REAL-TIME PHASE-ISOLATION ALGORITHM FOR SPEECH SEPARATION

David Ayllón [1], *Avram Levi* [2], *and Harvey Silverman* [2]

[1]Dept. of Signal Theory and Communications
University of Alcala
28871 Alcalá de Henares, Madrid, Spain
email: david.ayllon@uah.es

[2]Laboratory for Engineering Man/Machine Systems (LEMS)
Brown University
Providence, RI 02912, USA
email: hfs@lems.brown.edu
avram_levi@brown.edu

ABSTRACT

Beamforming techniques are applied to microphone arrays with the aim of separating sources and improving intelligibility, by means of spatial filtering. The non-stationary nature of speech implies the use of adaptive beamformers and several solutions have been implemented. Furthermore, interfering signals coming from the same direction as the target signal, cannot be filtered by the beamformer. The method presented in this paper is an alternative to adaptive beamforming, combining a simple delay-and-sum beamformer with a time-frequency masking method based on phase information. The beamformer is steered to the desired source and a function related to the phase differences between the steered signals at the microphones is evaluated to reject any interference that passed through the beamformer. Thus, the algorithm does not need to constantly adapt the filter coefficients and takes advantage of both beamforming properties and time-frequency separation techniques. The separation performance of the method has been evaluated in a noisy and reverberant environment using different arrays, talkers and scenarios. Real data are used to show the performance of the real-time algorithm when isolating one of the sources in the mixture.

1. INTRODUCTION

Isolating a speech source in a multi-talker environment has been a commonly-addressed problem for many years and it remains largely open and unsolved. Moreover, if the speech sources are in a real, closed space, such as a common room or office, their mixture is also contaminated by different types of background noise and echoes due to reflections, making the problem of separation more difficult. Finally, if the separation system is working in real-time, the algorithms applied for separation must be computable with minimal latency and thus be relatively simple.

Beamforming techniques and Blind Source Separation (BSS) methods are two different approaches for the speech separation problem. The former approach takes advantage of the spatial resolution given by a microphone array when sampling a mixture of sources, while the latter uses the knowledge of speech signal properties for separation, sometimes also with the aid of the spatial resolution of a microphone array.

Beamforming performs spatial filtering with the signals gathered by a microphone array using them to modify its beam pattern, enhancing the signal coming from the desired Direction Of Arrival (DOA), attenuating the others. This is very useful for speech separation when the beamformer is steered to the target source, attenuating the interfering sources coming from other directions such as reverberations or background noise. The spatial filter is achieved by introducing different attenuation and delay values in each of the M channels of the array and combining all of them. The beamformer design consists of calculating the best parameters for the M microphone channels. Beamforming was originally applied to narrowband signals and later adapted to wideband signals, for instance splitting the frequency spectrum into frequency bands and applying a different beamformer to each band, thus implying an FIR filter for each channel. In the case of non-stationary wideband signals, such as speech, the coefficients of these filters must be adapted constantly to track the changes of the signal (Adaptive Beamforming), increasing notably the computational cost. The Generalised Sidelobe Canceler (GSC) [1] is a typical linear structure for Adaptive Beamforming.

Different types of BSS methods exist: statistical-based algorithms, such as the one called 'Independent Component Analysis' (ICA) [2]; methods that relay on Computational Auditory Scene Analysis (CASA) [3]; or Time-Frequency masking methods [4], often referred to as the Degenerate Unmixing Estimation Technique (DUET) [5]. Time-Frequency masking methods compute a binary mask for the separation of one source. The mask is applied to the mixture signal and usually attempts to spectrally subtract all the time-frequency points belonging to interfering sources.

Beamforming techniques usually demand lower computational cost than BSS algorithms, and thus may be more suitable for real-time implementations. However even when beamforming is able to noticeably reduce the background noise and correlated interferences from different DOA's, its ability to reduce cross-talk interference is poor. The method described in this paper combines a simple beamforming technique with time-frequency masking in order to reject all types of interference present in a noisy, real-time environment. First, a simple delay-and-sum beamformer is steered to the target source, enhancing the signal coming from that direction. After that, the remaining interfering signal is removed by a binary mask that is computed for each time-frequency point according to a discriminant function that decides whether the point contains a high level of interference or not. The discriminant function is based on only the phase information of the M aligned channels of the beamformer, resulting in a algorithm that is computable in teal time.

