, denoted as 'ANC' [14]. 'ANC' is essentially a post-processing scheme based on least squares optimisation, which is proposed to remove cross interference in subband BSS algorithm. In our tests, the subband BSS based on the Informax algorithm [5] as used in [14] seems to work well for instantaneous or anechoic mixtures. However, the convergence of this algorithm was numerically unstable when tested on reverberant mixtures such as those used in our experiments. For this reason, we directly combined Mandel's method with ANC, in order to provide a fair comparison between the post-processing methods that we considered. In the seventh algorithm, we combined 'Proposed' with 'ANC', denoted as 'Proposed-ANC' for further interference reduction. First, the 'Proposed' and the 'ANC' methods are applied in parallel to the BSS outputs, to obtain two independent streams of enhanced target estimates. Then we obtain the final enhanced output by the integration of the above two enhanced target estimates as follows: we copy the TF points spanned by the blocks detected as interference from the enhanced target output via 'Proposed' to the final output, and copy the other TF points from the enhanced target output via 'ANC' to the final output.



(a)

(b)

Fig. 8. The signal-to-interference ratio (SIR) comparison for four different rooms (a) without and (b) with 10 dB Gaussian white noise corruption. The higher the SIR, the better the performance. The thin solid line represents the performance of the proposed algorithm.

To evaluate the performance of the BSS methods, we used the objective signal-to-interference ratio (SIR) and the perceptual evaluation of speech quality (PESQ) [51] as evaluation metrics.

Results Comparison and Analysis: We first show two examples of the enhanced audio spectrum after the interference removal scheme in Fig. 7. Our algorithm has successfully detected most of the interference block and attenuated them. The interference has been considerably reduced as compared to Fig. 2(b) and Fig. 2(d).

Then we demonstrated the SIR comparison as shown in Fig. 8. From this figure, we can observe that the 'Proposed' algorithm greatly improves the results obtained by 'Mandel', confirming that the visual information is effective in interference reduction for the separated speech. The spectral subtraction based method seems to be less effective as compared with adaptive filtering approach. This can be seen from the results by the competing 'ANC' algorithm [14] which achieves the second best results. It is, however, worth noting that the 'ANC' approach is computationally more expensive as compared with the spectral subtraction technique. We quantitatively evaluated the running time for both 'Proposed' and 'ANC', for each test, 10 s long speech data were used, and average time consumptions are about 0.6 s versus 42.8 s respectively (CPU: 3 GHz, RAM: 3.71 GB, MATLAB R2012a for Linux operating systems). We have also considered the combination of the visual-VAD based technique with the 'ANC', i.e. 'Proposed-ANC' which tends to give the best SIR results for all room types. In general, 'Proposed', 'ANC' and 'Proposed-ANC' achieve similar results for the above reverberant rooms, which are much better than the other baseline algorithms. The adaptive filtering techniques are the most popular techniques for reducing interference, our proposed algorithm achieves competitive results, which confirms the benefit of using visual modality to speech enhancement. Also, due to the easy-to-implement and efficient strategy, our proposed method provides the potential of real-time processing, as suggested by the average processing time presented earlier in this paragraph.

The 'ANC' method aims to attenuate the interference for each TF point of the BSS output. The 'Proposed' method, on the other hand, only attenuates the interference on certain TF points, i.e. in those TF blocks that are detected as interference based on VAD. In other words, the 'Proposed' method effectively reduces the interference when its residual in the BSS outputs remains relatively strong, but retains the TF points if they are detected as non-interference. For the TF blocks that are detected as being dominated by interference, the 'ANC' approach is less effective in removing the interference residual as compared with the 'Proposed' approach, since the update of the coefficients of the adaptive filter used in 'ANC' is dependent on the previous frames where the interference level may be low, thus leading to inaccurate estimation of the interference level of the current frame. We also notice that 'ANC' outperforms 'Proposed-ANC' for small angles for the following reason. In small-input-angle situations, the interference residual in the BSS outputs is much stronger as compared with the large-input-angle conditions due to the overlap of the binaural cues. As a result, more TF blocks are detected as interference, which include some TF blocks where the contribution from the target source is non-trivial and actually essential to speech quality. Such important information originating from the target speech, however, might get attenuated after applying Equation (19) once the associated TF block is detected as interference. This un-wanted over-subtraction results in slightly lower performance as compared to 'ANC'. Also, SIR evaluations of 'Proposed-ANC' and 'ANC' are not significantly different in Rooms A and B where the reverberation level is relatively low. This implies that the advantage of using spectral subtraction over the adaptive filtering technique is less significant for the attenuation of the interference residuals in the interference dominant TF blocks in low-reverberation conditions. However, for highly-reverberant environments, such as for Rooms C and D, the advantage of 'Proposed-ANC' over 'ANC' becomes more significant as shown in Fig. 8.

