Improving Time-Frequency Sparsity for Audio Spatialization by Time-Adaptive Windowing

Pradeep Gaddipati, Nipun Dave, Preeti Rao, and Rajbabu Velmurugan
Department of Electrical Engineering
Indian Institute of Technology Bombay
Mumbai, India
{light, nipundave, prao, rajbabu}@ee.iitb.ac.in

Abstract – **We propose a new time-adaptive windowing technique to obtain a sparse time-frequency representation for audio signals. This transformation helps in providing better source separation from stereo mixtures for improved subsequent spatial rendering over headphones. We start with standard stereo audio recordings, transform them to a sparse representation and then estimate the mixing parameters to be used for source separation. The performance of the new representation is compared with existing methods via the accuracy of mixing parameters estimation for a test dataset of multi-speaker stereo mixtures.**

Keywords – audio spatialization; data-adaptive time-frequency representation; sparse time-frequency representation; blind source separation

I. INTRODUCTION

Spatial audio refers to the rendering of realistic auditory experience with auditory immersion. Surround sound, an outcome of the extensive research on spatial audio, refers to the use of multiple loudspeakers to envelope a person watching a movie or listening to music, making them feel as if they are in the middle of the action or the concert. The surround sound tracks enable the audience to hear sounds coming from all around them, contributing to the sensation of what movie-makers call 'suspended disbelief'. Such a technique is only applicable in the case when the playback devices are placed at a considerable distance from the listener. The same audio signals are not as effective when headphones are used for listening. Hence this surround audio needs to be specially processed to obtain a signal that can generate the effect of auditory immersion over headphones.

Given a multi-channel audio mixture as input in any available format, audio spatialization or virtualization is the process of realistic spatial rendering of audio in the desired listening configuration (e.g. over headphones) as shown in Figure 1. One approach to this problem involves separating the individual sources from the multi-channel audio mixture, and then creating the desired listener-end mixtures by suitable recombination of the individual spatialized sources. The success of this approach hinges on achieving the proper separation of sources from the input multi-channel mixtures. In the present work we address the problem of separating two or more source signals from stereo mixtures.

When presented with a set of observations from the sensors such as microphones, the process of extracting the underlying sources is called source separation. Doing so without strong additional information about the individual sources or constraints on the mixing process is called blind sources separation (BSS). BSS algorithms can be classified according to the assumptions they make about the mixing model (discussed in section II). Based on the mixing model one can classify them as either instantaneous case (where sources arrive instantly at the sensors but with differing signal intensity), anechoic case (extension of the previous case, where arrival of delays between sensors are also considered), or echoic case (where the anechoic case is further extended by considering multiple paths between each source and each sensor). Also depending on the number of mixtures, m and sources, n one can classify them as either under-determined ($m < n$), even-determined ($m = n$) or over-determined ($m > n$) [4].

The under-determined case is the most common and to make the problem tractable, one usually needs to make certain assumptions about the nature of the sources such as statistical independence of sources, or sparsity of sources in some time-frequency representation. One such method which assumes statistical independence of the sources uses independent component analysis (ICA) [12] to solve the BSS problem for instantaneous under-determined and even-determined mixtures. A modification of ICA is complex independent component analysis [13] which is used for unmixing under-determined anechoic mixtures. Another approach for BSS problem which assumes sparser expansion of sources with respect to some basis solves under-determined instantaneous mixtures. In this case, one can formulate the source extraction problem as a constrained l^1 minimization problem, which typically yields a convex program [14]. Yet another approach for BSS is degenerate unmixing estimation technique (DUET) algorithm [1] which is applicable for anechoic under-determined mixtures; this algorithm uses sparsity of speech signals in the short-time Fourier transform (STFT) domain to construct binary time-frequency masks. These masks are then used to extract several sources from only two mixtures.

In this paper, we address the blind source separation problem by the DUET approach. We investigate a suitable time-frequency representation that enhances the sparseness and disjointness of the sources in the mixture, thus facilitating their separation. Section II reviews the DUET algorithm and presents the standard assumptions about the source characteristics and mixing model. In Section III, the problem of mixing parameter estimation is discussed. Time-frequency representations that could be useful in mixing parameter estimation are described in Section IV. Experiments to evaluate the performance of the various time-frequency representations are described in Section V. These experiments were carried out on the TIMIT speech database. Finally in Section VI we present the results and conclusions.

