Synchronizing Speech Mixtures in Speech Separation Problems under Reverberant Conditions

Cosme Llerena, Roberto Gil-Pita, Lorena Álvarez, and Manuel Rosa-Zurera

Polytechnic School, University of Alcalá,
Ctra. Madrid-Barcelona Km 33.200, 28850 Alcalá de Henares, Spain
{cosme.llerena,roberto.gil,lorena.alvrezp,manuel.rosa}@uah.es
http://portal.uah.es/portal/page/portal/politecnica/

Abstract. Blind Source Separation (BSS) techniques aim at recovering unobserved source signals from observed mixtures (typically, the outputs of an array of sensors). Practically all classical BSS techniques do not work properly under reverberant conditions and therefore, it still remains an open problem. In this sense, we propose in this document the use of synchronization of speech mixtures in order to improve the results of classical BSS techniques. Specifically, we have applied the synchronization of mixtures combined with one of the most well-known and robust BSS algorithms that works under non-reverberant conditions, the Degenerate Unmixing Estimation Technique (DUET). In the aim of synchronizing speech mixtures prior to the speech source separation, the suitability of working with seven Time Delay Estimation (TDE) techniques has been analyzed. Results show the feasibility of using synchronization since the results of DUET are improved and additionally, it has been observed what is the most useful TDE algorithm in this framework.

Keywords: Speech Source Separation, Time-Delay Estimation, Convolutive Mixing Model, Reverberant Conditions.

1 Introduction

Blind Source Separation (BSS) [1], which was firstly proposed in [2], consists in recovering unobserved source signals from observed mixtures received at a set of sensors. This problem is named as "blind" since: 1) the mixing process is unknown and, 2) there is not much information about the characteristics of the source signals. In order to compensate this lack of information, different techniques and assumptions about the nature of the sources are made. There is a powerful technique underlying BSS algorithms, which is based on spatial diversity. Put it very simple, spatial diversity is a property of sensor arrays that relies on the fact of having more than one sensor and has been exploited in many applications such as, wireless communications [3]. With respect to the different assumptions, the mutual statistical independence of the source signals is broadly supposed [4]; the Independent Component Analysis (ICA) method [5] being a

L. Rutkowski et al. (Eds.): ICAISC 2013, Part I, LNAI 7894, pp. 568–579, 2013.

good example of a BSS algorithm working under this assumption. Apart from this realistic hypothesis, sparsity, which is another property of source signals, is commonly used. Sparsity has different definitions [6] and it is commonly assumed that a signal is sparse when all its energy is concentrated in just one coefficient and all others are zero (or almost zero). In the particular case at hand, since the signals correspond to speech sources[1], an appropriate transformation must be carried out aiming at achieving an adequate sparse representation of them. In this regard, it is well-known that a speech signal represented in the Time-Frequency (T-F) domain can be considered as sparse, since the energy due to speech is contained in a reduced number of time-frequency points and, in general, these points do not overlap with points due to other sources. With this in mind, the Short-Time-Fourier-Transform (STFT) may be applied to the speech sources.

In this sense, the popular Degenerate Unmixing Estimation Technique (DUET) [7] is a good example of a BSS algorithm that makes use of the STFT and aims at assigning each T-F point to one source. In the effort of associating each T-F point with one source or another, it calculates a binary mask that helps the algorithm decide whether a point belongs to a source or the other. These masks are obtained by means of two different ratios computed from the STFT. Being more explicit, these measures include the Inter-sensor Level Difference (ILD) and Inter-sensor Time Difference (ITD). From a mathematical point of view, let us suppose two mixtures ($\mathbf{x}_1$ and $\mathbf{x}_2$) and their STFTs ($X_1(\omega, k)$ and $X_2(\omega, k)$), the mentioned ratios are calculated as shown in Equations (1) and (2)

$$ILD = a_{21} = \frac{|X_2(\omega, k)|}{|X_1(\omega, k)|} \tag{1}$$

$$ITD = \delta_{21} = -\frac{1}{\omega} arg\left(\frac{X_2(\omega, k)}{X_1(\omega, k)}\right) \tag{2}$$