Also, the ICA-based AV-BSS methods 'AV-LIU' and 'Rivet' do not work very well, since ICA algorithms are limited in reverberant environments. The reverberant impulse responses are much longer than the FFT size in these room mixtures. As a result we could not accurately estimate the demixing filters. Interestingly, in Room A and C when the input angle is very small, 'AV-LIU' shows very modest improvement over the TF masking method in noisy environments. Overall, 'AV-LIU' outperforms 'Rivet', but neither of them cope very well with a high reverberation or noise level.

Comparing the results of 'Proposed' and 'Ideal-VAD', we could evaluate the VAD accuracy on the interference removal scheme. Since the VAD algorithm proposed by Aubrey *et al.* [27] achieved similar results as our VAD method, it was not combined with the residual removal scheme. We found that the curve of our 'Proposed' method almost overlap with the curve of 'Ideal-VAD'. There are two reasons behind this phenomenon. First, less than 10% error rate is involved for our proposed visual VAD algorithm as compared to the ground-truth, and there is approximately 3-second silence for each of the 10-second speech, which means that less than 300 ms silence is misclassified as active. However, the temporal length of each of the half-overlapping blocks spans more than 300 ms. In addition, most of the misclassified frames are scattered at different time points. Consequently, many block energy ratios are affected, but the influence is very small that the interference detection is not affected. Second, the VAD attenuation to the spectra changes only the block energy ratios, but not the block correlations. The interference detection, however, depends on both. In the misclassified frames, the spectra might have low correlation ratios. In that case, whether or not the frame is detected as silence, it will not be classified as interference. Due to the above reasons, the final estimates of 'Proposed' and 'Ideal-VAD' remain similar.

It is worth noting that the quality of the visual VAD is dependent on lipreading results, which will inevitably be affected by the quality and resolution of the images in the video signals used for liptracking. Head rotations and the distance between the subject and the camera can also affect the lipreading results. Despite the fact that we have used fairly good quality video signals in our experiments, the visual VAD assisted interference reduction algorithm is fairly reliable and robust against the inaccurate visual VAD results. The influence of the video quality on the BSS performance can be evaluated equivalently by assessing the influence of visual VAD accuracy on the BSS performance. For this reason, we synthetically generated inaccurate VAD results, by alternating between one-second silence and one-second activity. The overall accuracy of such a VAD is therefore 50%, which given we only have two classes is no better than chance. The inaccurate VAD results were then applied to the proposed interference removal scheme. Compared to 'Proposed', the SIR results with inaccurate VAD suffers a [1.0, 1.1, 1.2, 1.2] dB degradation for the four different room types respectively. Still, it outperforms Mandel's BSS algorithm with [1.2, 1.6, 1.5, 1.1] dB improvement. This is because there are still interference blocks being detected and removed, even though the number of detected blocks is smaller than the situation when the accurate VAD cues are applied.