In the next section we will describe the Phase-Isolation algorithm [7] and the parameters used to evaluate its performance. Section three shows the results of the real-time implementation of the algorithm, in different scenarios. Finally, section four gives some conclusions obtained from this work.

2. METHODS

2.1 The Phase-Isolation Algorithm

Consider an array λ of M microphones in a multi-talker, noisy and reverberant environment where we desire to isolate the speech source $s(t)$. We can model the signal at microphone j as:

$$m_j(t) = a(t-\tau_{js}) + i_j(t) \tag{1}$$

where $a(t)$ is an attenuated version of the direct-path target source $s(t)$, τ_{js} is the delay of $s(t)$ at microphone j, and $i_j(t)$ is the sum of all interfering signals at microphone j. These interfering signals can be either correlated or uncorrelated, depending on their origin. We can consider three different sources of interference: echoes of the target source due to the reverberation of the room, that are attenuated and delayed copies of $s(t)$, so they are correlated interferences; the speech signals, both direct-path and echoes, coming from the remaining talkers, that are uncorrelated sources of interference; and background noise, also uncorrelated with our target signal.

Let us assume that we know the position of the desired source and those of the microphones in the array, so we can calculate the time delays τ_{js} for each channel and steer the M microphones to the target source:

$$m_j^s(t) = m_j(t+\tau_{js}) = a(t) + i_j(t+\tau_{js}) \tag{2}$$

The previous expression can be rewritten in the time-frequency domain, for the lth frequency and mth frame, as follows:

$$M_j^s(l,m) = A(l,m) + I_j(l,m)e^{i\Omega\tau_{js}} \tag{3}$$

where Ω is the discrete angular frequency. Then, the output of the uniformly-weighted and normalized delay-and-sum beamformer is:

$$M_\lambda^s(l,m) = \frac{1}{M}\sum_{j=1}^{M} M_j^s(l,m) = A(l,m) + I_\lambda^s(l,m) \tag{4}$$

where $A(l,m)$ is the desired signal and $I_\lambda^s(l,m)$ is the interfering signal left at the output of the beamformer. Combining the M aligned signals results in a constructive addition of the target source and a destructive addition of the interferences as the interfering signal at each microphone is different. The beamformer has reduced some interfering signal that comes from other directions, but there remains a high level of cross-talk as well as noise and echoes coming from the steered direction. Thus, we can define the Signal-to-Interference Ratio (SIR) for the lth frequency in the mth frame as:

$$SIR(l,m) \equiv \frac{|A(l,m)|}{|I_\lambda^s(l,m)|} \tag{5}$$

The problem to tackle now is to identify those l frequency points in each frame m with low SIR and remove them from the output. For this purpose, we are going to use the Generalized Cross-Correlation with Phase Transform weighting function (GCC-PHAT) proposed in [6]. The GCC-PHAT function between microphones j and k is computed as follows:

$$\psi_{jk}^s(l,m) = \frac{M_j^s(l,m)M_k^{s*}(l,m)}{|M_j^s(l,m)||M_k^s(l,m)|} \equiv e^{i(\phi_j^s(l,m)-\phi_k^s(l,m))} \equiv e^{i\phi_{jk}^s(l,m)} \tag{6}$$

where $\phi_j^s(l,m)$ and $\phi_k^s(l,m)$ are the phases of the steered signals at the microphone j and k respectively, and $\phi_{jk}^s(l,m)$ their difference. This phase difference is directly related to the *SIR* function defined in (5): a phase difference of zero degrees implies $SIR(l,m) \to \infty$, and when this phase difference increases, the *SIR* decreases. In the method proposed in [7], the Steered Response Power (SRP-PHAT) of the beamformer was calculated from the GCC-PHAT function and used as discriminant function. In this work we took the real part of the sum of the GCC-PHAT functions for only unique pairs of microphones, instead of for all possible pairs, reducing the number of operations. Then, the function to evaluate is:

$$\gamma_\lambda^s(l,m) = \frac{2}{N(N-1)} Re\{\sum_{j=1}^{N}\sum_{k=j+1}^{N} \psi_{jk}^s(l,m)\} \tag{7}$$

and combining (6) with (7) we obtain:

$$\gamma_\lambda^s(l,m) = \frac{2}{N(N-1)} Re\{\sum_{j=1}^{N}\sum_{k=j+1}^{N} e^{i\phi_{jk}^s(l,m)}\} \tag{8}$$