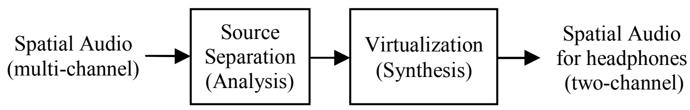

Figure 1. Audio spatialization block diagram.

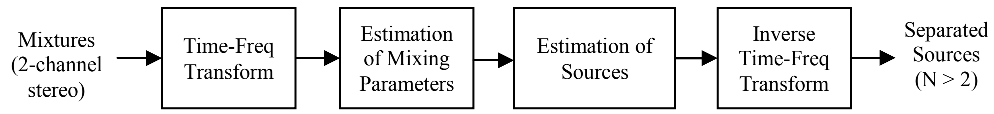

Figure 2. A block diagram of the DUET algorithm.

II. Review of DUET Algorithm

DUET algorithm, proposed by Jourjine, Rickard, and Yilmaz [1], relies on the assumption that the sources are sparse and disjoint in some time-frequency representation. Though speech signals violate this assumption and are only approximately sparse and disjoint, this technique has been shown to achieve good de-mixing [2], [3]. This algorithm is capable of separating N (greater than 2) sources from two available mixtures, i.e., this technique is applicable for unmixing under-determined anechoic mixtures.

A block diagram representation of DUET algorithm is shown in Figure 2. The various steps involved in this algorithm are: 1) transformation of the available mixtures into some sparse time-frequency representation 2) estimation of mixing parameters by clustering the ratios of these time-frequency representations of the mixtures obtained in the previous step 3) the estimates of mixing parameters obtained from the previous step are used to partition the time-frequency representation of one of the mixtures to obtain the estimates of the sources in the time-frequency domain and 4) finally these time-frequency partitions are inverted back to time domain signals using an appropriate inverse time-frequency transformation to recover the original sources. The present work is restricted to investigations related to the first two blocks in Figure 2.

A. Mixing Model

Consider the mixtures of N source signals, $s_j(t)$, $j = 1,\dots,N$ being received at a pair of microphones where only the direct path is present. In this case, without loss of generality we can absorb the attenuation and delay parameters of the first mixture, $x_1(t)$, into the definition of the sources. The two anechoic mixtures can thus be expressed as,

$$x_1(t) = \sum_{j=1}^{N} s_j(t) \tag{1}$$

$$x_2(t) = \sum_{j=1}^{N} a_j s_j(t-\delta_j) \tag{2}$$

where, δ_j is the arrival delay between the microphones resulting from the angle of arrival, a_j is the relative attenuation factor corresponding to the ratio of the attenuations of the paths between source and microphones. a_j and δ_j are referred to as the mixing parameters of the mixing model.

B. Local Stationarity

Windowed Fourier transform of a signal $s(t)$ is obtained as

$$F^{w}\{s(\cdot)\}(\omega,\tau) = \int_{-\infty}^{\infty} s(t)W(t-\tau)e^{-i\omega t}dt \tag{3}$$

will be referred as $S^{W}(\omega,\tau)$ where appropriate. Using (3) and the Fourier transform pair,

$$s_j(t-\delta) \leftrightarrow e^{-i\omega\delta}S_j(\omega) \tag{4}$$

we have

$$F^{W}\{s_j(\cdot-\delta)\}(\omega,\tau) = e^{-i\omega\delta}F^{W}\{s_j(\cdot)\}(\omega,\tau) \tag{5}$$

when $W(t) \equiv 1$. However, when $W(t)$ is a windowing function, (5) is not necessarily true. This can be thought of as a form of a narrowband assumption in array processing [11], but this label is perhaps misleading in that speech is not narrowband and local stationarity seems a more appropriate moniker. For DUET it is necessary that equation (5) holds for all δ, $|\delta| \leq \Delta$, even when $W(t)$ has finite support [10]. Here Δ is the maximum time difference possible in the mixing model (the microphone separation divided by the speed of sound signal propagation).