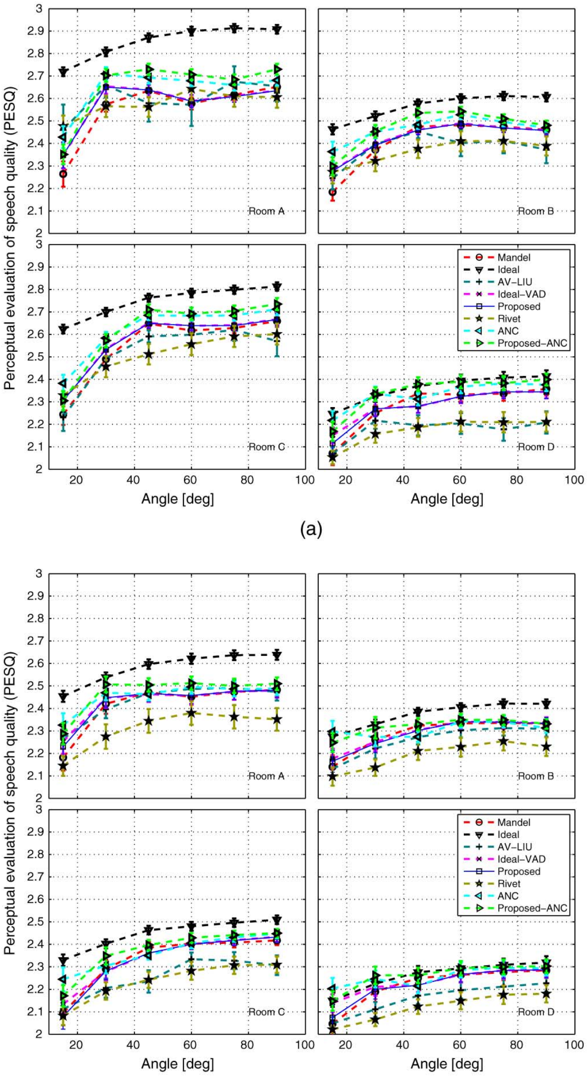


Fig. 9. The perceptual evaluation of speech quality (PESQ) comparison for four different rooms (a) without and (b) with 10 dB Gaussian white noise corruption. The higher the PESQ, the better the performance. The thin solid line represents the performance of the proposed algorithm.

We have also quantitatively evaluated the effect of the 2D Gaussian smoothing on the interference reduction method. To do this, we calculate the average SIR results in Room A for both noise-free and 10 dB noisy situations, by varying the standard deviations δ_1 and δ_2 of the 2D filters in Equation (11). δ_1 and δ_2 are scaled by the same factor to their default values as introduced after Equation (11), and the factor is chosen from the candidates [0, 0.5, 1, 1.5, 2], where 0 means no spectral smoothing and 1 the default smoothing. The average result for noisy and noise-free situations is obtained as: [13.72, 14.98, 16.31, 16.43, 15.35] dB respectively. It can be observed that the best results are obtained for the default values and their 1.5-scaled versions, which further confirms the benefit of applying Gaussian filters before the interference detection. If the smoothing process is not applied before the interference reduction, the performance of the proposed interference reduction scheme will be approximately 3 dB lower.

Finally we compared the perceptual metric using PESQ. From Fig. 9, we found that our VAD-assisted BSS method achieves better perceptual performance, for both noise-free and noisy room mixtures. The PESQ evaluations overall are consistent with the SIR results, even though the improvement is not as obvious as the SIR evaluations. When two speakers are near, the improvement is higher. However, some artificial distortion is introduced in our interference removal scheme, which degrades the accuracy to some extent. This is especially the case when the audio-domain method already successfully recovers the sources, for example, when two sources are far away from each other, i.e. large input angle. Therefore, the improvement is modest in that situation. In the most reverberant room D with noise, also the most adverse condition, our proposed algorithm can recover the source signals almost as good as the ideal masking, which shows the effectiveness of our method in real-world auditory scenes.

In our work, the use of high-resolution coloured video sequences, is mainly to demonstrate the proof of concept that the visual information of lipreading is very helpful for interference reduction of the associated audio utterance, and therefore can be used to improve the quality of speech sources separated by an audio-domain BSS algorithm. The proposed algorithm can be further improved when operated in a more realistic scenario, with potentially larger variations in e.g. image resolution, illumination, head movement, and/or the distance between the speaker and the video camera, which, however, is out of the scope of this work.

