A SOURCE REASSIGNMENT TECHNIQUE FOR TIME-FREQUENCY MASKING AUDIO SEPARATION

Maximo Cobos, Jose J. Lopez and Jan O. Hinz

Institute for Telecommunications and Multimedia Applications (iTEAM)
Universidad Politécnica de Valencia
Camino de Vera s/n, Valencia, Spain

ABSTRACT

A neighborhood-based source reassignment technique is proposed for being used on time-frequency masking audio source separation methods. This technique identifies all the time-frequency clusters that form the separation masks in the Short-Time Fourier Transform (STFT) domain, and labels each time-frequency bin with a value that denotes the size of their corresponding clusters. The bins corresponding to clusters of small size are reassigned to the source which has maximum likelihood in its time-frequency neighborhood. An example is described using the DUET algorithm for under-determined mixtures, showing that this technique improves substantially the isolation of the estimated signals.

***Index Terms*—** Audio source separation, time-frequency masking, speech enhancement

1. INTRODUCTION

For several years, Blind Audio Source Separation (BASS) has been receiving increasing attention. BASS deals with the problem of recovering the source signals from their mixtures when the mixing process is unknown. The term blind comes from the fact that very little information is needed to carry out the separation, although some assumptions are always necessary. In this context, several techniques for solving the BASS problem have been developed, such as Independent Component Analysis (ICA) [1], Computational Auditory Scene Analysis (CASA) [2] or Sparse Decompositions [3]. As audio signals are not sufficiently sparse in time domain, some transformations are applied for dealing with underdetermined mixtures, i.e. systems with more sources than sensors. Rickard et al. [4] showed that speech signals are sparsely distributed in time-frequency (TF) representations. In fact, time-frequency masking methods have shown to provide better performance in the under-determined condition than other methods based on ICA. The reason is that speech signals in the TF domain only overlap in few points, being approximately orthogonal to each other. This property is usually referred as the W-Disjoint Orthogonality property of speech signals. The use of TF masks enables to emphasize regions of the TF spectrum that are dominated by a specific source and attenuate regions dominated by the other sources, resulting in a better intelligibility of the separated sound [5] [6]. In this work we propose a source reassignment technique that can be applied as post-processing to the binary masks obtained by TF masking algorithms. This technique allows to reassign isolated and small clusters of non-zero elements in the masks to the source which has maximum likelihood in the TF neighborhood of the elements. The proposed approach allows reducing the residuals from other sources in the mixture.

The organization of the paper is as follows. Section 2 explains the conventional separation approach used by TF masking methods. Section 3 presents the proposed reassignment technique. Experimental results and conclusions are given in Section 4 and Section 5, respectively.

2. TIME-FREQUENCY MASKING

Next we describe the convolutive mixing model. Assuming N sources and M sensors, this can be mathematically expressed as:

$$x_i(t) = \sum_{j=1}^{N} \sum_{l} h_{ij}(l) s_j(t-l) \quad i = 1, \ldots, M, \tag{1}$$

where $x_i(t)$ are the observation mixture signals, $s_j(t)$ are the source signals and $h_{ij}(l)$ is the impulse response from source j to sensor i. The goal of BASS algorithms is to obtain the separated signals $y_j(t)$ that are estimations of $s_j(t)$. Sometimes, it is sufficient to estimate the image of source j in sensor i:

$$s_{ij}(t) = \sum_{l} h_{ij}(l) s_j(t-l). \tag{2}$$

In the context of under-determined source separation, the Short Time Fourier Transform (STFT) has been widely used. In the STFT domain, the model of Eq.1 becomes:

$$X_i(k,m) \approx \sum_{j=1}^{N} H_{ij}(k) S_j(k,m) \quad i = 1, \ldots, M \tag{3}$$

where $H_{ij}(k)$ is the frequency response from source j to sensor i, and $S_j(k,m)$ is the STFT of the source $s_j(t)$. The indices k and m denote the frequency index and the time index, respectively. In general terms, the advantages of working with STFT representations are twofold. First, convolutive mixtures can be approximated as instantaneous mixtures at each frequency. The second one is the fact that the sparseness is higher under this representation.

Time-frequency masking attempts to construct a set of masks that can be applied to the mixtures in order to obtain the estimates of the sources:

$$Y_{ij}(k,m) = M_j(k,m)X_i(k,m), \tag{4}$$

being $Y_{i,j}(k,m)$ the STFT of the image of s_j in sensor i and $M_j(k,m)$ is the separation mask. The estimates of the sources in the time domain are obtained applying the inverse STFT operator.