C. Microphone Spacing

Additionally, one crucial issue is that DUET is based on the extraction of attenuation and delay parameters estimates from each time-frequency point. We will utilize the local stationarity assumption to turn the delay in time into a multiplicative factor in time-frequency. Of course, this multiplicative factor $e^{-i\omega\delta}$ uniquely specifies δ only if $|\omega\delta| < \pi$ as otherwise we have an ambiguity due to phase-wrap [5]. So we require,

$$|\omega\delta_j| < \pi, \quad \forall\omega, \forall j \tag{6}$$

to avoid phase ambiguity. This is guaranteed when the microphones are separately by less than $\pi c/w_m$ where w_m is the maximum frequency present in the sources and c is the speed of sound.

D. W-disjoint Orthogonality

Given a windowing function $W(t)$, we call two functions $s_j(t)$ and $s_k(t)$ W-disjoint orthogonal if the supports of the windowed Fourier transforms of $s_j(t)$ and $s_k(t)$ are disjoint. The W-disjoint orthogonality assumption can be stated concisely as,

$$S_j^W(\omega,\tau)S_k^W(\omega,\tau) = 0, \ \forall j \neq k, \forall\omega,\tau \tag{7}$$

This assumption is the mathematical idealization of the condition that it is likely that every time-frequency point in the mixture with significant energy is dominated by the contribution of one source. W-disjoint orthogonality is crucial to DUET because it allows for the separation of a mixture into its component sources using a binary mask.

III. Mixing Parameter Estimation

The assumptions of anechoic mixing and local stationarity allow us to rewrite the mixing equations (1) and (2) in the time-frequency domain as,

$$\begin{bmatrix} X_1^W(\omega,\tau) \\ X_2^W(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 & \cdots & 1 \\ a_1 e^{-i\omega\delta_1} & \cdots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} S_1^W(\omega,\tau) \\ \vdots \\ S_N^W(\omega,\tau) \end{bmatrix} \tag{8}$$

With the further assumption of W-disjoint orthogonality, at most one source is active at every (ω,τ), the mixing process can be described for each (ω,τ) and for some j as,

$$\begin{bmatrix} X_1^W(\omega,\tau) \\ X_2^W(\omega,\tau) \end{bmatrix} = \begin{bmatrix} 1 \\ a_j e^{-i\omega\delta_j} \end{bmatrix} S_j^W(\omega,\tau) \tag{9}$$

here j is the index of the source active at (ω,τ).

Now, we can calculate the relative attenuation and delay parameters associated with one source, using

$$(\hat{a}_j, \hat{\delta}_j) = \left(\left\| \frac{X_2^W(\omega,\tau)}{X_1^W(\omega,\tau)} \right\|, \Im\left(\log\left(\frac{X_1^W(\omega,\tau)}{X_2^W(\omega,\tau)}\right)\right)/\omega \right) \tag{10}$$

for some j, where $\Im$ denotes taking imaginary part. Using (10), every (ω,τ) yields an estimate pair for the relative attenuation-delay parameter associated with each source. For W-disjoint orthogonal signals, if we calculate the attenuation-delay estimates from a number of time-frequency points, we would expect to see clusters around the true mixing parameters for each source.

If we now construct a two dimensional weighted histogram using the attenuation-delay estimates, the number of peaks found would be the estimate of the number of sources, and the peak centers would be the attenuation-delay estimates associated with each source (see Figure 5). From these estimates of mixing parameters we then construct the time-frequency masks which de-mix the mixtures.

The main observation that DUET leverages is that the ratio of the time-frequency representations of the mixtures does not depend on the source components but only on the mixing parameters associated with the active source component. Thus it can be seen that, the successful extraction of mixing parameters relies on the sparsity of speech in the time-frequency domain.

IV. Time-Frequency Representations

Time-frequency representations describe signals in terms of their frequency content at a given time. These representations are useful for analyzing signals varying both in time and frequency. For speech and music signals where we have continuously time-varying frequency content, frequency domain representations cannot be used because they only give spectral information and no time information i.e. they fail to convey when, in time, the different events are occurring in the signal. The short-time Fourier transform is one of the most widely used approaches to time-frequency analysis.

A. Short-Time Fourier Transform

The short-time Fourier Transform (STFT) of the signal *x(t)* is defined as in (3), where *W(t)* is the window function. *W(t)* can be considered as a window that selects a particular portion of the signal centered around the given time location, and the Fourier transform of the windowed signal yields the frequency content of the signal at the given time.

