Mixture Models Applied to the Estimation of Mixing Parameters in Multi-channel Blind Source Separation Algorithms

César Clares-Crespo, Roberto Gil-Pita, Manuel Rosa-Zurera, Joaquín García-Gómez, and Inma Mohíno-Herranz
University of Alcalá, Department of Signal Theory and Communications, 28805-Alcalá de Henares, Madrid, Spain
Email: {cesar.clares, roberto.gil}@uah.es

***Abstract*—The Sound Source Separation (SSS) problem is treated depending on the premises that characterize a specific separation problem. Within some that carry out a blind separation the limitation come from the acoustic scene with the reverberation. It is a need to look for solutions focusing on these limitations. For methods based on a time-frequency approach where an estimation of the parameters of the mixture is required, we study different ways of estimation based on different probability density functions that can perform better in more disadvantaged acoustic scenes.**

***Index Terms*—audio signal processing, sound source separation, microphone array, mixture model, speech enhancement**

I. Introduction

Sometimes humans want to hear a desired sound source in the presence of other unwanted ones. One example of this occurs in a room where there are many people talking at the same time. That is why the sound source separation problem is commonly called "cocktail party problem".

Sound Source Separation (SSS) algorithms are present in a wide variety of application, from surveillance or medical applications to economic studies. Many of them are focused to speech and audio signals for purposes like noise suppression, speech enhancement, etc.

Within Blind Sound Separation (BSS) algorithms, DUET (Degenerate Unmixing Estimation Technique) described in [1], is a well-known algorithm that deals with the undetermined separation problem. Assuming an anechoic mixing model, it is able to recover any number of sources from only two mixtures.

Its main characteristic is to exploit the sparsity of the speech sources in the time-frequency domain, a property that assumes low probability of two sources having energy in the same frequency at the same time. It makes source separation trough binary masking feasible, as described in [2]. The construction of these binary mask is carried out once the mixing parameters have been estimated.

The mixing matrix is estimated by means of clustering the relative attenuation-delay coefficients between the two microphones for each source. In DUET, a clustering method by means of a two-dimensional histogram is proposed. Because of several studies, such as [3] and [4], this approach shows limitations because of reverberation and noise. Several works, such as [5] and [6], have been presented with alternative clustering methods based on the use of probability distributions that aim to shape the distribution of the mixing parameters.

The main problem of the DUET algorithm is that the estimation of the mixing parameters should be carried out with a two-dimensional histogram where the peaks of the clusters are the mixing parameters. In reverberant environments clusters are widened and so, some peaks can be hidden beneath other clusters.

This work tries to solve the estimation problem using Laplacian Mixture Models and Generalized Gaussian Mixture Models to shape the clusters of the mixing parameters. First, the mixture model distribution for the current problem must be defined. Then, an optimization function must be executed.

The goal is to find the best solution for the mixture model that shapes the distribution of the attenuation-delay coefficients obtained from the mixtures.

II. A Classical Clustering Method for the Estimation of the Mixing Parameters

In an acoustic scenario where there are only two sensors and S sound sources, assuming an anechoic mixing model, the mixture signals are composed by a combination of an attenuated and delayed version of each source signal.

Ignoring noise components, the anechoic mixing model in the time-frequency domain can be expressed as

$$\begin{bmatrix} Y_1(k,l) \\ Y_2(k,l) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-iw_k\delta_1} & \cdots & a_S e^{-iw_k\delta_N} \end{bmatrix} \begin{bmatrix} X_1(k,l) \\ \cdots \\ X_S(k,l) \end{bmatrix} \quad (1)$$

where a_s and d_s are the level and time differences between both microphones for the s-th source, respectively.

