2011

The Synchronized Short-Time-Fourier-Transform: Properties and Definitions for Multichannel Source Separation.

Ruairí de Fréin
Technological University Dublin, ruairi.defrein@tudublin.ie

Scott Rickard
Citadel

Recommended Citation

de Fréin, R. & Rickard,S.T. (2011) "The Synchronized Short-Time-Fourier-Transform: Properties and Definitions for Multichannel Source Separation," in *IEEE Transactions on Signal Processing,* vol. 59, no. 1, pp. 91-103, Jan. 2011. doi: 10.1109/TSP.2010.2088392

Funder: Science Foundation Ireland

The Synchronized Short-Time-Fourier-Transform: Properties and Definitions for Multichannel Source Separation

Ruairí de Fréin, *Student Member, IEEE*, and Scott T. Rickard, *Senior Member, IEEE*

***Abstract*—This paper proposes the use of a synchronized linear transform, the synchronized short-time-Fourier-transform (sSTFT), for time-frequency analysis of anechoic mixtures. We address the short comings of the commonly used time-frequency linear transform in multichannel settings, namely the classical short-time-Fourier-transform (cSTFT). We propose a series of desirable properties for the linear transform used in a multichannel source separation scenario: stationary invertibility, relative delay, relative attenuation, and finally delay invariant relative windowed-disjoint orthogonality (DIRWDO). Multisensor source separation techniques which operate in the time-frequency domain, have an inherent error unless consideration is given to the multichannel properties proposed in this paper. The sSTFT preserves these relationships for multichannel data. The crucial innovation of the sSTFT is to locally synchronize the analysis to the observations as opposed to a global clock. Improvement in separation performance can be achieved because assumed properties of the time-frequency transform are satisfied when it is appropriately synchronized. Numerical experiments show the sSTFT improves instantaneous subsample relative parameter estimation in low noise conditions and achieves good synthesis.**

***Index Terms*— Signal analysis, source separation.**

I. Introduction

THE authors of [1] show that partitions of a time-frequency representation of a mixture of speech signals exist which can be used to demix mixtures of several speech signals. This is because speech is sparse in the time-frequency domain. The degenerate unmixing estimation technique (DUET) algorithm, proposed in [1], demixes an arbitrary number of sources from a two channel observation of the mixture using masks obtained from relative attenuation and delay estimates. However, the authors of [1] report a bias in their parameter estimates. This error is due to the application of unsynchronized time-frequency transforms on each channel. Previously, the authors of [2] identified a similar bias in magnitude-squared coherence estimation which was also due to misalignment of the signals; they suggested a realignment prior to coherence estimation. In this paper, we query the candidature of the cSTFT as a time-frequency transform for generating sparse representations of a general multichannel anechoic mixture. We propose a revised set of properties the appropriate transform should have. These properties supersede those proposed in [1]. Our proposed properties have broader scope than the DUET setting, and may be selectively applied where similar assumptions are made to facilitate demixing. As an embodiment of such a transform, we introduce the synchronized short-time-Fourier-transform (sSTFT) which satisfies our properties and makes performance gains possible in the class of algorithms of interest. This class of algorithms consists of supervised or unsupervised multichannel direction-of-arrival (DOA) and source separation algorithms, with a convolutive mixing model, which use co-information between channels, for example, [1], [3]–[8]. As an example of co-information, relative delay estimates are commonly used to determine the underlying sources in an anechoic environment.

This paper discusses the attributes of the sSTFT, assuming that the appropriate synchronization is known *a priori*. A companion paper [9] deals with the practical implementation of the sSTFT and has its novelty in that it shows how the synchronization parameter may be learned and then used for basis adaptation. Moreover, [9] shows how the sSTFT maybe applied in a multisource setting and a new algorithm called Iterative DUET is proposed. Iterative DUET incorporates contextual information available via the sSTFT into the algorithm, which facilitates gains in separation performance. This partitioning of our work into: 1) the properties of the sSTFT (in this paper) and 2) how to synchronize the sSTFT, allows for a more focussed discussion the sSTFT as a general stand-alone contribution.

This paper is organized as follows. In Section II, we review the properties (p1, p2, p3, and p4) introduced in [1]. Section III introduces the notation used for fractional sample delay of discrete signals. Section IV reviews time-frequency analysis and the role of windows functions. The sSTFT is presented in Section V. The sSTFT is mathematically defined in Section V-A and graphically motivated in Section V-B. The short comings of the cSTFT analysis windows are discussed in Section V-C. Section V-D, Section V-E and Section V-F illustrate new properties of the sSTFT which make it appealing for relative attenuation and delay estimation, and also, the fractional delay problem. Section VI demonstrates the bin-wise improvement achievable via linear transform synchronization, and, the suitability of a range of window functions for use with the sSTFT, specifically, the effect of subsample delay error. We discuss the structure of candidate window functions for use with the sSTFT and conclude with a thematic review of the paper in Section VI-C.

Manuscript received December 14, 2009; accepted September 27, 2010. Date of publication October 18, 2010; date of current version December 17, 2010. The associate editor coordinating the review of this paper and approving it for publication was Dr. Danilo Mandic. This work was supported by the Science Foundation Ireland by Grant 05/YI2/I677.

The authors are with the Sparse Signal Processing Group, University College Dublin, Dublin, Ireland (e-mail: rdefrein@gmail.com; scott.rickard@ucd.ie).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2010.2088392

II. Properties of Time-Frequency Linear Transforms for Multichannel Analysis

In this section we describe the assumptions DUET makes about the properties of time-frequency analysis in [1] and how these assumptions are typical of a whole suite of algorithms.

Given two mixtures, namely $x_1(t) = \sum_{j=1}^{N_s} s_j(t)$ and $x_2(t) = \sum_{j=1}^{N_s} \alpha_j s_j(t-\delta_j)$; where α_j and δ_j are the relative attenuation and delay due to the propagation of the j^{th} source to the second sensor x_2 and N_s is the number of sources; the DUET algorithm attempts to recover the various sources $s_j(t)$. DUET relies on four assumptions given below with their accompanying explanations. These assumptions involve a linear transform T on the set of sources $\mathcal{S} \in L^2(\mathbb{R})$, (that is we assume sources to be square integrable), which we assume to be a vector space. The linear transform $T\{s_j(t)\}(\lambda)$ maps $s_j(t) \in \mathcal{S}$ to $S_j(\lambda)$, where examples of appropriate transforms will be discussed later. The four assumptions are now rigorously stated as follows:

p1) $T^{-1}\{T\{s_j(t)\}(\lambda)\}(t) = s_j(t), \forall s_j(t) \in \mathcal{S}$: T is invertible.

p2) $\Lambda_j \cap \Lambda_k = \emptyset$ for $j \neq k$, where Λ_j is the support of $S_j(\lambda)$, i.e., $\Lambda_j \doteq \operatorname{supp} S_j(\lambda) \doteq \{\lambda : S_j(\lambda) \neq 0\}$: the images of different sources under T have disjoint supports.

p3) $\operatorname{supp} T\{s_j(t-\delta)\}(\lambda) = \operatorname{supp} T\{s_j(t)\}(\lambda)$ for any $s_j(t) \in \mathcal{S}, \forall |\delta| < \Delta_n$, where Δ_n is some appropriate bound.

p4) For every $j \in \{1, \ldots N_s\}$ there exist two operators F_j and G_j such that

$$\alpha_j = F_j\left(T\left\{x_1(t)\right\}(\lambda), T\left\{x_2(t)\right\}(\lambda)\right),$$
$$\delta_j = G_j\left(T\left\{x_1(t)\right\}(\lambda), T\left\{x_2(t)\right\}(\lambda)\right)$$

As an example, consider, T to be the Fourier transform, $\hat{s}(\omega)$. As it is invertible, it clearly satisfies p1. Looking to p2, for two source signals, $T\{s_1(t)\}(\omega) = \hat{s}_1(\omega)$ and $T\{s_2(t)\}(\omega) = \hat{s}_2(\omega)$, p2 is satisfied if the signals are disjoint in frequency, i.e., $\hat{s}_1(\omega)\hat{s}_2(\omega) = 0, \forall \omega$. It is a property of the Fourier transform that a delay in time is a phase shift in frequency, and thus a delay by δ does not change the support of a signal in the frequency domain, and the Fourier transform thus satisfies p3. For the functions F and G of p4, DUET uses

$$F\left(T\left\{x_1(t)\right\}(\omega), T\left\{x_2(t)\right\}(\omega)\right) = \left|\frac{\hat{x}_2(\omega)}{\hat{x}_1(\omega)}\right|,$$
$$G\left(T\left\{x_1(t)\right\}(\omega), T\left\{x_2(t)\right\}(\omega)\right) = \frac{-1}{\omega}\angle\left(\frac{\hat{x}_2(\omega)}{\hat{x}_1(\omega)}\right)$$

which extract the relative attenuation and delay from the mixtures for each source, and all four conditions are satisfied.

Of course, if $T\{\cdot\}(\omega)$ were in practice the Fourier transform, condition p2 that the sources are disjoint in frequency is quite restrictive and not likely to be satisfied for many interesting classes of signals. DUET uses, therefore, the Windowed Fourier Transform [10], [11] which greatly increases the set of applicable signals. However, condition p2, due to properties of time-frequency analysis, cannot be satisfied in a strict sense for signals such as speech signals. Nevertheless, DUET replaces the equality in p1–p4 with appropriately defined approximate equality and the robustness and success of the demixing results provides evidence that these approximations are valid. The purpose of this paper is to investigate the following caveat; mixing parameter estimates in [1] exhibit a bias.

Signals that satisfy p2 using the Windowed Fourier Transform were termed windowed disjoint orthogonal (WDO) signals by Jourjine *et al.*, in [12], however, WDO time-frequency representations of speech are typically sparse representations [13]. Sparsity is commonly assumed in multichannel anechoic mixing source separation algorithms, thus the scope of applications for a time-frequency transform that improves the signal representation (in the spirit of p1–p4) is potentially broad. As a first example, properties p3 and p4 resonate strongly with the assumptions made by the generalized cross correlation (GCC) algorithm where signal dominance in a frequency bin is emphasized by an appropriate weighting scheme. Time delay estimation via the GCC algorithm can be used as an approach for source separation or localization, see [14] or the more recent approach taken by Benesty *et al.* in [15].