Another viewpoint pioneered by Gabor [9] provides insight into the STFT, fundamental to the adaptive time-frequency representation discussed in the following sections. The modulated window function, $W(t-\tau)e^{-i\omega t}$ is concentrated in time-frequency around the location (ω, τ). The STFT projects the signal onto a non-orthogonal basis formed by a set of these functions; the projection, or inner product, with a particular $W(t-\tau)e^{-i\omega t}$ represents the time-frequency content of the signal at (ω, τ). Ideally, the projection function should be an impulse in time-frequency. Gabor found that Gaussian signals $e^{-i\omega_o(t-t_o)+c(t-t_o)^2}$ achieve minimum time-frequency uncertainty, which implies that they are the closest approximation to an impulse in time-frequency; hence, time-shifted and frequency-modulated Gaussian functions appear to be the best basis in a projection-based time-frequency representation such as the STFT.

The choice of the window considerably affects the signal concentration in the STFT. In fact the STFT performs well in terms of concentration and resolution of a given component when a properly chosen window is used. For signals composed of several different components occurring at different instants in time-frequency, the best window differs for each component. The fact that different windows are appropriate for different signal components suggests the use of a data-dependent time-and-frequency-varying window function for analysis to achieve a high concentration and resolution of any signal component present at any time-frequency location.

B. Data-adaptive Time-Frequency Representations

An adaptive time-frequency representation (ATFR) was developed by Jones, and Parks [6] for signal visualization. The window function used in ATFR is Gaussian. The ATFR differs from the STFT with a Gaussian window in that the Gaussian parameter may vary with time and frequency. The basic idea behind the ATFR is that the extra degree of freedom at every time-frequency location can improve the performance over that of a fixed-window STFT.

Most real world signals are essentially stationary over a short interval of time. Consequently, a sparse representation could be obtained by analysing each frame with a window that has been optimized for that frame. Long windows give sparser representation for frames containing steady frequency components than when shorter windows are used. On the contrary, the time-frequency representation of impulses is sparser with short windows. The stronger time-frequency components, especially, should be resolved as precisely as possible while reducing the leakage into adjacent windows. Consequently, a window which resolves the stronger time-frequency components might be sufficient for analysing the entire frame to provide a sparse representation of the signal. Thus, a time-adaptive representation could also be expected to provide a sparse representation for a signal [6], [8].

The time-adaptive representation (TAR) of the signal *x(t)* is obtained as follows,

$$TAR(\omega,\tau) = \int_{-\infty}^{\infty} x(t) \left[\frac{-2C_\tau}{\pi}\right]^{\frac{1}{4}} e^{C_\tau (t-\tau)^2} e^{-i\omega t} dt \qquad (11)$$

The signal *x(t)* is projected onto the unit-energy Gaussian basis elements

$$\left[\frac{-2C_t}{\pi}\right]^{\frac{1}{4}} exp(C_t(\tau - t)^2) e^{-i\omega\tau} \qquad (12)$$

It should be noted here that this is very similar to the method used for obtaining the ATFR [6]. The difference in this case is that the Gaussian parameter C_τ is adapted only with time. Here C_τ is chosen such that the time-frequency representation of the signal frame has the maximum sparsity. Any measure quantifying the peakiness of a distribution such as kurtosis, Gini index or entropy can be used to measure sparsity or concentration.

The set of normalized Gaussian windows used in obtaining the TAR are shown in Figure 3. It shows windows with durations ranging from 20 ms to 64 ms in steps of 4 ms.

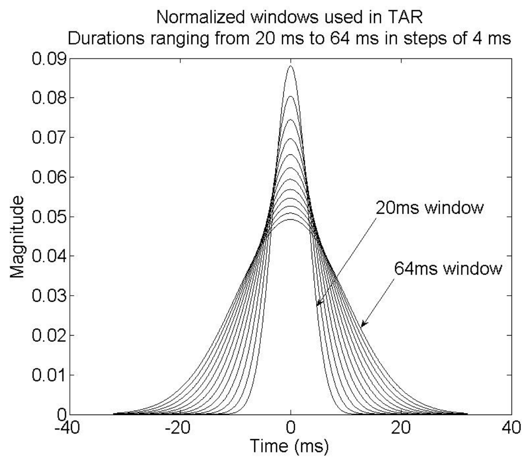

Figure 3. Normalized Gaussian windows used in TAR.

The duration of the Gaussian window is defined by the time between the points where the Gaussian function has died down to less than 1% of its maximum value.