The function $\gamma_\lambda^s(l,m)$ depends on $\psi_{jk}^s(l,m)$, which in turn depends on the SIR, so $\gamma_\lambda^s(l,m)$ is also related to the SIR. If the values of the phase difference $\phi_{jk}^s(l,m)$ are close to zero (high SIR), then the function $\gamma_\lambda^s(l,m)$ tends to 1, when $\phi_{jk}^s(l,m)$ increases (low SIR), $\gamma_\lambda^s(l,m)$ tends to 0. Figure 1 shows this effect for a mixture of real data. In order to discriminate points with low SIR, we set a threshold of SIR, that corresponds with a determined value R of the function γ_λ^s. Then we can use the γ_λ^s function as discriminator between points with high SIR and points with low SIR.

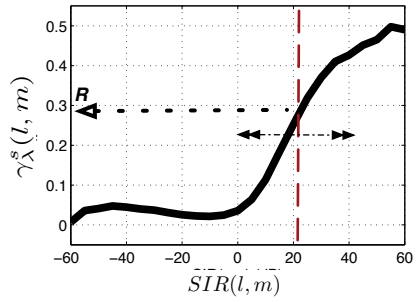


Figure 1: Mean values of $\gamma_\lambda^s(l,m)$ with relation to the SIR for real data. The value of R depends on the SIR threshold value chosen.

If the previous discriminant function is combined with noise spectral subtraction to remove the background noise, we can obtain a good separation of the target speech source from the rest of the talkers in an adverse environment.

The complete Phase-Isolation algorithm is summarized in the next steps:

1. Estimate the background noise spectrum, $M_\lambda^n(l,m)$, from a period of silence in the real environment.
2. Compute the Short-Time Fourier-Transform for the current frame.
3. Steer the microphones to the target direction and compute the output of the delay-and-sum beamformer.
4. Calculate the discriminant function $\gamma_\lambda^s(l,m)$.
5. Attenuate time-frequency points with low SIR at the output of the beamformer, evaluating the next two expressions:

$$\gamma_\lambda^s(l,m) > R \tag{9}$$

$$\frac{|M_\lambda^s(l,m)|^2}{|M_\lambda^n(l,m)|^2} > \rho \tag{10}$$

   If (9) and (10) are satisfied, the point has high SIR so the output is taken from the beamformer, and $\hat{S}(l,m) = M_\lambda^s(l,m)$. Otherwise the point has low SIR and must be attenuated to a small value μ, having $\hat{S}(l,m) = \mu$.
6. Reconstruct the signal in the time domain, $\hat{s}(t)$.

2.2 Algorithm Implementation

The algorithm was implemented in real time in C++, running on a PC Intel(R) Core (TM) i7 CPU 860 @ 2.80GHz with 3.49 GB of RAM operated with Microsoft Windows XP professional. For real-time data acquisition, an M-Audio Fast Track Ultra 8R sound card was used, allowing simultaneous sampling of 8 input channels and having also 8 output channels. The sound card is connected via USB interface to the PC. The sound driver user was ASIO for Windows.

The sample rate was 48 kHz and the frame length was 1024 samples. Thus, the speech was processed every 21.3 ms and played back to the output.

The interfering threshold R was set to 0.4 and the noise threshold ρ to 10 dB. The value of μ was equal to 0.001 to avoid musical noise. These parameters were set once for the specific room and used for all the experiments. Tests were carried out in a very noisy and reverberant room, with a T60 of around 350 ms. Our experimental system allowed the real-time monitoring of its output through headphones attached to the sound card.

