# Visual Based Reference for Enhanced Audio-Video Source Extraction

Jack Harris, Naqvi Syed Mohsen, Bertrand Rivet, Jonathon Chambers, Christian Jutten

# Visual Based Reference for Enhanced Audio-Video Source Extraction

Jack Harris*†, Syed Mohsen Naqvi†, Bertrand Rivet*, Jonathon Chambers†, Christian Jutten*
*GIPSA-Lab, CNRS UMR5216, Université de Grenoble, France
{jack.harris,bertrand.rivet,christian.jutten}@gipsa-lab.grenoble-inp.fr
†School of Electronic, Electrical and Systems Engineering, Loughborough University, UK
{j.a.chambers, s.m.r.naqvi}@lboro.ac.uk

**Abstract**

This paper addresses the problem of source extraction in a complex scene where only moving audio sources are present. An algorithm using a unique yet simple method avoiding higher-order statistics has been developed. The principle idea of the algorithm is to use a video camera array for locating a moving source whose position is used to isolate a noise reference, and thus allowing noise subtraction from the mixture based on the widely-known Widrow adaptive filtering method, that only uses second-order statistics. This adaptive approach provides an alternative to traditional methods particularly when there is need for a real time implementation.

## I. Introduction

The Cocktail Party Problem was first proposed in [1], which describes a problem where there are multiple human speakers talking simultaneously within an enclosed environment where it is required that each speaker's voice is isolated (separated) from the other present voices, similar to the manner in which a human sensory system can identify individual speakers in a situation such as a party, hence the name of the problem. Frequently such a problem is addressed using higher order statistics, a common solution is to perform independent component analysis (ICA) [2]. In some applications, particularly with moving sources in real time, it is desirable to use a more efficient method as higher-order methods can consume large amounts of system resources and require longer periods of time to produce accurate estimates.

Previous work into audio source separation using video has been limited. In [3], a method that exploits silent periods in speech is described. Whilst using video to identify these silent periods by tracking lip movements, it uses an ICA approach to separate speech. The authors in [4] present work that discusses video tracking and voice separation in meeting room environments with multiple speakers. A scenario which is much closer to the method in this paper is described in [5], however a modified version of the FastICA algorithm (rather than the adaptive separation method described here) is used in conjunction with speaker location information provided by an array of video cameras.

The proposed method uses video signals to identify the location of a 'target' source and uses this as a priori information to orientate a microphone array, so that the target source is equidistant from two microphones which work as a pair. By exploiting the equidistance property as well as some additional processing of the detected microphone signals, the noise reference (from a second source which is not equidistant to the pair of microphones) is isolated and used as a noise reference in an adaptive filtering scheme. The noise reference may be multiple speakers or background noise and the position is not critical to the functionality of the algorithm.

Experimental results are based on a set of audio recordings from a low reverberation environment, with sources located on a arc pattern in relation to an array of microphones. In this work, we simulate and evaluate the video cancellation performance by considering locations of the target source when it is equidistant to each microphone, and that the interference source does not match this equidistance property. Errors due to (low) reverberation or approximate positioning of the target are considered by shifts on the grid, breaking the equidistance property. Finally, performance of the 2-stage source enhancement algorithm can be evaluated in different situations, showing the advantages of the method.

## II. Algorithm

In this paper, we restrict the problem to the two-microphone and two-source case, where one source is the target source, $s_1(t)$, the other being the interference source, $n(t)$. The output of each microphone is then a convolutive mixture of target and interference source:

$$x_i(t) = \big(h_{i1} * s_1\big)(t) + \big(h_{i2} * n\big)(t) \tag{1}$$

on microphone $i$ ($i = \{L, R\}$), where $*$ denotes convolution, $h_{ij}$ is the impulse response between the jth source and the ith microphone, $t$ is the discrete time index, $n(t)$ is the original noise interference signal and $x_i(t)$ is the received signal at each microphone. Figure II shows how each room IR relates to each source and microphone.