Property p2 is reminiscent of the assumption made by the class of algorithms which seeks to separate and localize latent sources by projecting them onto a representative time-frequency signal dictionary, typically in a supervised manner [4], [16], [17]. Two sources cannot physically inhabit the same location, and thus while sparsity is the stated assumption, WDO is implied. Similarly, independent component analysis (ICA) approaches commonly leverage the parsimonious nature of speech in time-frequency to perform separation [18], [19]. Moreover, approaches that solve the related instantaneous mixing model, [20]–[22], leverage properties p1 and p2. The contribution in [23] defines a well-posed nonnegative matrix factorization (NMF) as being a member of a separable factorial articulation family, which implicitly links sparsity and WDO as being crucial for a unique NMF solution. The approximate WDO condition [24] is more representative of speech (and other signals of interest) mixtures than the WDO and, therefore, will be one of the key assumptions underpinning the results in this paper, however, the approximate WDO condition is more demanding than simply requiring that the sources are sparse. In summary, the WDO measure of the constituents of the mixture in [1] is a good indicator of the attainable success possible via DUET demixing, thus, a time-frequency representation that boosts the measure of WDO of the sources is appealing for source separation applications.

III. Notation and Definitions for Subsample Delay

Delaying discrete signals by a noninteger number of samples is a challenging problem in array and multirate signal processing [25]. Anticipating possible ambiguity in the notation, we use a signal delay/interpolation problem to define our notation and our benchmark method. Accurately delaying a signal by a subsample delay in discrete time and in discrete time-frequency is crucial to the definition of our synchronized linear transform.

A continuous time signal $s(t)$ is denoted by $s[n] = s(nT)$ in the discrete time domain where T is the sampling period and $n = 0, 1, 2, \ldots Q$. This continuous time signal, delayed by $\delta \in \mathbb{R}$ seconds, is given by $s(t-\delta)$. Similarly, the discrete signal, $s[n]$, can be delayed by an integer d number of samples giving

$s[n-d]$. When $d=\delta/T$ is not necessarily an integer, we use the notation

$$s_j^{\delta}[n] = s_j(nT-\delta) \tag{1}$$

to indicate that the signal is discrete but that the delay in samples could in fact be noninteger. Explicitly, the sample values of $s(t)$ for noninteger sample delay are given by

$$\begin{aligned} s^{\delta}[n] &= s(nT-\delta) \\ &= \int_{-\infty}^{+\infty} s(t-\delta) \sum_{k=-\infty}^{\infty} \delta_p(t-kT) \\ &\qquad \times \mathbf{1}_{[-\frac{T}{2},\frac{T}{2}]}(t-nT)dt \end{aligned} \tag{2}$$

where $\delta_p(\cdot)$ indicates a Dirac pulse, δ is delay in seconds and the indicator function $\mathbf{1}_{[-(T/2),(T/2)]}(t)=1$ when $|t|<(T/2)$ and 0 otherwise. Using sinc interpolation, given that the signal is bandlimited and sampled at a sufficiently high sampling rate, results in

$$s^{\delta}[n] = \sum_{k=-\infty}^{+\infty} s(kT) h_T(t-kT) \tag{3}$$

when $t=nT-\delta$ and $h_T(t)=\sin(\pi t/T)/(\pi t/T)$. This follows from the shift and convolution properties of the Fourier transform

$$\mathcal{F}\left\{s(t-\delta)\right\}(\omega) = \operatorname{rect}\left(\frac{\omega T}{2\pi}\right) e^{-j\omega\delta} \cdot T \sum_{k=-\infty}^{\infty} s(kT) e^{-j\omega kT} \tag{4}$$

where the rectangular function, $\operatorname{rect}(x)=1$ when $|x|<(1/2)$ and 0 when $|x|>(1/2)$. In practice a finite length approximation of the sinc function leads to error in the estimate of $s^{\delta}[n]$.

Noninteger sample delay of a bandlimited signal sampled above the Nyquist rate can also be *approximated* by multiplying the discrete Fourier transform of $s[n]$

$$\operatorname{DFT}\left\{s[n]\right\} = S[k] = \sum_{n=0}^{N-1} s[n] W^{kn} \tag{5}$$

where $k=0,1,\ldots,N-1$ and $W=e^{-j(2\pi/N)}$, by a linear phase term W^{kd}. This corresponds to a circular shift of the signal by d samples when $d\in\mathbb{Z}$. Using a functional notation, we define the zero-padding function

$$\begin{aligned} s_z[q] &= \operatorname{ZP}(b, s[n], e) \\ &= \begin{cases} 0 & 0 \le q < b \\ s[n], n = q-b & b \le q < Q+b \\ 0 & Q+b \le q < Q+b+e \end{cases} \end{aligned} \tag{6}$$

which appends b and e zeros to the beginning and end of the signal, respectively. The inverse-pad function $\operatorname{IP}(b, s[n], e)$ removes b and e samples from the beginning and end of the signal. Zero-padding by $\lceil d \rceil$, where $\lceil\cdot\rceil$ is the ceiling function; taking the DFT; multiplying by the linear phase term; taking the IDFT; and inverse-padding gives the desired result. We define $\operatorname{IDFT}\{S[k]\} = (1/N)\sum_{k=0}^{N-1} S[k] W^{-kn}$ to be the inverse DFT. The *frequency domain method* in (7) is the benchmark method used for the remainder of this paper.

$$\hat{s}_{\star}^{\delta}[n] = \operatorname{IP}\left(\lceil d \rceil, \operatorname{IDFT}\left\{\operatorname{DFT}\left\{\operatorname{ZP}(\lceil d \rceil, s[n], \lceil d \rceil)\right\} W^{kd}\right\}, \lceil d \rceil\right). \tag{7}$$

The error measured by the Euclidean distance $\|s^{\delta}[n] - \hat{s}_{\star}^{\delta}[n]\|_2$ is considered to be sufficiently small, although it is still greater than machine error.

IV. The Classical Short-Time-Fourier-Transform

In this section we outline the principles of time-frequency analysis and window selection as a background for the sSTFT. The window most frequently associated with the Gabor transform [26] is the Gaussian bell, $w_a(t) = (1/2\sqrt{\pi\alpha}) e^{-t^2/4\alpha}$. The Gabor transform of $s(t)\in L^2(\mathbb{R})$, an inner product of the signal with weighted exponential basis functions, is

$$S(\omega,\tau) = \int w_a(t-\tau) s(t) e^{-j\omega t} dt \tag{8}$$

where $\{\omega,\tau\} = \{k\Omega, mT\}, \{k,m\} \in \mathbb{Z}$ and $\Omega T \le 2\pi$. One interpretation of (8) is that the lowpass filter (LPF), $w_a(t)$, is modulated and shifted in time such that the signal is filtered with a set of bandpass filters (BPF) yielding time-frequency coefficients. The inverse Gabor transform of the time-frequency representation $S(\omega,\tau)$ is defined as

$$s(t) = \sum_{k\in\mathbb{Z}} \sum_{m\in\mathbb{Z}} w_s(t-mT) S(k\Omega, mT) e^{jk\Omega t} \tag{9}$$

where the synthesis window $w_s(t) = w_a(t)$. Bastiaans generalized the above expressions in [27] via the Zak transform. We define $S^{c,0}[k,m]$ to be the discrete "Classical" short-time-Fourier-transform (cSTFT) of $s^0[n]\in L^2(\mathbb{Z})$

$$\begin{aligned} S^{c,0}[k,m] &= \operatorname{cSTFT}\left\{s^0[n]\right\} \\ &= \sum_{n=mR}^{N-1+mR} s^0[n] w_a[n-mR] W^{k(n-mR)} \end{aligned} \tag{10}$$

positioned at sample mR where $w_a[n]\in\mathbb{R}^N$ is the analysis window function and R is the number of window hop-size samples, e.g., the rational oversampling factor. The notation $s^0[n]$ and $S^{c,0}[k,m]$ denotes the reference signal. The discrete-time representations of $s^0[n]$ and $s^{\delta}[n]$ are related by a relative delay, δ. In discrete-time-frequency these signals are denoted by $S^{c,\delta}[k,m]$ and $S^{0,c}[k,m]$. [k,m] are the discrete frequency and time indices, respectively. N is the DFT size. The cSTFT is inverted using the synthesis window $w_s[n]\in\mathbb{R}^N$ and overlap and add (OLA) resynthesis given R.

$$\begin{aligned} \operatorname{cISTFT}\left\{S^{c,0}[k,m]\right\} &= \sum_{m\in\mathbb{Z}} w_s[n-mR] \\ &\quad \times \sum_{k\in\mathbb{Z}} S^{c,0}[k,m] W^{-k(n-mR)}. \end{aligned} \tag{11}$$

Regarding window selection, if both the root mean square duration and bandwidth of the window $w_a(t)$, Δ_{w_a}, and

Δ_{W_a} respectively, are finite then $w_a(t)$ is a time-frequency window. Time-frequency windows satisfy the property, $\Delta_{w_a}\Delta_{W_a} \geq (1/2)$, thus, they are localized in both time and frequency domains. For example, $\Delta_{w_a}\Delta_{W_a} = 1/2$ when $w_a(t)$ and $W_a(\omega)$ are Gaussian windows in the time and frequency domain, respectively. For the classic text on analysis windows see [28] and the subsequent comments in [29]. The set of discrete analysis functions based on the analysis window w_a is defined as

$$\{w_{a,k,m}[n]\} = \left\{ w_a(n - mR) e^{-\frac{j2\pi kn}{N}} | \{k, m\} \subset \mathbb{Z} \right\} \quad (12)$$

which is a discrete set of signals obtained by shifting and modulating the elementary analysis window w_a. The locally static nature of these basis functions is the underlying problem when they are used for multichannel anechoic observations.

V. The sSTFT

We define the sSTFT mathematically and then motivate it graphically. The appealing properties of the sSTFT are defined in the following subsections; they supersede the properties of the cSTFT in both accuracy and scope. The term classical STFT (cSTFT) or unsynchronized STFT refers to the unsynchronized time-frequency analysis in (10) and the term synchronized STFT (sSTFT) refers to our synchronized time-frequency transform.

A. Definition of the sSTFT

Definition 1: The reference analysis and synthesis windows are nonzero for $N/2$ samples and zero-padded by $N/2$ zeros and are defined as

$$\begin{aligned} w^0_{az}[n] &= \mathrm{ZP}\left(N/4, w_a[2n], N/4\right), \\ w^0_{sz}[n] &= \frac{1}{2}\mathrm{ZP}\left(N/4, w_s[2n], N/4\right). \end{aligned} \quad (13)$$

They form a pair of windows of length N samples.