2.3 Measurements

The method was tested with three different microphone arrays of 8 elements each, combined with several different positions of two sources and a total of 14 different settings. For each of these settings, 5 different measurements have been performed using different speech sources, selected randomly from the TIMIT database and mixing both male and female voices.

The three different types of arrays are shown in Figure 2. In ARRAY1, microphones are placed in a rectangular fashion, while in ARRAY2 in a equidistant line. ARRAY3 is similar to ARRAY1 but the distances between microphones are halved.

Figure 3 shows two different scenarios for the position of the sources and the array. Let us consider α_T and α_I be the angles that the target and the interference source forms with the center of the array. In the first scenario, the sources are placed forming the same angle with respect to the center

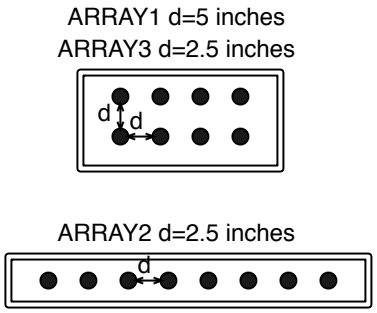


Figure 2: Microphone Arrays

of the array ($\alpha_T = \alpha_I$). In order to keep the reverberation times about constant and create a few different settings, only the distance d was varied and L kept constant. In the second scenario, the target source was located just opposite the array, having $\alpha_T = 0$ and α_I was varied, again keeping the distance L between sources and the array constant.

Table 1 summarizes the three different pairs of angle values that were tested for each scenario totaling 6 different configurations. ARRAY1 and ARRAY2 were tested with the 6 configurations while ARRAY3 only with configurations 1 and 4 in order to compare with ARRAY1

2.4 Quality evaluation

The W-Disjoint Orthogonality (WDO) quality factor introduced in [8] is a measure of the disjointness between the speech sources contained in a mixture and can be also applied to evaluate the quality of the separation for time-frequency methods. The WDO factor is computed from the Preserved-Signal Ratio (PSR) and the Signal-to-Interference Ratio (SIR). The PSR has the ideal value of 1 and represents the ratio of the energy of the desired signal that has been preserved by the masking algorithm. PSR is calculated from the expression:

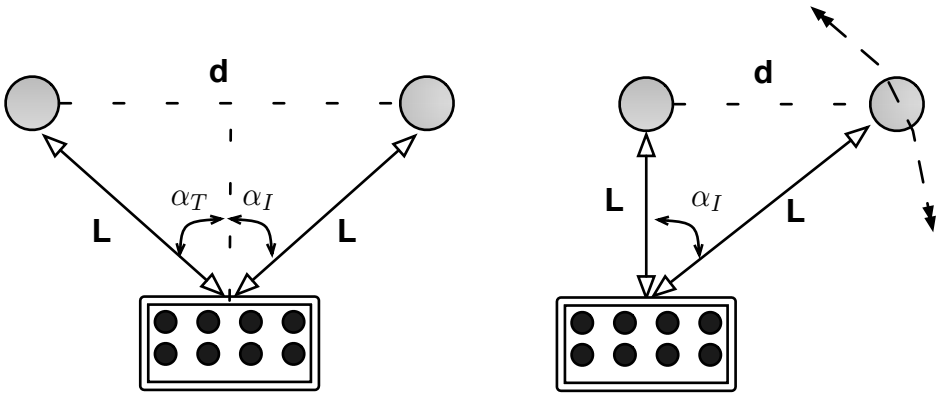
$$PSR_T = \frac{\sum_{l=1}^{L}\sum_{m=1}^{M}|M_T(l,m) * S_T^S(l,m)|^2}{\sum_{q=1}^{L}\sum_{r=1}^{M}|S_T^S(q,r)|^2} \tag{11}$$

where $M_T(l,m)$ is the binary mask computed for the separation of the target source and $S_T^S(l,m)$ is the non-mixed target speech source.

The SIR parameter measures the energy ratio of the target signal to the interference speech signal after separation, so the higher the value the better. In our case, we only have

Table 1: Different configurations of array-source positions

	α_T	α_I	d
CONF1	30	30	100 cm
CONF2	45	45	140 cm
CONF3	60	60	172 cm
CONF4	0	30	50 cm
CONF5	0	45	71 cm
CONF6	0	60	87 cm



(a) Scenario 1 (CONF1, CONF2, CONF3). Both speakers form the same angle with the array.