VII. Conclusion

In this paper, we have presented a system for the enhancement of BSS-separated target speech, incorporating the speaker-independent visual voice activity detector obtained in the off-line training stage. We have proposed a novel interference removal scheme to mitigate the residual distortion of traditional BSS algorithms, where the VAD cues are integrated to suppress the target speech in silence periods. Experimental results show that the system improves the intelligibility of the target speech estimated from the reverberant mixtures, in terms of both signal-to-interference (SIR) ratio and perceptual evaluation of speech quality (PESQ). However, due to the fixed block analysis, which is not flexible to the size variation in different phonemes, our algorithm cannot perfectly detect and remove all the residual. But using the same principle, combined with more advanced TF analysis techniques such as wavelet analysis, our method could be further improved for interference suppression, which is the priority of our future work.

Acknowledgment

The authors would like to thank the Associate Editor and the anonymous reviewers for their contributions toward improving the quality of the paper.

References

[1] E. C. Cherry, “Some experiments on the recognition of speech, with one and with two ears,” *J. Acoust. Soc. Amer.*, vol. 25, no. 5, pp. 975–979, 1953.

[2] C. Jutten and J. Herault, “Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture,” *Signal Process.*, vol. 24, no. 1, pp. 1–10, Jul. 1991.

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

[4] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non-Gaussian signals," in *Proc. Inst. Elect. Eng. F.*, Dec. 1993, vol. 140, no. 6, pp. 362–370.

[5] A. Bell and T. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, no. 6, pp. 1129–1159, 1995.

[6] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sep. 2002.

[7] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Appl. Signal Process.*, vol. 2003, pp. 1135–1146, Jan. 2003.

[8] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2000, vol. 5, pp. 2985–2988.

[9] H. Sawada, S. Araki, and S. Makino, "Underdetermined convolutive blind source separation via frequency bin-wise clustering and permutation alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 3, pp. 516–527, Mar. 2011.

[10] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.

[11] A. Cichocki, S. L. Shishkin, T. Musha, Z. Leonowicz, T. Asada, and T. Kurachi, "EEG filtering based on blind source separation (BSS) for early detection of Alzheimer's disease," *Clinical Neurophysiol.*, vol. 116, no. 3, pp. 729–737, 2005.

[12] S. Vorobyov and A. Cichocki, "Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis," *Biological Cybern.*, vol. 86, no. 4, pp. 293–303, 2002.

[13] J.-M. Valin, J. Rouat, and F. Michaud, "Microphone array post-filter for separation of simultaneous non-stationary sources," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 221–224.

[14] S. Y. Low, S. Nordholm, and R. Togneri, "Convolutive blind signal separation with post-processing," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 539–548, Sep. 2004.

[15] R. Aichner, M. Zourub, H. Buchner, and W. Kellermann, "Post-processing for convolutive blind source separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 5.

[16] K.-S. Park, J. S. Park, K. S. Son, and H.-T. Kim, "Postprocessing with wiener filtering technique for reducing residual crosstalk in blind source separation," *IEEE Signal Process. Lett.*, vol. 13, no. 12, pp. 749–751, Dec. 2006.

[17] C. Choi, G.-J. Jang, Y. Lee, and S. R. Kim, "Adaptive cross-channel interference cancellation on blind source separation outputs," in *Proc. ICA*, 2004, vol. 3195, pp. 857–864.

[18] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2008, pp. 45–48.

[19] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T Recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, no. 9, pp. 64–73, Sep. 1997.

[20] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[21] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[22] H. Mcgurk and J. Macdonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.

[23] D. A. Bulkin and J. M. Groh, "Seeing sounds: Visual and auditory interactions in the brain," *IEEE Trans. Neural Netw.*, vol. 16, no. 4, pp. 415–419, Aug. 2006.

[24] J.-L. Schwartz, F. Berthommier, and C. Savariaux, "Seeing to hear better: Evidence for early audio-visual interactions in speech identification," *Cognition*, vol. 93, no. 2, pp. B69–B78, 2004.

[25] P. Liu and Z. Wang, "Voice activity detection using visual information," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, May 2004, vol. 1.

[26] 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," *J. Acoust. Soc. Amer.*, vol. 125, no. 2, pp. 1184–1196, Feb. 2009.

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

[28] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Statist.*, vol. 37, no. 6, pp. 1554–1563, 1966.