Definition 2: Locally translated and dilated versions of these analysis and synthesis windows are defined as

$$\begin{aligned} w^\delta_{az}[n] &= \mathrm{IDFT}\left\{\mathrm{DFT}\left\{w^0_{az}[n]\right\} W^{kd}\right\} \\ w^\delta_{sz}[n] &= \frac{1}{2}\mathrm{IDFT}\left\{\mathrm{DFT}\left\{w^0_{sz}[n]\right\} W^{kd}\right\}. \end{aligned} \quad (14)$$

Definition 3: For a delay of $|\delta|/T < \Delta_s$ samples and $\delta/T \in \mathbb{Z}$, the synchronized STFT $S^\delta_j[k, m]$ of $s^\delta_j[n]$, is

$$\begin{aligned} S^\delta_j[k, m] &= \mathrm{sSTFT}\left\{s^\delta[n], \delta\right\} \\ &= \sum_{n=mR}^{N-1+mR} s^\delta_j[n] w^\delta_{az}[n - mR] W^{k(n-mR)} \end{aligned} \quad (15)$$

where the window hop-size or oversampling factor is $R = N/4$ samples. By convention $\Delta_s = N/4$. Similar to the cSTFT, the analysis basis functions are obtained by shifting and modulating the elementary signal $w^0_{az}[n]$

$$\left\{w^0_{az,k,m}[n]\right\} = \left\{ w^0_{az}[n - mR] e^{-\frac{j2\pi kn}{N}} | \{k, m\} \subset \mathbb{Z} \right\}. \quad (16)$$

However, the structure of the reference analysis window $w^0_{az}[n]$ allows the windowed signal $w^0_{az}[n]s^0[n]$ to be shifted by $|\delta|/T < N/4$ samples such that the circular shift property still holds for each local windowed version of $s^0[n]$ without wrap-around in time. Consequently, $w^0_{az}[n]$, can be synchronized with the delayed source signal using an additional local synchronization parameter, δ, so that the same samples of $s^0[n]$ and $s^\delta[n]$ are weighted by the same samples of $w^\delta_{az}[n]$. Thus, the analysis basis functions have an additional flexibility over the functions in (12) due to the synchronization parameter δ

$$\left\{w^\delta_{az,k,m}[n]\right\} = \left\{ w^\delta_{az}[n - mR] e^{-\frac{j2\pi kn}{N}} | \{k, m\} \subset \mathbb{Z} \right\}. \quad (17)$$

Definition 4: The inverse synchronized STFT is defined as

$$\begin{aligned} \mathrm{sISTFT}\left\{S^\delta[k, m], \delta\right\} &= \sum_{m\in\mathbb{Z}} w^\delta_{sz}[n - mR] \\ &\quad \times \sum_{k\in\mathbb{Z}} S^\delta[k, m] W^{-k(n-mR)}. \end{aligned} \quad (18)$$

B. Graphical Motivation of the sSTFT

Consider a discrete time source signal $s^0_j[n] \in L^2(\mathbb{Z})$ which is delayed by $\delta \in \mathbb{R}$ seconds as it propagates to sensor x_i yielding $s^\delta_j[n]$. Without loss of generality we neglect propagation attenuation effects such that the direct path attenuation is 1. $S^{c,\delta}_j[k, m]$ (or $S^\delta_j[k, m]$) and $S^{c,0}_j[k, m]$ (or $S^0_j[k, m]$) are the cSTFT (or sSTFT) of the delayed and reference signals, e.g. $s^\delta_j[n]$ and $s^0_j[n]$ respectively. When the windows are not synchronized with the observed signals but with some absolute clock time across the channels, the estimated relative attenuation and delay between the channels is typically inaccurate. Fig. 1(a), (d), (g), and (j) illustrates the inaccuracies of the cSTFT in a multichannel setting where instantaneous relative parameter information [Fig. 1(g), (j)] is desired between two windowed observations [Fig. 1(a), (d)]. In contrast, the windows in Fig. 1(b) and (e) are locally synchronized to each signal. The relative attenuation and delay estimates are accurate [Fig. 1(h), (k)]. However, the window used in Fig. 1(b) and (e) is suitable due to the support of the signal; it is unreasonable to assume every signal is zeropadded, and thus, for more general signals the synchronized windows in Fig. 1(b) and (e) are not appropriate.

Definition 5: The term *FFT-support* describes the frame/vector of N samples which is analyzed. Typically, this is the set of signal indices $\{mR \leq n \leq N - 1 + mR\}$.

Definition 6: Window-support describes the samples of the signal which are not attenuated to zero by the window, for example, for the cSTFT $\mathrm{supp}\, w_a[n - mR] \doteq \{n : w_a[n - mR] \neq 0\}$.

Fig. 1(c) and (f) illustrates a practical sSTFT implementation where the FFT-support is the same as the FFT-support of the cSTFT. Window dilation and zeropadding facilitate local synchronization of $w^0_{az}[n]$ which extends the applicability of the sSTFT to multichannel settings without misalignment for a range interesting signals as the signal-window product can be circularly shifted without wrap-around in time. Previously the physical displacement of the sensors combined with the global

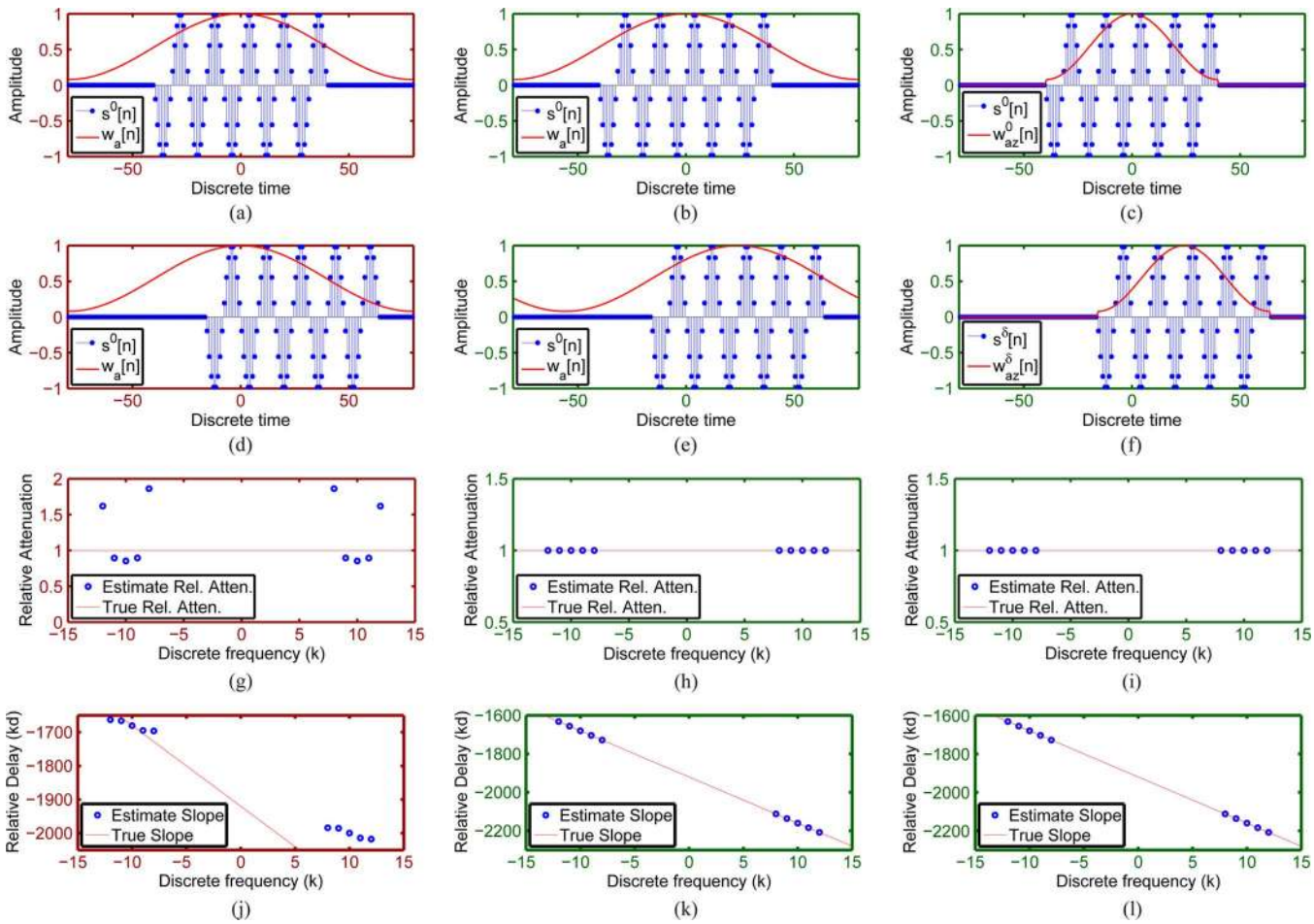

Fig. 1. cSTFT versus sSTFT analysis. $s^0[n]$, is shown (stems) in (a), (b), (c). $s^\delta[n]$, observed at a physically displaced sensor is shown (stems) in (d), (e), (f). A Hamming window $w_a[n]$ (solid line) is positioned at the same global position in (a), (d). The Hamming window (solid line) in (b), (e) is synchronized to $s^0[n]$ and $s^\delta[n]$ respectively. (c), (f) show a practical synchronized window for more general signals. In (g), (j) estimation using the cSTFT of the observations in (a), (d) does not give the correct $\{\alpha, \delta\}$. (h), (k) show the $\{\alpha, \delta\}$ estimates using the sSTFT shown in (b), (e). The ideal and the estimated $\{\alpha, \delta\}$ match exactly for the sSTFT in (h), (k), (i), (l). A subset of bins with significant signal power is used to illustrate the estimates. (a) cSTFT analysis (Ref. Sig.). (b) sSTFT analysis (Ref. Sig.). (c) sSTFT analysis (Ref. Sig.). (d) cSTFT analysis (Del. Sig.). (e) sSTFT analysis (Del. Sig.). (f) sSTFT analysis (Del. Sig.). (g) cSTFT instantaneous Rel. Atten.. (h) sSTFT instantaneous Rel. Atten.. (i) sSTFT instantaneous Rel. Atten.. (j) cSTFT instantaneous Rel. Del. (k) cSTFT instantaneous Rel. Del. (l) sSTFT instantaneous Rel. Del.

window placement of the cSTFT conspired to violate the properties set out by the authors of [1], unless the propagation path or the signal had special properties, e.g., periodicity or both observations of the signal had an equal propagation distance, however, now local STFT synchronization removes this error. In summary the window-support, the set $\{mR+\delta/T+\Delta_s \leq n \leq N-1+mR-\Delta_s+\delta/T\}$, and the FFT-support of the sSTFT, $\{mR \leq N-1+mR\}$, are different. Combining the local synchronization parameter with the zero-padded structure of the window, $w^\delta_{az}[n]$, allows the windowed signal to be shifted arbitrarily, locally to the reference signal within the FFT-support of the window up to $\delta/T < \Delta_s$ without wrap-around in time. Consequently, the same portions of the test signal are scaled by the appropriate samples of the analysis function on all channels and multichannel time-frequency error is removed. Assuming the constituent signals of the mixture are WDO—the underpinning assumption made by [1], [4]–[6], [16]—we can synchronized the sSTFT for each source in the knowledge that in a subset of the time-frequency points, Λ_j, source j is dominant, thus, the synchronized kernel for each time-frequency bin is appropriate for the dominant source in that set of time-frequency bins.

