Blind Separation of Speech Mixtures via Time-Frequency Masking

Özgür Yılmaz and Scott Rickard

***Abstract*—Binary time-frequency masks are a powerful tool for the separation of sources from a single mixture. Perfect demixing via binary time-frequency masks is possible provided the time-frequency representations of the sources do not overlap, a condition we call *W-disjoint orthogonality*. We introduce here the concept of *approximate W-disjoint orthogonality* and present experimental results demonstrating the level of approximate W-disjoint orthogonality of speech in mixtures of various orders. The results demonstrate that ideal binary time-frequency masks exist which can separate several speech signals from one mixture. While determining these masks blindly from just one mixture is an open problem, we show that we can approximate the ideal masks in the case where just two anechoic mixtures are provided. Motivated by the maximum likelihood mixing parameter estimators, we define a power weighted two-dimensional histogram constructed from the ratio of the time-frequency representations of the mixtures which is shown to have one peak for each source with peak location corresponding to the relative attenuation and delay mixing parameters. The histogram is used to create time-frequency masks which partition one mixture into the original sources. Experimental results on speech mixtures verify the technique. Example demixing results can be found online: http://www.princeton.edu/~srickard/bss.html**

I. Introduction

The goal in blind source separation is to determine the original sources given mixtures of those sources. When the number of sources is greater than the number of mixtures, the problem is degenerate in that traditional matrix inversion demixing cannot be applied. However, when a representation of the sources exists such that the sources have disjoint support in that representation, it is possible to partition the support of the mixtures and obtain the original sources. One solution to the problem of degenerate demixing is thus to (1) determine an appropriate disjoint representation of the sources and (2) determine the partitions in this representation which demix. In this paper, we show that the Gabor expansion (i.e., the discrete short-time (or windowed) Fourier transform) is a good representation for demixing speech mixtures. Specifically, we show that partitions of the time-frequency lattice exist that can demix mixtures of up to ten speech signals from one mixture. Determining the partition blindly from one mixture is an open problem, but, given a second mixture, we describe a method for partitioning the time-frequency lattice which separates the sources.

Özgür Yılmaz (oyilmaz@math.umd.edu) is with the Department of Mathematics, University of Maryland.

Scott Rickard (srickard@princeton.edu) is with the Program in Applied and Computational Mathematics, Princeton University and is also with Siemens Corporate Research in Princeton, NJ.

Submitted to IEEE Transactions on Signal Processing, November 4, 2002.

Formally, let $\mathcal{S}$ be the family of signals of interest. Typically $\mathcal{S}$ will be some collection of square integrable bandlimited functions. Suppose there exists some linear transformation $T : s_j \in \mathcal{S} \mapsto S_j$ (where T maps the set $\mathcal{S}$ to another family of functions) with the following properties:

(i) T is invertible on $\mathcal{S}$ (i.e., $T^{-1}(Ts) = T(T^{-1}s) = s, \forall s \in \mathcal{S}$).

(ii) $\Lambda_j \cap \Lambda_k = \emptyset$ for $j \neq k$, where Λ_j is the support of S_j, i.e., $\Lambda_j = \text{supp } S_j := \{\lambda : S_j(\lambda) \neq 0\}$.

For example, we can consider the case where $\mathcal{S}$ is a collection of square integrable functions with mutually disjoint supports in the Fourier domain; any two functions s_1 and s_2 in $\mathcal{S}$ satisfy $\hat{s}_1(\omega)\hat{s}_2(\omega) = 0$ for all ω, where $\hat{s}_j$ denotes the Fourier transform of s_j. Then if we define T on $\mathcal{S}$ as $Ts := \hat{s}$, it is clear that T satisfies (i) and (ii).

For any T with properties (i) and (ii), we can demix a mixture x_1 of signals in $\mathcal{S}$, $x_1(t) = \sum_{j=1}^{N} s_j(t)$, via

$$s_j = T^{-1}(1_{\Lambda_j} T x_1) \tag{1}$$

where 1_Λ is the indicator function of the set Λ. Going back to our example above, this corresponds to

$$s_j = (1_{\Lambda_j} \hat{x}_1)\check{} \tag{2}$$

which is certainly true since the functions in $\mathcal{S}$ satisfy (ii). Here $(1_{\Lambda_j}\hat{x}_1)\check{}$ denotes the inverse Fourier transform of $1_{\Lambda_j}\hat{x}_1$.

Suppose now that we have another mixture $x_2(t) = \sum_{j=1}^{N} a_j s_j(t - \delta_j)$, which is the case in anechoic environments when we have two microphones. In the mixing, a_j and δ_j are the attenuation and delay parameters respectively corresponding to the j^{th} source. Assume

(iii) supp $Ts(\cdot - \delta) =$ supp Ts for any $s \in \mathcal{S}$, $\forall |\delta| < \Delta$, and

(iv) there exist functions F and G such that $a_j = F(Tx_1(\lambda), Tx_2(\lambda))$ and $\delta_j = G(Tx_1(\lambda), Tx_2(\lambda))$ for $\lambda \in \Lambda_j$ for $j = 1, \ldots, N$,

where Δ is the maximum possible delay between mixtures due to the distance separating the sensors. Using (iii) and (iv), we can label each $\lambda \in$ supp Tx_1 with the pair $(F(Tx_1(\lambda), Tx_2(\lambda)), G(Tx_1(\lambda), Tx_2(\lambda)))$, and Λ_j is exactly the set of all points with the label (a_j, δ_j). It follows that given the mixtures $x_1(t)$ and $x_2(t)$, we can demix via

$$s_j = T^{-1}(1_{\Lambda_j} T x_1). \tag{3}$$

Clearly, (iii) will be satisfied for the example above since the Fourier transform of $s(\cdot - \delta)$ will be just a modulated version of the Fourier transform of s and thus it will have the same support as s. As to the existence of functions F and G, one can show that $F(\hat{x}_1(\omega), \hat{x}_2(\omega)) = |\hat{x}_2(\omega)/\hat{x}_1(\omega)|$ and $G(\hat{x}_1(\omega), \hat{x}_2(\omega)) = -1/\omega \sphericalangle \hat{x}_2(\omega)/\hat{x}_1(\omega)$ where $\sphericalangle z$ denotes the phase of the complex number z taken between $-\pi$ and π, satisfies (iv). The above described F and G are the DUET attenuation and delay estimators for the special case where the windowing function $W \equiv 1$. The DUET estimators are discussed in Section III-A.

The general algorithm explained above mainly depends on two major points: (a) the existence of an invertible transformation T that transforms the signals to a domain on which they have disjoint representations (properties (i), (ii), and (iii)), and (b) finding functions F and G that provide the means of labeling on the transform domain (property (iv)). Note that in the description above we required F and G to yield the exact mixing parameters. Although this is desired since the mixing parameters provide the perfect labels, and they can also be used for various other purposes (e.g., direction-of-arrival determination), it is not necessary for the demixing algorithm to work. Some function that provides a unique labeling on the transform domain is sufficient. Moreover, requirement (ii) that the transformation T is "disjoint" is very strong. In practice, one is usually more interested in transforms that satisfy (ii) in some approximate sense. Therefore, we are interested in transforms that result in sparse representations for the signals of interest.

There are many examples in the literature that do use this type of approach with various choices of T for various mixing models and demixing methods[1–9]. The mixing model in [1–3, 5, 7, 8] is "instantaneous" (sources have different amplifications in different mixtures) while [4, 6, 9] use an anechoic mixing model (sources have different amplifications and time delays in different mixtures). [1–3, 8] consider the time domain sampling operator as T. The general assumption in these is that at any given time at most one source is non-zero. [4–7, 9] use the short-time Fourier transform (STFT) operator as T. Condition (ii) is satisfied in this case, at least approximately, because of the sparsity of the time-frequency representations of speech signals. Empirical support for this can be found in [10], and a more extensive discussion is given in Section II-A. In principle, [1–9] all use some clustering algorithm for estimating the mixing parameters, although there are several different approaches to demixing. [1, 3, 4, 6–8] use a labeling scheme based on the estimated mixing parameters and thus demix in the above described way by creating binary masks in the transform domain corresponding to each source. That is, given the mixtures x_1 and x_2, demixing is done by grouping the clusters of points in (Tx_1, Tx_2) space, although different techniques are used to detect these clusters. For example, [4, 6, 7] demix essentially by constructing binary time-frequency masks that partition the time-frequency plane such that each partition corresponds to the time-frequency points that "belong" to a particular source. The fact that such a mask exists has been observed also in [11] in the context of BSS of speech signals from *one* mixture, and in [12] in the context of source localization. In [2, 8, 9], the demixing is done making additional assumptions on the statistical properties of the sources and using a maximum a posteriori (MAP) estimator. [5, 8] demix by assuming that the number of sources active in the transform domain at any given point is equal to the number of mixtures. They then demix by inverting the now non-degenerate M-by-M mixing matrices and appropriately combining the outputs. The above comparison is summarized in Figure 1. Alternative approaches to degenerate blind source separation include [13–15].

mixing model	T operator	demixing
instantaneous [1–3, 5, 7, 8]	sampling [1–3, 8]	masking [1, 3, 4, 6–8]
anechoic [4, 6, 9]	STFT [4–7, 9]	MAP [2, 8, 9]
		matrix masking [5, 8]

Fig. 1. A comparison of degenerate demixing methods using disjoint representations.

In this paper, we consider the short-time Fourier transform (STFT) and Gabor expansions (the discrete version of the STFT) of speech signals. We present extensive empirical evidence that speech signals indeed satisfy (ii) when T is the STFT with an appropriate window function. Based on this, we extend the DUET algorithm, originally presented in [4] for sources with disjointly supported STFTs, to anechoic mixtures of speech signals. The algorithm we propose relies on estimating the mixing parameters via maximum likelihood (ML) motivated estimators, and constructing binary time-frequency masks using these estimates. Thus the method presented here: (1) uses an anechoic mixing model, (2) uses the STFT as T, and (3) performs demixing via masking.

In Section II we introduce a way of measuring the degree of "approximate" W-disjoint orthogonality, WDO$_M$, of a signal in a given mixture for a given mask M. We construct a family of time-frequency masks, Φ^x, that correspond to the indicator functions of the time-frequency points in which one source dominates the others by x dB. We test the demixing performance of these masks experimentally and illustrate that WDO$_{\Phi^x}$ is indeed a good measure of the demixing performance of the masks Φ^x. The results show that binary time-frequency masks exist that are capable of demixing up to ten speech signals from just a single mixture.