Manuscript received January 14, 2019; revised July 20, 2019.

doi: 10.18178/ijsps.7.3.85-91

The property of sparsity in speech sources is an approximation since the sources are quasi-sparse, in consequence, this property can be expressed as

$$X_s(k,l)X_j(k,l) \simeq 0,\ s \neq j \quad (2)$$

where $X_s(k,l)$ and $X_j(k,l)$ are time-frequency representations of two sources. Then, considering the validity of the sparseness of speech sources in the time-frequency domain, that is, only the j-th source is active in a point (k,l), Equation (1) results

$$\begin{bmatrix} Y_1(k,l) \\ Y_2(k,l) \end{bmatrix} \simeq \begin{bmatrix} 1 \\ a_j e^{-iw_k\delta_j} \end{bmatrix} [X_j(k,l)] \quad (3)$$

It is considered that the ratios between the mixtures do not depend on the source itself, but directly on the mixing parameters related to that source.

$$R(k,l) = \frac{Y_2(k,l)}{Y_1(k,l)} = a_j e^{-iw_k\delta_j} \quad (4)$$

Then, the local mixing parameters for each time-frequency point (k,l) are estimated through

$$\hat{a}(k,l) = |R(k,l)| \quad (5)$$

$$\hat{\delta}(k,l) = -\frac{1}{\omega_k}\angle\big(R(k,l)\big) \quad (6)$$

Please note that the local mixing parameter would be the mixing parameters only if the speech sources were strictly sparse, however, since sparsity is an approximated assumption, the local mixing parameters will cluster around the mixing parameters.

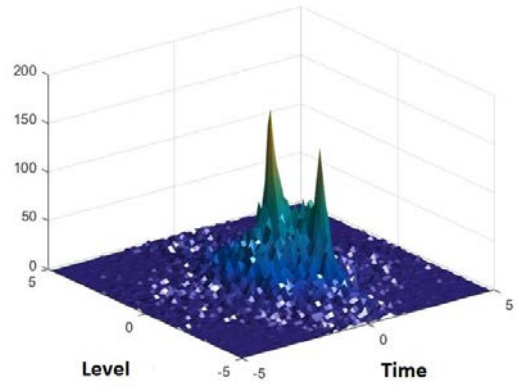

Figure 1. Two sources in the presence of noise.

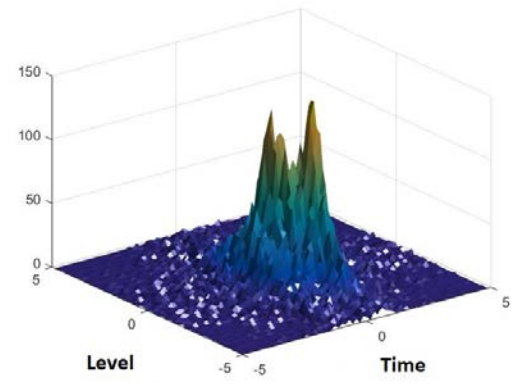

Figure 2. Two sources in the presence of reverberation.

The clustering process is carried out by means of a two-dimensional smoothed and weighted histogram. Fig. 1 and Fig. 2 show the two-dimensional histogram for different scenarios where two sources are active.

III. An Alternative Estimation Method Based on Mixture Models

The DUET algorithm does not propose any peak identification method for the two-dimensional histogram. Thus, in this section we propose some methods to identify the centers of the clusters. Also, the DUET algorithm is limited to work with only two sensors. So, another purpose of this work is to present a solution for a greater number of sensors.

A. Multi-dimensional Histogram

The separation problem can be carried out by using more than two microphones. Then, the mixing model expressed by Equation (1) can be generalized for any number of sensors. Due to the sparseness of the speech sources, the mixing model in a point (k,l), where only the source $X_j(k,l)$ is working, can be expressed as

$$\begin{bmatrix} Y_1(k,l) \\ \vdots \\ Y_m(k,l) \\ \vdots \\ Y_M(k,l) \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ a_{mj}e^{-iw_k\delta_{mj}} \\ \vdots \\ a_{Mj}e^{-iw_k\delta_{Mj}} \end{bmatrix} X_j(k,l) \quad (7)$$

So, the mixing parameters can be estimated for any number of microphones by means of a multidimensional histogram, defined in [3]. This multi-dimensional histogram is built of ratios between the mixtures $Y_m(k,l)$, $\forall m = 2,\dots,M$ and the reference one $Y_1(k,l)$. In order to simplify the mathematical formulation, the use of a vector $v(k,l)$ is proposed, to contain these ratios for each time-frequency point

$$v(k,l) = \left[\frac{Y_2(k,l)}{Y_1(k,l)} \quad \cdots \quad \frac{Y_m(k,l)}{Y_1(k,l)} \quad \cdots \quad \frac{Y_M(k,l)}{Y_1(k,l)}\right] \quad (8)$$

If speech sources are sparse in the time-frequency domain and simplifying this expression, the vector $v(k,l)$results

$$v(k,l) = \left[a_{2j}e^{-iw_k\delta_{2j}} \quad \cdots \quad a_{Mj}e^{-iw_k\delta_{Mj}}\right] \quad (9)$$

It can be observed that for one time-frequency point (k,l) where the j-th source is present, $v(k,l)$ will contain the mixing parameters of the j-th column of the mixing matrix.