C. Limitations of Classical Time-Frequency Windows

The applicability of the cSTFT is limited for accurate use in multichannel applications due to an inherent paradox in classical time-frequency window construction. Edge effects due to the finite support of the windowed signal, and the inherent uncertainty about the data lying outside of the window, are contributing factors in the relative delay error between the windowed reference and delayed signal when the cSTFT is used in Fig. 1. This effect is typically coupled with signal scaling due to window misalignment and bell-shaped structure. The scaling effect adversely affects relative delay and attenuation unless the signal is a unit impulse or has similar characteristics. Coupling of the scaling and edge effects motivates the structure of the sSTFT window. Fig. 2 illustrates a reference signal in row 1, $s^0[n]$ and a delayed version of this test signal in row 2, $s^\delta[n]$. A Kaiser window is overlayed on both of these signals in a cSTFT-like manner. The Kaiser window is tuned so that, first, for the window-supported set of the signal, the weights are approximately 1, and second, for the window-supported set of the signal the weights are bell-shaped. The flat window (Kaiser with $\beta = 1$) seems to be ideal as the signal is periodic for a region and then zero elsewhere, thus,

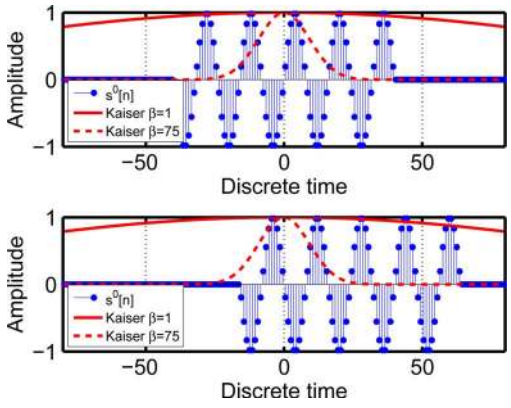

Fig. 2. Scaling and edge effect tradeoff. Translatable and dilatable windows preserve the co-information between two relatively delayed observations. Neither a flat nor bell-shaped window has the desired properties.

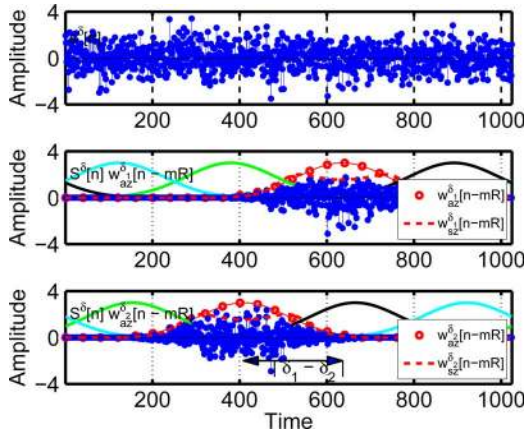

Fig. 3. Local stationary invertibility: the sSTFT using either synchronization parameter (row 2, δ_1, or row 3, δ_2) is invertible.

the signal samples are scaled by $\approx$1 on both channels (there are no relative miscaling effects). The disadvantage of this window is that any change in the characteristics of the signal at the bounds of the frame can affect the co-information between both observations of the windowed signal quite dramatically. Undesirable innovations in the delayed signal are weighted, without prejudice, by an equal amount as the desired portion. The bell-shaped window would appear then to be ideal as it focusses on a narrower range of samples, and discriminates against the undesirable samples. Conversely, the bell-shaped window ($\beta = 75$) significantly alters the spectrum of the windowed delayed signal relative to the reference signal, due to the global positioning of both windows. Different samples on the second channel are weighted compared to the first channel.

Inappropriate scaling due to misalignment of the window is inherently dependent on these edge effects as new information comes into view as the signal is shifted relative to the reference signal. Uncoupling these effects necessitates knowledge of the form of the data appearing outside of the frame. In a multichannel anechoic setting we can estimate or sometimes assume that prior knowledge has informed us of the structure of the signal in the near future, and thus, choose the appropriate window. Taking the sSTFT of two signals which have a relative delay between them, we assume we know more than just the statistics of the reference signal lying within the time frame under observation. Consequently, we dilate the analysis windows and zero-pad them. By translating the window within an acceptable range we uncouple the edge and scaling effects. Consequently, the sSTFT now assumes the role of supplying the "prior knowledge" (cf. [9]).

D. Properties of the sSTFT

We now motivate and define the appealing properties of the sSTFT. An injective mapping to the related properties in Section II for the Fourier transform is not intended because windowing issues do not arise there. We define each property for integer sample delays; for subsample delay the accuracy is typically sufficiently good so that approximate equality can be assumed.

P̂1a) **Local Stationary Invertibility**
$\text{sISTFT}\{\text{sSTFT}\{s^\delta[n],\delta_1\},\delta_1\} = s^\delta[n] = \text{sISTFT}\{\text{sSTFT}\{s^\delta[n],\ \delta_2\},\delta_2\}$ for $|\delta_1|/T < \Delta_s$ and $\delta_1/T \in \mathbb{Z}$ and $|\delta_2|/T < \Delta_s$ and $\delta_2/T \in \mathbb{Z}$.

Proposition 1: The start time of the global clock, which aligns the windows in time, is irrelevant as long as the appropriately positioned synthesis window is used which corresponds to the translated analysis window.

Graphically, Fig. 3 row 1 shows a random signal, $s^\delta[n]$, drawn from a normal distribution. Row 2 shows a windowed portion of the signal using the window $w_{az}^{\delta_1}[n-mR]$ with $\delta_i = \delta_1$, where δ_1 is an arbitrary synchronization parameter, and also the train of window functions positioned at multiples of R. The appropriate synthesis window for $w_{az}^{\delta_1}[n-mR]$, e.g. $w_{as}^{\delta_1}[n-mR]$ is overlayed over the frame of the illustrated signal portion of interest, $s^\delta[n]w_{az}^{\delta_1}[n-mR]$. Similarly, row 3 shows a windowed portion of the same signal but positioned using $\delta_i = \delta_2$, e.g., $s^\delta[n]w_{az}^{\delta_2}[n-mR]$. The train of analysis windows positioned at multiples of R is also shown along with the appropriately positioned synthesis window for $s^\delta[n]w_{az}^{\delta_2}[n-mR]$, e.g. $w_{as}^{\delta_2}[n-mR]$. Either synchronization parameter can be used to linearly transform the data as long as the appropriate synthesis window is used.

Lemma 1:

$$\text{sSTFT}\left\{s^\delta[n],\delta_1\right\} \neq \text{sSTFT}\left\{s^\delta[n],\delta_2\right\} \text{ when } \delta_1 \neq \delta_2 \tag{19}$$

unless, for example, the test signal $s^\delta[n]$ is periodic and an appropriate δ_1 and δ_2 are chosen.

In effect the relative shift of the two window functions causes the linear transform to consider two different observations of the same signal. The window-supports for both frames of data are not the same whereas the FFT-supports are the same.

$$\begin{aligned}\text{supp}\, w_{az}^{\delta_1}[n] \neq \text{supp}\, w_{az}^{\delta_2}[n],& \\ \left\{\frac{N}{4}-1-\delta_1/T,\ldots,3\frac{N}{4}-1-\delta_1/T\right\}& \\ \neq \left\{\frac{N}{4}-1-\delta_2/T,\ldots,3\frac{N}{4}-1-\delta_2/T\right\}&.\end{aligned} \tag{20}$$

For $\delta_1 \in \mathbb{Q} \setminus \mathbb{Z}$ the analysis window $w_{az}^{\delta_1}[n]$ is typically asymmetric ($\mathbb{Q} \setminus \mathbb{Z}$ is the complement of set $\mathbb{Z}$ relative to set $\mathbb{Q}$). Generally, even windows are desirable as they are linear phase functions in the frequency domain, however, $w_{az}^{\delta_1}[n]$ with $\delta_1 \in \mathbb{Q} \setminus \mathbb{Z}$ is an approximately linear phase window as a subsample

shift can be used to center this window such that it has a real spectrum.

$\hat{\text{P}}1\text{b}$) **Relative Delay**
$\text{sSTFT}\{s^\delta[n],\delta\} = \text{sSTFT}\{s^0[n],0\}W^{k(\delta/T)}$ when $(\delta/T)\in\mathbb{Z}$ and $|\delta|/T<\Delta_s$

Proposition 2: Local shifts of a partitioned signal introduce error to the globally shifted signal unless each frame is invariant to circular shifting for a given delay, i.e., by invariant we mean that the circular shift of each frame is a conventional linear shift. Integer sample signal delay can be implemented on a frame by frame basis if it is segmented using a sSTFT type segmentation with the appropriate analysis parameters.

Lemma 2: If the window-support of the analysis windows used with the sSTFT satisfies the condition

$$\begin{aligned}&\text{supp}\, w^\delta_{az}[n-mR] \equiv \text{supp}\, w^0_{az}[n-mR]-\delta/T\\ &\forall m\in\mathbb{Z}, \frac{\delta}{T}\in\mathbb{Z}, |\delta|/T<\Delta_s \text{ and } mR\le n\le N-1+mR \quad (21)\end{aligned}$$

the windows are not wrapped around in time.

Lemma 3:

$$\begin{aligned}&s^\delta[n]w^\delta_{az}[n-mR]\\ &\quad=\text{circ}\left\{s^0[n]w^0_{az}[n-mR],\delta\right\}\\ &\quad=\text{circ}\left\{s^0[n],\delta\right\}\text{circ}\left\{w^0_{az}[n-mR],\delta\right\} \quad (22)\end{aligned}$$

$\forall m\in\mathbb{Z}, (\delta/T)\in\mathbb{Z}, |\delta|/T<\Delta_s$ and $mR\le n\le N-1+mR$, and not wrapped around in time.