where ω is the index over the frequency bins and k labels the one over the time frames. In this point, it is highlight to mention a certain problem arising in this context when the mixtures are delayed more than the length of a time frame, what basically involves the T-F points do not coincide and then, the information extracted from the abovementioned ratios is wrong. Aiming at overcoming this problem, we propose in this paper, prior to the speech source separation problem carried out by means of DUET algorithm, to firstly synchronize the speech mixtures captured at the set of sensors (microphones in this case). In this sense, in [8], it is studied how clock synchronization affects the performance of sound source separation with a distributed microphone array.

The first step to synchronize the mixtures is to identify the delays. In the particular case at hand, speech mixtures can experiment two different delays. The first one is the propagation delay which involves the time required for the signal to propagate from the source to the microphones and, the second one is due to the synchronization of the microphones since, in a real study-case, it

[1] The task of recovering speech sources from audio mixtures is the so-called Blind Audio Source Separation (BASS) in the literature.

seems clear to note that the microphones involved in a set of sensors will not start the recording of the signal at the same time. Note that this latter delay should not exceed the length of a time frame. An example of a scheme to overcome the synchronization problem of distributed audio capture devices is shown in [9].

In this regard, the synchronization of single speech signals by means of Time Delay Estimation (TDE) algorithms has been widely studied in the literature [10,11]. Put it very simple, TDE is the process of determining the relative time shift between a reference signal and a delayed signal and lies at the core of many modern signal-processing algorithms. Different TDE algorithms have been proposed in both time and frequency domain. In this paper, we will focus on a set of very well-known TDE algorithms proposed for time domain.

Within these algorithms, the cross-correlation-based TDE algorithms are the most popular ones. In this kind of algorithms, the goal is to search the maximum value of the cross-correlation, since that value indicates when a signal and the shifted version of another signal have the maximum similarity. Aiming at enhancing the performance of these methods, a large number of improvements have been proposed [12] and they basically consist in introducing a filter or weighting function in the expression of the cross-correlation. These algorithms are known as Generalized Cross-Correlation (GCC) methods [13]. The objective of these algorithms is to make easier the search of the aforementioned maximum value. Examples of GCC methods include the Phase Transform Algorithm (PHAT) or the Roth Processor (ROTH), which both have been studied in this paper. Apart from these two methods, it has been also explored here the use of other algorithms such as, for instance, the Average Square Difference Function (ASDF) method or an adaptive algorithm like the Maximum Likelihood (ML) method.

In the speech signals framework, it is important to point out that the vast majority of the aforementioned TDE algorithms aim at estimating the delay under the assumption of *single* source signals, or in other words, only one speech source is contained in the signal or, at most, the speech signal with a signal due to noise. In the problem at hand, multiple signals are presented in the mixtures, and consequently, speech mixtures are more complex. Note that TDE algorithms for speech mixtures has seen little treatment in the literature so far. For illustrative purposes, in [14] can be found a TDE algorithm working with speech mixtures. It must also be mentioned a interesting work [15], where a very efficient scheme of synchronization combined with a BSS [16] method is proposed.

To sum up, we propose in this paper the synchronization of speech mixtures aiming at improving the results obtained with DUET algorithm in scenarios of convolutive mixtures, paying special attention to situations under reverberant effects. To be more precise, the speech mixtures are firstly synchronized by means of TDE algorithms and after that, the DUET algorithm is carried out. In order to evaluate the feasibility of using the study-case TDE algorithms, we have made use of the so-called signal-to-noise-ratio (SNR) between source signals and the estimated ones as will be shown in the numerical results.

The remainder of this paper is organized as follows. In Section 2, the speech separation problem is described. Section 3 contains the description of the TDE

methods that have been implemented in this paper. In Section 4, the experimental setup and the database used for the experiments are explained, along with the results obtained. Finally, Section 5 summarizes the conclusions of this work.