In Section III we introduce a mixture model based on the results of Section II and demonstrate that given a second anechoic mixture, we can approximate these demixing masks blindly. To construct the masks, we first derive the maximum likelihood estimators for the delay and attenuation coefficients. We compare the performance of these with other estimators motivated by the maximum likelihood estimators. The modified delay and attenuation estimators are weighted averages of the instantaneous time-frequency delay and attenuation estimates. The delay and attenuation estimators can be combined and we show that a weighted two-dimensional histogram can be used to enumerate the sources, determine the mixing parameters, and demix the sources. The number of peaks in the histogram is the number of sources, the peak locations reveal the mixing parameters, and the mixing parameters can be used to partition the time-frequency representation of one of the mixtures to obtain estimates of the original sources.

In Section IV, we verify the method presenting demixing results for speech signals mixed synthetically and in both anechoic and echoic rooms.

II. W-Disjoint Orthogonality

In this section, we focus on showing that binary time-frequency masks exist which are capable of separating multiple speech signals from one mixture. Our goal is, given a mixture

$$x_1(t) = \sum_{j=1}^{N} s_j(t) \tag{4}$$

of sources $s_j(t)$, $j = 1, \ldots, N$, to recover the original sources. In order to accomplish this, we assume the sources are pairwise W-disjoint orthogonal.

We call two functions s_1 and s_2 **W-disjoint orthogonal** (W-DO) if, for a given a window function W, the supports of the windowed Fourier transforms of s_1 and s_2 are disjoint[4]. The windowed Fourier transform of s_j is defined

$$F^W(s_j(\cdot))(t,\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} W(\tau - t) s_j(\tau) e^{-i\omega\tau} d\tau \tag{5}$$

which we will refer to as $\hat{s}_j(t,\omega)$ where appropriate. For a detailed discussion of the properties of this transform consult [16]. The W-disjoint orthogonality assumption can be stated concisely

$$\hat{s}_1(t,\omega)\hat{s}_2(t,\omega) = 0, \forall t, \omega. \tag{6}$$

The two limiting cases for W, namely $W = 1$ and $W(t) = \delta(t)$, result in interesting sets of W-DO signals. In the $W = 1$ case, the t argument in (6) is irrelevant as the windowed Fourier transform is simply the Fourier transform. The condition is satisfied by signals which are frequency disjoint, such as frequency division multiplexed signals. In the other extreme, signals which are time disjoint such as time-division multiplexed signals satisfy the condition. In general, for window functions which are localized in time and frequency, the W-disjoint orthogonality condition is the goal of frequency-hopped multiple access systems. Indeed, the method presented here could be applied to time-domain multiplexed, frequency domain multiplexed, or frequency-hopped multiple access signals, however, in this paper we exclusively consider speech signals.

Unfortunately, (6) will not be satisfied for simultaneous speech signals as the time-frequency representation of active speech is rarely zero. However, speech is sparse in that a small percentage of the time-frequency coefficients in the Gabor expansion of speech capture a large percentage of the overall power. In other words, the magnitude of the time-frequency representation of speech is often small. The goal of this section is to show that speech signals satisfy a weakened version of (6) and are thus approximately W-DO. The higher the degree of approximate W-disjoint orthogonality, the better separation results are possible. Figure 2 illustrates that speech signals have sparse time-frequency representations and satisfy a weakened version of (6), in that the product of their time-frequency representations is almost always small.

A condition similar to (6) is considered in [17], the only difference being that the time-frequency transform used was the Wigner distribution. Signals satisfying (6) for the Wigner distribution were called “time-frequency disjoint.”

The approximate W-disjoint orthogonality of speech has been described as the “sparsity” and “disjointness” of the short-time Fourier transform of the sources[5], “when one source has large energy the other does not” and “harmonic components” which “hardly overlap”[6], “when a datapoint is large the most likely decomposition is to assume that it belongs to a single source”[9], “spectra [that] are non-overlapping”[11], and “useful” time-frequency points containing a “contribution of one speaker...significantly higher than the energy of the other speaker”[18]. A quantitative measure of approximate W-disjoint orthogonality is discussed later in this section.

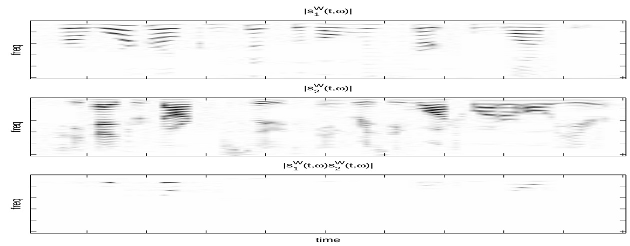


Fig. 2. A picture of W-disjoint orthogonality. The three figures are gray scale images of $|\hat{s}_1(t,\omega)|$, $|\hat{s}_2(t,\omega)|$, and $|\hat{s}_1(t,\omega)\hat{s}_2(t,\omega)|$ for two speech signals $s_1(t)$ and $s_2(t)$ normalized to have unit energy. A Hamming window of length 64 ms was used as $W(t)$ and all signals had length 3 seconds. $|\hat{s}_1(t,\omega)\hat{s}_2(t,\omega)|$ contains fewer large components than $|\hat{s}_1(t,\omega)|$ or $|\hat{s}_2(t,\omega)|$. Further analysis of these signals reveals that the time-frequency points that contain 90% of the energy of s_1 contain only 1.1% of the energy of s_2. Similarly, the time-frequency points that contain 90% of the energy of s_2 contain only 0.6% of the energy of s_1. Thus we claim that the speech signals approximately satisfy the W-disjoint orthogonality condition.

We can rewrite the model from (4) in the time-frequency domain

$$\hat{x}_1(t,\omega) = \hat{s}_1(t,\omega) + \ldots + \hat{s}_N(t,\omega). \tag{7}$$

Assuming the sources are pairwise W-DO, at most one of the N sources will be non-zero for a given (t,ω), and thus

$$\hat{x}_1(t,\omega) = \hat{s}_{J(t,\omega)}(t,\omega) \tag{8}$$

where $J(t,\omega)$ is the index of the source active at (t,ω).

To demix, one creates the time-frequency mask corresponding to each source and applies the each mask to the mixture to produce the original source time-frequency representations. For example, if $M_j := 1_{\{J(t,\omega)=j\}}$ is the indicator function for the support of source j, one obtains source j’s time-frequency representation from the mixture via

$$\hat{s}_j(t,\omega) = M_j(t,\omega)\hat{x}_1(t,\omega), \forall t, \omega. \tag{9}$$

A. Measuring the W-Disjoint Orthogonality of Speech

Clearly, the W-disjoint orthogonality assumption is not satisfied for our signals of interest. We introduce here a measure of approximate W-disjoint orthogonality based on the demixing performance of time-frequency masks created using knowledge of the instantaneous source and interference time-frequency powers of speech mixtures. Experiments on speech mixtures reveal that speech is approximately W-DO. In order to measure W-disjoint orthogonality for a given mask, we combine two important performance criteria: (1) how well the mask preserves the source of interest, and (2) how well the mask suppresses the interfering sources. These two criteria, the PSR and SIR, are introduced below.

First, given a time-frequency mask M such that $0 \leq M(t,\omega) \leq 1$ for all (t,ω), we define PSR_M, the preserved-signal-ratio of the mask M as

$$\mathrm{PSR}_M = \frac{\|M(t,\omega)\hat{s}_j(t,\omega)\|^2}{\|\hat{s}_j(t,\omega)\|^2} \quad (10)$$

which measures the percentage of energy of source j remaining after demixing using the mask. Note that $\mathrm{PSR}_M \leq 1$ with $\mathrm{PSR}_M = 1$ only if supp $M_j \subseteq$ supp M.

Now, we define

$$y_j(t) = \sum_{\substack{k=1 \\ j \neq k}}^{N} s_k(t) \quad (11)$$

so that $y_j(t)$ is the summation of the sources interfering with source j. Then, we define the signal-to-interference ratio of time-frequency mask $M(t,\omega)$

$$\mathrm{SIR}_M = \frac{\|M(t,\omega)\hat{s}_j(t,\omega)\|^2}{\|M(t,\omega)\hat{y}_j(t,\omega)\|^2} \quad (12)$$

which is the output signal-to-interference ratio after using the mask to demix.

We now combine the PSR_M and SIR_M into one measure of approximate W-disjoint orthogonality. We propose the normalized difference between the signal energy maintained in masking and the interference energy maintained in masking as a measure of W-disjoint orthogonality:

$$\begin{aligned} \mathrm{WDO}_M &= \frac{\|M(t,\omega)\hat{s}_j(t,\omega)\|^2 - \|M(t,\omega)\hat{y}_j(t,\omega)\|^2}{\|\hat{s}_j(t,\omega)\|^2} && (13) \\ &= \mathrm{PSR}_M - \mathrm{PSR}_M/\mathrm{SIR}_M. && (14) \end{aligned}$$

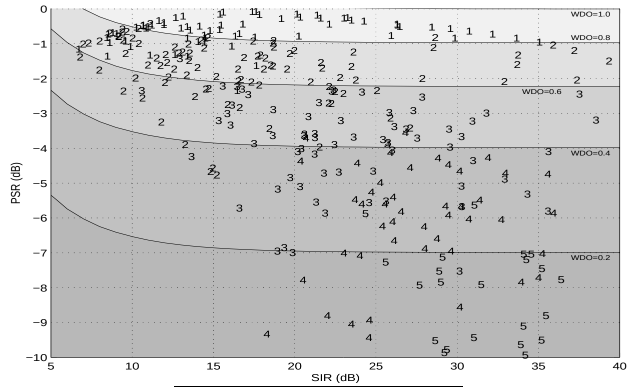
Using the mask $M(t,\omega) = 1_{\{J(t,\omega)=j\}}$, for signals which are W-DO we note that $\mathrm{PSR}_M = 1$, $\mathrm{SIR}_M = \infty$, and $\mathrm{WDO}_M = 1$. Moreover, $\mathrm{WDO}_M = 1$ implies that $\mathrm{PSR}_M = 1$, $\mathrm{SIR}_M = \infty$, and that (6) is satisfied. That is, $\mathrm{WDO}_M = 1$ implies that the signals are W-DO.