(b) Scenario 2 (CONF4, CONF5, CONF6). The target speaker is set in front of the array.

Figure 3: Schema of the two different scenarios for measurements. The circles represent the speakers and the rectangle represents the array. L is the distance from the speakers to the center of the array, and d is the distance between speakers.

one interference source, so the SIR is calculated from the expression:

$$SIR_T = \frac{\sum_{l=1}^{L}\sum_{m=1}^{M}|M_T(l,m) * S_T^S(l,m)|^2}{\sum_{q=1}^{L}\sum_{r=1}^{M}|M_T(q,r) * S_I^S(q,r)|^2} \quad (12)$$

where $S_I^S(l,m)$ is the non-mixed interference speech source.

The WDO factor is computed using the next equation from the PSR and the SIR and has an ideal value of 1.

$$WDO_T = PSR_T - \frac{PSR_T}{SIR_T} \quad (13)$$

To calculate the WDO factor, we need the non-mixed target source and the non-mixed interference source. These two non-mixed signals must be contaminated by the same interfering signals that the mixture. In order to have a good approximation of these two signals, we recorded the output of the beamformer when only one source was played. Thus, for the non-mixed target source, only the target signal is played and recorded by the beamformer aimed to that direction. The same procedure was followed for the interference signal, with the beamformer again steered to the target direction.

3. RESULTS

Table 2 shows the PSR, SIR and WDO parameters for the different configurations and arrays used to evaluate the algorithm. The three parameters were averaged, in every configuration, over 5 measurements with different speech sources. For calibration, according to subjective listening tests carried out in this work, a WDO value lower than 0.4 is noisy and unintelligible, a value around 0.5 is somewhat intelligible but still noisy, 0.6 usually means that the quality is quite acceptable, and WDO values above 0.7 corresponds with clean and intelligible speech. Generally high (> 0.7) WDO values are obtained for all configurations and arrays. Comparing ARRAY1 with ARRAY2, we can see that ARRAY1 obtains better results for the scenario 2, while ARRAY2 obtains better separation for the scenario 1. ARRAY3 obtains worse WDO values than ARRAY1 for the two configurations compared.

Table 2: Quality Results for Separation of the Real-Time Phase-Isolation Algorithm in the different configurations for 3 different types of array

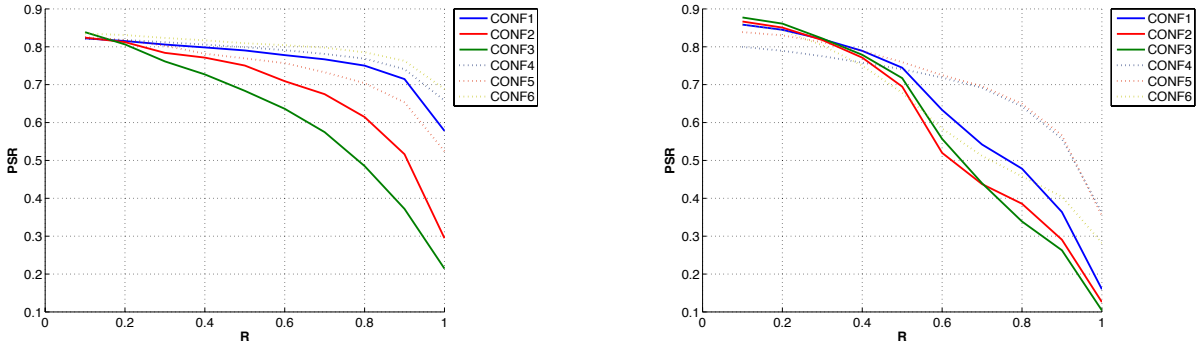
	CONF	*PSR*	*SIR*	WDO
ARRAY1	CONF1	0.795	28.523	0.774
	CONF2	0.770	17.206	0.717
	CONF3	0.728	17.664	0.682
	CONF4	0.805	15.087	0.755
	CONF5	0.787	20.867	0.750
	CONF6	0.814	67.208	0.812
ARRAY2	CONF1	0.788	80.006	0.762
	CONF2	0.776	99.167	0.764
	CONF3	0.776	45.734	0.750
	CONF4	0.759	28.208	0.712
	CONF5	0.787	56.346	0.757
	CONF6	0.758	44.499	0.731
ARRAY3	CONF1	0.787	16.447	0.712
	CONF4	0.815	8.694	0.627

This is likely due to the fact that, given they have the same arrangement, ARRAY1 has twice the aperture of ARRAY3.