If speech sources are strictly sparse, the mixing parameters will be in $v(k,l)$ but speech sources are quasi-sparse and so, the elements of $v(k,l)$will contain local mixing parameters that are clustered around the mixing parameters. So, our objective is to estimate the center of the clusters which will be the best approximation to the mixing parameters.

The vector $v(k,l)$ contains $M-1$ complex coefficients, each one of them has a level component (attenuation) and

a phase component (delay). Since attenuation and delay represent two different dimensions and their estimators are obtained from the separation of these components, the work in [3] proposes a new vector that contains $2(M-1)$ elements. Each one of these represents one dimension in the multi-dimensional histogram, consequently, there will be $2(M-1)$ dimensions. In each dimension there will be concentrations of points that we denote local mixing parameters.

$$w(k,l) = [a_{2s} \dots a_{Ms} \quad e^{-iw_k\delta_{2s}} \dots e^{-iw_k\delta_{Ms}}] \qquad (10)$$

B. Mixture Model Based on Multivariate Probability Distributions

Before going into the content of this section, in order to simplify the mathematical expressions, let us refer to the set of $K \cdot L$ time-frequency points by a new index $x = 1, \dots, N$ with $N = K \cdot L$, where each x represents univocally a time-frequency point (k, l). Thereby, the vector w(k, l) can be referenced in terms of the new variable as w(x). Furthermore, instead of identifying the frequency component, the variable $k = 1, \dots, K$ will be the index of the dimension of the histogram where K=2(M-1).

Within the set of points that belong to the s-th source in each dimension it will be observed that the values follow a certain Probability Density Function (PDF). Then, each PDF $f_{ks}(x_k)$ shapes the cluster of local mixing parameters
for the s-th source in the k-th dimension.

Since the components of w(x) result of the ratios between pairs of mixtures $Y_1(x)$ and $Y_m(x)$ with $m \neq 1$, the mixing model that has generated them will determine how the events are distributed in each dimension. To have a function that models these events across all dimensions we propose to use a multivariate probability distribution, defined in [7], [8]. We will define the joint probability density function $f_s(x_1, \dots, x_K)$ that models the multivariate probability distribution for the s-th source. And so, if the observations between two dimensions are independent, it can be expressed as

$$f_s(x_1, \dots, x_K) = \prod_{k=1}^{K} f_{ks}(x_k), \ k = 1, \dots, K \qquad (11)$$

where $f_{ks}(x_k)$ is the random variable of the s-th cluster in the k-th dimension.

In addition, it has been observed that in presence of different sources, different clusters will appear along each dimension. It will appear many clusters as sources are present in the mixtures. This S clusters will shape the local mixing parameters of the S sources, and them can be modelled as a sum of random variables. Therefore, the expression that models the entire multi-dimensional histogram will be

$$f(x_1, \dots, x_K) = \frac{1}{S} \sum_{s=1}^{S} f_s(x_1, \dots, x_K), \ s = 1, \dots, S \qquad (12)$$

where $f_s(x_1, \dots, x_K)$ is the multivariate probability distribution associated with the s-th source. Replacing equation (11) in (12), and considering that the sources could have different weight in the final expression, we define the generic expression with $s = 1, \dots, S$ and $k = 1, \dots, K$. The term W_s is the weighting factor of the s-th source in the multi-dimensional histogram.

$$f(x_1, \dots, x_K) = \frac{1}{\sum_{s=1}^{S} W_s} \sum_{s=1}^{S} W_s \prod_{k=1}^{K} f_{ks}(x_k\} \qquad (13)$$

It has been observed that an approximated symmetry occurs in the cluster along a dimension k. So, it can be considered that using probability density functions which have symmetry respect of the mean μ, a proper performance of the Equation (13) will be obtained when the mean μ_{ks} of the PDF's $f_{ks}(x_k)$ matches with the center of the clusters. Thus, our objective is to find such means μ_{ks}.