The circular shift operator is defined for $0\le n\le N-1$

$$\begin{aligned}\text{circ}\{x[n],\delta\} &\doteq \text{IDFT}\left\{\text{DFT}\{x[n]\}W^{k(\delta/T)}\right\}\\ &=\sum_{m=0}^{N-1}x[m]D[(n-m) \bmod N] \quad (23)\end{aligned}$$

where $D[n]=\text{sinc}[n-d]$ is the delayed unit impulse for $d=\delta/T\in\mathbb{Z}$, and $D[n]$ and $x[n]$ are length N sequences. The Relative Delay property $\hat{\text{P}}1\text{b}$ allows signal delay to be performed accurately in the time-frequency domain (even with overlapping windows) as a local shift of each individual frame of the windowed data followed by reconstruction via the appropriate synthesis windows. Fig. 4(a) shows an N sample frame of a windowed signal $s^0[n]w^0_{az}[n]$ in row 1 (where $m=0$). The signal is not wrapped around in time (22) when it is shifted in time using the circular shift operator (23) if the delay is less than Δ_s. In Fig. 4(a) row 2 $s^\delta[n]w^\delta_{az}[n]$ is linearly shifted by 150 samples relative to $s^0[n]w^0_{az}[n]$. A suitable synthesis window $w^\delta_{as}[n]$ can be used to resynthesize the signal if it is synchronized with the analysis window $w^\delta_{az}[n]$ [Fig. 4(a) row 3] and consequently generate the appropriate contribution towards the delayed signal $s^\delta[n]$.

In comparison, each frame of the observed data taken by the cSTFT is not guaranteed to satisfy (22) as $w_a[n]$ is not zero-padded.

$$\begin{aligned}&s^\delta[n]w_a[n-mR]\neq\text{circ}\left\{s^0[n]w_a[n-mR],\delta\right\}\\ &\forall m\in\mathbb{Z}, \frac{\delta}{T}\in\mathbb{Z}, |\delta|/T<\Delta_s \text{ and } mR\le n\le mR+N-1. \quad (24)\end{aligned}$$

If the frame of data in Fig. 4(b) was part of a longer signal and each frame of data was locally delayed similar to Fig. 4(a), and then combined using OLA, the wrap-around in time in each frame would cause significant distortion to the resynthesized signal.

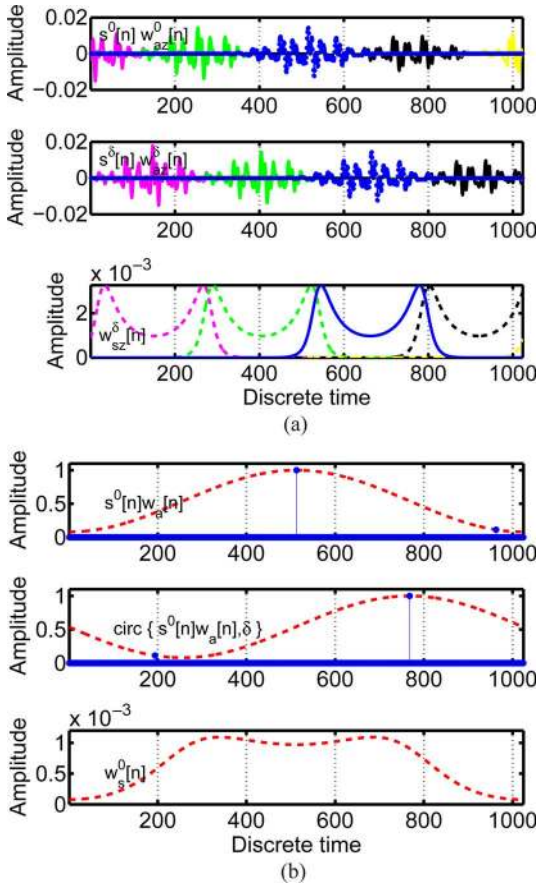

Fig. 4. The whole signal can be analyzed, delayed locally in time-frequency and then resynthesized frame-by-frame: In (a) $s^0[n]$ is windowed with $w^0_{az}[n]$ (row 1 stems) and circularly shifted (23) by 150 samples (without time wrap-around), $s^\delta[n]w^\delta_{az}[n]$ (row 2 stems). $s^\delta[n]w^\delta_{az}[n]$ is resynthesized with $w^\delta_{sz}[n]$, (row 3 full-line). Other windowed segments of the signal (row 1, 2) give the context of the locally delayed signal portion. (b), row 1 shows $s^0[n]$ (2 delta pulses) and $w_a[n]$. When $s^0[n]w_a[n]$ is circularly shifted, $w_a[n]$ causes distortion when the whole signal is resynthesized using OLA. (a) Relative delay: sSTFT. (b) Relative delay: cSTFT.

E. Window Dependence for Fractional Delay

Property $\hat{\text{P}}1\text{b}$ becomes a window dependent approximation for fractional or subsample delay, however, the performance of the sSTFT for the sequence of operations; analysis and then resynthesis—for subsample delay—is sufficiently good for windows of practical interest.

Proposition 3: Delaying the signal in the time-frequency domain using property $\hat{\text{P}}1\text{b}$ is exact for integer sample delay, $\delta/T\in\mathbb{Z}$, and an approximation for subsample delay $\delta\in\mathbb{Q}\setminus\mathbb{Z}$.

Let the reference signal be an infinitely long sinc function, $s^0[n]=\text{sinc}[n]$. The analysis window, $w^0_{az}[n]$, is N samples long and an N sample FFT is used. We approximate the infinitely long sinc with an M-sample sinc where $M\gg N$, ideally $M\to\infty$. We delay the sinc exactly by a subsample number of samples δ/T using its closed form $\text{sinc}[n-\delta/T]$. We window $\text{sinc}[n-\delta/T]$ by taking a truncated windowed portion of length N samples, positioned at $-N/2$ in discrete time. This forms the left-hand side (LHS) of (25) and is the ideal way to analyze the

delayed signal, with respect to the sSTFT analysis paradigm. Alternatively, the right-hand side (RHS) of (25) is another way of computing the frequency domain representation of this fractionally delayed windowed signal segment. The reference signal is windowed and then delayed on the RHS. The inequality in (25) explains the error inherent in subsample sSTFT processing.

$$\begin{aligned}\operatorname{sinc}\left[n-\delta/T+\left(\frac{M-N}{2}\right)\right]w_{az}^{\delta}[n]\\ \neq \mathrm{IDFT}\left\{\left(\sum_{n=0}^{N-1}\operatorname{sinc}\left[n+\left(\frac{M-N}{2}\right)\right]\right.\right.\\ \left.\left.\times w_{az}^{0}[n]W^{kn}\right)W^{k(\delta/T)}\right\}\end{aligned}\tag{25}$$

for $0 \leq n \leq N-1$. The sSTFT assumes that the LHS of (25) is equivalent to the RHS of (25). For the more general case of the signal $s[n]$, we write (26).

Lemma 4: The discrete sSTFT is inaccurate for fractional delay because, for finite length sinc filters

$$\begin{aligned}\sum_k s[k]\operatorname{sinc}(nT-\delta-kT)\cdot\sum_k w_{az}^{0}[k]\operatorname{sinc}(nT-\delta-kT)\\ \neq\sum_k s[k]w_{az}^{0}[k]\operatorname{sinc}(nT-\delta-kT).\end{aligned}\tag{26}$$

Conversely, for integer sample shifts, $d=\delta/T\in\mathbb{Z}$, (25) reduces to a delayed delta pulse times a delayed window yielding equality in this relationship

$$\begin{aligned}\sum_{m=0}^{N-1}s[m]D\left[(n-m)\bmod N\right]\cdot\sum_{m=0}^{N-1}w_{az}^{0}[m]D\left[(n-m)\bmod N\right]\\ =\sum_{m=0}^{N-1}s[m]w_{az}^{0}[m]D\left[(n-m)\bmod N\right].\end{aligned}\tag{27}$$

$D[n]$ is a unit impulse delayed by d samples. $D[n]$, $s[n]$ and $w_{az}^{0}[n]$ are length N signals. In summary, the cSTFT suffers from misalignment error which increases as integer sample delay gets larger, whereas the sSTFT is correct for integer sample delay and inaccurate for subsample delay.

In Fig. 5(a) we highlight the inaccuracy in (26) by using a rectangular window. The rectangular window does not taper the signal thus the effect of subsample delay on the circular shift assumption of the sSTFT is emphasized. The rectangular window suffers from severe Gibbs effect and high side lobes. A tapered window reduces these side lobe levels but spreads the main lobe width, and thus, decreases the resolution. Firstly, signal delay using the expression $\operatorname{circ}\{s^{0}[n],\delta\}\operatorname{circ}\{w_{az}^{0}[n],\delta\}$ is illustrated with squares. The signal is an element-wise non-negative signal drawn from a rectified Gaussian distribution (illustration convenience). There is uncertainty as to the structure of the signal in the near future and hence appropriate synchronization is critically important. Secondly, signal delay using the method $\operatorname{circ}\{s^{0}[n]w_{az}^{0}[n],\delta\}$ is illustrated using circles. Finally, the original windowed signal, $s^{0}[n]w_{az}^{0}[n]$, is illustrated using dot-stems. The discrepancy between each of the implementations can be improved via tapered windows.