[29] Q. Liu, W. Wang, and P. Jackson, "A visual voice activity detection method with adaboosting," in *Proc. Sensor Signal Process. Defence Conf.*, 2011.

[30] Y. Freund and R. Schapire, "A short introduction to boosting," *Japanese Soc. AI*, vol. 14, no. 5, pp. 771–780, 1999.

[31] J. Woodruff and D. Wang, "Binaural localization of multiple sources in reverberant and noisy environments," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 5, pp. 1503–1512, Jul. 2012.

[32] Q. Liu, W. Wang, P. Jackson, and M. Barnard, "Reverberant speech separation based on audio-visual dictionary learning and binaural cues," in *Proc. IEEE Statist. Signal Process. Workshop*, Aug. 2012, pp. 664–667.

[33] Q. Liu and W. Wang, "Blind source separation and visual voice activity detection for target speech extraction," in *Proc. 3rd Int. Conf. Awareness Sci. Technol.*, 2011, pp. 457–460.

[34] E.-J. Ong and R. Bowden, "Robust lip-tracking using rigid flocks of selected linear predictors," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recog.*, 2008.

[35] A. Liew, S. Leung, and W. Lau, "Lip contour extraction using a deformable model," in *Proc. Int. Conf. Image Process.*, Sep. 2000, vol. 2, pp. 255–258.

[36] K. S. Jang, "Lip contour extraction based on active shape model and snakes," *Int. J. Comput. Sci. Netw. Security*, vol. 7, pp. 148–153, Oct. 2007.

[37] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vision Pattern Recog.*, 2001, vol. 1, pp. 511–518.

[38] I. Titze, *Principles of Voice Production*. Englewood Cliffs, NJ, USA: Prentice Hall, 1994.

[39] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP–27, no. 2, pp. 113–120, Apr. 1979.

[40] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[41] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York, NY, USA: Springer, 2005, ch. 12, pp. 181–197.

[42] J. C. Platt, Microsoft Research, "Sequential minimal optimization: A fast algorithm for training support vector machines," Tech. Rep. MSR-TR-98-14, 1998, vol. Advances in kernel methods - support vector learning.

[43] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Cultural factors in the regression of non-verbal communication perception," in *Proc. IEEE Int. Conf. Comput. Vision*, Nov. 2011, pp. 1242–1249.

[44] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Jan. 2008 [Online]. Available: http://www.csie.ntu.edu.tw/cjlin/liblinear/

[45] J. Magarey and N. Kingsbury, "Motion estimation using a complex-valued wavelet transform," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 1069–1084, Apr. 1998.

[46] C. Hummersone, "A psychoacoustic engineering approach to machine sound source separation in reverberant environments," Ph.D. dissertation, Univ. of Surrey, Surrey, U.K., Feb. 2011.

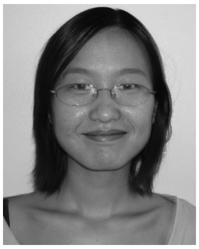
[47] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, 1999, pp. 72–77 [Online]. Available: http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb/

[48] G. E. Peterson and I. Lehiste, "Duration of syllable nuclei in English," *J. Acoust. Soc. Amer.*, vol. 32, no. 6, pp. 693–703, 1960.

[49] Q. Liu, W. Wang, and P. Jackson, "Use of bimodal coherence to resolve the permutation problem in convolutive BSS," *Signal Process.*, vol. 92, no. 8, pp. 1916–1927, Aug. 2012.

[50] B. Rivet, L. Girin, C. Serviere, D.-T. Pham, and C. Jutten, "Using a visual voice activity detector to regularize the permutations in blind separation of convolutive speech mixtures," in *Proc. Int. Conf. Digit. Signal Process.*, 2007, pp. 223–226.

[51] *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Recommendation P.862, International Telecommunication Union, Geneva, Switzerland, 2001 [Online]. Available: http://www.itu.int/rec/T-REC-P.862/en



Qingju Liu received the B.Sc. degree in electronic information engineering from Shandong University, Jinan, China, in 2008, and the Ph.D. degree in signal processing, under the supervision of Dr. W. Wang, from the Machine Audition Group at the Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, U.K., in 2013. Since October 2013, she has been working as a research fellow in CVSSP. Her current research interests include audio-visual signal processing, spatial audio reproduction, and machine learning.