2 Speech Separation Problem

2.1 The Mixing Model

Fig. 1 illustrates the particular speech separation problem explored in this paper. As shown, $N = 2$ speech sources and $M = 2$ microphones are presented, what involves an even-determined case. Although it will be better understood later on, we can say in advance that this figure depicts the typical scenario in which two people are speaking in a room.

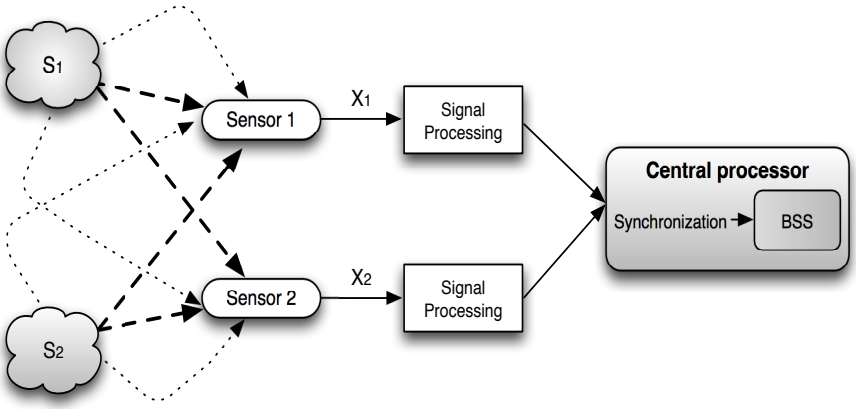

Fig. 1. An illustrative representation of the particular speech separation problem explored in this paper, that is, convolutive mixing model with noise and reverberation effects. Note that prior to the speech separation, mixtures are synchronized in a central processor by means of TDE methods.

In order to carry out the speech source separation, it is necessary to previously understand the way the mixing process happens. In our particular study-case, we suppose a convolutive mixing model in a noisy and reverberant environment. Convolutive mixing process refers here to the fact that the differences of delays that a speech source suffers among the different microphones are taken into account. Regarding the noise, it has been assumed an additive Gaussian noise with mean equals to zero and variance equals to σ^2. In addition, echoes of the target's reflected waves also have been considered due to the reverberant conditions.

Put it in a more mathematical way, it is assumed that at the discrete-time t, a set of N sources signals, that is, $\mathbf{s}(t) = [s_1(t), \ldots, s_N(t)]$ is received at M sensors that are part of an array sensor, $\mathbf{x}(t) = [x_1(t), \ldots, x_M(t)]$ being thus the received mixtures at the time t. This can be clearly observed in Equation (3)

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) * s_n(t), \ \forall \ m = 1, \ldots, M, \tag{3}$$

where any mixed signal is a linear mixture of filtered versions of each source signal and $a_{mn}(t)$ label the mixing filter coefficients which basically depend on the position of sources and microphones. And since noise and reverberation effects are considered, Equation (3) can be rewritten as shown in Equation (4)

$$x_m(t) = \sum_{n=1}^{N} a_{mn}(t) * s_n(t) + i_m(t), \ \forall \ m = 1, \ldots, M. \tag{4}$$

$i_m(t)$ being the sum of interfering signals at the discrete time t and at the microphone m. These interference signals may occur because of 1) the background noise and/or 2) echoes of the sources due to reverberation phenomena, which result in attenuated and delayed copies of the sources $s_n(t), \forall \ n = 1, \ldots, N$.

As stated in the Introduction, the studied BSS algorithms work in the T-F domain in order to obtain a sparse representation of the source signals and Equation (4) is thus re-written as depicted in Equation (5)

$$X_m(\omega, k) = \sum_{n=1}^{N} A_{mn}(\omega) \cdot S_n(\omega, k) + I_m(\omega, k), \tag{5}$$

where $X_m(\omega, k)$ and $S_n(\omega, k)$ represent the STFT for the ω-th frequency bin and m-th time frame of $x_m(t)$ and $s_n(t)$, respectively.