The microphones within the array are physically linked on a pivot point. Thus, when the target source is located the microphones are orientated so that they are equidistant to the target speaker, it is assumed that a mechanical system is available

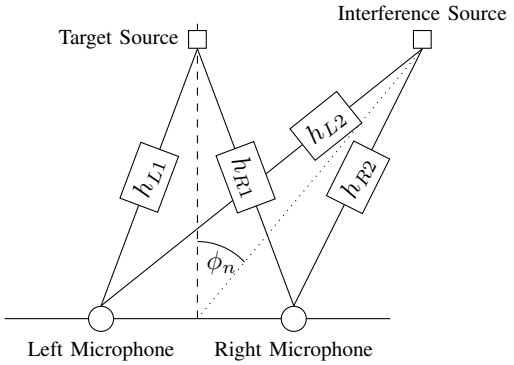



Fig. 1: Example scenario, where the variations of $h$ are impulse responses representing the room. The signal detected at each microphone is a convolutive mixture of each source. The dashed line indicates the imaginary line where, at all positions, the target source is equidistant to both microphones. $\phi_n$ denotes the angle of the interference source.

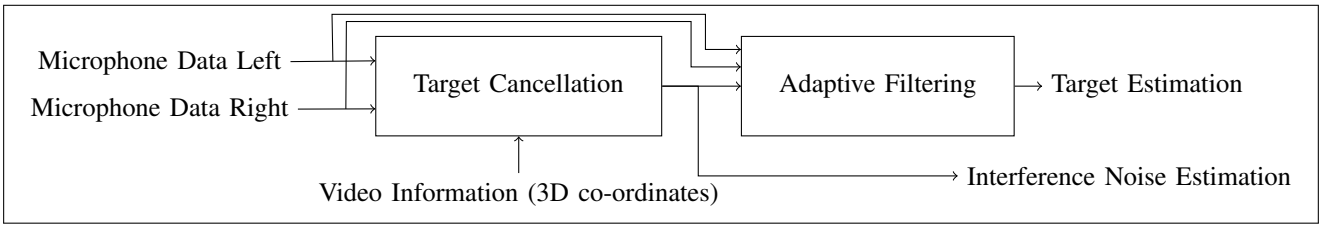



Fig. 2: System overview that shows the two main stages to the algorithm; target cancellation that produces a noise reference, and the adaptive filtering stage which recovers the target source. The work in this paper assumes that the location and the 3D co-ordinates of the speaker are known.

to orientate the microphone array. Then, subtracting signals from the two microphones, sample-by-sample, cancel the target source and thus provides a noise reference of the interference. Practically, the video-based cancellation is not perfect due to source localisation errors from the video and room reverberation.

The algorithm is based on two-stage architecture; (i.) a target source audio-video cancellation stage, to isolate the noise reference, then (ii.) an adaptive filtering stage to provide an estimate of the target source using the noise interference signal from the previous stage (Fig. 2). For computational simplicity the algorithm is implemented in the frequency domain, it is assumed that the original microphone signals have been converted using a short time Fourier transform (STFT).

Within the camera array it is assumed that there is a minimum of two cameras which are able to extract the 3D co-ordinates of a speaker's lips or head, similar to the method described in [5]. Although the algorithm described in this paper does not deal with identifying a speaker and extracting a set of 3D co-ordinates, which is covered in [5], [6], it is assumed that this is known and is simulated by placing a loudspeaker at specific distances and angles with respect to a microphone array.

### *A. Target Cancellation*

In an ideal scenario (when it is assumed that there is no reverberation in the room, there are no positioning errors and that the two microphones are equidistant to the target source) it is possible to subtract the microphone outputs on a sample-by-sample basis to isolate the interference source (i.e. $h_{R1}(t) - h_{L1}(t) = 0$), however in practice this is hard to achieve. The noise reference ($z(t)$) is found by:

$$z(t) = \big(h_{L1}(t) - h_{R1}(t)\big) * s_1(t) + \big(h_{L2}(t) - h_{R2}(t)\big) * n(t) \tag{2}$$

The goal of the target cancellation stage is to find an equalising filter $w(t)$ to ensure that; $\big(h_{L1}(t) - w(t) * h_{R1}(t)\big) = 0$.