C. Concentration Measure

The concentration measure we use for quantifying the peakiness of the distribution is kurtosis. The kurtosis κ is calculated as follows,

$$\kappa = \frac{\mu_4}{\sigma_4} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})^2\right)^2} - 3 \quad (13)$$

where n is the number of samples, x_i are the samples and $\bar{x}$ is the sample mean. In our application, samples are the coefficients of the time-frequency representation of the signal i.e. coefficients obtained from equation (11). Some observed properties of kurtosis are as follows; the smaller the number of peaks presents in the distribution, the higher is the kurtosis measure and when two distributions have the same number of peaks, the distribution having sharper peaks has higher kurtosis. Thus, the sparser the distribution, the higher is its kurtosis measure. For a particular frame we select the window, from the set of windows available, which provides the highest kurtosis value (i.e. highest sparsity).

Figure 4 shows the time-frequency representation of a speech signal using TAR. The dashed line shows the adaptation of the window length based on the sparsity measure (i.e. kurtosis). The window length corresponding to the highest sparsity (i.e. highest kurtosis value) is selected from the set of windows (20 ms to 64 ms in steps of 4 ms).

V. Experiments

For illustrating the higher sparsity exhibited by speech signals in TAR domain as compared to STFT domain, 10, three-second speech signals from the TIMIT speech database were chosen. The kurtosis measure for the time-frequency representation of each of these speech signals were calculated and plotted. The results for this are shown in Figure 5. We observe that TAR has improved performance as compared to STFT in terms of sparsity, i.e. higher kurtosis value.

The performance of the STFT and the TAR in estimation of mixing parameters is evaluated for ten sets of signals created from the TIMIT speech database. Each set consists of 3, three-second speech signals spoken by different speakers each saying different sentences. The sampling frequency of these signals is 16 kHz. A 2-channel stereo mixture (1), (2) is created using a set of signals by localizing the three speaker signals to three distinct locations given by θ_{pan} (14) on the horizontal plane using attenuation panning technique. The constant power panning law [15] is used to obtain the attenuation parameters a_1 and a_2 (15) for each of the sources. The speakers for the stereo reproduction are assumed to be at 30° on either side of the human head.

$$\theta_{pan} = \frac{\theta_\circ - \theta}{2\theta_\circ} 90^\circ, \quad \theta_\circ = 30^\circ \quad (14)$$

$$a_1 = cos\theta_{pan} \;\; and \;\; a_2 = sin\theta_{pan} \quad (15)$$

If the sources are assumed to be placed at angles *(θ)* -10°, 0° and 10° in the stereo mixture, then these correspond to panning angles *(θ_{pan})* of 60°, 45° and 30° respectively and the corresponding attenuation parameters would be (cos60°, sin60°), (cos45°, sin 45°) and (cos30°, sin30°) respectively. Figure 6 shows the reproduction model for two-channel stereo.

The panning angles for creating the mixtures are selected from the set {10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°} and the delay parameters (number of samples) are selected from a range of -10 to 10. For each set of three source signals, 3360 (336x10) distinct mixtures were generated, with 336 ($_8P_3$) distinct combinations of attenuation parameters (i.e. the panning angles) and 10 distinct combinations of delay parameters. So the performance of STFT and TAR is evaluated on a total of 33600 mixtures (10 sets each of 3360 number of mixtures).

There are two aspects to this evaluation. The first is the percentage of mixtures in which the number of sources is correctly estimated to be three and the second aspect is the accuracy of the estimated mixing parameters viz. attenuation and delay parameters in the cases where the number of sources was estimated correctly.

VI. Results and Conclusions

Figure 7 shows a 2-D weighted (energy weighted) histogram of the attenuation-delay estimates (for two mixtures and three sources) obtained from the time-frequency representations of the mixtures. The attenuation parameters (panning angles) were (30°, 45°, 60°) and the corresponding delay parameters (number of samples) were (0, -6, +3). The positive delay parameter *(+d)* implies a delay of *d* samples with respect to the first source and a negative delay parameter *(-d)* implies an advance of *d* samples with respect to the first