2.2 Source Demixing

As succinctly mentioned in the Introduction, DUET makes use of a time-frequency mask ($M_{\omega k}$) to separate speech sources in the T-F domain and this mask is calculated from Inter-sensor Level Differences and Inter-sensor Time Differences as explained in [7]. From a mathematical point of view, this mask is used as follows:

$$\widehat{S}_n(\omega, k) = M_{\omega k} \cdot X_m(\omega, k), \tag{6}$$

where $\widehat{S}_n(\omega, k)$ is the estimation of the n-th source and $X_m(\omega, k)$ labels the mixture at ω-th frequency bin and k-th frame for the m-th microphone.

Regrettably, in the case of speech mixtures, this mask may not work properly since the sparse property is not always correct because of the fact that there are contributions of different sources, echoes of these sources and so on.

3 Time Delay Estimation

The study-case TDE algorithms are explained in a detailed way in the paragraphs that follow. As previously mentioned, they have been chosen because they are well-known and robust methods for estimating delays between different

kinds of signals. In Section 4, we will explore how well these algorithms work in estimating delays for the case of speech mixtures. In order to explain the methods, it is worth mentioning that we denote the two speech mixtures of our study-case by ($\mathbf{x}_1$ and $\mathbf{x}_2$) and the delay between them by D_{12}.

3.1 Cross-Correlation (CC) Method

The cross-correlation between speech mixtures is calculated. If the length of the mixtures is T, the expression of the cross-correlation is shown in Equation (7)

$$R_{x_1x_2}(\tau) = E\left[x_1(t)x_2(t-\tau)\right],\ \forall\ 1 \leq t \leq T. \tag{7}$$

It is well-known that the delay between both mixtures can be obtained from the position of the maximum peak of the cross-correlation [12].

3.2 Phase Transform (PHAT) Method

This algorithm has been chosen since it has been widely used for estimating delays between acoustic signals arriving at spatially distributed microphones. PHAT method can be classified into the group of Generalized Cross-Correlation (GCC) methods, or in other words, a weighting function (ψ_p) is introduced in the expression of the cross-correlation, as it can be observed in Equation (8)

$$R_{x_1x_2}(\tau) = \int_{-\infty}^{\infty} \psi_p(f) G_{x_1x_2}(f) e^{j2\pi f\tau} df \tag{8}$$

where $G_{x_1x_2}(f)$ labels the cross-spectrum of the received signals and the weighting function responds to Expression (9)

$$\psi_p(f) = \frac{1}{\left|G_{x_1x_2}(f)\right|}. \tag{9}$$

This new weighting function can be very useful since it aims to sharpen the peaks of the cross-correlation by means of whitening the input mixtures, making easier to find the location of the maximum peak. Having a look at Expression (9), it seems clear to note that the information related to phase is preserved.

3.3 Modified Phase Transform (PHAT-β) Method

This modified version [17] of PHAT algorithm has been also studied. It has been shown that it provides very good results in estimating delays when signals are corrupted by both independent noise and reverberation effects. Within this algorithm, the weighting function is very similar to that of PHAT algorithm but in this case, a new parameter (β) is taken into account. The expression of this new weighting function can be observed in Equation (10)

$$\psi_{p\beta}(f) = \frac{1}{\left|G^{\beta}_{x_1x_2}(f)\right|}. \tag{10}$$

This parameter allows us to control the degree of whitening and limit the amount of degradation from the independent noise. Please note that β is a real number ranging from 0 to 1. If β is equal to 0, the algorithm is equivalent to CC method and if β is set to be 1, the algorithm is equivalent to PHAT method. In the case of intermediate values, a process of partial whitening occurs.