Andrew J. Aubrey (S'04–M'08) received the B.Eng.(Hons.) degree in electronic engineering, the M.Sc. degree in electronic engineering, and the Ph.D. degree in audio-visual signal processing from Cardiff University, Cardiff, U.K., in 2002, 2003, and 2008, respectively.

He joined the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K., in July 2009 as a Postdoctoral Research Associate, where he worked in the area of modelling and animating human face dynamics and also perception of human expressions. From November to December 2011, he was a visiting researcher at the Cognitive Systems Lab, Korea University, Seoul, South Korea, and in August 2012 he was a guest scientist in the Human Perception, Cognition and Action Department, Max Plank Institute for Biological Cybernetics, Tübingen, Germany. Since August 2013, he has been with 3dMD LLC, Atlanta, Georgia, USA, as an R&D Engineer working on 3D image and video capture systems. His current research interests include audio visual speech processing, modelling and animation human face dynamics, speech separation, voice activity detection and analysis, and perception of human expressions and emotions.

Dr. Aubrey has been a reviewer for several IEEE journals, including IEEE TRANSACTIONS ON MULTIMEDIA, as well as several IEEE international conferences. He currently holds an Honorary Research Fellow position at the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K.

Wenwu Wang (M'02-SM'11) was born in Anhui, China. He received the B.Sc. degree in automatic control, the M.E. degree in control science and control engineering, and the Ph.D. degree in navigation guidance and control from Harbin Engineering University, Harbin, China, in 1997, 2000, and 2002, respectively.

He joined Kings College, London, U.K., in May 2002, as a postdoctoral research associate and transferred to Cardiff University, Cardiff, U.K., in January 2004, where he worked in the area of blind signal processing. In May 2005, he joined the Tao Group Ltd. (now Antix Labs Ltd.), Reading, U.K., as a DSP Engineer working on algorithm design and implementation for real-time and embedded audio and visual systems. In September 2006, he joined Creative Labs, Ltd., Egham, U.K., as an R&D Engineer, working on 3D spatial audio for mobile devices. In spring 2008, he was a visiting scholar at the Perception and Neurodynamics Lab and the Center for Cognitive Science, Ohio State University, Columbus, OH, USA. Since May 2007, he has been with the Centre for Vision Speech and Signal Processing, University of Surrey, Guildford, U.K., where he is currently a Senior Lecturer and a Co-Director of the Machine Audition Lab. He has authored and co-authored over 130 publications, including two books, *Machine Audition: Principles, Algorithms and Systems* (IGI Global, 2010) and *Blind Source Separation* (Springer, 2014). His current research interests include blind signal processing, sparse signal processing, audio-visual signal processing, machine learning and perception, machine audition (listening), and statistical anomaly detection.

Dr. Wang has been a member of the Ministry of Defence (MoD) University Defence Research Collaboration (UDRC) in Signal Processing since 2009, a member of the BBC Audio Research Partnership since 2011, and an associate member of Surrey Centre for Cyber Security since 2014. He has been a Principal Investigator or Co-Investigator on a number of research grants funded by the U.K. governmental bodies (such as the Engineering and Physical Sciences Research Council (EPSRC), Ministry of Defence (MoD), Defence Science and Technology Laboratory (DSTL), Home Office (HO), and the Royal Academy of Engineering (RAEng)), as well as the U.K. industry (such as BBC and Samsung (U.K.)). He has been a regular reviewer for many IEEE journals including IEEE TRANSACTIONS ON SIGNAL PROCESSING and IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and an associate editor of *The Scientific World Journal: Signal Processing*. He has also been a Chair, Session Chair, or Technical/Program Committee Member on a number of international conferences, including Local Arrangement Co-Chair of MLSP 2013, Session Chair of ICASSP 2012, Area and Session Chair of EUSIPCO 2012, and Track Chair and Publicity Co-Chair of SSP 2009. He was a Tutorial Co-Speaker on ICASSP 2013.