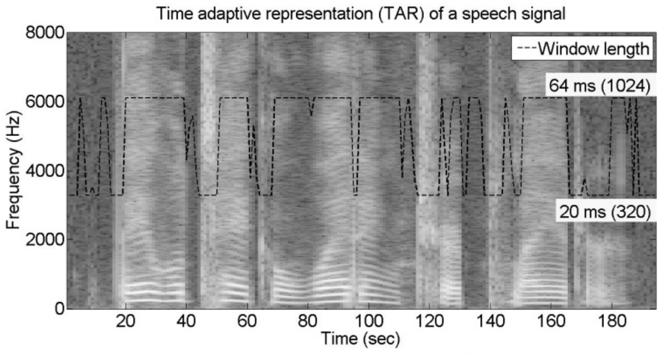

Figure 4. Time adaptive representation (TAR) of a speech signal.

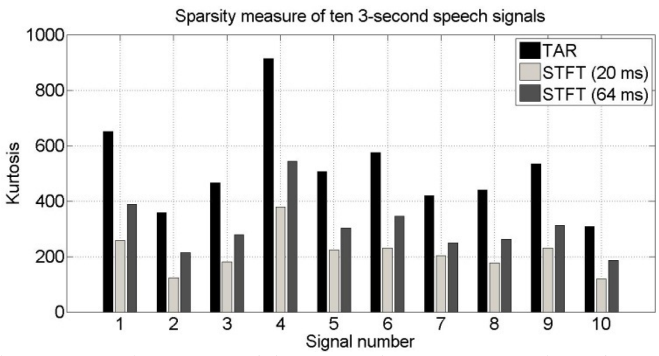

Figure 5. Sparsity measure of the TAR and STFT representations of ten 3-second speech signals. The TAR yields sparser representations of the data (i.e. higher kurtosis value).

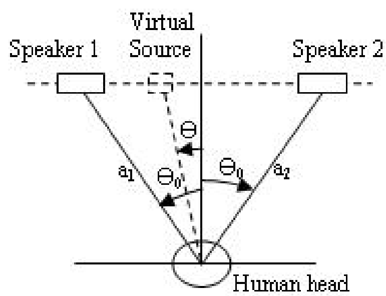

Figure 6. Reproduction model for two-channel stereo.

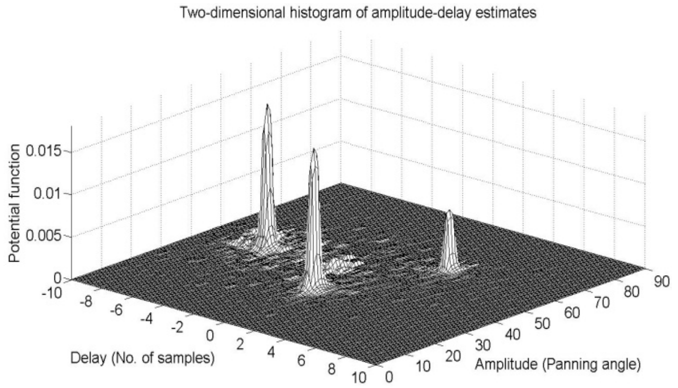

Figure 7. 2-D histogram of attenuation-delay estimates from two mixtures and three sources. The attenuation parameters (panning angle) were (30°, 45°, 60°) and the corresponding delay parameters (number of samples) were (0, -6, +3). Time-frequency representation used is TAR.

TABLE I. PERFORMANCE OF STFT AND TAR IN ESTIMATION OF MIXING PARAMETERS

T-F Representation	Cases with correct estimation of number of sources (%)	Error in estimation of attenuation parameters	Error in estimation of delay parameters
TAR	**71.9**	**0.5871**	**1.4547**
STFT (32ms)	67.3	0.7158	1.5017
STFT (64ms)	65.4	0.8053	1.6415

$$Error = \frac{1}{N}\sum\left(\left|\frac{original\ value - estimated\ value}{original\ value}\right| \times 100\right),\ N = No.\ of\ observations$$

source. The time-frequency representation was obtained using TAR. It should be noted here that the peaks correspond to the mixing parameters associated with the three sources present in the mixtures.

We detect the peaks in the histogram above a certain threshold (20% of the maximum value) in order to avoid the detection of the spurious peaks, this sometimes leads in missing the peaks with smaller attenuation (but the peak is prominently visible). This happens because the energy content of a source may be very less compared to that of others. So in such cases, manual identifications of the peaks is preferred.