3.4 Maximum Likelihood (ML) Method

ML method [18] is also included in GCC methods and it has been selected since it works in systems where multipath effects are considered. It tends to obtain maximum likelihood solutions for TDE problems. Within this method, the weighting function responds to the expression shown in Equation (11)

$$\psi_{ML}(f) = \frac{1}{|G_{x_1x_2}(f)|} \frac{|\gamma_{x_1x_2}(f)|^2}{1 - |\gamma_{x_1x_2}(f)|^2} \tag{11}$$

where $|\gamma_{x_1x_2}(f)|^2$ is the magnitude squared coherency and it responds to Equation (12)

$$|\gamma_{x_1x_2}(f)|^2 = \frac{|G_{x_1x_2}(f)|^2}{G_{x_1x_1}(f) \cdot G_{x_2x_2}(f)}. \tag{12}$$

The ML function aims at increasing the accuracy of the calculation of the delay. It can be observed that the greater weight is assigned to frequency bands that give near-unity coherence. In the same line of reasoning as that in the previous methods, the maximum of the cross-correlation must be computed.

3.5 Roth Processor (ROTH)

ROTH processor [19] has been chosen since it has been proven to be very efficient in scenarios where additive noise is presented [13], by means of suppressing the frequency regions where noise is clearly presented. Within this algorithm, the weighting function has been found to be as follows:

$$\psi_{roth}(f) = \frac{1}{G_{x_1x_1}(f)}. \tag{13}$$

3.6 Smoothed Coherence Transform (SCOT)

SCOT method [20] has been used in many TDE applications where the presence of noise is important. In this case, the expression of the weighting function is as indicated in Equation (14)

$$\psi_{scoth}(f) = \frac{1}{\sqrt{G_{x_1x_1}(f) \cdot G_{x_2x_2}(f)}}. \tag{14}$$

It can be considered as a pre-whitening filter followed by a process of cross-correlation. Having a look at Equation (13), it seems clear to note that if $G_{x_1x_1}(f) = G_{x_2x_2}(f)$, SCOT method is equivalent to ROTH algorithm.

3.7 Average Square Difference Function (ASDF) Method

ASDF [21] method does not belong to GCC methods, since instead of using the cross-correlation function, it uses a difference function what involves lower usage of computational load, since multiplications are not needed. This difference function is the square error between the signals as shown in Equation (15)

$$R_{ASDF}(\tau) = \frac{1}{T}\sum_{t=0}^{T-1} |x_1(t) - x_2(t-\tau)|^2 . \tag{15}$$

By searching the minimum of the previous function, the delay between the signals is determined from its corresponding τ.

To sum up, it can be mentioned that these classical TDE algorithms have been chosen because they have demonstrated to have several advantages in classical TDE problems, not only in terms of computational cost, but also in robustness against the presence of noise, reverberations or multipath effects, etc. Then, we are interested in exploring their performances in our BSS problem.

4 Results

4.1 Experimental Setup

The sound database has been created from TIMIT database [22]. TIMIT database includes a total of 630 speakers (70 % male and 30 % female) of American English. The signals are 16-bit with a sampling frequency of 16000 Hz. From these speech signals, signals of different lengths have been obtained (0.25, 0.5, 1, 2, 4, 8 and 16 seconds). Frame size of the STFT (L_f) has also been set to different values (128, 256, 512, 1024 and 2048 samples), aiming at exploring the performance of DUET using the study-case TDE algorithms.

To carry out the experiments, we have set up a simple scenario that simulates the situation of two people talking simultaneously in a room of dimensions $6 \times 6 \times 3$ m. A 2-microphone array has been used and in order to simulate its response, the model that we use is the so-called Mirror Image model [23], which performs the microphone impulse response including room impulse response calculation. It considers both directivity pattern of the microphone and attenuation due to distance. For a number of N sources, the mentioned model considers that there are $(2 \cdot N+1)^3$ virtual sources to simulate the echoes of the speech sources. In this model, we have modified the reflection coefficient (C_r), from 0 (non-reverberant environment) to 0.2 in steps of 0.1 (reverberant environments).

To evaluate the performance of BSS algorithms, we have chosen a metric that considers the quality of the separated signals, to be more precise, the SNR between original and separated sources.