In this work, the use of an optimization algorithm is proposed to find these parameters searching the maximum of the likelihood function.

C. Optimization Function

Due to the complexity of the problem and, not being the purpose of this paper to develop an optimization algorithm we choose an existing one. Matlab provides useful tools to solve optimization problems. Here we will use the *fmincon* [9] which is part of the Matlab Optimization Toolbox.

This function has the purpose of searching the minimum of a function of several variables. This multi-variable function will be the function that defines our mixture model.

It should be highlighted that constraints can be added to the input parameters that the function evaluates. For example, it is possible to establish a lower and upper limit to the μ_{ks} terms. Thus, we can avoid the optimization to converge at a point where these terms would be outside the limits that allow the separation.

Another advantage is that by reducing the number of values that the optimization function must evaluate, we can significantly reduce the time spent on performance.

D. Laplacian Mixture Model (LMM)

It has been observed that, in conditions close to a real case, that is, within a convolutive mixing model and in the presence of noise, the shape adopted by the clusters has a certain similarity with a Laplace distribution. Consequently, we proposed the use of a mixture model based on Laplace distributions as the one presented on Equation (14)

$$f_{ks}(x_k) = \frac{1}{2b_k} exp\left(-\frac{|x_k - \mu_{ks}|}{b_k}\right) \qquad (14)$$

where b_k controls the width of the Laplacian distributions in the k-th dimension.

Let us define a Laplacian Mixture Model (LMM) with S centers of K dimensions, so that the center s-th has coordinates $[\mu_{1s}, \dots, \mu_{Ks}]^T$. Then, the probability density function $f(x_1, .., x_K)$ generated by the LMM can be obtained through particularizing the Equation (14) in Equation (13). Assuming that there is a cloud of N points contained in a matrix $X = [x_1, \dots, x_N]$ with K parameters of each point so $x_n = [x_{1n}, \dots, x_{Kn}]^T$. The likelihood function defines the joint probability density of all

observations and under independence conditions it can be expressed as

$$L(w,b,\mu)|x_{1n},\dots,x_{Kn}) = \prod_{n=1}^{N} f(x_{1n},\dots,x_{Kn}|w,b,\mu) \quad (15)$$

In practice, the logarithm of the likelihood function is usually used. Then, the objective is to determine the parameters (w, b, μ) that maximize the logarithm of the likelihood function. Hence, the Expression (16) is applied

$$MLL = argmax_{(w,b,\mu)}\ LL(w,b,\mu) \quad (16)$$

The parameters to be optimized are the S weight factors W_s, the K width factors b_k, and the S centers μ_{ks} with K values each one. So, in total, the number of parameters to be optimized is K+S+K·S.

Fig. 3 and Fig. 4 show in a two-dimensional case how the LMM can obtain the center of the clusters.

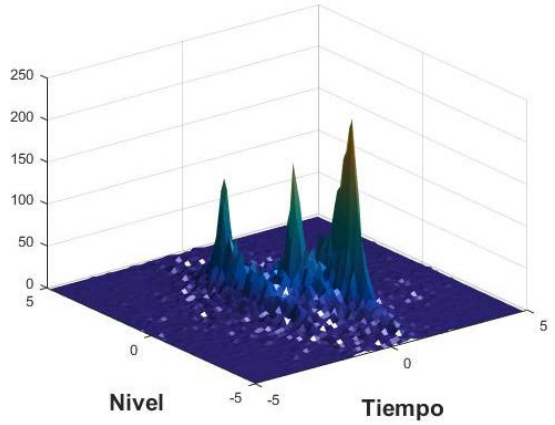

Figure 3. Two-dimensional histogram with three active sources.