Now we establish that binary time-frequency masks exist which are capable of demixing speech signals from one mixture and detail their performance in relation to the three presented measures. Consider the following family of time-frequency masks

$$\Phi_j^x(t,\omega) = \begin{cases} 1 & 20\log(|\hat{s}_j(t,\omega)| \,/\, |\hat{y}_j(t,\omega)|) \geq x \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

which is the indicator function for the time-frequency points where source j dominates the interference in the mixture by x dB. We will use $\mathrm{PSR}_j(x)$ and $\mathrm{SIR}_j(x)$ as shorthand for $\mathrm{PSR}_{\Phi_j^x}$ and $\mathrm{SIR}_{\Phi_j^x}$, respectively.

To determine the demixing ability of the above mask type, the masks for various x were applied to speech mixtures of various order and the demixing performance measures, $\mathrm{PSR}_j(x)$ and $\mathrm{SIR}_j(x)$, were determined. The demixed speech was then rated by the authors as falling into one of five subjective categories. The speech signals were selected from 16 male and 16 female continuous speech segments of 3 seconds taken from the TIMIT and normalized to unit energy. The time-frequency representation of the 16kHz sampled data was created using a Hamming window of 1024 samples with adjacent window centers separated by 512 samples. The results of the 333 listening tests are displayed in Figure 3. We note that there is a fairly accurate relationship between the WDO performance measure and the subjective ratings listed in the table under the figure.

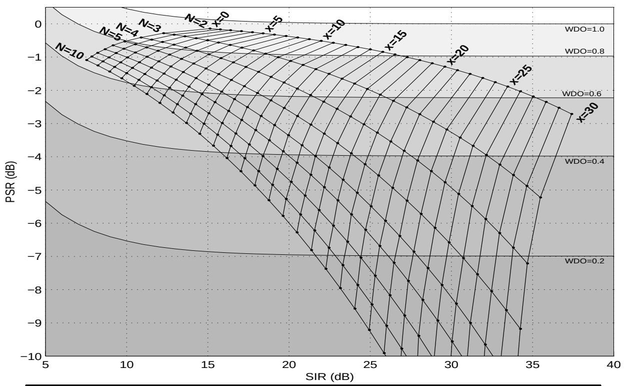


rating	meaning
1	perfect
2	minor artifacts or interference
3	distorted but intelligible
4	very distorted and barely intelligible
5	not intelligible

Fig. 3. Results of subjective listening test. For example, $.8 > \mathrm{WDO} \geq .6$ implies a “minor artifacts or interference” rating or better

Now that we have some idea how PSR, SIR, and WDO map to demixing performance, we analyze the demixing performance of the masks described in (15). Figure 4 shows plots of $\mathrm{PSR}_j(x)$ versus $\mathrm{SIR}_j(x)$ and a table of $(\mathrm{PSR}_j(x), \mathrm{SIR}_j(x))$ pairs averaged for groups of speech mixtures of different orders. For $N = 2$, each source was compared against each of the remaining 31 sources, resulting in $32 \times 31 = 992$ tests being averaged for each data point. For larger N, each source was compared against a random mixing of $N-1$ of the remaining 31 sources. This was done 31 times per source in order to keep the number of tests per data point constant at 992. As we tested mixtures from $N = 2$ to $N = 10$, a total of $9 \times 992 = 8928$ mixtures were created to generate the data for Figure 4. Note, we can average the $\mathrm{PSR}_j(x)$'s for different sources together because all sources have been normalized. In general, however, for sources with different powers, this averaging would not make sense as the $\mathrm{PSR}_j(x)$'s would be different for each source. The same is true of $\mathrm{SIR}_j(x)$. Figure 4 demonstrates that time-frequency masks exist which exhibit excellent demixing performance.

Now that we know that good time-frequency masks exist, we wish to determine the dependence of these performance measures on the window function $W(t)$ and window size. For this task, we examine the performance of the 0 dB mask, Φ_j^0. Figure 5 shows PSR, SIR, and WDO for pairwise mixing for various window sizes and types. Each data point in the figure represents the average of the results for 992 mixtures. In all measures, the Hamming window of size 1024 samples performed the best. Note, however, that the performance of the other masks (with the exception of the rectangle) was extremely similar and exhibited better than 90% W-disjoint orthogonality for pairwise mixing across a wide range of window sizes (from



N	$x = 0$	$x = 5$	$x = 10$	$x = 15$
2	$(0.97, 15.12)$	$(0.94, 18.39)$	$(0.89, 21.96)$	$(0.83, 25.76)$
3	$(0.94, 12.26)$	$(0.88, 15.64)$	$(0.80, 19.32)$	$(0.69, 23.25)$
4	$(0.91, 10.88)$	$(0.83, 14.32)$	$(0.72, 18.11)$	$(0.59, 22.12)$
5	$(0.89, 9.87)$	$(0.79, 13.37)$	$(0.65, 17.23)$	$(0.51, 21.33)$
10	$(0.78, 7.52)$	$(0.62, 11.23)$	$(0.44, 15.33)$	$(0.28, 19.62)$

Fig. 4. Time-Frequency Mask Demixing Performance. Plot contains $\mathrm{PSR}_j(x)$ (in dB) versus $\mathrm{SIR}_j(x)$ (in dB) for $x = 0, 1, \ldots, 30$ for $N = 1, 2, \ldots, 10$. Table contains $(\mathrm{PSR}_j(x), \mathrm{SIR}_j(x))$ in (%,dB) for $N = 2, 3, 4, 5, 10$ for $x = 0, 5, 10, 15$ dB. The different gray regions correspond to different regions of approximate W-disjoint orthogonality as determined by the lines of constant WDO. For example, using the $x = 5$ dB mask in mixtures of four sources yields 14.32 dB output SIR while maintaining 83% of the desired source power. This $(\mathrm{PSR}, \mathrm{SIR}) = $ (83%,14.32 dB) pair results in WDO $= 80\%$, which from Figure 3 implies perfect demixing performance. In other words, if we can correctly map time-frequency points with 5 dB or more single source dominance to the correct corresponding output partition, we can recover 83% of the energy of each of the original sources and produce demixtures with 14.32 dB output SIR from a mixture of four sources.

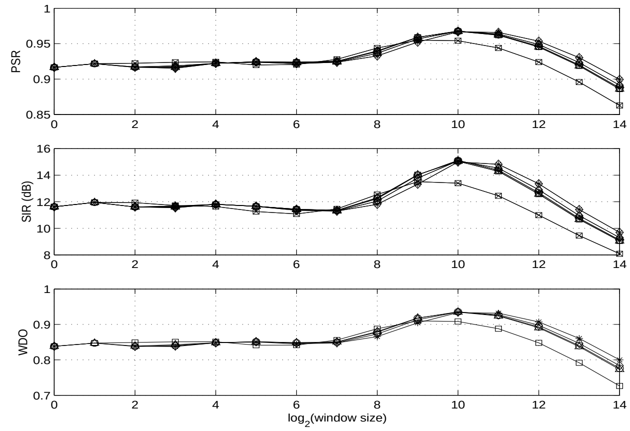


Fig. 5. Window size and type comparison. Hamming (○), Blackman (∗), Hann (⋄), Triangle (△), and Rectangle (□). PSR, SIR, and WDO for the 0 dB mask for window size $= 1, 2, 4, \ldots, 16384$ samples for various window types for pairwise mixing of speech signals sampled at 16 kHz. The Hamming window of size 1024 has the best performance.

N	2	3	4	5	6	7	8	9	10
WDO	93.6	88.0	83.4	79.2	75.6	72.3	69.3	66.6	64.0

Fig. 6. Percentage WDO for the 0 dB mask for mixtures of various order.

roughly 500 to 4000 samples). Other mixture orders and masks (i.e., $x > 0$) exhibited similar performance and in all cases the Hamming window of size 1024 had the best performance. A similar conclusion regarding the optimal time-frequency resolution of a window for speech separation was arrived at in [6].

Note that even when the window size is 1 (i.e., T is sampling), the mixtures still exhibit a high level of PSR, SIR, and WDO. This fact was exploited by those methods that used the time-disjoint nature of speech[1–3, 8]. However, Figure 5 clearly shows the advantage of moving from the time domain to the time-frequency domain: the speech signals are more disjoint in the time-frequency domain.

The approximate W-disjoint orthogonality of speech is a result of the sparsity of the Gabor representation of speech. Sparsity, in the strict sense, implies that most of the Gabor coefficients are zero[19]. However, signals of practical interest exhibit only a weakened definition of sparsity, in that most of the energy of the signal is captured by a small number of the coefficients. For different speech signals, it is unlikely that these coefficients coincide, which leads to approximate W-disjoint orthogonality. We close this section by proposing WDO_M with $M = \Phi_j^0$ as the measure of W-disjoint orthogonality.

Figure 6 shows a table of $\mathrm{WDO}_{\Phi_j^0}$ values for mixtures of various order. Again, each data point represents the average measurement over 992 mixtures. It can be shown using (13) that the 0 dB mask, Φ_j^0, maximizes WDO, and thus the 0 dB mask line represents the upper bound of WDO for any mask. We thus say that, for example, *speech signals in pairwise mixtures are* 93.6% *W-disjoint orthogonal.*

III. Parameter Estimation and Demixing

In this section, we will present a demixing algorithm that separates an arbitrary number of sources using two mixtures. We start by describing our anechoic mixing model. Suppose we have N sources $s_1(t), \ldots, s_N(t)$. Let $x_1(t)$ and $x_2(t)$ be the mixtures such that

$$x_k(t) = \sum_{j=1}^{N} a_{kj} s_j(t - \delta_{kj}), \; k = 1, 2 \tag{16}$$

where parameters a_{kj} and δ_{kj} are the attenuation coefficients and the time delays associated with the path from the j^{th} source to the k^{th} receiver. Without loss of generality we set $a_{1j} = 1$ and $\delta_{1j} = 0$ for $j = 1, \ldots, N$, for simplicity we rename a_{2j} as a_j and δ_{2j} as δ_j. In addition we assume that the windowed Fourier transform of any source function, $F^W[s_j](t, \omega)$ satisfies the narrowband assumption for array processing, i.e.,

$$\begin{aligned} F^W[s_j(\cdot - \delta)](t, \omega) &= \exp(-i\omega\delta) \cdot F^W[s_j](t - \delta, \omega) \\ &\approx \exp(-i\omega\delta) \cdot F^W[s_j](t, \omega). \end{aligned} \tag{17}$$

This assumption is realistic as long as the window function W is chosen appropriately. A detailed discussion about this assumption can be found in [20].