4.2 Numerical Results

In the aim of demonstrating the advantages of synchronizing the input speech mixtures in DUET algorithm, different experiments are carried out considering a

large number of parameter combinations (length of mixtures, STFT frame size, reflection coefficient, ...). Due to the large number of parameter combinations, all the results cannot be shown, nevertheless, the most important ones are presented. For example, it has been observed that the longer the length signal is, the better the results obtained are and this is the reason why the results for signals of 16 seconds length are shown in Table 1 and Table 2. Specifically, these tables show the mean SNR of 60 experiments between the separated and original sources in a non-reverberant and in a reverberant environment, respectively.

Deepening a little more in the results depicted in Table 1, the first row shows the values obtained by using DUET algorithm without synchronizing the speech mixtures, in the scenario proposed in Section 4.1. As illustrated, these values basically range from 3.26 to 3.54 dB, which are low in terms of speech quality and motivate us to explore the performance of synchronizing the speech mixtures. In the rest of rows in Table 1, the outcomes achieved thanks to the combination of the synchronization of the speech mixtures and DUET algorithm are shown, for different STFT frame sizes (L_f). Note that the TDE algorithms used in the synchronization process are explained in Section 3. It is also worth mentioning that for PATH-β algorithm, β is varied from 0.1 to 0.9 in steps of 0.1, although Table 1 only shows the cases in which the highest SNR is obtained, that is, for $\beta = 0.1, 0.2, 0.3, 0.4$ and 0.9. Looking at the SNRs obtained, it seems clear to note that an important increase of the SNR has been obtained when compared to those values obtained without synchronization, leading to reach values of SNR higher than 7 dB, what represents significant improvements. For the cases of shorter STFT frame sizes, especially for $L_f = 128$ and 256, an improvement of more than 70 % is obtained, reaching more than 100 % of improvement when longer STFT frame sizes are used, like, for example, for $L_f = 1024$ and 2048. Then, it is clear to note that the longer the STFT frame size is, the better the SNR obtained is and roughly speaking, this increase of SNR for longer frame sizes occurs with all the study-case TDE methods. For illustrative purposes, PATH-0.2 obtains a SNR equals to 5.80 dB for $L_f = 128$, whereas it reaches a SNR equals to 7.39 dB for $L_f = 2048$. It is interesting to note that PATH-β obtains in general very good results for all the frame sizes for low values of β (from 0.1 to 0.4), what it makes sense since PATH-β is especially designed for cases in which reverberation effects and noise are presented. Note that ASDF method decreases drastically its performance as the STFT frame size increases.

Table 2 illustrates very interesting information since a speech separation problem in a room under reverberation effects ($C_r = 0.2$, a typical reflection coefficient) is considered. Speech separation in reverberant conditions still remains an open problem since, due to its complexity, the vast majority of BSS algorithms do not achieve good results. Table 2 represents the same information as Table 1 but for a reverberant case. Looking at the first row of Table 2, DUET algorithm without synchronizing speech mixtures obtains lower SNRs than for the same situation without reverberation, these values ranging from 2.17 to 2.67 dB. It is important to note that, despite reverberation effects, an improvement close to 65 % has been obtained for the shorter STFT frame sizes and reaching an

Table 1. Mean SNR obtained by DUET without (first row) and using (the rest of rows) the different TDE algorithms, for the even-determined convolutive case of two mixtures and two sources, with noise and without reverberation effects ($Cr = 0$). 60 speech separation experiments have been carried out per each combination of parameters.