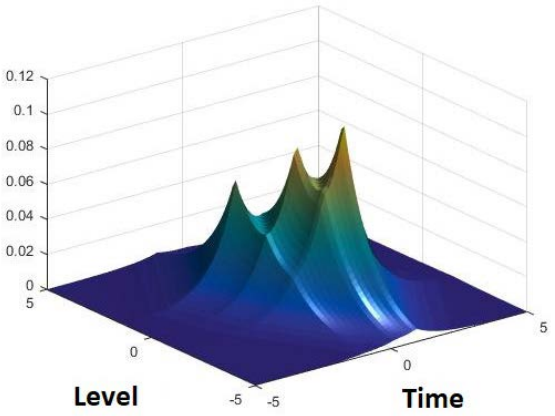

Figure 4. Two-dimensional LMM solution with three active sources.

E. Generalized Gaussian Mixture Model (GGMM)

As expected, the variations of noise level and reverberation in the acoustic scene have consequences in the distribution of the histogram. The clusters of the local mixing parameters can take forms that differ from Laplacian distributions. According to this, it is appropriate to define a model that allows some adaptability in conditions such as those described.

In this way, taking one more step in the approximation of the histogram, we propose to use a very similar model to the LMM. The novelty of this one is to include the shape parameter β. Then, our new model is based on Generalized Gaussian Distributions

$$f_{ks}(x_k) = \frac{\beta_k}{2\alpha_k \Gamma(1/\beta_k)} exp\left(-\left(\frac{|x_k-\mu_{ks}|}{\alpha_k}\right)^{\beta_k}\right) \quad (17)$$

where α_k controls the scale in the k-th dimension and β_k is the shape parameter in the k-th dimension. Note that in case of $\beta = 1$, $f_{ks}(x_k)$ is a Laplacian distribution. n case of $\beta = 2$, $f_{ks}(x_k)$is a Normal distribution.

Starting from a Multivariate Generalized Gaussian Distribution (MGGD) and bearing in mind the fact of different sources existing in the K dimensions, we need to define the Generalized Gaussian Mixture Model (GGMM) with S center of K dimensions that the center s-th has coordinates $[\mu_{1s}, \dots, \mu_{Ks}]^T$.

Therefore, the probability density function $f(x_1, \dots, x_K)$ generated by the GGMM can be obtained through particularizing Equation (17) in Equation (13). Then the likelihood function of this model must be obtained through

$$L(w,\alpha,\beta,\mu \mid x_{1n},\dots,x_{Kn}) = \\ = \prod_{n=1}^{N} f(x_{1n},\dots,x_{Kn} \mid w,\alpha,\beta,\mu) \quad (18)$$

In the same way that in the LMM, the objective is to estimate the parameters (w, α, β, μ) that maximize the logarithm of the likelihood function. Hence, is applied the expression

$$MLL = argmax_{(w,\alpha,\beta,\mu)}\ LL(w,\alpha,\beta,\mu) \quad (19)$$

The parameters to optimize are the S weight factors W_s, the K scale parameters α_k, the K shape parameters β_k and the S centers $\mu_{ks}$with K values each one. So, in total, the number of parameters to be optimized is 2K+S+K·S.

Fig. 3 and Fig. 5 show in a two-dimensional case how the GGMM can obtain the center of the clusters.

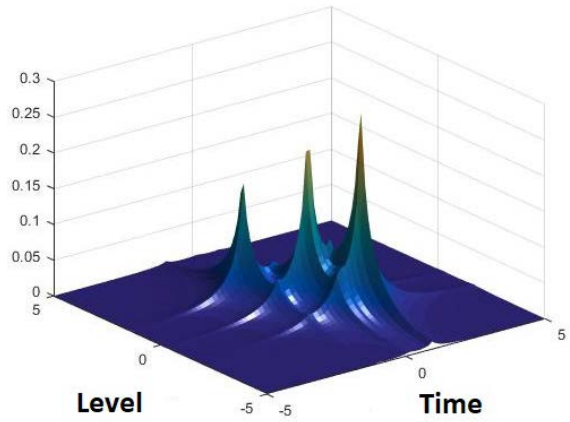

Figure 5. Two-dimensional GGMM solution with three active sources.

IV. Experimental Work

This section explains how will be evaluated the performance of the presented methods. In the first section the simulation process is detailed. Also, is presented the speech database used in the experiments. Then, two objective measurements are presented to quantify the quality of the separated speech signals.