Separation methods based on TF masking usually construct the masks from the Interaural Level Difference (ILD) and Interaural Time Difference (ITD) of the mixture channels $x_i(t)$. Although TF masking methods using more than two microphones are available [7], stereo separation methods are usually employed. In the case of two-channel separation, the ILD and ITD are estimated as:

$$\text{ILD}(k,m) = \left|\frac{X_2(k,m)}{X_1(k,m)}\right|, \tag{5}$$

$$\text{ITD}(k,m) = -\frac{N}{2\pi k}\angle\frac{X_2(k,m)}{X_1(k,m)}. \tag{6}$$

where N is the FFT size. The clustering of TF bins to sources is usually done by means of a likelihood function $L_j(k,m)$ that represents the closeness of a local estimation of the mixing parameters in each bin to the estimation with highest support given by the rest of TF points. Distances to peaks observed from a two-dimensional histogram analysis or to the centroids of K-means clustering have been shown to be powerful [7].

3. NEIGHBORHOOD-BASED REASSIGNMENT

The estimation of the separation masks is not always perfect in the sense that they differ from the ideal binary masks. Assuming $E_i(k,m)$ the energy of source i in TF-bin (k,m) and $N_j(k,m)$ the energy of the j-th interfering signal in this TF-bin, the ideal binary mask $I_i(k,m)$ for target source i and a threshold of 0 dB is defined as:

$$I_i(k,m) = \begin{cases} 1 & \text{if} \quad E_i(k,m) - N_j(k,m) > 0 \quad \forall j \\ 0 & \text{else} \end{cases} \tag{7}$$

In Figure 1a, the ideal binary mask (black represents zero and white represents one) for the extraction of one source in a two-channel anechoic mixture of three speech sources is represented. Figure 1b. shows the corresponding estimated mask using the DUET algorithm [4]. In the figure it is clear to see that, whereas most of the non-zero elements in the ideal mask are robustly clustered around harmonic partials and uniform zones, the estimated mask has much more small elements scattered around these areas. These scattered small clusters contribute to musical noise and audible residuals from the other sources when the mask is applied to recover the source.

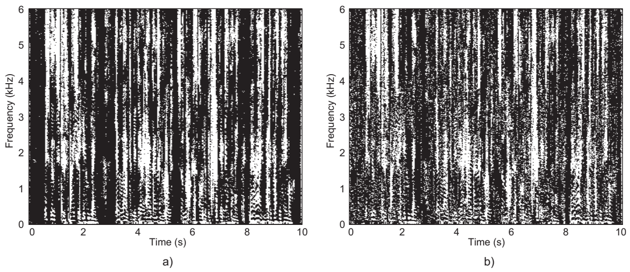

Fig. 1. (a) Ideal binary mask. (b) DUET binary mask

In this paper, we propose to identify and reassign these small elements using a neighborhood based criterion. In the following subsections we will describe in detail how to carry out the proposed reassignment.

3.1. Cluster Labelling

In the first step, the estimated masks corresponding to the sources in the mixtures are analyzed for finding clusters of non-zero elements in them. The clusters are formed by grouping 4-connected or 8-connected objects in the binary masks, as described in [8]. When a cluster C_{nj} is found in a mask $M_j(k,m)$, the TF-bins in the mask that form the cluster are labeled with the cluster size, forming a TF cluster map:

$$C(k,m) = N_{C_{nj}}\big|_{(k,m)\in C_{nj}}, \tag{8}$$

where $N_{C_{nj}}$ denotes the number of elements in the cluster C_{nj}. This way, isolated TF bins in a mask will be labeled as 1, while points in a cluster of 100 connected points will have a label of 100. Figure 2a shows the cluster map of the example mixture. It can be observed that TF bins with big labels are predominant. Figure 2b. shows the TF bins that form clusters with no more than 3 elements. As we will explain in the next subsection, these elements will be the candidates for being reassigned in the final masks.

3.2. Source Reassignment

From the observation of Figure 1a and Figure 1b, it is possible to see that TF bins forming small clusters are more likely to appear in the estimated masks than in an ideal binary mask. This fact is also graphically depicted in the histogram of Figure 3. The histogram shows the number of TF bins in $C(k,m)$

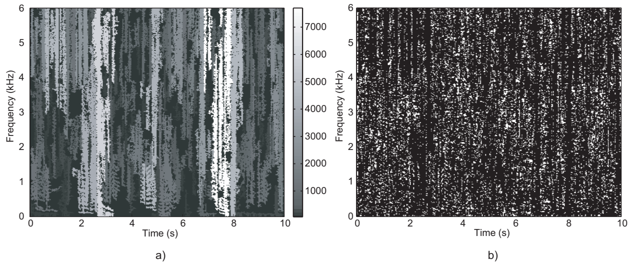

Fig. 2. (a) TF-bins labeled with the size of their corresponding cluster. (b) TF-bins with label lower than 4.