TDE	L_f=128	L_f=256	L_f=512	L_f=1024	L_f=2048
-	3.31	3.33	3.26	3.30	3.54
CC	5.79	5.82	6.11	6.84	7.37
PHAT	5.61	5.73	5.88	6.47	6.93
PHAT-0.1	5.79	5.82	6.11	6.84	7.35
PHAT-0.2	5.80	5.78	6.02	6.97	7.39
PHAT-0.3	5.85	5.78	6.03	6.84	7.24
PHAT-0.4	5.80	5.76	6.01	6.80	7.18
PHAT-0.9	5.56	5.54	5.85	6.58	6.92
ML	5.47	5.69	5.75	6.47	6.93
ASDF	5.79	5.82	6.29	5.32	4.39
ROTH	5.44	5.52	5.62	6.23	6.69
SCOT	5.72	5.66	6.00	6.50	7.01

Table 2. Mean SNR obtained by DUET without (first row) and using (the rest of rows) the different TDE algorithms, for the even-determined convolutive case of two mixtures and two sources, with noise and reverberation effects ($Cr = 0.2$). 60 speech separation experiments have been carried out per each combination of parameters.

TDE	L_f=128	L_f=256	L_f=512	L_f=1024	L_f=2048
-	2.39	2.17	2.41	2.67	2.56
CC	3.81	3.85	3.90	3.96	4.66
PHAT	3.84	3.89	3.92	3.99	4.47
PHAT-0.1	3.81	3.83	3.86	3.94	4.68
PHAT-0.2	3.84	3.83	4.03	3.97	4.65
PHAT-0.3	3.84	3.83	4.03	3.97	4.65
PHAT-0.4	3.81	3.81	3.96	4.04	4.64
PHAT-0.9	3.84	3.89	3.92	3.99	4.47
ML	3.83	3.88	3.89	4.01	4.46
ASDF	3.81	3.85	4.09	3.55	3.43
ROTH	3.73	3.68	3.81	3.83	4.52
SCOT	3.87	3.87	3.74	3.99	4.42

improvement of approximately 80 % for STFT frames of $L_f = 2048$. Unexpectedly, for the particular case of $L_f = 1024$, the improvement is lower, being about 50 %. Note that when $C_r = 0.2$, there is not a most appropriate TDE algorithm, since the results depend on the STFT frame size. As the reader can note, despite that DUET algorithm does not work properly for reverberant problems as the one proposed here, its results have been significantly increased (reaching more than 4.5 dB for the best cases), what leads to think about the idea of applying synchronization of speech mixtures with other BSS algorithms.

5 Conclusions

This paper focuses on applying synchronization of speech mixtures prior to the speech separation problem for BSS algorithms that use ILDs and ITDs, DUET being a very representative example. We have studied a convolutive mixing case with additive Gaussian noise and with or without reverberation effects, specifically, we have implemented a problem in a room using the Mirror Image Model to simulate the reverberation and multipath effects. We pay special attention to speech separation problems under reverberation effects due to its difficulty.

We have tested seven TDE methods in order to synchronize speech mixtures and different results have been obtained depending on some parameters as the reflection coefficient, STFT frame size, etc. Both in the non-reverberant case as in the reverberant one, an important improvement of the SNR has been obtained.

In the case without reverberation, a considerable increase of the SNR has been achieved, in some cases, *doubling* the value of SNR. According the STFT frame size increases, the SNR increases, for example, the 7.39 dB obtained by PATH-0.2 for a STFT frame size of 2048. We also realize that broadly, PATH-β method for values of β equal to 0.1, 0.2 and 0.3, achieves the better results, while ASDF method performs worse results with longer STFT frame sizes. The rest of TDE methods work achieving similar results. With reverberation effects, we have also improved the outcomes of the DUET algorithm, increasing the SNR close to 70% when longer STFT frame sizes. Unlike the non-reverberant case, all the algorithms achieve very similar results except ASDF method.

Therefore delays such as, the propagation delay of the sources or the delay due to synchronization of the microphones, do not affect the results of our BSS algorithm. To sum up, these results point out to a new filed of research in the jointly use of TDE and better adapted BSS algorithms to reverberant cases.

Acknowledgments. This work has been funded by the Spanish Ministry of Science (project TEC2012-38142-C04-02) and the Spanish Ministry of Defence (DEFENSA2011-10032110035).