A similar scheme to the one found in [7] is used to find $w(t)$, where equalising filters are pre-learnt by filtering 10 seconds of white noise through the previously known real room IR (RIR) which is then used to find two equalising filters using a normalised least mean squared (LMS) algorithm in the frequency domain. The filter $w(t)$ is found in the frequency domain, by:

$$W_i(k, l+1) = W_i(k, l) + \mu \frac{X_i(k, l)}{|X_i(k, l)|^2} \big[X_j(k, l) - W_i(k, l) X_i(k, l)\big] \tag{3}$$

where $k$ is the frequency bin index, $l$ is the time block index, $W$ are the frequency domain equalising filter weights, $X$ is the STFT of $x(t)$, $\mu$ is the step size parameter, $i = \{L, R\}$ and $j = \{R, L\}$ (i.e. when $i = L$, $j = R$ and vice versa). Once these filters are found they are used to perform (at each microphone) the cancellation:

$$Z_i(k, l) = X_i(k, l) - W_j(k, l) X_j(k, l) \tag{4}$$

where $Z$ is the noise reference signal for each channel and $i = \{L, R\}$, $j = \{R, L\}$. The overall noise reference, $Z$, is an average across two microphones, so that: $Z = (Z_R + Z_L)/2$.

RIRs that are used for training the equalisation filter have been calculated from data recorded in an anechoic room which is discussed in Section III-A.

### B. Adaptive Filtering: LMS Algorithm

The Least Mean Square algorithm is a widely known algorithm [8], it is ideal for real time applications due to its linear update equation in the time and frequency domains. Here it is implemented in the frequency domain due to the reduced computational complexity. At each frequency bin, $k$:

$$Q(k, l+1) = Q(k, l) - \mu e(k, l)\bar{Z}(k, l) \tag{5}$$

where $Q$ is the complex filter weight, $E$ is the error and $\bar{Z}$ indicates the complex conjugate of the noise reference. The error ($E$) is calculated by: $E(l) = X_i(l) - Y_i$, where $Y_i$ is the filtered output of each channel. As with the noise reference the average of the left and right outputs can be calculated.

## III. Results

The algorithm was tested using a 10 second speech utterance and a 10 second noise vector as the target signal and interference respectively, which are played through standard PC speakers. It is assumed the transfer function of the soundcard, loadspeaker and microphones used have a minimal effect on the RIRs. The RIRs were calculated based on recordings taken at GIPSA-Lab. STFTs where used with a window length of 1024 samples with 50% overlap. It is assumed that the signal-to-noise ratio of the interference signal and the target signal (to be cancelled) is 0dB.

### A. Anechoic Room Recordings

To test the algorithm, it necessary to have a set of RIRs suitable to this application. A variety of room recordings were taken using an array of 2 microphones, placed 0.06m apart, with omnidirectional response and a computer speaker placed at different positions in an anechoic room with low reverberation the room is the same as in [9]. At each position in the room a 50Hz - 10kHz linear frequency sweep of 5 seconds was recorded. In order to obtain the impulse responses between each microphone and the source (loudspeaker) position the cross correlation of the original frequency sweep and recorded version was calculated, using: $\gamma_{x_i s}(\tau) = h_i(\tau) * \gamma_{ss}(\tau)$ where, $x_i(\tau)$ is the recorded version of the frequency sweep, $\gamma_{x_i s}(\tau)$ is the cross correlation of the original frequency sweep and recorded version, $\gamma_{ss}(\tau)$ is the auto correlation of the original frequency sweep (which resembles a unit impulse response) and $h_i(\tau)$ is the impulse response. The graph in Fig. 3 is a typical impulse response, Fig. 4 shows the corresponding transfer function.