Figure 4 shows the effect of varying the threshold R on the PSR and SIR parameters. By increasing R, we can see how we get a higher SIR, but also a smaller PSR, due to the discriminant function rejecting more points of signal of any type, resulting in an unintelligible output signal. It is clear that we must select a value of R that is a suitable compromise for a particular room and, perhaps, microphone/talker arrangement.

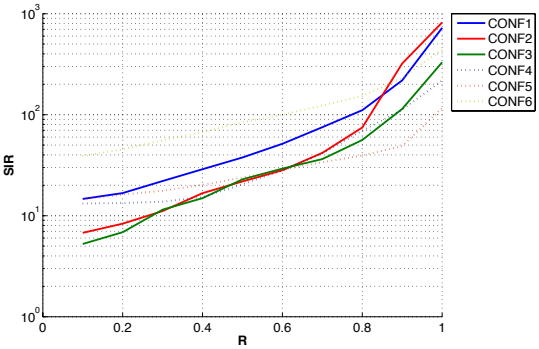
4. CONCLUSIONS

In this paper we implemented a modified version of the algorithm described in [7] in real time and tested the real-time version using real data and many settings. Our tests used a single interferer, but in a very noisy and reverberant room. To establish validity, the algorithm was tested in different scenarios, varying the positions of the sources, the array ge-

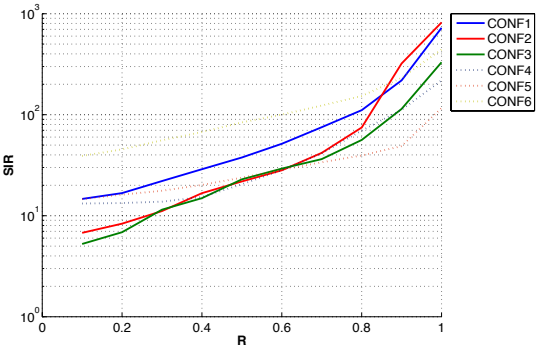


(a) PSR vs R for ARRAY1

(b) PSR vs R for ARRAY2



(c) SIR vs R for ARRAY1



(d) SIR vs R for ARRAY2

Figure 4: Effect of varying the threshold 'R'

ometries, and with different male and female speech sources. While it is difficult to convey speech quality to a reader of a paper, we assert the method significantly enhanced the isolation of the desired talker consistently and without introducing much distortion. The WDO numbers verify this assertion. We expect that the phase-isolation algorithm will find some important applications.

5. ACKNOWLEDGES

This work was carried out during a research visit of D. Ayllón at Brown University. During this period, the author has been funded by the Spanish Ministry of Science and Innovation, under project TEC2009-14414-C03-03, and the scholarship AP2009-3932.

REFERENCES

[1] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *EEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

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

[3] DeLiang Wang and Guy J. Brown, *"Computational auditory scene analysis: Principles, Algorithms, and Applications"*, IEEE Press/Wiley-Interscience, 2006.

[4] D. Wang, "Time-Frequency Masking for Speech Separation and Its Potential for Hearing Aid Design," *Trends in Amplification*, vol. 12, no. 4, pp. 332–353, 2008.

[5] S. Rickard, *"Blind Speech Separation"*, chapter 8: 'The DUET Blind Source Separation Algorithm", pp. 217–241, SpringerLink, 2007.

[6] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320 – 327, 1976.

[7] A. Levi and H.F. Silverman, "An alternate approach to adaptive beamforming using srp-phat," in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2726 –2729.

[8] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, 2004.