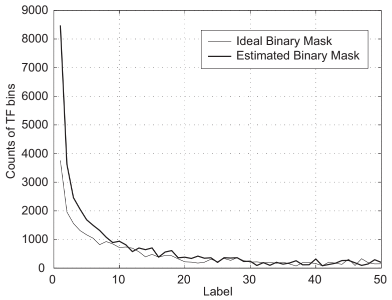

Fig. 3. Histogram showing the number of TF bins with low labels in an estimated mask and the corresponding ideal mask.

labeled with small numbers for the two masks shown in Figure 1. In the case of the DUET mask, a higher number of TF bins form small clusters of nonzero elements.

In order to reduce the number of scattered nonzero elements, TF bins labeled with small numbers are selected as candidates for being reassigned to a different source. A threshold κ can be defined for setting the minimum cluster size considered as having reliable non-zero elements in the mask. Although this threshold can be modified, experimental results suggest that good values for κ are from $\kappa = 1$ to $\kappa = 5$. This can be also inferred from the histogram of Figure 3, which shows that the distribution of elements with label below 5 are more easily found in an estimated mask than in an ideal mask.

The reassignment of the selected TF bins is carried out exploring the TF neighborhood of the bin, i.e. its N-neighbors in the TF plane. The span of the neighborhood in rows and columns is defined by the γ parameter. This approach is powerful in the sense that, as small clusters are not easily found in ideal masks, it is probable that the points of small clusters belong to the source with maximum likelihood in their TF neighborhood. The likelihood function should be chosen in agreement with the TF method used. For example, the maximum likelihood function used by DUET enables to assign each time-frequency point to the source with the mixing parameters that best explains the mixtures. However, if other clustering techniques are used, such as K-means, the likelihood function can be chosen as the distance of the data to final cluster centroids. Denoting $L_j(k,m)$ the likelihood matrix related to source s_j, the reassignment algorithm can be described as follows.

3.2.1. Reassignment algorithm

Inputs: TF cluster map $C(k,m)$, estimated separation masks $M_j(k,m)$, source likelihoods $L_j(k,m)$, maximum cluster size allowed κ and neighborhood span γ.

1. Initialize the final masks F_j with the value of the current masks: $F_j = M_j$. Start to explore each TF point (k,m).
2. If $C(k,m) \leq \kappa$ go to 3, else go to 6.
3. Find the source j to which the point (k,m) belongs, i.e. $M_j(k,m) = 1$.
4. Find the source q with maximum likelihood in the neighborhood of (k,m): $q = \arg\max_j \{L_j(k_q,m_q)\}$. The span of the neighborhood is defined by γ:
   $$k - \gamma \leq k_q \leq k + \gamma, \quad k_q \neq k$$
   $$m - \gamma \leq m_q \leq m + \gamma, \quad m_q \neq m.$$
5. Set $F_j(k,m) = 0$ and $F_q(k,m) = 1$.
6. If all the points were explored: end. Else, update (k,m) to the next point and go to 2.

Outputs: Final masks $F_j(k,m)$.

4. EXPERIMENTS

In this section we evaluate the proposed approach over the masks obtained from a synthetic anechoic mixture using the DUET algorithm. This mixture consists of three speech sources mixed with different time delays and amplitude gains. The source signals were obtained from the development data used in the Stereo Audio Source Separation Evaluation Campaign [9], coded as 16-bit 16 kHz audio. For comparison purposes, we will use the same objective performance measures used in the mentioned evaluation. These measures are the Source to Distortion Ratio (SDR), the Source to Interference Ratio (SIR) and the Source to Artifacts Ratio (SAR) [10].

The STFT of the mixtures was performed using time windows of length 1024 and 50% overlap. The reassignment of the masks was carried out for different combinations of κ and γ. Table 1 shows the SDR, SIR and SAR values obtained for the different sources using different combinations of the parameters. The last combination with $\gamma = 0$ means that no reassignment is performed and the candidate points are just eliminated. The best average performance (taking into account the results obtained for all the sources) was found for $\kappa = 3$ and $\gamma = 1$, obtaining a maximum SIR gain of 2.2 dB and 0.5 dB in SDR. The original masks and the reassigned masks for this case are shown in Figure 4. Similar improvements are found for a convolutive mixture with $\mathrm{RT}_{60} = 50$ ms. Although both listening tests and objective

Table 1. Performance Evaluation Measures.