In DUET algorithm, the separation of sources depends on mixing parameters and the sparseness of the time-frequency representation. Hence this improved accuracy in estimation of mixing parameters and the higher sparseness of TAR is expected to lead to a better separation of sources and would further result in better virtualization.

Table I summarizes the experimental results on the performance of STFT and TAR in the estimation of mixing parameters. It can be seen from the results that using TAR improves the accuracy in estimation of mixing parameters. This is due to the higher sparseness and disjointness exhibited in the time-frequency representation obtained by TAR as compared to STFT. It has also been observed that, the mixtures in which one or more sources have panning angles at the extremes i.e. close to 0° or 90° primarily contribute to errors in the estimation of mixing parameters.

It should also be noted that the time-adaptive transformation used to obtain the time-frequency representation is non-linear i.e., the time-adaptive representation for the sum of signals might not be equal to the sum of the time-adaptive representation of the individual signals. This linearity property is vital during the separation of sources.

In this paper, we proposed a time-adaptive representation (TAR) to improve audio spatialization. Our evaluation of the proposed approach was based on the accuracy of mixing parameter estimation. With this the TAR approach provides better estimates compared to the STFT approach. A complete evaluation would include comparing the estimated sources with the actual sources, and testing the spatialized audio created from these separated sources. This will be part of future research.

ACKNOWLEDGMENT

We thank Nokia, India for providing financial support and technical inputs for the work reported here.

REFERENCES

[1] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: demixing n sources from 2 mixtures," in *Proceedings of the 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Istanbul, Turkey, June 5-9 2000, vol. 5, pp. 2985-2988.

[2] S. Rickard, R. Balan, and J.Rosca, "Real-time time-frequency based blind source separation," in *3rd International Conference on Independent Component Analysis and Blind Source Separation,* San Diego, CA, December 9-12 2001.

[3] S. Rickard, and O. Yilmaz, "On the approximate w-disjoint orthogonality of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* Orlando, Florida, USA, May 13-17 2002, pp. 529-532.

[4] R. Saab, O. Yilmaz, M. J. McKeown, and R. Abugharbieh, "Underdetermined sparse blind source separation with delays," *Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS05)*, Rennes, France, 2005.

[5] S. Makino, T. Lee, and H. Sawada, *Blind Speech Separation,* chapter 8, Published by Springer, 2007.

[6] D. L. Jones, and T. Parks, "A high resolution data-adaptive time-frequency representation," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* December 1990, vol. 38.

[7] J. J. Burred, "From sparse models to timbre learning: new methods for musical source separation," *Ph.D. dissertation,* Technical University of Berlin, September 2008.

[8] N. Dave, "Sparsity-based source separation for audio spatialization," *Dual degree dissertation*, Department of Electrical Engineering, I.I.T. Bombay, Mumbai, India, June 2009.

[9] D. Gabor, "Theory of communication," *J. Inst. Elec. Eng.*, vol. 93, pp. 429-441, 1946.

[10] R. Balan, J. Rosca, S.Rickard, and J. O'Ruanaidh, "The influence of windowing on time delay estimates," in *Proceedings of the 35th Annual Conference on Information Sciences and Systems (CISS2000),* Princeton, NJ, March 15-17 2000, vol. 1, pp. WP1-(15-17).

[11] H. Krim and M. Viberg, "Two decades of array signal processing research, the parametric approach," *IEEE Signal Processing Magazine,* pp. 67-94, July 1996.

[12] C. Jutten, J. Herault, P. Comon, and E. Sorouchiary, "Blind separation of sources, parts i, ii, and iii," *Signal Processing,* vol. 24, pp. 1-29, 1991.

[13] J. Anemuller, T. Sejnowski, and S. Makeig, "Complex independent component analysis of frequency domain electroencephalographic data," in *Neural Networks,* 2003, pp. 16:1311-1323.

[14] Y. Li, A. Cichocki, and S. Amari, "Sparse component analysis for blind source separation with less sensors than sources," in *Proceedings of 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003),* Ricken. Kyoto, Japan: ICA, Apr. 2003, pp. 89-94.

[15] S. L. Lee, K. Y. Han, S. R. Lee, and K. M. Sung, "Reduction of sound localization error for surround sound system using enhanced constant power panning law," *IEEE Transactions on Consumer Electronics,* August 2004, vol. 50, pp. 941-944.