A. Simulation and Database

This paper is intended to cover the separation problem applied to speech signals. So, a set of experiments has been carried out using the TIMIT database [10]. This database contains recordings of sentences said by different people. The signals have a length of 4 seconds with a sampling frequency of 10 kHz. Also, the RMS power of the speech signals has been normalized to 0 dB.

This experiment has been simulated in acoustic scenarios that consist of rectangular rooms where the array of microphones is at the edges of the room in random positions.

Similarly, the sources are randomly distributed but throughout the space of the room. Rooms of different sizes have been simulated, from 4x3x2 m to 20x20x5 m, also with different levels of reverberation and noise.

The microphone signals have been obtained filtering the source signals by the room impulse response obtained with the RIRG function, free implementation is given in [11]. This function simulates the reflections of a room based on the image method, which was first proposed in [12]. Then the additive noise is added to each microphone signal.

B. Speech Quality Measurements

To evaluate the quality of the processed speech two different measures has been proposed based on a comparison of the separated source signal and the original one.

1) Signal to Noise Ratio (SNR). This measures the degradation of the separated source from its original version. The process is carried out through a sample to sample comparison between both signals. In summary, this measure shows the proportion between the power of the original signal and the power of the error signal.
2) Short Time Objective Intelligibility (Time). This popular measure is proposed in [13] and it is a popular function that estimates a correlation coefficient between the envelopes of the clean and processed signal. It works by means of a segmentation of the signal in time frames of approximately 400 ms. A Short Time Fourier Transform is applied to the signals. Then, grouping the DFT bins in one-third octave band it performs a comparison between both signal for each frame. Finally, it provides the average value of the correlation coefficient along all the frames.

V. Results

Once the separation process has been carried out in our proposed acoustic scenarios, a set of results is introduced here. As mentioned above, different reverberation levels have been tested depending of the reflection coefficient of the walls Cr. These are one without reverberation (Cr = 0), one with low reverberation (Cr = 0.3) and one with high reverberation (Cr = 0.6). Also, the experiments have been carried out with different noise levels.

Table I and Table II show the average values of STOI and SNR, respectively, with -30 dB of additive noise in the microphone signals.

Having a look at the STOI measured with two mixes, the GGMM is up to the DUET performance. Even under reverberant conditions in large rooms, which we can proclaim as the rooms where the separation is more difficult by the greater time differences of the reflections.

In the case of three active sources is where GGMM can obtain slight improvements over DUET in all the experimented environments. However, LMM does not show such good results but it keeps close results to the other two algorithms.

Having three sensors instead of two has not meant a remarkable improvement, as well as it keeps close results to the others. Furthermore, increasing the number of microphones can cause the algorithm to work badly, since the number of parameters to optimize increases exponentially. So, we decided to use only three sensors.

The speech signals have been normalized to 0 dB in the source location. The received power of the signal will always be lower than 0 dB in the microphones due to propagation loss. Thus, assuming there was only one source active the SNR of the signal would be lower than -30 dB. However, we should be aware of the degradation that the algorithm includes. So, on the SNR measure we will always have worst SNR's for two or more sources due to the separation process.

The GGMM also shows improvements over DUET according to the SNR measure. In a similar way as with the STOI, greater improvements with respect to DUET are obtained when there are three active sources. However, the LMM does not show SNR improvements over DUET except in specific cases.

With the intention of studying the effect of the additive noise in the separation process, we have varied the level of noise in the microphone signals. Table III shows the average values of STOI obtained in the experiments with additive noise of -10 dB power.

A general degradation of the quality is present in all scenarios. But the differences of values between the three methods are approximately the same that in the previous examples. We can remark that, in the case of having two sources the LMM algorithm has obtained better results that the GGMM and DUET. Even in the worst conditions, that is, large rooms with high reverberation level.

The GGMM remains the best method for the estimation of the mixing parameters when there are three active sources.

We would like to highlight that in noise-dominated environment the worsening due to the increase of reverberation is less noticeable.