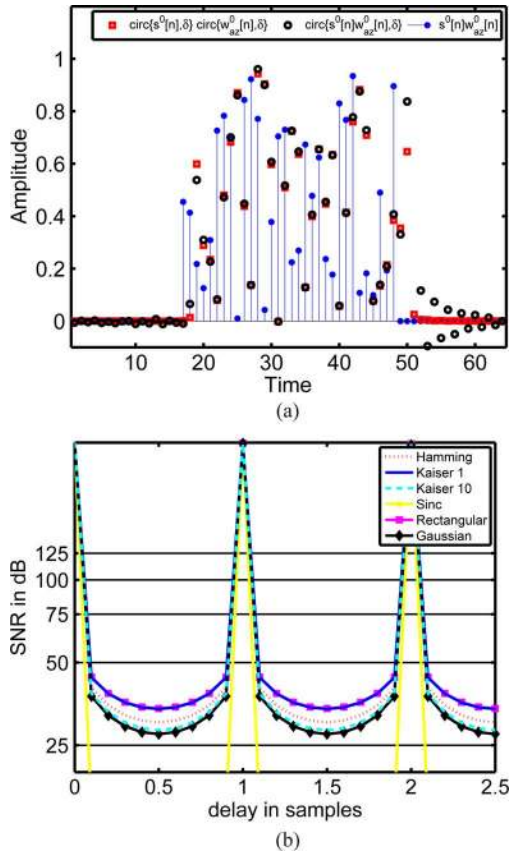

Fig. 5. Comparing fractional delay of a windowed signal: A rectified Gaussian distributed signal $s^{0}[n]$ and rectangular window $w_{az}^{0}[n]$ ($N/2$ samples long and zero-padded either-side by $N/4$) is used. Dot-stems denote $s^{0}[n]w_{az}^{0}[n]$. Estimates using the LHS (squares) and RHS (circles) of (26) for a 1.75 sample delay are shown, e.g., $\operatorname{circ}\{s^{0}[n],\delta\}\operatorname{circ}\{w_{az}^{0}[n],\delta\}$ and $\operatorname{circ}\{s^{0}[n]w_{az}^{0}[n],\delta\}$. Smooth windows reduce the difference. The SNR between the LHS and RHS of (28) demonstrates a dependence on the window for a sinc signal in Fig. 5(b). Kaiser 1 (1 denotes curvature) is best; the sinc window is the worst. (a) sSTFT error for fractional delays. (b) sSTFT dependence on the window.

F. Limitations of Synchronized Time-Frequency Windows

The error in the approximation in (28), measured by the SNR between the LHS and RHS, is illustrated for a range of window functions (flat to bell-shaped to pulse-like) for subsample and integer sample delays in Fig. 5(b).

$$\begin{aligned}\sum_k\operatorname{sinc}(kT)\operatorname{sinc}(nT-\delta-kT)\cdot\sum_k w_{az}^{0}[k]\operatorname{sinc}(nT-\delta-kT)\\ \approx\sum_k\operatorname{sinc}(kT)w_{az}^{0}[k]\operatorname{sinc}(nT-\delta-kT).\end{aligned}\tag{28}$$

A sinc function is used as an analysis window function $w_{az}^{\delta}[n]$ as an example of inappropriate window dilation. Naturally, the SNR deteriorates the fastest for the sinc window for subsample delay. The approximately flat Kaiser window, for $\beta=1$, and the Rectangular window give the best performance for subsample delay for this particular signal. The more bell-shaped the window becomes, e.g., the Hamming, Kaiser with $\beta=10$ and Gaussian window, the worse the approximation becomes for subsample delay. Fig. 5(b) gives an intuition of the effect of 1) truncation before or truncation after delaying the signal and 2) weighted truncation before or after delaying the signal. For this particular

signal, truncation without weighting is best. For subsample delay the sSTFT is window and signal dependent. Integer delay effectively gives equality in (28), up to machine error. Subsample delay gives a slightly poorer approximation. The performance using the cSTFT is not plotted as its deterioration as a function of delay would necessitate a significantly larger dynamic range in SNR in Fig. 5(b). The variation in performance supports the hypothesis that for subsample shifts the sSTFT is window dependent. Although the subsample performance is a function of both the window and the signal, the dynamic range of the SNR is small for reasonable window functions, e.g., bell-shaped windows. Note, the Gaussian window performs the worst for subsample delay. We will book-end this discussion which began in Section V-C, with a complementary discussion in Section VI-C. To conclude, signal delay in the time-frequency domain is exact for integer sample delay. For subsample delay, the approximation is sufficiently good for signals of practical interest, due to the mitigation of a suitable taper.

$\hat{\mathrm{P}}2$) **Relative Attenuation**

$\hat{\mathrm{P}}2\mathrm{a}$) $|\mathrm{sSTFT}\{s^{\delta}[n],\delta\}| \equiv |\mathrm{sSTFT}\{s^{0}[n],0\}|$ for $\delta/T < N/4$ and $\delta/T \in \mathbb{Z}$.

Proposition 4: The sSTFT preserves the relative attenuation between the received signals, $s^0[n]$ and $s^\delta[n]$, at two spatially displaced sensors, up to a relative delay of Δ_s samples, when $\delta/T \in \mathbb{Z}$.

Property $\hat{\mathrm{P}}2$ requires that the relative delay property $\hat{\mathrm{P}}1\mathrm{b}$ is satisfied. Given that the window-support of the time-frequency transform is not wrapped around with relative delay, as specified by Lemma 2, (27) shows that delaying a signal by an integer number of samples, by first windowing it and subsequently delaying the window-signal product, is equivalent to delaying the window and delaying the signal separately and then taking the product. If these two paradigms lead to the same signal, then the magnitude of the signals delivered by both methods are equivalent. The magnitude is not necessarily preserved when the cSTFT is applied as the window scales different parts of the delayed signal.

In the case of subsample delay we propose

$\hat{\mathrm{P}}2\mathrm{b}$)
$|\mathrm{sSTFT}\{s^{\delta}[n],\delta\}| - |\mathrm{sSTFT}\{s^{0}[n],0\}| = \epsilon(s[n], w[n], \delta)$ for $\delta/T \in \mathbb{Q} \setminus \mathbb{Z}$ where $\epsilon(\cdot)$ is a small error which is a function of the signal, the window and the delay.

Taking the difference between the absolute value of the DFT of both sides of (26) gives an expression for the error in the relative scaling between both observations of the signal due to subsample delay. Empirical trials are undertaken in [30] to determine the best window one should use for speech. Furthermore, Section VI-B shows that the window that best minimizes some function of the error in $\hat{\mathrm{P}}2\mathrm{b}$ is not necessarily the same window that minimizes the corresponding error for relative delay.

WDO is a crucial assumption in multichannel anechoic mixing source separation algorithms, however, the cSTFT reduces the measure of WDO of two sources observed at different physical locations. The authors of [1] do not measure and compare the quality of the WDO approximation on both channels, as a function of relative delay. A first empirical evaluation of this distortion in the special case of the DUET algorithm is demonstrated in [30] for speech where this property is a necessity for good separation performance and empirical evaluation is more meaningful.

$\hat{\mathrm{P}}3$) **Delay Invariant Relative WDO**

$\Lambda_j^\delta \bigcap \Lambda_k^\delta = \emptyset$ for $j \neq k$ where Λ_j^δ is the support of $S_j^\delta[k,m]$, i.e. $\Lambda_j^\delta = \mathrm{supp}\ S_j^\delta[k,m] = \{[k,m] : S_j^\delta[k,m] \neq 0\}$, moreover, $\Lambda_j^\delta \equiv \Lambda_j^0\ \forall \delta/T \in \mathbb{Z}$ and $\delta/T < N/4$.

Proposition 5: The images of different sources (from the class of signals of interest) under the synchronized time-frequency transform, sSTFT, have disjoint support.

Although the WDO property assumed by DUET is only approximately true for real signals, such as speech, it is typically better realized in the synchronized time-frequency domain when the analysis is synchronized to the sources. The redefinition of WDO (p2) as DIRWDO accounts for the relative delay between the channels. The sSTFT window removes relative window misalignment effects from the images of the signals whereas the cSTFT introduces new components to the signal observed at some displaced sensor relative to the reference and may increase the likelihood of erroneous time-frequency bins being activated, and worse, overlap. As DUET uses multiple channels to obtain the masks Λ_j, the WDO assumption should acknowledge this multichannel dependence. DIRWDO naturally extends p2 to the multichannel case and is more suited to multichannel scenarios. Consequently, the relationship $\Lambda_j^\delta \equiv \Lambda_j^0$ typically deteriorates as a function of relative delay when the cSTFT is used. The superscript on the mask Λ_j^δ indicates the set of time-frequency bins where the source $s_j^\delta[n]$ is dominant. The source $s^\delta[n]$ is delayed by δ seconds relative to the same source $s_j^0[n]$ observed at another physical location. The mask Λ_j^0 is associated with $s^0[n]$ and should be the same set of time-frequency bins. For notational completeness, $\Lambda_j^{c,\delta} = \mathrm{supp}\ S_j^{c,\delta}[k,m] = \{[k,m] : S_j^{c,\delta}[k,m] \neq 0\}$ denotes the set of time-frequency bins which comprise the mask for the j^{th} source delayed by δ using the cSTFT linear transform. In summary, DIRWDO $(\hat{P}3)$ is an approximation for both the cSTFT and sSTFT when $|\delta/T| < 1$. Nevertheless, the sSTFT gives better approximation accuracy outside of this range whereas the cSTFT deteriorates badly as a function of relative delay. DIRWDO is better realized when the data is linearly transformed using the sSTFT because the mask for signal separation is obtained using the same portion of the signal from multichannel scenarios. Moreover, DIRWDO specifies that the linear transform used should allow either observation to be used to demix the sources; choosing the inappropriate cSTFT transformed channel may reduce the level of separation possible.

Returning to the theme of the relationship between the sparse constraint and the WDO constraint, we infer from $\hat{\mathrm{P}}3$ that using the appropriate sSTFT on each channel will also preserve the signal sparsity across the channels. Consider the scenario of a signal which comprises of a pure tone, observed on multiple spatially displaced sensors. An interfering, yet disjoint signal consists of a more broadband signal. If the windowing is globally positioned, e.g., in a cSTFT fashion, smearing could reduce the level of the disjoint support of the two signals at the different sensors. The appropriate synchronization preserves the joint or relative sparsity across the observations in this scenario.