Source	DUET			$\kappa = 3, \gamma = 1$			$\kappa = 1, \gamma = 1$			$\kappa = 5, \gamma = 0$		
	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
$\hat{s}_1$	7,6	24,4	7,7	7,4	24,4	7,5	7,5	24,6	7,5	7,3	25,8	7,4
$\hat{s}_2$	4,3	13,6	5,1	4,8	15,8	5,3	4,5	14,3	5,2	4,7	16,1	5,1
$\hat{s}_3$	7,3	20,5	7,6	7,1	20,7	7,4	7,2	20,7	7,5	6,7	21,7	7,0

measures only show a slight improvement after the reassignment, we think that further work on new reassignment criteria will make possible to obtain a more significative increase in the quality of the separated sources. Examples can be found at http://personales.upv.es/macoser1/iwaencdemos.html.

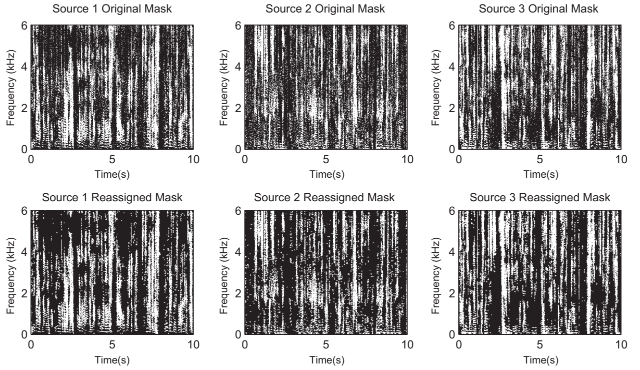

Fig. 4. DUET masks and Reassigned masks with $\kappa = 3$ and $\gamma = 1$.

5. SUMMARY AND CONCLUSIONS

Time-Frequency Masking has been shown to be a powerful method for underdetermined Sound Source Separation. Although very acceptable results have been obtained using TF masking algorithms, the estimated masks are sometimes corrupted by scattered nonzero points that cause a noticeable degradation of the extracted sources. In this paper we have proposed a source reassignment technique that can be applied as post-processing to the binary masks obtained by TF masking algorithms. This technique allows to reassign isolated and small clusters of non-zero elements in the masks to the source which has maximum likelihood in the TF neighborhood of the elements. A separation example using the DUET algorithm has been presented, showing that the proposed approach allows reducing the residuals from other sources in the mixture.

6. ACKNOWLEDGMENT

This work was supported by the Ministry of Education and Science under the project TEC2006-13883-C04-01 and FEDER funds.

7. REFERENCES

[1] J. F. Cardoso, “Blind signal separation: Statistical principles,” in *Proccedings of the IEEE*. October 1998, vol. 86, pp. 2009–2025, IEEE Computer Society Press.

[2] D. L. Wang and G. J. Brown, Eds., *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley, 2006.

[3] M. G. Jafari, M. D. Abdallah, M. D. Plumbey, and M. E. Davies, “Sparse coding for convolutive blind audio source separation,” in *ICA 2006*, Charleston, SC, USA, March 2006, pp. 132–139, Spriger-Verlag.

[4] 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, July 2004.

[5] S. Schulz and T. Herfet, “Binaural source separation in non-ideal reverberant environments,” in *Proceedings of the Int. Conference on Digital Audio Effects (DAFx-07)*, Bordeaux, France, September 2007.

[6] M. Cobos and J. J. Lopez, “Improving isolation of blindly separated sources using time-frequency masking,” *IEEE Signal Processing Letters*, accepted for publication.

[7] S. Araki, H. Sawada, R. Mukai, and S. Makino, “A novel blind source separation method with observation vector clustering,” in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC ’05)*, Eindhoven, 2005, p. 117120.

[8] R. M. Haralick and G. Shapiro Linda, *Computer and Robot Vision*, vol. I, pp. 28–48, Addison-Wesley, 1992.

[9] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J. P. Rosca, “First stereo audio source separation evaluation campaign: data, algorithms and results,” in *ICA 2007*, London, UK, September 2007.

[10] E. Vincent, R. Gribonval, and C. Févotte, “Performance measurement in blind audio source separation,” *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.