TABLE I. AVERAGE STOI OF THE SOURCES WITH PNOISE=-30 DB

STOI			Small room			Medium room			Large room		
Mixes	Sources	Method	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6
2	2	DUET	0.83	0.81	0.73	0.85	0.81	0.73	0.83	0.79	0.70
		LMM	0.82	0.79	0.74	0.79	0.80	0.73	0.82	0.78	0.70
		GGMM	0.87	0.83	0.76	0.85	0.82	0.74	0.81	0.79	0.71
	3	DUET	0.70	0.67	0.60	0.71	0.68	0.59	0.70	0.65	0.55
		LMM	0.68	0.65	0.58	0.68	0.65	0.58	0.67	0.64	0.54
		GGMM	0.76	0.71	0.62	0.77	0.71	0.61	0.73	0.68	0.57
3	2	LMM	0.80	0.76	0.70	0.78	0.71	0.66	0.65	0.63	0.57
		GGMM	0.84	0.81	0.76	0.82	0.80	0.74	0.77	0.74	0.69
	3	LMM	0.68	0.65	0.57	0.66	0.65	0.54	0.61	0.55	0.48
		GGMM	0.77	0.71	0.62	0.73	0.68	0.60	0.65	0.63	0.56

TABLE II. AVERAGE SNR OF THE SOURCES WITH PNOISE=-30 DB

SNR			Small room			Medium room			Large room		
Mixes	Sources	Method	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6
2	2	DUET	10.74	8.01	4.59	11.04	8.19	4.89	10.46	8.12	5.07
		LMM	10.30	7.47	4.73	9.48	7.71	4.93	9.48	7.48	5.05
		GGMM	12.08	8.56	5.01	11.43	8.40	5.16	10.15	8.18	5.42
	3	DUET	6.13	4.50	2.56	6.38	4.69	2.76	6.32	4.67	2.88
		LMM	5.11	4.14	2.66	5.39	4.31	2.85	5.25	4.31	2.87
		GGMM	7.24	4.74	2.75	7.32	4.97	2.95	6.59	4.82	3.12

TABLE III. AVERAGE STOI OF THE SOURCES WITH PNOISE=-10 DB

STOI			Small room			Medium room			Large room		
Mixes	Sources	Method	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6	*Cr*0.0	*Cr*0.3	*Cr*0.6
2	2	DUET	0.74	0.72	0.67	0.66	0.63	0.60	0.54	0.52	0.48
		LMM	0.73	0.72	0.69	0.66	0.65	0.63	0.57	0.55	0.52
		GGMM	0.71	0.70	0.67	0.64	0.62	0.60	0.54	0.52	0.48
	3	DUET	0.62	0.60	0.55	0.56	0.54	0.50	0.48	0.46	0.40
		LMM	0.60	0.58	0.54	0.55	0.54	0.50	0.47	0.45	0.41
		GGMM	0.62	0.60	0.56	0.57	0.55	0.51	0.49	0.46	0.41

VI. CONCLUSIONS

After reviewing the results, it would be correct to say that each method may be suitable for specific scenarios. It has been shown that in general the GGMM algorithm works better that DUET in the presence of reverberation when the noise level is low. We could say that it tends to overcome limitations due to reverberation, although it does not show huge results.

We should never say that GGMM is better that DUET since it would depend on the context in which we are speaking. For example, this work only deals with a speech separation problem. Anyway, the start from the property of sparsity in speech signals is something common to both algorithms. Therewith, different approaches to the performance of the algorithms are no treated in this work, such as, the computational cost of each algorithm and how it affects the time spent in the separation in a specific device.

This paper also shows how the acoustic scenario affects this type of algorithm, either by the room dimension or by the reverberation and noise conditions. This in cases that try to characterize in a very general way spaces of common dimensions that usually find the people in their day to day.

The use of a specifically configured optimization function has also predetermined a performance of the algorithms based in mixing models in some inherent way. The using of constraints in the optimization function allowed to obtain better results since the optimization is limited to the desired domain of points for the different variables. However, the algorithm does not converge always to the best solution.

ACKNOWLEDGMENT

This work has been funded by the Spanish Ministry of Economy and Competitiveness-FEDER under Project TEC2015-67387-C4-4-R, and by the University of Alcalá under Project CCGP2017-EXP/060.

REFERENCES

[1] 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.

[2] S. Araki, "Blind sparse source separation with spatially smoothed time-frequency masking," in *Proc. IWAENC 2006*, 2006.

[3] C. L. Aguilar, *et al.*, "Multi-channel speech separation in reverberant environments". Ph. D thesis, Dept. Signal Theory and Comms., Univ. of Alcalá, Madrid, 2016.