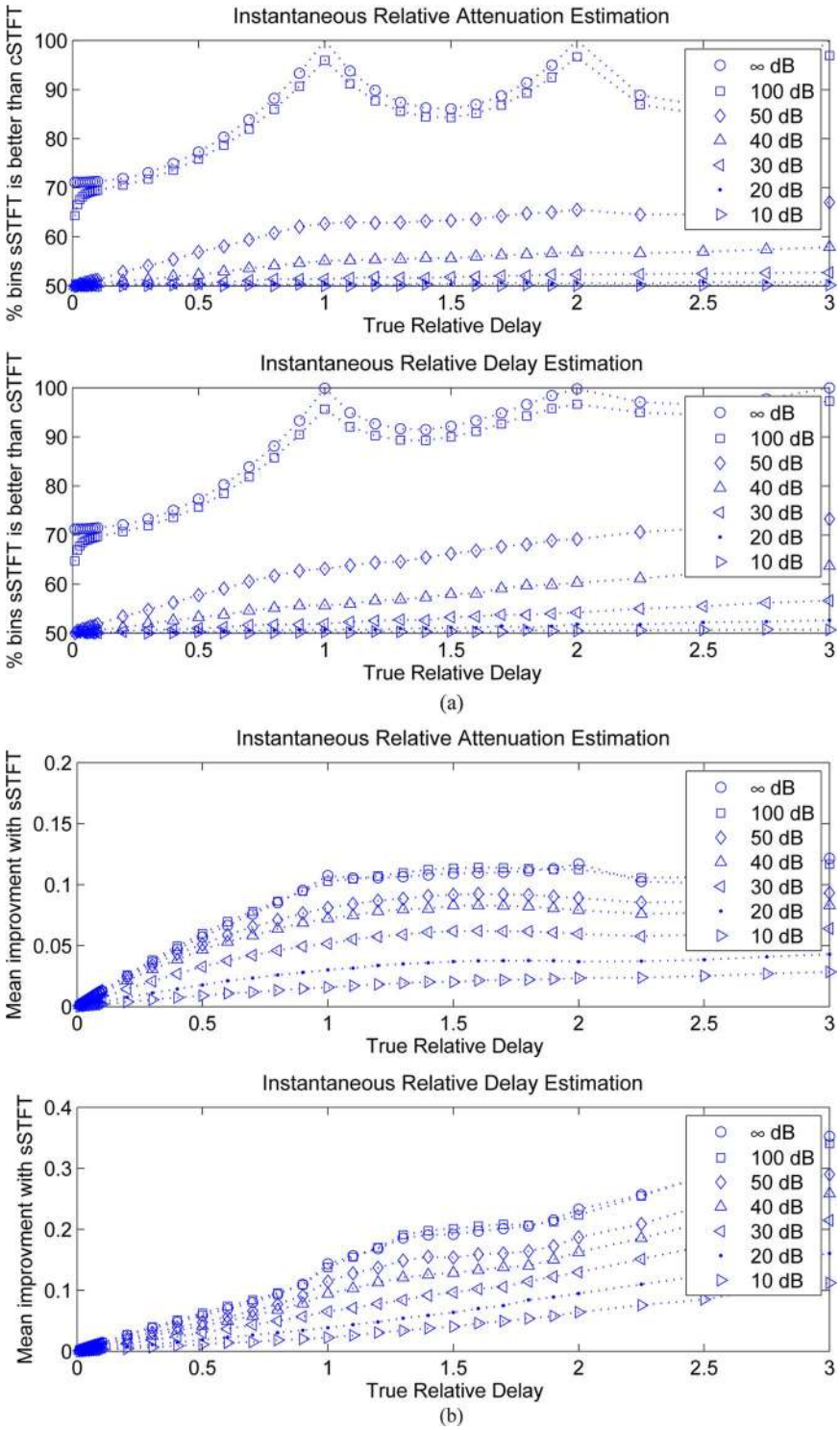

Fig. 6. Improvements in instantaneous parameter estimation via time-frequency synchronization vs. signal delay. (a) % bins the sSTFT improves Rel. Parameter Est. (b) Mean improvement.

VI. Window Comparison

We demonstrate the effect of linear transform synchronization by comparing bin-wise parameter estimation for the sSTFT and cSTFT. We then show that for subsample delay the sSTFT exhibits dependence on the window function used.

A. Better Bin-Wise Parameter Estimation via the sSTFT

We demonstrate the percentage of time-frequency bins that give improved instantaneous relative parameter estimates when the sSTFT is used. Speech from the TIMIT database, sampled at 16 kHz and analyzed with the cSTFT and sSTFT with a Hamming window of length $N = 2^{10}$ samples, is observed at two sensors. Additive white Gaussian noise is mixed on each channel consecutively with SNRs of $\{\infty$, 100, 50, 40, 30, 20, 10$\}$ dB. The target source is consecutively relatively delayed in steps from 0.01-to-3 samples. Prior synchronization knowledge is assumed for the sSTFT. Instantaneous relative attenuation and delay estimation is performed in each time-frequency bin for

the cSTFT and sSTFT. Fig. 6(a) demonstrates the percentage of bins the sSTFT gives better parameter estimates than the cSTFT. Because phase-wrap-around bins are discarded, the percentage increases for the estimated relative delay as the true delay increases. The percentage is greater than 50% for all noise levels. Even though we use rudimentary estimators—instantaneous estimates are sensitive to noise—that do not exploit speech source dominance in a few of the time-frequency bins, nor denoising, the sSTFT improves parameter estimation. Moreover, Fig. 6(b) illustrates the mean improvement achieved by the sSTFT in comparison with the cSTFT estimates. This improvement, defined as the mean of the difference between the absolute error for both classical and synchronized estimators increases as a function of relative integer delay for all SNR levels (but degrades slightly for subsample delay). In short, the sSTFT improves the instantaneous relative parameter estimates which can then be used in weighted estimators which are more robust to noise. Noise is ubiquitous in real applications, however, similar to source separation, leveraging the sparsity of speech, particularly its dominance in the formant frequencies could yield further improvement.

B. Illustrating sSTFT Subsample Delay Window Dependence

Due to the inaccuracy of the sSTFT assumption

$$\operatorname{circ}\left\{s^0[n]w^0_{az}[n-mR],\delta\right\} = \operatorname{circ}\left\{s^0[n],\delta\right\}\operatorname{circ}\left\{w^0_{az}[n-mR]\right\}$$

when $|\delta/T| < 1$, we illustrate the dependence of the sSTFT on the curvature of the window for subsample delay. We empirically investigate properties $\hat{\mathrm{P}}1\mathrm{b}$ and $\hat{\mathrm{P}}2$ and determine the window that yields the smallest estimation error from a small set of candidate windows. The degree to which improvement in subsample parameter estimation is possible based on window choice is an open problem, however. We use two observations of a white Gaussian noise (WGN) signal measured at spatially displaced sensors so that all frequency bins contribute equally to the estimate. The signal is 4096 samples long, and experiences consecutive intersensor delays. We analyze both observations using the appropriately synchronized sSTFT, with an FFT of 4096 bins, and analysis window $w^{\delta}_{az}[n]$ of length 4096 samples. We vary the Kaiser window curvature used with (14) to construct $w^{\delta}_{az}[n]$. We also vary the relative delay for a given curvature. We perform 100 Monte Carlo experiments for each $\{d,\beta\}$ and average the results.

On average, for a WGN fractionally delayed signal, relative attenuation and relative delay are best estimated using a slightly curved window in Fig. 7. Fig. 7(a) and (b) shows that the best on average approximation of the relative attenuation and delay is given when $\beta \approx 5$. The variance is lowest for both estimates for $\beta \approx 5$. The metric of comparison for each set $\{d,\beta\}$ is the mean instantaneous relative attenuation and relative delay over each frame. The ideal relative attenuation and delay are 1 and 0.6 samples, respectively.

In Fig. 7(c) and (d) we demonstrate that performance deteriorates for a given curvature as a function of relative delay. For the particular cases of relative delay of 0.2 and 0.5 samples, the

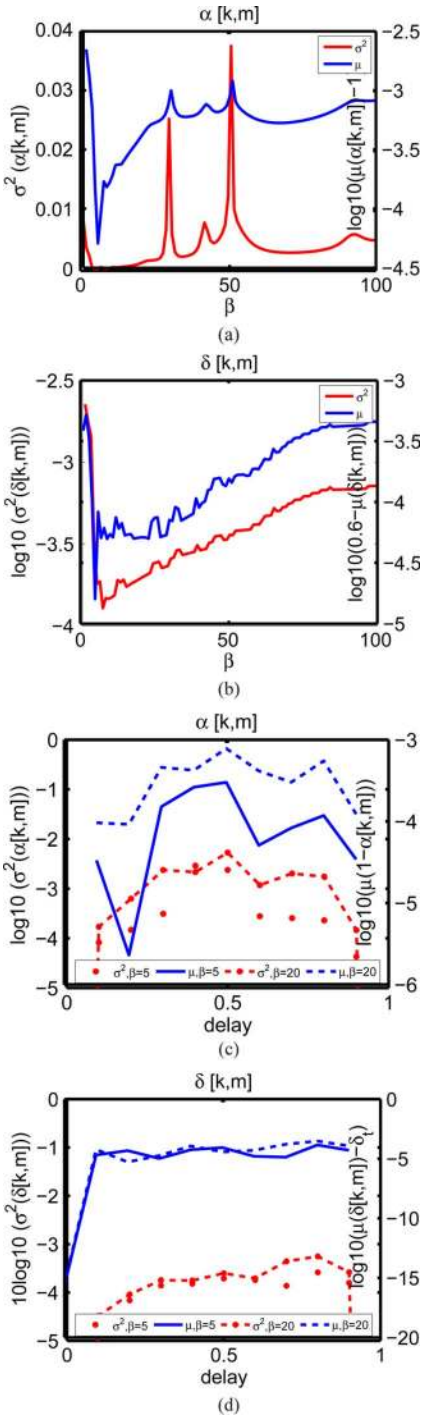

Fig. 7. (a) and (b): Mean error and variance $(\mu(\cdot),\sigma^2(\cdot))$ of the relative attenuation, $\alpha[k,m]$, and delay, $\delta[k,m]$, estimates versus window curvature (β) given the true parameters $(\{\alpha_t,\delta_t\}=\{1,0.6\}$ samples). The dynamic range of the error for a good (in the mean $\beta=5$ is best) and bad $(\beta=20)$ window curvature is small. (c) and (d): Estimated mean error in $\alpha[k,m]$ and the difference between the $\delta[k,m]$ and δ_t versus signal delay δ_t. The best window for $\alpha[k,m]$ and $\delta[k,m]$ for $\{\alpha_t,\delta_t\}=\{1,0.2\}$ and $\{\alpha_t,\delta_t\}=\{1,0.5\}$ is different. $\beta=5$ is best for $\alpha[k,m]$ and $\beta=20$ is best for $\delta[k,m]$. (a) Rel. Atten. versus window curvature β. (b) Rel. Del. versus window curvature β. (c) Rel. Atten. versus delay for $\beta=5$ and $\beta=20$. (d) Rel. Del. versus Delay for $\beta=5$ and $\beta=20$.

window curvature giving the least instantaneous relative attenuation error is $\beta=5$ whereas the window curvature giving the least instantaneous relative delay estimation is $\beta=20$. In summary, the accuracy of subsample relative parameter estimates is

dependent on the structure of the signal, the relative delay and the curvature of the window. The variance of the instantaneous estimates gives an indication of the improvement possible although this improvement is small compared to the that gained by synchronizing the linear transform. Nevertheless, the deviation from the true value is apparent for both relative attenuation and relative delay estimation for fractional delay.