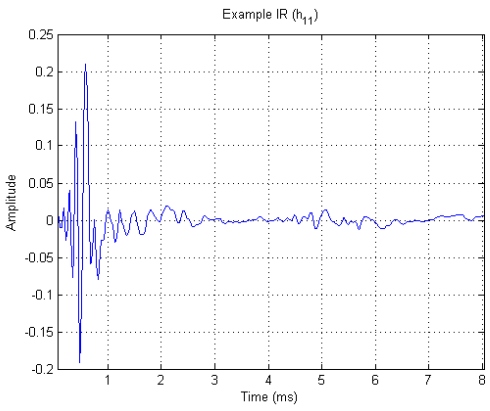



Fig. 3: Example RIR between the target source and the right microphone at a distance of 0.5m from the microphone array

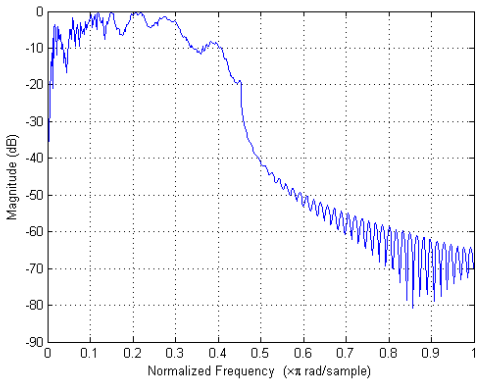


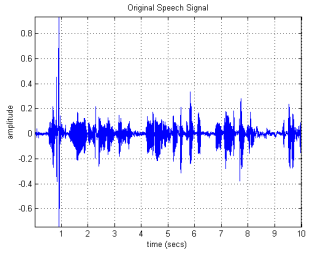

Fig. 4: Corresponding transfer function

### B. Algorithm Performance

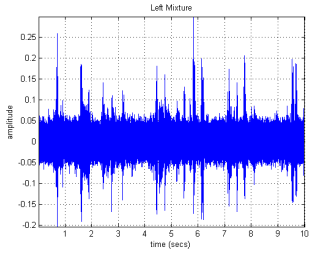
The algorithm performance is determined using source to distortion ratio (SDR) and the source to interference ratio (SIR) as described in [10]. Performance was recorded in the situation where the target signal was 0.5m in front of the microphones and the interference source was 0.6m away from the microphones at various angles (the angle of the interference source is denoted $\phi_n$) (Table I). The table shows that the performance of the algorithm is decreased when the target source and interference

| | Target Source 0 deg | | Target Source 5 deg | | Target Source 40 deg | | Target Source 60 deg | |
|---|---|---|---|---|---|---|---|---|
| $\phi_n$ (degrees) | SDR (dB) | SIR (dB) | SDR (dB) | SIR (dB) | SDR (dB) | SIR (dB) | SDR (dB) | SIR (dB) |
| 0 | -0.7 | 9.5 | 10.3 | 11.4 | 4.6 | 9.7 | -5.9 | -5.8 |
| 5 | 10.0 | 11.1 | 5.0 | 6.1 | 2.7 | 7.7 | -3.8 | 2.9 |
| 10 | 6.9 | 7.8 | 9.1 | 10.1 | 2.4 | 7.7 | -5.3 | -5.2 |
| 20 | 9.2 | 10.2 | 13.2 | 14.6 | 6.4 | 12.3 | -1.7 | 5.3 |
| 30 | 8.1 | 9.2 | 12.4 | 13.7 | 5.4 | 11.7 | 1.0 | 8.6 |
| 40 | 7.7 | 9.0 | 12.6 | 14.1 | -0.2 | 12.8 | -1.4 | 7.8 |
| 50 | 9.2 | 10.3 | 13.4 | 14.8 | 6.8 | 13.2 | -1.6 | 7.0 |
| 60 | 7.7 | 8.9 | 9.4 | 10.8 | 3.6 | 10.5 | -0.7 | 10.1 |

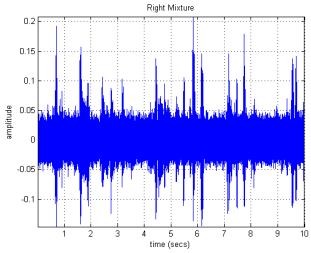
TABLE I: This table shows that the algorithm performs well at all angles of the interference source particularly when the target source is at either 0 or 5 degrees. The interference source angle is denoted by $\phi_n$.