[4] D. A. Alvarez, M. R. Zurera, and R. G. Pita, "Speech enhancement algorithms for audiological applications," Ph. D thesis, Dept. Signal Theory and Comms., Univ. of Alcalá, Madrid, 2013.

[5] D. Ayllón, R. Gil-Pita, P. Jarabo-Amores, and M. Rosa-Zurera, "Speech source separation using a generalized mean shift algorithm," *Signal Processing*, vol. 92, no. 9, pp. 2248-2252, 2012.
[6] M. I. Mandel, R. J. Weiss, and D. P. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 2, pp. 382-394, 2010.
[7] S. Kotz, N. Balakrishnan, and N. L. Johnson, *Continuous Multivariate Distributions, Volume 1: Models and Applications*, New York: John Wiley & Sons, 2004.
[8] T. W. Anderson and E.-U. Mathematicien, *An Introduction to Multivariate Statistical Analysis*, vol. 2, New York: Wiley, 1958.
[9] M. D. Center, *Optimization Toolbox, Constrained Optimization, fmincon*, 2018.
[10] V. Zue, S. Sene, and J. Glass, "Speech database development at mit: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351-356, 1990.
[11] S. G. McGovern, "Fast image method for impulse response calculations of box-shaped rooms," *Applied Acoustics*, vol. 70, no. 1, pp. 182-189, 2009.
[12] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943-950, 1979.
[13] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125-2136, 2011.

César Clares-Crespo was born in Cuenca, Spain, in 1993. He received the B.Eng. Degree in Telecommunications Systems from the University of Alcalá, Madrid, Spain, in 2018. Since 2017 he is a Researcher in the Signal Theory and Communications Department at University of Alcalá. He is currently pursuing a M.Eng. Degree in Telecommunications Engineering, in the University of Alcalá, Madrid. His research interests include speech signal processing focusing in sound source separation and sound source localization.

Roberto Gil-Pita received the M.Eng. degree in Telecommunication Engineering and the Ph.D. degree (with hons.) in electrical engineering from the University of Alcalá, Madrid, Spain, in 2001 and 2006, respectively. From 2001, he has worked at the Signal Theory and Communications Department in the University of Alcalá. His research interests include pattern recognition and signal processing, focusing on sound source separation, hearing aids, and emotional speech. He is project manager of several projects with public and private fundings, including the 2- year ATREC project for the real-time analysis of combat stress, funded by the Spanish Ministry of Defense, and the SSPressing-Colist project for smart audio processing, funded by the Spanish Ministry of Economy and Competitiveness.

Manuel Rosa-Zurera received the B.Eng. degree (with hons.) in Technical Telecommunication Engineering from the University of Alcalá, Madrid, Spain, in 1990, the M.Eng. degree in Telecommunication Engineering from the Technical University of Madrid, Spain, in 1995, and the Ph.D. degree (with hons.) from the University of Alcalá, Madrid, Spain, in 1998. Since 1997, he has worked at the Signal Theory and Communications Department in the University of Alcalá, where he is Full Professor since 2010. He has been Head of the department from 2004 to 2010, and Dean of the Polytechnic School from 2010 to 2017. His research interests include statistical signal processing, signal models, source coding, speech and audio signal processing, and radar signal processing

Joaquín García-Gómez received the B.Eng. Degree in Telecommunications Technologies Engineering, and the M. Eng. Degree in Telecommunications Engineering, from the University of Alcalá, Madrid, Spain, in 2015 and 2017, respectively. Since 2017 he is a Researcher and PhD student in the Signal Theory and Communications Department at University of Alcalá. His research interests include pattern recognition and audio signal processing, focusing on event sound detection.

Inma Mohino-Herranz received the Ph.D degree (with hons.) in Information Technologies and Communications in 2017, in 2015 the M.Sc. Degree in Information Technologies and Communications, graduated in 2012 as M. Eng. degree in Electronics engineer (with hons.), and graduated in 2010 as B.E. degree in Technical Telecommunication Engineering, from the University of Alcalá (Spain). From 2012 she is a Researcher in the Signal Theory Department at University of Alcalá. Her research interests include emotion recognition, acoustic signal processing, speech processing, biological signal processing and automatic speech recognition to classify.