C. Discussion: Translatable and Dilatable Windows

Regarding window selection for the sSTFT, we propose real even, element-wise nonnegative, even length windows constructed using (14) in this paper. The DFT of an even length window $w_a[n] = w_a[-n] \in \mathbb{R}^N$ has a real spectrum times a linear phase term. A half sample shift as well as a $N/2$ sample shift is needed to center the window on zero due to the definition of the window indices, e.g., $n = 0, \ldots N-1$. Defining the analysis window using (13) means there is a discontinuity when the window transitions from the window-support region to the zero-padding on either side at indices $\{N/4-1, 3N/4-1\}$. We have considered analysis windows of the form $f(x) \longrightarrow f((x-\mu)/\sigma)$ that smoothly capture the spirit of the (14) but without discontinuities. A suitable choice of the parameters μ and σ translates and dilates the window so that it approximates $w^0_{az}[n]$ and $w^\delta_{az}[n]$. The resultant function goes to a small value in the appropriate region and goes to zero in the limit. For example, a Gaussian window maybe parameterized so that its structure is similar to $w^0_{az}[n]$ generated using a Hamming window (13). However, there is an inherent tradeoff between the linear phase criterion and the Gibbs effect due to discontinuities. Translated and dilated Gaussian windows are no longer even and symmetric due to truncation. Moreover, the ratio of the reference and delayed window is typically numerically unstable as the Fourier transform of the Gaussian's standard deviation is inversely proportional to the standard deviation in the time domain.

Given prior knowledge of the true relative delay, one might consider what the effect of upsampling the test signal in the experiments above would be, such that fractional delay becomes integer delay before performing parameter estimation. This would remove the subsample dependence described above. We have evaluated the mean processing error introduced by interpolation and concluded that the error—coloring of the signal—introduced by interpolation is large irrespective of the delay [30]. Thus, for the WGN signal above we rely on the raw data as interpolation preprocessing degrades the signal.

Regarding phase-wrap-around and the applicability of the sSTFT, the maximum relative delay in [1] is $|d| \leq 5$ samples. In practice separation is successful for $|d| \leq 2.5$. However, tiled DUET [31] is robust for large relative delays $|d| \leq 170$ samples. As the focus of this paper is the introduction of the sSTFT and not specific source separation algorithms our experiments consider instantaneous relative delays which are $|d| \leq 1$ sample. Naturally, the larger the permissable relative delay without phase-wrap-around, the greater the potential benefits of time-frequency synchronization [4], [16], [17].

In summary, time-frequency domain multichannel anechoic mixing algorithms, typically rely on the existence of an invertible transform which transforms the signals into a domain where they are sparse. Assumptions pertaining to the mixing parameters are typically not satisfied when the cSTFT is used. Implicit in all four of the properties proposed in [1] is the notion that a global shift (or delay) and attenuation of one source signal observed at a sensor x_2 relative to that same signal observed at another sensor, x_1, can be estimated from windowed segments of both observations. In this paper we explain that an error is introduced to the attenuation and delay between multichannel observations unless a synchronized linear transform is used to transform the signal. A general framework for blind sSTFT synchronization is proposed in [9] which makes this result applicable to related array processing techniques, for example, [3], [5], and [7].

VII. Conclusion

In this paper we identified the source of the estimation error in the original DUET paper [1]. This error was due to a misynchronization of the time-frequency analysis and as a result we proposed a new approach that synchronized the transform locally to the signal and not to a global clock, namely the sSTFT. We then introduced a series of properties that this transform has which we believe would be of interest to a wider audience in sparse and multichannel signal processing. We evaluated its application in the case of subsample delays and we concluded that depending on the level of accuracy required when the delay was subsample, that approximate equality could be assumed in the properties depending on the window used. We demonstrated that the sSTFT improves relative parameter estimation in many of the time-frequency bins. What was not discussed in this paper was how one would practically synchronize the window to the signal. This is the main topic of a companion paper [9] where we demonstrate that it is possible to learn the synchronization blindly and we propose a new variant of the DUET algorithm which offers the potential for improvement in the estimation step.

Acknowledgment

The authors thank the reviewers for their detailed and thoughtful comments. The authors would also like to acknowledge and thank Dr. K. Drakakis, Dr. G. McDarby, and Prof. B. A. Pearlmutter.

References

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

[2] G. Carter, "Bias in magnitude-squared coherence estimation due to misalignment," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 1, pp. 97–99, Feb. 1980.

[3] R. Roy and T. Kailath, "ESPRIT—Estimation of signal parameters via rotational invariance techniques," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 7, pp. 984–995, Jul. 1989.

[4] R. de Fréin, S. Rickard, and B. Pearlmutter, "Constructing time-frequency dictionaries for source separation via time-frequency masking and source localisation," in *Independent Component Analysis and Signal Separation*, ser. Lecture Notes in Comput. Sci., T. Adali, C. Jutten, J. Romano, and A. Barros, Eds. Berlin/Heidelberg: Springer, 2009, vol. 5441, pp. 573–580.

[5] H. Sawada, S. Araki, and S. Makino, "A two-stage frequency-domain blind source separation method for underdeterminded convolutive mixtures," in *Proc. IEEE Workshop on Appl. Signal Process. to Aud. and Acoust.*, Oct. 2007, pp. 139–142.

[6] M. I. Mandel, R. J. Weiss, and D. P. W. Ellis, "Model-based expectation maximization source separation and localization," *IEEE Audio, Speech, Lang. Process.*, Feb. 2010.
[7] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[8] F. Abrard and Y. Deville, "A time-frequency blind signal separation method applicable to underdetermined mixtures of dependent sources," *Signal Process.*, vol. 85, no. 7, pp. 1389–1403, Jul. 2005.
[9] R. de Fréin and S. T. Rickard, "The synchronized STFT: iDUET," *IEEE Trans. Signal Process.*, 2010, to be submitted.
[10] B. Porat, *A Course in Digital Signal Processing*. New York: Wiley, 1996.
[11] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 1989, 07458.
[12] A. Jourjine, S. Rickard, and O. Yilmaz, "Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Jun. 2000, vol. 5, pp. 2985–2988.
[13] S. Rickard, "Sparse sources are separated sources," in *Proc. EUSIPCO'06*, Sep. 2006.
[14] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 24, no. 4, pp. 320–327, Aug. 1976.
[15] J. Benesty, J. Chen, and Y. Huang, "Time-delay estimation via linear interpolation and cross correlation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 509–519, Sep. 2004.
[16] D. Model and M. Zibulevsky, "Signal reconstruction in sensor arrays using sparse representations," *Signal Process.*, vol. 86, no. 3, pp. 624–638, 2006.
[17] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.
[18] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 2, pp. 109–115, Mar. 2003.
[19] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 530–538, Sep. 2004.
[20] B. A. Pearlmutter and R. K. Olsson, "Linear program differentiation for single-channel speech separation," in *Proc. IEEE Int. Workshop on Mach. Learn. Signal Process. (MLSP'06)*, Sep. 2006.
[21] P. D. O'Grady and B. A. Pearlmutter, "The LOST algorithm: Finding lines and separating speech mixtures," *EURASIP J. Adv. in Signal Process.*, 2008.
[22] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, no. 5, pp. 1457–1469, Nov. 2004.
[23] D. Donoho and V. Stodden, "When does non-negative matrix factorization give correct decomposition into parts?," in *Proc. 17th Ann. Conf. Neur. Inf. Process. Syst. (NIPS)*, 2003, MIT Press.
[24] R. Balan and J. Rosca, "Statistical properties of STFT ratios for two channel systems and applications to blind source separation," Independent Component Analysis Helsinki, 2000, pp. 429–434.
[25] T. I. Laakso, V. Välimäki, M. Karjalainen, and U. K. Laine, "Splitting the unit delay—Tools for fractional dealy filter design," *IEEE Signal Process. Mag.*, vol. 13, no. 1, pp. 30–60, Jan. 1996.
[26] D. Gabor, "Theory of communications," *J. Inst. Elect. Eng.*, vol. 93, no. 3, pp. 429–457, 1946.
[27] M. Bastiaans, "Gabor's signal expansion and the Zak transform," *Appl. Opt.*, vol. 33, no. 23, pp. 5421–5455, 1994.
[28] F. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proc. IEEE*, vol. 66, no. 1, pp. 51–83, Jan. 1978.
[29] A. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Trans. Acoust., Speech Signal Process.*, vol. 29, no. 1, pp. 84–91, Feb. 1981.
[30] R. de Fréin, "Adapting bases using the synchronized short-time-Fourier-transform and non-negative matrix factorization," Ph.D. dissertation, Univ. College Dublin, Dublin, 2010.
[31] S. Rickard, *Blind Speech Separation*. New York: Springer, 2007, ch. The DUET Blind Source Separation Algorithm, pp. 217–237.

Ruairí de Frein (S'04) received the B.Eng. degree in electronic engineering from University College Dublin (UCD) in 2004. He received the Ph.D. degree in electronic engineering and is affiliated with the Complex and Adaptive Systems Laboratory in UCD.

His research interests include time-frequency analysis applied to signal processing, blind source separation, network management, and distributed signal processing.

Scott T. Rickard (SM'03) received the S.B. degree in mathematics in 1992, the S.B. degree in computer science and engineering in 1993, and the S.M. degree in electrical engineering and computer science, also in 1993, all from the Massachusetts Institute of Technology (MIT), Cambridge. He received the M.A. and Ph.D. degrees in applied and computational mathematics from Princeton University, Princeton, NJ, in 2000 and 2003, respectively.

From 1991 to 1993, he was a Research Assistant with the Charles Stark Draper Laboratory, MIT, and worked on a prototype analog neural network computer, designed neural networks for mine detection from sonar images, and designed large sets of frequency-hopped waveforms with nearly ideal ambiguity properties for sonar applications. From 1993 to 2003, he was a Member of Technical Staff at Siemens Corporate Research, Princeton. From 1995 to 1996, he was with the Neural Networks Group, Siemens, Munich, Germany. While with Siemens, he developed and applied machine learning technology to industrial problems such as vehicle navigation, automated image analysis, biomedical signal classification, and industrial plant state prediction. He is currently the Director of the Complex and Adaptive Systems Laboratory at University College Dublin, Ireland. His research for the past several years has focused on the application of time-frequency methods and sparse signal processing for the blind separation of more sources than sensors. His research interests include time-frequency/scale analysis applied to signal processing, wireless communications, blind source separation, multicarrier communication systems, and Costas arrays.