References

1. Cao, X.R., Liu, R.: General approach to blind source separation. IEEE Transactions on Signal Processing, 562–571 (1996)
2. Hérault, J., Jutten, C., Ans, B.: Détection de grandeurs primitives dans un message composite par une architecture de calcul neuromimétique en apprentissage non supervisé. In: 10 Colloque sur le Traitement du Signal et Des Images, France (1985)
3. Diggavi, S.N., Al-Dhahir, N., Stamoulis, A., Calderbank, A.R.: Great expectations: The value of spatial diversity in wireless networks. Proceedings of the IEEE, 219–270 (2004)
4. Cichocki, A., Georgiev, P.: Blind source separation algorithms with matrix constraints. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, 522–531 (2003)

5. Te-Won, L.: Independent component analysis: theory and applications. Kluwer Academic Publishers, Boston (1998)
6. Hurley, N., Rickard, S.: Comparing measures of sparsity. IEEE Transactions on Information Theory, 4723–4741 (2009)
7. Yilmaz, O., Rickard, S.: Blind separation of speech mixtures via time-frequency masking. IEEE Transactions on Signal Processing, 1830–1847 (2004)
8. Zicheng, L.: Sound source separation with distributed microphone arrays in the presence of clock synchronization errors. In: Proc. Int. Workshop Acoustic Echo and Noise Control, IWAENC (2008)
9. Lienhart, R., Kozintsev, I., Wehr, S., Yeung, M.: On the importance of exact synchronization for distributed audio signal processing. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, vol. 4, pp. IV-840–IV-843. IEEE (2003)
10. Brandstein, M.S., Adcock, J.E., Silverman, H.F.: A practical time-delay estimator for localizing speech sources with a microphone array. Computer Speech and Language, 153–170 (1995)
11. Yegnanarayana, B., Prasanna, S.R.M., Duraiswami, R., Zotkin, D.: Processing of reverberant speech for time-delay estimation. IEEE Transactions on Speech and Audio Processing, 1110–1118 (2005)
12. Carter, G.C.: Coherence and time delay estimation. Proceedings of the IEEE, 236–255 (1987)
13. Knapp, C., Carter, G.: The generalized correlation method for estimation of time delay. IEEE Transactions on Acoustics, Speech and Signal Processing, 320–327 (1976)
14. Emile, B., Comon, P., Le Roux, J.: Estimation of time delays with fewer sensors than sources. IEEE Transactions on Signal Processing, 2012–2015 (1998)
15. Wehr, S., Kozintsev, I., Lienhart, R., Kellermann, W.: Synchronization of acoustic sensors for distributed ad-hoc audio networks and its use for blind source separation. In: Proceedings of the IEEE Sixth International Symposium on Multimedia Software Engineering, pp. 18–25. IEEE (2004)
16. Francourt, C., Parra, L.: The coherence function in blind source separation of convolutive mixtures of non-stationary signals. In: IEEE Workshop on Neural Networks for Signal Processing, pp. 303–312 (2001)
17. Donohue, K.D., Agrinsoni, A., Hannemann, J.: Audio signal delay estimation using partial whitening. In: Proceedings of the IEEE SoutheastCon, pp. 466–471. IEEE (2007)
18. Saarnisaari, H.: ML time delay estimation in a multipath channel. In: Proceedings of the IEEE 4th International Symposium on Spread Spectrum Techniques and Applications, pp. 1007–1011. IEEE (1996)
19. Roth, P.R.: Effective measurements using digital signal analysis. IEEE Spectrum 8, 62–70 (1971)
20. Carter, G.C., Nuttall, A.H., Cable, P.G.: The smoothed coherence transform. Proceedings of the IEEE, 1497–1498 (1973)
21. Jacovitti, G., Scarano, G.: Discrete time techniques for time delay estimation. IEEE Transactions on Signal Processing, 525–533 (1993)
22. Seneff, S., Zue, V.: Transcription and alignment of the timit database, TIMIT CD-ROM Documentation (1998)
23. McGovern, S.G: A model for room acoustics, `http://www.2pi.us/rir.html`