Now we go back to discussing the mixing model, described in (16). We take the windowed Fourier transform of x_1 and x_2 with an appropriate choice of W. Using the assumptions discussed above, the mixing model (16) reduces to

$$\begin{bmatrix} \hat{x}_1(t, \omega) \\ \hat{x}_2(t, \omega) \end{bmatrix} = \begin{bmatrix} 1 & \ldots & 1 \\ a_1 e^{-i\omega\delta_1} & \ldots & a_N e^{-i\omega\delta_N} \end{bmatrix} \begin{bmatrix} \hat{s}_1(t, \omega) \\ \vdots \\ \hat{s}_N(t, \omega) \end{bmatrix}. \tag{18}$$

A. Parameter Estimation and Demixing for W-DO sources

To motivate the Degenerate Unmixing Estimation Technique (DUET) algorithm which we will describe in the next section, we first consider the case where the sources are W-DO, i.e.,

$$\hat{s}_i(t,\omega)\hat{s}_j(t,\omega) = 0,\ \forall(t,\omega),\ \forall i \neq j. \tag{19}$$

This condition is the idealization of the properties of speech signals discussed in Section II. We now construct the parameter estimators and the demixing algorithm for W-DO signals. Clearly, when the sources are W-DO, at most one source will be active at any time-frequency point (t,ω), in particular for any (t,ω) at which $\hat{x}_1(t,\omega) \neq 0$, there exists $j = J(t,\omega)$ such that $\hat{s}_j(t,\omega) \neq 0$ and $\hat{s}_i(t,\omega) = 0$ for $i \neq j$. Define the time-frequency mask M_j as

$$M_j(t,\omega) = \begin{cases} 1 & \text{if } \hat{s}_j(t,\omega) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

Note that $M_j(t,\omega)M_i(t,\omega) = 0$ for all (t,ω) if $j \neq i$. From (18) one can easily deduce that

$$\hat{s}_j = M_j \hat{x}_1. \tag{21}$$

This shows that one can demix an arbitrary number of sources from only one of the mixtures if one can construct the corresponding mask M_j for each source. Next we will describe how to construct the masks M_j using the mixtures x_1 and x_2.

Let j be arbitrary, and define $\Omega_j = \{(t,\omega) : M_j(t,\omega) = 1\}$ so that $M_j = 1_{\Omega_j}$. Note that the Ω_j are pairwise disjoint. Now consider

$$R_{21}(t,\omega) = \frac{\hat{x}_2(t,\omega)}{\hat{x}_1(t,\omega)}. \tag{22}$$

Clearly, on Ω_j

$$R_{21}(t,\omega) = a_j e^{-i\delta_j \omega}. \tag{23}$$

In this case $|R_{21}(t,\omega)| = a_j$ and $-1/\omega \sphericalangle R_{21}(t,\omega) = \delta_j$, where $\sphericalangle z$ denotes the phase of the complex number z taken between $-\pi$ and π.

The observation above yields a way of constructing the sets Ω_j and thus a demixing algorithm: we simply label each time-frequency point (t,ω) with the pair $(|R_{21}(t,\omega)|, -1/\omega \sphericalangle R_{21}(t,\omega))$. Since the sources are W-DO, there will be N distinct labels. By grouping the time-frequency points (t,ω) with the same label, we construct the sets Ω_j, thus the masks $M_j = 1_{\Omega_j}$.

The above described demixing algorithm is the motivation behind DUET. Note that the algorithm separates the sources without inverting the mixing matrix, which makes it possible to deal with mixtures of an arbitrary number of sources. Aside from demixing, it also yields the mixing parameters: the labels $(|R_{21}(t,\omega)|, -1/\omega \sphericalangle R_{21}(t,\omega))$ which we used to construct the masks are exactly the mixing parameters a_j and δ_j. Motivated by this fact we define the instantaneous DUET attenuation and delay parameter estimators as

$$\tilde{a}(t,\omega) := |R_{21}(t,\omega)| \tag{24}$$

$$\tilde{\delta}(t,\omega) := -\frac{1}{\omega}\sphericalangle R_{21}(t,\omega) \tag{25}$$

respectively. We will use these estimators in the next section.

In summary, the **DUET algorithm for demixing W-DO sources** is thus,

1) From mixtures $x_1(t)$ and $x_2(t)$ construct time-frequency representations $\hat{x}_1(t,\omega)$ and $\hat{x}_2(t,\omega)$.
2) For each non-zero time-frequency point, calculate $(\tilde{a}(t,\omega), \tilde{\delta}(t,\omega))$.
3) Take the union of the $(\tilde{a},\tilde{\delta})$ pairs, $S = \bigcup\{(\tilde{a}(t,\omega),\tilde{\delta}(t,\omega)) \mid \forall(t,\omega)\}$. Note S will be equal to $\{(a_j,\delta_j) \mid j = 1,\ldots,N\}$.
4) For each (a_j,δ_j) in S, $j = 1,\ldots,N$, construct $\hat{s}_j(t,\omega) = 1_{\{(\tilde{a}(t,\omega),\tilde{\delta}(t,\omega))=(a_j,\delta_j)\}}(t,\omega)\hat{x}_1(t,\omega)$ for (t,ω) with $\hat{x}_1(t,\omega) \neq 0$ and $\hat{s}_j(t,\omega) = 0$ otherwise. Note, $\hat{s}_j(t,\omega)$ will be the time-frequency representations of one of the original sources. The numbering of the sources is arbitrary.
5) Convert each $\hat{s}_j(t,\omega)$ back into the time domain.

Remark 1: Note that the instantaneous DUET delay estimator yields a meaningful estimate at a time-frequency point $(t,w) \in \Omega_j$ only if

$$|\omega\delta_j| < \pi. \tag{26}$$

This follows from the periodicity of the complex exponential. For $(t,w) \in \Omega_j$, we have $-1/\omega \sphericalangle R_{21}(t,\omega)) = r(\omega\delta_j)\delta_j$, with $r(u) := \frac{<u>}{u}$ where $< u > := (u+\pi) \pmod{2\pi} - \pi$. When (26) is not satisfied, the delay estimate obtained using the instantaneous DUET estimator will be a fraction of its true value. Let $\omega_{\max}$ be the element of the set $\cup_j \Omega_j$ with the largest modulus, which is the maximum frequency present in the sources, and denote by ω_s the sampling rate. Let $\delta_{\text{jmax}} = \max_j |\delta_j|$. Clearly, (26) is guaranteed for all j and for all $\omega \in \cup_j \Omega_j$ if

$$\omega_{\max}\delta_{\text{jmax}} < \pi. \tag{27}$$

Now define $\delta_{\omega\text{max}} := \pi/\omega_{\max}$. Any delay parameter with modulus less than $\delta_{\omega\text{max}}$ can be estimated correctly. Clearly, (27) is equivalent to the condition $\delta_{\text{jmax}} < \delta_{\omega\text{max}}$. If $\omega_{\max} = \omega_s/2$, the Nyquist frequency, then this means that the maximum delay, $\delta_{\omega\text{max}} = \frac{2\pi}{\omega_s}$, is exactly equal to the sampling period. In other words, as long as the delay between the two microphone readings is less than a sample, the estimated phase will be accurate. While the $\omega_{\max}$ is determined by the characteristics of speech signals, the maximum physically possible delay, which we will denote by δ_{dmax}, is determined by the microphone spacing. For two microphones separated by a distance d, $\delta_{\text{dmax}} = d/c$ where c is the speed of sound. Clearly, we have $\delta_{\text{jmax}} < \delta_{\text{dmax}}$, and therefore (27) will be satisfied if $\delta_{\text{dmax}} < \delta_{\omega\text{max}}$. This suggests that one can guarantee (27) simply by choosing d, and thus δ_{dmax}, sufficiently small. For example, for a sampling rate $\omega_s/(2\pi) = 16$ kHz, assuming $\omega_{\max} = \omega_s/2$ and $c = 344\frac{m}{s}$, we obtain that $\delta_{\text{dmax}} \leq \delta_{\omega\text{max}}$ as long as $d \leq 2.15$ cm. If we knew, however, that $\omega_{\max}/(2\pi) = 4$ kHz, then this distance would be increased by a factor of 4 to 8.60 cm. The smaller the largest frequency present in the signal, the larger the allowable microphone separation (or equivalently the larger we can choose δ_{jmax}) that guarantees accurate phase parameter estimates.

B. Parameter Estimation for Approximately W-DO sources

1) Maximum Likelihood Estimators for Delay and Attenuation Coefficients: In Section II, we have illustrated that the time-frequency representations of speech signals are sparse, and one can indeed recover a speech signal from one mixture of an arbitrary number of sources if one can construct an appropriate time-frequency mask. This suggests that a weakened W-DO condition holds for speech signals: if at a time-frequency point one of the sources has considerable power, the contribution of all the other sources at that time-frequency point is likely to be small. This observation is the key to the demixing algorithm we propose in this section. First we shall discuss how to estimate the mixing parameters.

Instead of the continuous windowed Fourier transform, we use the equivalent discrete counterpart[1]

$$\hat{s}_j[k,l] = \hat{s}_j(kt_0, l\omega_0) \tag{28}$$

where t_0 and ω_0 are the time-frequency lattice spacing parameters.