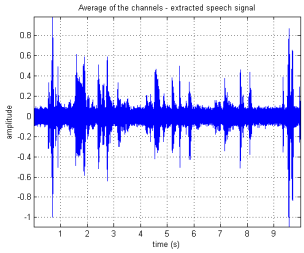

(a) Original speech signal provided to the source




(b) Left channel mixture




(c) Right channel mixture




(d) Output from the algorithm (enhanced target signal)

Fig. 5: Example outputs showing the key stages of the algorithm where the target signal is at 50cm and 0 degrees, and the interference source is at 60cm and 40 degrees.

source are in-line with each other (shown in grey in Table I). It is surprising to note that when the target is at 5 degrees it performs better that at 0 degrees, this may be due to positioning errors, however this is consistent across different sets of RIRs that were recorded. An improvement can be seen when the mixtures are compared to the output of the algorithm (Fig. 5). Most importantly, a difference of approximately $\pm 5$ degrees gives a good performance.

## IV. Conclusion

This paper discusses an audio-video source separation method for convolutive moving sources by using second-order statistics, which is achieved by using target source cancellation to provide a noise interference reference to an adaptive filtering (LMS) algorithm. By using adaptive filtering techniques based on simpler second-order statistics rather than higher order statistics, the audio-visual method has the potential to track a mobile target source and to save system resources when implemented in real-time. Future works include implementation of the audio-video canceller, extension to more than one interference source, the situation when the interference and target source cross paths (are inline with each other) and complexity analysis. Finally, in addition to this method, there is potential to use visual characteristics of a speaker (e.g. lip motions) for enhancing the source extraction.

## References


[1] E. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.

[2] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 2010.

[3] B. Rivet, L. Girin, and C. Jutten, "Visual Voice Activity Detection as a Help for Speech Source Separation from Convolutive Mixtures," *Speech Communication*, vol. 49, no. 7-8, pp. 667–677, 2006.

[4] D. Gatica-Perez, G. Lathoud, J. Odobez, and I. McCowan, "Audiovisual Probabilistic Tracking of Multiple Speakers in Meetings," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 2, pp. 601–616, 2007.

[5] S. Naqvi, Y. Zhang, and J. Chambers, "Multimodal Blind Source Separation for Moving Sources," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 125–128.

[6] S. Naqvi, M. Yu, and J. Chambers, "A Multimodal Approach to Blind Source Separation of Moving Sources," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 4, no. 5, pp. 895–910, 2010.

[7] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-Stage Binaural Speech Enhancement with Wiener Filter Based on Equalization-Cancellation Model," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA '09.*, 2009, pp. 133 –136.

[8] M. Dentino, J. McCool, and B. Widrow, "Adaptive Filtering in the Frequency Domain," *Proceedings of the IEEE*, vol. 66, no. 12, pp. 1658–1659, 1978.

[9] A. Van Hirtum and Y. Fujiso, "Insulation Room for Aero-Acoustic Experiments at Moderate Reynolds and Low Mach Numbers," *Applied Acoustics*, vol. 73, no. 1, pp. 72–77, 2012.

[10] C. Févotte, R. Gribonval, and E. Vincent, "BSS EVAL Toolbox User Guide," http://www.irisa.fr/metiss/bss_eval/, IRISA Technical Report 1706, Rennes, France, Tech. Rep. 1706, 2005.