We will say that s_j is *dominant* at $[k,l]$ if $|\hat{s}_j[k,l]| \geq |\hat{y}_j[k,l]|$, where $\hat{y}_j$ is as in (11). Note, the 0 dB mask, Φ_j^0, in (15) is the indicator function for the dominant time-frequency points for source j. Let us concentrate on one source, say s. Let Ω be the set of time-frequency points $[k,l]$ at which s is dominant in the sense described above. On Ω, we model the mixtures x_1 and x_2 as follows:

$$\begin{aligned} \hat{x}_1[k,l] &= \hat{s}[k,l] + n_1[k,l] \\ \hat{x}_2[k,l] &= ae^{-i\delta l\omega_0}\hat{s}[k,l] + n_2[k,l] \end{aligned} \tag{29}$$

where n_1 and n_2 are i.i.d. white complex Gaussian noise with zero-mean and variance σ^2. Here n_1 and n_2 model the contributions of other sources at the time-frequency points where s is the dominant source. We model the interfering sources as independent Gaussian noise in order to obtain simple closed-form source and mixing parameter estimators. In reality, the interference in the different mixtures will be correlated and may not be Gaussian distributed. For the model in (29), we want to employ a ML estimate to find the parameter pair $(a,\delta) \in \mathbb{R}^2$ as well as $\hat{s}[k,l]$ that maximize $P(\hat{x}_1, \hat{x}_2|a,\delta)$. To that goal, we define the likelihood , L_0, of $(\mathbf{s}, a, \delta)$, where $\mathbf{s} = (\hat{s}[k,l])_{(k,l)\in\Lambda}$ with each $\hat{s}[k,l] \in \mathbb{C}$ for some index set $\Lambda \subset \mathbb{Z}^2$, given the data $\hat{x}_1[k,l]$ and $\hat{x}_2[k,l]$, by

$$\begin{aligned} L_0(\mathbf{s},a,\delta) &:= p(\mathbf{x}_1,\mathbf{x}_2|\mathbf{s},a,\delta) \\ &= \prod_{(k,l)\in\Lambda} f_{N_1,N_2}(\hat{x}_1[k,l]-\hat{s}[k,l], \hat{x}_2[k,l]-ae^{-i\delta l\omega_0}\hat{s}[k,l]) \\ &= C\exp\Bigg(-\frac{1}{2\sigma^2}\sum_{(k,l)\in\Lambda}|\hat{x}_1[k,l]-\hat{s}[k,l]|^2 + \\ &\qquad \Big|\hat{x}_2[k,l]-ae^{-i\delta l\omega_0}\hat{s}[k,l]\Big|^2\Bigg) \end{aligned} \tag{30}$$

where $\mathbf{x}_i = (\hat{x}_i[k,l])_{(k,l)\in\Lambda}$. The last equality holds because we assume i.i.d. complex Gaussian noise. Clearly, maximizing L_0 is equivalent to maximizing

$$L(\mathbf{s},a,\delta) := -\sum_{(k,l)\in\Lambda}|\hat{x}_1[k,l]-\hat{s}[k,l]|^2 + \Big|\hat{x}_2[k,l]-ae^{-i\delta l\omega_0}\hat{s}[k,l]\Big|^2. \tag{31}$$

[1] The equivalence is nontrivial and only true for appropriately chosen window functions W with sufficiently small t_0 and ω_0. An illustrative discussion can be found in [16].

For this purpose, we want to solve the equations $\frac{\partial L}{\partial\alpha[k,l]} = 0$, $\frac{\partial L}{\partial\beta[k,l]} = 0$ for all $(k,l) \in \Lambda$, $\frac{\partial L}{\partial a} = 0$ and $\frac{\partial L}{\partial\delta} = 0$ simultaneously, where $\alpha[k,l]$ and $\beta[k,l]$ denote the real and imaginary parts of $\hat{s}[k,l]$ respectively. We start with $\frac{\partial L}{\partial\alpha[k,l]}$. For any $(k,l) \in \Lambda$, we have

$$\begin{aligned} \frac{\partial L}{\partial\alpha[k,l]} &= \frac{\partial}{\partial\alpha[k,l]}\Big(|\hat{x}_1[k,l]-\alpha[k,l]-i\beta[k,l]|^2 + \\ &\qquad \Big|\hat{x}_2[k,l]-ae^{-i\delta l\omega_0}(\alpha[k,l]+i\beta[k,l])\Big|^2\Big). \end{aligned} \tag{32}$$

We solve then $\frac{\partial L}{\partial\alpha[k,l]}\Big|_{\alpha[k,l]=\alpha^*[k,l]} = 0$ for $\alpha^*[k,l]$ and obtain

$$\alpha^*[k,l] = \mathbf{Re}\left\{\frac{\hat{x}_1[k,l]+ae^{i\delta l\omega_0}\hat{x}_2[k,l]}{1+a^2}\right\}. \tag{33}$$

Similarly, solving $\frac{\partial L}{\partial\beta[k,l]}\Big|_{\beta[k,l]=\beta^*[k,l]} = 0$ for $\beta^*[k,l]$ yields

$$\beta^*[k,l] = \mathbf{Im}\left\{\frac{\hat{x}_1[k,l]+ae^{i\delta l\omega_0}\hat{x}_2[k,l]}{1+a^2}\right\} \tag{34}$$

which we combine with (33) to get the ML estimate $\mathbf{s}^*$ for $\mathbf{s}$:

$$\mathbf{s}^*[k,l] = \frac{\hat{x}_1[k,l]+ae^{i\delta l\omega_0}\hat{x}_2[k,l]}{1+a^2}. \tag{35}$$

Next, we consider $\frac{\partial L}{\partial\delta}$. We have

$$\begin{aligned} \frac{\partial L}{\partial\delta} &= \frac{\partial}{\partial\delta}\left(\sum_{(k,l)\in\Lambda}\Big|\hat{x}_2[k,l]-ae^{-i\delta l\omega_0}\hat{s}[k,l]\Big|^2\right) \\ &= 2a\sum_{(k,l)\in\Lambda} l\omega_0\mathbf{Im}\left\{\hat{x}_2[k,l]\overline{\hat{s}[k,l]}e^{i\delta l\omega_0}\right\}. \end{aligned} \tag{36}$$

We now plug in $\hat{s}[k,l] = \mathbf{s}^*[k,l]$ in (36), which yields

$$\begin{aligned} \frac{\partial L}{\partial\delta} &= \frac{2a}{1+a^2}\sum_{(k,l)\in\Lambda} l\omega_0\,|\hat{x}_1[k,l]|^2\,\mathbf{Im}\left\{R_{21}[k,l]e^{i\delta l\omega_0}\right\} \\ &= \frac{2a}{1+a^2}\sum_{(k,l)\in\Lambda} l\omega_0\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|\sin(\sphericalangle R_{21}[k,l]+\delta l\omega_0) \end{aligned} \tag{37}$$

where $R_{21}[k,l] := \frac{\hat{x}_2[k,l]}{\hat{x}_1[k,l]}$. We define the instantaneous DUET delay estimate for the parameter δ, the discrete version of (25)

$$\tilde{\delta}[k,l] = -\frac{1}{l\omega_0}\sphericalangle R_{21}[k,l]. \tag{38}$$

and for convenience, define $\tilde{\delta}[k,l] = 0$ if $\hat{x}_1[k,l] = 0$ or $\hat{x}_2[k,l] = 0$. We assume that $|\sphericalangle R_{21}[k,l]+\delta l\omega_0| = \Big|l\omega_0(\delta-\tilde{\delta}[k,l])\Big|$ is small which is reasonable because we are considering only the $[k,l]$ where s is dominant and we make the approximation

$$\sin(\sphericalangle R_{21}[k,l]+\delta l\omega_0) \approx \sphericalangle R_{21}[k,l]+\delta l\omega_0. \tag{39}$$

After plugging (39) into (37), we solve the equation $\frac{\partial L}{\partial\delta}\Big|_{\delta=\delta^*} = 0$ for δ^* and obtain

$$\delta^* = \frac{\sum_{(k,l)\in\Lambda}\tilde{\delta}[k,l]l^2\omega_0^2\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|}{\sum_{(k,l)\in\Lambda}l^2\omega_0^2\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|}. \tag{40}$$

Note that δ^*, the ML estimate for the parameter δ, is a weighted average of the instantaneous DUET delay estimates, with each estimate weighted by the product magnitude of the mixtures as well as $(l\omega_0)^2$. We also observe that the ML estimate δ^* does not depend on the attenuation parameter a.

Finally, we will solve the equation $\frac{\partial L}{\partial a}\big|_{a=a^*}$ for a^*. We have

$$\frac{\partial L}{\partial a} = \frac{\partial}{\partial a}\left(\sum_{(k,l)\in\Lambda}\left|\hat{x}_2[k,l] - ae^{-i\delta l\omega_0}\hat{s}[k,l]\right|^2\right)$$
$$= \sum_{(k,l)\in\Lambda} 2a\left|\hat{s}[k,l]\right|^2 - 2\mathbf{Re}\left\{\hat{x}_2[k,l]\overline{\hat{s}[k,l]}e^{i\delta l\omega_0}\right\}. \quad (41)$$

After setting $\hat{s}[k,l] = \mathbf{s}^*[k,l]$ and some algebra we get

$$a^* - \frac{1}{a^*} = \frac{\sum_{(k,l)\in\Lambda}\left|\hat{x}_1[k,l]\hat{x}_2[k,l]\right|\left(\tilde{a}[k,l] - 1/\tilde{a}[k,l]\right)}{\sum_{(k,l)\in\Lambda}\mathbf{Re}\left\{\hat{x}_2[k,l]\overline{\hat{x}_1[k,l]}e^{i\delta^* l\omega_0}\right\}} \quad (42)$$

where

$$\tilde{a}[k,l] = |R_{21}[k,l]| \quad (43)$$

is the discrete version of the instantaneous DUET attenuation estimate (24). For convenience, we define $\tilde{a}[k,l] = 1$ if $\hat{x}_1[k,l] = 0$ or $\hat{x}_2[k,l] = 0$.

The estimate for $a^* - \frac{1}{a^*}$ is symmetric in $\mathbf{x}_1$ and $\mathbf{x}_2$: swapping the mixture labels will only result in a sign change of this quantity (i.e., $(1/a) - (1/(1/a)) = -(a - 1/a)$). In the original presentation of DUET [4], the logarithm of the attenuation estimates was used solely because it has the same property (i.e.,$\log(1/a) = -\log(a)$). However, motivated by its appearance in the ML estimator (42), we will replace the role of the logarithm with the *DUET symmetric attenuation estimator* defined as $\tilde{a}[k,l] - 1/\tilde{a}[k,l]$.

Remark 2: Although the estimate given in (42) is not a weighted average, it is interesting to note that if we replace δ^* in (42) with the instantaneous DUET phase estimates $\tilde{\delta}[k,l]$ which satisfy

$$e^{i\tilde{\delta}[k,l]l\omega_0} = \frac{R_{21}[k,l]}{|R_{21}[k,l]|} \quad (44)$$

we obtain that

$$\mathbf{Re}\left\{\hat{x}_2[k,l]\overline{\hat{x}_1[k,l]}e^{i\tilde{\delta}[k,l]l\omega_0}\right\} = |\hat{x}_1[k,l]\hat{x}_2[k,l]| \quad (45)$$

and in this case, (42) becomes a weighted average of $\tilde{a}[k,l] - 1/\tilde{a}[k,l]$.

2) Experimental Evaluation of the ML Estimator: In this section, we experimentally evaluate the ML estimators as well as other estimators motivated by the previous section. In order simulate mixtures, we use the model in (29) and adjust the noise power to model the different number of interfering sources. The model in (29) is valid for the dominant time-frequency points of one source s. In order to determine the set of dominant time-frequency points Λ, a speech signal taken from the TIMIT database was compared to a random mixture of 1, 2, 4, or 9 TIMIT speech signals to model the $N = 2, 3, 5, 10$ mixture orders, and in each case, the time-frequency points corresponding to the 0-dB mask were selected. The mixtures of interfering sources were only used to determine Λ and were discarded after the dominant time-frequency points were identified. In order to simulate the presence of interfering sources, i.i.d. Gaussian white noise was added to the dominant time-frequency points of source s on both channels. The added noise was amplified to produce a 15.12 dB, 12.26 dB, 9.87 dB, or 7.52 dB SNR so as to model mixing of order $N = 2, 3, 5,$ or 10, respectively. These SNR's were selected to model different mixture orders because they match the average SIR's for the 0-dB mask from Figure 4. That is, 15.12 dB, 12.26 dB, 9.87 dB, and 7.52 dB are the expected SIR's after applying the 0-dB mask to mixtures of order $N = 2, 3, 5,$ and 10, and thus in order to model these mixture orders for the dominant time-frequency points of one source, we add noise to one source to produce the corresponding SNR's. Note that the dominate time-frequency points are precisely the support of the source's 0 dB mask. Thus the performance of the estimator evaluated with this model at these SNR's should approximate the true performance of the estimator in speech mixtures of order $N = 2, 3, 5,$ and 10. All results in the remainder of this section will be obtained using this model.

We choose to experimentally evaluate the estimators using the model as described above as opposed to creating synthetic mixtures of multiple speech signals because we (1) wanted to prevent the results from depending on the specific choice of mixing parameters of the interfering sources and (2) wanted to evaluate the estimators using the model that motivated them. The disadvantage of modeling the presence of interfering sources in this way is that the interference should be correlated and this correlation is lost when the interference is modeled as independent noise. Our desire in this section is to explore the qualitative performance of a family of estimators to motivate the demixing algorithm, and modeling the interference as noise is sufficient for this purpose.

Figure 7 shows the ML estimate δ^* from (40) versus δ as δ ranges linearly from -5 samples to 5 samples with $a = 1$. We can see that the ML delay estimator exhibits bad performance outside the -1 to 1 sample range, and biased performance inside this range. The bad performance for larger delays is due to the phase wrap around problem discussed in Remark 1 in Section III-B.1. The squared frequency weighting factor in the ML delay estimator accentuates this problem. In addition, such a frequency weighting would make signals with higher frequency content have higher likelihood estimates of their delay parameters. In the next sections, we will be using these weightings to construct weighted histograms for source separation and it is undesirable to assign more likelihood to one set of parameters simply because their associated source contains higher frequencies. While methods for unwrapping the phase do exist, these methods are inappropriate for our purposes as different sources may be active from one frequency to the next. In order to see if we could reduce the bias, eliminate the wrap-around effect, and remove the high frequency weighting, we removed the squared frequency weighting factor in the ML delay estimator and considered estimators of the following form

$$\delta_j^{(p)} = \frac{\sum_{(k,l)\in\Lambda}\left|\hat{x}_1[k,l]\hat{x}_2[k,l]\right|^p\tilde{\delta}[k,l]}{\sum_{(k,l)\in\Lambda}\left|\hat{x}_1[k,l]\hat{x}_2[k,l]\right|^p}. \quad (46)$$

Note that with $p = 1$, this estimator is the ML delay estimator with the squared frequency weighting factor removed.

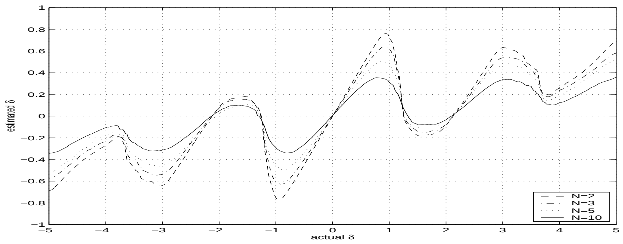


Fig. 7. Maximum Likelihood Delay Estimator. Plot compares estimated δ versus true δ for the ML estimator for δ ranging linearly from -5 to 5 samples when $a = 1$ for mixture models of 2, 3, 5, and 10 sources.

The estimator in (46) for $p = 1/2, 1, 2$ is compared with the ML estimator in Figure 8. For this test, δ ranges linearly from -5 samples to 5 samples while $a - 1/a$ ranges linearly from 0.15 to -0.15 and the SNR and Λ were selected as before to model mixture orders of 2, 3, 5, and 10. The $p = 1/2$ estimator suffers similar deficiencies as the ML estimator. The $p = 1$ estimator is clearly biased outside the -1 to 1 delay range, but is monotonic with increasing δ and exhibits good (although still biased) performance inside -1 to 1 sample delay. The $p = 2$ estimator exhibits near perfect estimates.

Figure 8 also shows $a^* - 1/a^*$ versus $a - 1/a$ for the same data as was used for the delay estimates. Similar to the delay case, in addition to the ML estimator, we consider estimators of the following form,

$$a^{(p)} - \frac{1}{a^{(p)}} = \frac{\sum_{(k,l)\in\Lambda} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p \left(\tilde{a}[k,l] - \frac{1}{\tilde{a}[k,l]}\right)}{\sum_{(k,l)\in\Lambda} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p}. \quad (47)$$

Note that with $p = 1$, this estimator is the ML symmetric estimator with substitution of (44) as described in Remark 2 previously. The ML and $p = 1/2$ symmetric attenuation estimators are clearly biased. The $p = 1$ symmetric attenuation estimator is also biased, although less so. The $p = 2$ symmetric attenuation estimator exhibits near perfect performance.

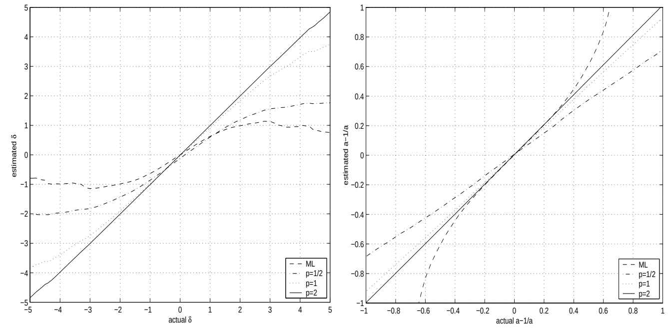
3) Multiple sources and the demixing algorithm: In Section III-B.1, we derived that the ML estimates for the mixing parameters a_j and δ_j, the attenuation and delay coefficients for source j, can be determined via certain weighted averages of the instantaneous DUET delay ($\tilde{\delta}[k,l]$) and DUET symmetric attenuation ($\tilde{a}[k,l] - 1/\tilde{a}[k,l]$) estimates, at the time-frequency points at which source j is dominant.

In Section III-B.2 we compared the performance of the set of weights suggested by the ML estimators with other empirically motivated weights, and we illustrated that the best estimates of (a_j, δ_j) are determined by

$$a_j^{(p)} - \frac{1}{a_j^{(p)}} = \frac{\sum_{(k,l)\in\Lambda_j} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p \left(\tilde{a}[k,l] - \frac{1}{\tilde{a}[k,l]}\right)}{\sum_{(k,l)\in\Lambda_j} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p} \quad (48)$$

$$\delta_j^{(p)} = \frac{\sum_{(k,l)\in\Lambda_j} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p \, \tilde{\delta}[k,l]}{\sum_{(k,l)\in\Lambda_j} |\hat{x}_1[k,l]\hat{x}_2[k,l]|^p} \quad (49)$$

when $p = 2$ where $\tilde{a}[k,l]$ and $\tilde{\delta}[k,l]$ are the instantaneous DUET estimates for the delay and attenuation coefficients, respectively. To employ these estimates however, we need to first construct the sets $\Lambda_j = \{[k,l] : |\hat{s}_j(k,l)| > |\hat{y}_j(k,l)|\}$ for each j. Note that these sets would be the discrete version of Ω_j of Section III-A if the sources are W-DO. In the W-DO case we used the instantaneous DUET estimates as labels for each time-frequency point (t,ω), and each Ω_j consisted of the points with identical labels. In the approximately W-DO case the instantaneous DUET estimates for the time-frequency points in Λ_j will not be identical anymore. However, we claim that we can still use the instantaneous DUET estimates as a means of labeling, and thus construct the sets Λ_j, at least approximately. Once we know the Λ_j, we demix simply by partitioning the support of $\hat{x}_1[k,l]$ using Λ_j and converting the resulting time-frequency representations back into the time domain. In order to determine the Λ_j, we rely on three observations which lead us to create a smoothed two-dimensional power weighted histogram of the $(\tilde{a}[k,l] - 1/\tilde{a}[k,l], \tilde{\delta}[k,l])$ pairs. Enumerating the peaks in this histogram estimates the number of sources, the peak centers estimate the mixing parameters, and the set of time-frequency points which contribute to a given peak provide an estimate for the associated Λ_j.



(a) delay (b) attenuation

Fig. 8. Delay and Attenuation Estimator Comparison. Plot (a) compares estimated δ versus true δ for the ML delay estimator and the weighted average DUET delay estimators with $p = 1/2, 1, 2$ as δ ranges linearly from -5 samples to 5 samples while $a - 1/a$ ranges linearly from 0.15 to -0.15 for a mixture model of 5 sources. Plot (b) compares estimated $a - 1/a$ versus true $a - 1/a$ for the ML and weighted average DUET symmetric attenuation estimators for the same experimental data that was used to generate plot (a).

Observation 1: The time-frequency points with instantaneous DUET estimates $(\tilde{a}[k,l] - 1/\tilde{a}[k,l], \tilde{\delta}[k,l])$ inside a small rectangle centered on the true mixing parameter pair $(a - 1/a, \delta)$ contain most of the source energy.

We wish to show that the time-frequency points which yield DUET estimates that are in close proximity to the true mixing parameters contain most of the energy of the source. Let,

$$M_{b,B}[k,l] = \begin{cases} 1 & \text{if } |(\tilde{a}[k,l] - 1/\tilde{a}[k,l]) - b| < B \\ 0 & \text{otherwise} \end{cases} \quad (50)$$

be the indicator function for time-frequency points with DUET attenuation estimate within B of b. We are interested in,

$$\mathrm{PSR}_{M_{b,B}} = \frac{\sum_{(k,l)} M_{b,B}[k,l] \, |s[k,l]|^2}{\sum_{(k,l)} |s[k,l]|^2} \quad (51)$$

when $b = a - 1/a$ which will show the percentage of energy of source s with DUET symmetric attenuation estimates within B of the true value $a - 1/a$. Plot (a) in Figure 9 shows $\text{PSR}_{M_{b,B}}$ averaged over 100 randomly selected speech signals taken from the TIMIT database. As before, the model in (29) was used to model the $N = 2, 3, 5, 10$ mixture orders. The curves represent the expected energy close to the true symmetric attenuation for mixtures of various orders. For example, with $B = 0.1$ we expect more than 60% of the source power to be located within 0.1 of the true $a - 1/a$ value in mixtures of 5 sources.

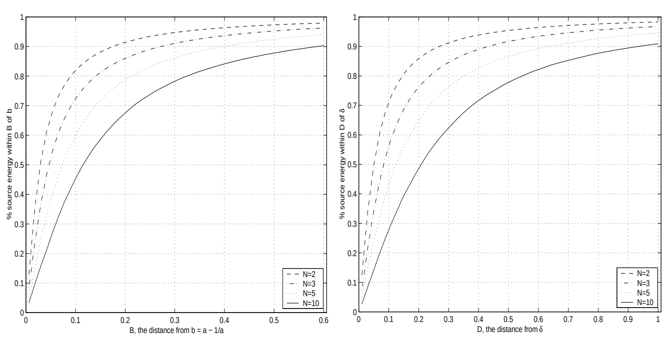
Similarly, for the delay, we define

$$M_{d,D}[k,l] = \begin{cases} 1 & \text{if } \left|\tilde{\delta}[k,l] - d\right| < D \\ 0 & \text{otherwise.} \end{cases} \tag{52}$$

Then we are interested in

$$\text{PSR}_{M_{\delta,D}} = \frac{\sum_{(k,l)} M_{\delta,D}[k,l] \left|s[k,l]\right|^2}{\sum_{(k,l)} \left|s[k,l]\right|^2} \tag{53}$$

which will show the percentage of energy of source s with DUET delay estimates within D of the true value δ. Plot (b) in Figure 9 shows $\text{PSR}_{M_{\delta,D}}$ as a function of D for various mixture orders. For example, 70% of the energy of the source is expected to be within 0.1 samples of the true δ estimate in pairwise mixing.



(a) attenuation

(b) delay

Fig. 9. Energy distribution of DUET estimates around the true mixing parameters. Note, for W-DO signals, the corresponding source power distribution would be 100% for all distances from a-1/a (and δ).

Figure 9 shows that a significant portion of a source's energy is contained in time-frequency points with instantaneous DUET symmetric attenuation estimate localized around $a - 1/a$ and instantaneous DUET delay estimate localized around δ. We now show that the source energy is localized simultaneously around $(a - 1/a, \delta)$. To do so, we look at

$$\text{PSR}_{M_{b,B} M_{\delta,D}} = \frac{\sum_{(k,l)} M_{b,B}[k,l] M_{\delta,D}[k,l] \left|s[k,l]\right|^2}{\sum_{(k,l)} \left|s[k,l]\right|^2} \tag{54}$$

which measures the percentage of source power for time-frequency points which yield estimate within B of $b = a - 1/a$ and D of δ. Before we examine $\text{PSR}_{M_{b,B} M_{\delta,D}}$, we need to determine the appropriate B to D ratio. Plotting (B, D) pairs for $\text{PSR}_{M_{b,B}} = \text{PSR}_{M_{\delta,D}}$ for the same mixture order reveals that the (B, D) lie essentially along a line. The least-mean-square fit of this line determines a ratio of $B/D = 1/1.7$ samples. This means that, for example, $\text{PSR}_{M_{b,B}} \approx .6$ for $B = .1$ for $N = 5$ implies that $\text{PSR}_{M_{\delta,D}} \approx .6$ for $D = 1.7 \times .1 = .17$ samples for $N = 5$, a property which can be verified from the data displayed in Figure 9. Figure 10 shows $\text{PSR}_{M_{B,D}}$ versus B and D for $1.7B = D$. Note that the B axis is at the bottom and the D axis is at the top. For example, 60% of the energy of the source is contained in a rectangle with dimensions .2-by-0.33 centered on $(a - 1/a, \delta)$ for mixtures of three sources. As the number of sources increases, the energy spreads over a wider area, but the source energy remains well localized around the source's mixing parameters.

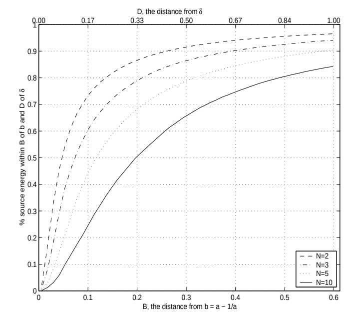


Fig. 10. Energy distribution of DUET estimates in a rectangle centered on the true mixing parameter pair $(a - 1/a, \delta)$.

Observation 2: Observation 1 is true for the individual sources in mixtures.

This observation is based on the fact that, from the experiments of speech mixtures (see Figure 4), we know that the time-frequency points when one source dominates maintain a significant percentage of the dominating source's energy. For $N = 2, 3, 4, 5$, and 10 the percentage source energy preserved when only considering dominant time-frequency points is $97\%, 94\%, 91\%, 89\%$, and 78%, respectively. Then, using the dominance model, considering the time-frequency points when one source dominates, Figure 10 show that the DUET estimates in rectangle centered on true estimates maintain a significant percentage of that source's power when considering only the dominant time-frequency points. For example with N=2 to N=10, with $(B, D) = (.33, .2\text{samples})$ the percentage ranges from 87% to 50%. Thus, we would expect in pairwise mixing for example, time-frequency points which yield DUET estimates $(\tilde{a}[k,l] - 1/\tilde{a}[k,l], \tilde{\delta}[k,l])$ inside a .2-by-.33 rectangle centered on $(a_1 - 1/a_1, \delta_1)$ to contain $87\% \times 97\% = 84\%$ of source 1's energy. Similarly, we would expect 84% of source 2's energy to come from time-frequency points which have DUET estimate pairs within a .2-by-.33 rectangle centered on $(a_2 - 1/a_2, \delta_2)$. As N increases, the source energy percentage we expect to see in a fixed size rectangle centered on each source's mixing parameters decreases (it is 39% for N=10), nevertheless, Observation 1 will hold for mixtures of sources.

Observation 3: The peaks in a smoothed two-dimensional power weighted histogram of the DUET estimates will be in one-to-one correspondence with the rectangle centers in Observation 2.

One way to determine the mixing parameters for multiple sources is to look to $M_{b,B}[k,l]M_{\delta,D}[k,l]\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|^p$ for some range of possible choices for $b = a - 1/a$ and δ, for some p. Figure 8 would suggest that we choose parameter $p = 2$, but we will see later that $p = 1/2, 1,$ or 2 all result in accurate estimates. If (B, D) is chosen large enough to capture a large portion of the source power, as determined by Figure 10, yet small enough so the (B, D) rectangle does not contain significant energy contributions from multiple sources, we would expect the local maxima to occur around the true mixing parameter pairs $(a_j - 1/a_j, \delta_j),\ j = 1, \ldots, N$. Therefore, one way of determining the mixing parameters would be to calculate $M_{b,B}[k,l]M_{\delta,D}[k,l]\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|^p$ for the range of interest of (b, δ) pairs and select the local maxima. A computationally efficient way of doing this is to construct a two-dimensional weighted histogram at a high resolution, and then smooth that histogram with a kernel of the dimensions of the desired (B, D) rectangle. We perform smoothing to capture the time-frequency points which are likely to correspond to one source. Recall that the estimators (48) and (49) averaged the instantaneous estimates over all time-frequency points where the source of interest was dominant. We know from the results shown in Figure 10 that that a rectangle centered on the true mixing parameters will capture most of the corresponding source's energy. By smoothing, we locate the rectangle centers that capture locally the largest amount of power contribution, and thus estimate the mixing parameters. Histograms have been used for parameter estimation of voice mixtures, for example, [21] clusters onset arrival difference to determine the time delays of the various sources.

Now we construct a two dimensional weighted histogram for $(\tilde{a}[k,l] - 1/\tilde{a}[k,l], \tilde{\delta}[k,l])$, where $\tilde{a}[k,l]$ and $\tilde{\delta}[k,l]$ are the instantaneous DUET estimates, with the weights $|\hat{x}_1[k,l]\hat{x}_2[k,l]|^p$ for some p. The weighted histogram with resolution widths β and Δ and weighting exponent p, is defined as

$$h^{(p)}_{\beta,\Delta}(b,\delta) = \sum_{k,l} M_{b,\beta/2}[k,l]M_{\delta,\Delta/2}[k,l]\,|\hat{x}_1[k,l]\hat{x}_2[k,l]|^p\,. \tag{55}$$

We will use the sampled version of this histogram

$$h^{(p)}_{\beta,\Delta}[m,n] = h^{(p)}_{\beta,\Delta}(m\beta, n\Delta) \tag{56}$$

which we will smooth with a kernel of size $(2n_b+1)$-by-$(2n_d+1)$, to produce smoothed histogram

$$H^{(p)}_{B,\beta,D,\Delta}[m,n] = \eta \sum_{u=-n_b}^{n_b} \sum_{v=-n_d}^{n_d} h^{(p)}_{\beta,\Delta}[m-u, n-v] \tag{57}$$

where $\eta = \frac{1}{(2n_b+1)(2n_d+1)}$, $n_b = \lceil \frac{B}{\beta} \rceil$, and $n_d = \lceil \frac{D}{\Delta} \rceil$.

Figure 11 shows an example histogram before and after smoothing generated using the dominant time-frequency points of a speech signal in a mixture of 5 sources. The presence of the interfering sources was modeled as additive Gaussian noise resulting in a 9.87 dB SNR on the dominant time-frequency points. For the mixing model, $a - 1/a$ was set to .2 and δ to .5, which match well with the peak location. The importance of the smoothing is clear in that it combines all the estimates power in a local region and results in a clear single peak and thus mixing parameter estimate.

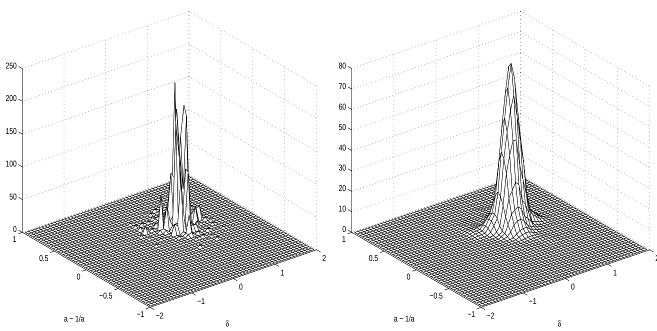
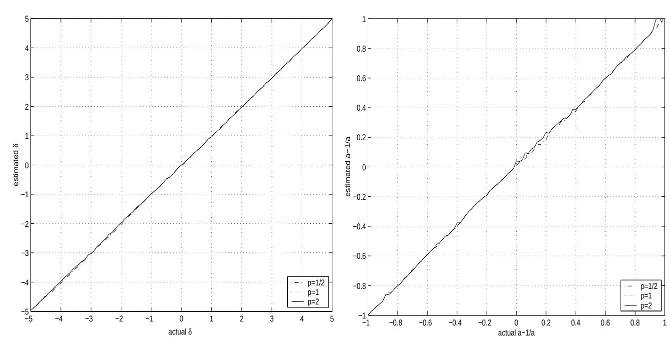


Fig. 11. Example raw and smoothed $p = 2$ power weighted histograms for one speech signal in a mixture of five. The peak location of the smoothed histogram corresponds to the mixing parameters $(a - 1/a, \delta) = (.2, .5)$.

In order to evaluate the usefulness of the histogram as a parameter estimator, a smoothed histogram was created for each of the tests used to generate Figure 8 and the peak location of the histogram was used as the symmetric attenuation and delay estimate. Figure 12 contains the results of these tests. Each estimator histogram consisted of 401-by-401 points with a delay range from -6 to 6 samples and symmetric attenuation from -1.2 to 1.2, and the smoothing kernel had parameters $(B, D) = (.12, .2)$. Comparing Figure 8 and Figure 12, we conclude that the histogram based estimators are superior to the previously considered ML motivated estimators.



(a) (b)

Fig. 12. Histogram Delay and Attenuation Estimator Comparison. Plot (a) compares estimated δ versus true δ for the smoothed histogram peak estimators with $p = 1/2, 1, 2$ for δ ranging linearly from -5 to 5 samples as $a - 1/a$ ranges linearly 0.15 to -0.15. Plot (b) compares estimated $a - 1/a$ to the true $a - 1/a$. Both plots were generated using a model of 5 source mixing.

The similar estimator performance for different choices of p in Figure 12 suggests that the choice of p should be driven by other concerns. Identifying the peaks in the histogram is the crucial step in the separation process. Two important criteria for the weighting exponent p selection are (1) the shape around the peak (the "peak shape") and (2) the relative peak heights. In order to aid in peak identification, we want the peak shape to be narrow and tall, and we want the peaks to be roughly of the same height. Figure 13 compares the histogram peak shapes for $p = 1/2, 1, 2$ for both the $a - 1/a$ and δ axes by taking the summation along the other axis. That is, Figure 13 contains 1-D weighted histograms for both $a - 1/a$ and δ. As p increases, the peak shape becomes narrower and taller. This would suggest that we should select p as large as possible. However, the larger we choose p, the more the peak heights depend only on the largest instantaneous product power time-frequency components of each source. If these components have different magnitude distributions for different sources, the resulting peaks heights can vary by several orders of magnitude making identification of the smaller peaks impossible. While $p = 2$ results in the best peak shape, smaller choices of p may result in easier peak identification. The choice of p is thus data dependent, however, motivated once again by the form of the ML estimators, we will suggest $p = 1$ as the default choice.

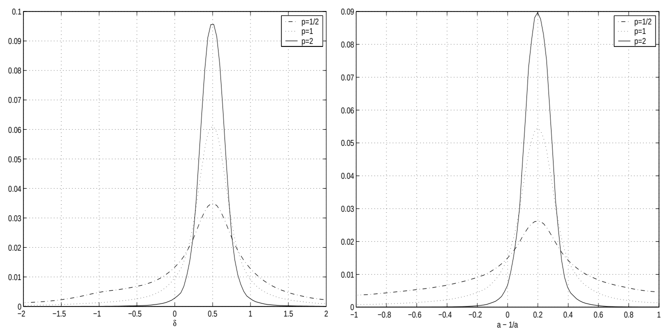


Fig. 13. Peakshape for p=1/2, 1, and 2.

C. Demixing Algorithm for Approximately W-DO Sources

Recall that in the W-DO case, sources were demixed using time-frequency masks that were constructed by grouping the time-frequency points that yield the same instantaneous parameter estimates. We demix in a similar way for approximately W-DO sources. First we estimate the mixing parameters, for example, using the histogram method described in the previous section. Then, we group time-frequency points that yield instantaneous parameter estimates that are "close" to these estimated mixing parameters. One natural definition of closeness is the instantaneous likelihood function for source j

$$L_j[k,l] := p(\hat{x}_1[k,l], \hat{x}_2[k,l] | a_j, \delta_j) = \frac{1}{2\pi\sigma^2} e^{-\frac{1}{2\sigma^2}\left|a_j e^{-i\delta_j l\omega_0}\hat{x}_1[k,l] - \hat{x}_2[k,l]\right|^2/(1+a_j^2)} \quad (58)$$

obtained by substituting the instantaneous ML source estimate (35) into the likelihood function in (30) modified to consider only time-frequency point $[k, l]$. $L_j[k, l]$ is, in a sense, the likelihood that source j is dominant at time-frequency point $[k, l]$. One way to demix the mixtures is to construct a time frequency mask for source j by taking those time-frequency points for which $L_j[k, l] \geq L_i[k, l], \forall i \neq j$. The time-frequency mask for demixing source j is thus

$$\tilde{M}_j = 1_{\{[k,l]: j = \arg\max_m L_m[k,l]\}} \quad (59)$$

$$= 1_{\{[k,l]: j = \arg\max_m \left|a_m e^{-i\delta_m l\omega_0}\hat{x}_1[k,l] - \hat{x}_2[k,l]\right|^2/(1+a_m^2)\}} \quad (60)$$

and defining

$$\tilde{\Lambda}_j = \{[k,l] : \arg\max_m \frac{\left|a_m e^{-i\delta_m l\omega_0}\hat{x}_1[k,l] - \hat{x}_2[k,l]\right|^2}{1 + a_m^2} = j\} \quad (61)$$

the estimate of the time-frequency points for which source j is dominant, we can relate this demixing mask to those that were used in the W-DO case. There are many other ways we can envision using these likelihoods, for example, some type of relative weighting resulting in fractional masks instead of the binary winner-take-all masks created by the scheme we have proposed. However, we have shown in Section II that the 0-dB binary masks exhibit excellent demixing performance and maximize the WDO performance measure so we consider exclusively binary time-frequency masks in this paper.

As before, we estimate the source by converting

$$\tilde{\hat{s}}_j[k,l] = \tilde{M}_j[k,l]\hat{x}_1[k,l] \quad (62)$$

into the time domain. Note, we could apply the mask to x_2 as well, and, could combine the two demixtures using the ML estimate of the source as in (35). However, in order to compare with the results obtained in Section II, the experimental results presented in the next section will use (62).

In summary, the **DUET algorithm for demixing Approximately W-DO sources** is,

1) From mixtures $x_1(t)$ and $x_2(t)$ construct time-frequency representations $\hat{x}_1[k, l]$ and $\hat{x}_2[k, l]$.
2) For each time-frequency point, calculate $(\tilde{a}[k, l], \tilde{\delta}[k, l])$.
3) Construct histogram and locate peaks:
   a) Construct a high resolution histogram as in (56)
   b) Smooth the histogram as in (57)
   c) Locate peaks in histogram. There will be N peaks, one for each source, with peak locations approximately equal to the true mixing parameter pairs, $\{(a_j - 1/a_j, \delta_j) \,|\, j = 1, \ldots, N\}$.
4) For the N pairs of (a, δ) estimates, construct the time-frequency masks corresponding to each pair using the ML partitioning as in (59) and apply these masks to one of the mixtures to yield estimates of the time-frequency representations of the original sources.
5) Convert each estimate back into the time domain.

IV. Experiments

In order to demonstrate the technique, we present results in this section for both synthetic and real mixtures. One issue that we have not addressed in the preceeding discusion is how the histogram peaks are automatically enumerated and identified. For the following demonstration, we used an ad hoc technique that iteratively selected the highest peak and removed a region surrounding the peak from the histogram. Peaks were removed as long as the histogram maintained a threshold percentage of its original weight. The threshold percentage and region dimensions had to be occasionally altered in the course of the tests to ensure the correct number of sources was found. Indeed, peak enumeration and identification remains a topic of future research.

A. Synthetic mixtures

Figure 14 shows the smoothed histogram (57) for a six source synthetic mixing example with histogram resolution widths $(\beta, \Delta) = (0.05, 0.12\,\text{samples})$ and smoothing kernel dimensions $(B, D) = (0.12, 0.2\,\text{samples})$. The six sources were taken from the TIMIT database and the (a, δ) the stereo mixture was created using mixing parameters pairs $(a, \delta) = (1, -2)$, $(3/2, -1)$, $(3/2, 1)$, $(1, 2)$, $(2/3, 1)$, and $(2/3, -1)$. It is clear given only the stereo mixture, one can determine how many sources were used to create the mixture by enumerating the peaks in the histogram. Using the ML partitioning, the first channel of the mixture was demixed and the SNR, PSR, and WDO measured; the results are shown in Figure 15. For comparison, WDO_{Φ^0}, the optimal WDO created using the 0 dB mask is shown in the last column. The demixtures average over 13 dB SNR gain and the WDO numbers indicate demixtures which would rate right on the border between “minor artifacts or interference” and “distorted but intelligible.” Note that even though the blind method performs reasonably well, the performance of the 0 dB mask shows that there exist time-frequency masks which would further improve the performance. Figure 16 shows the original six sources, the two mixtures, and the six demixtures.

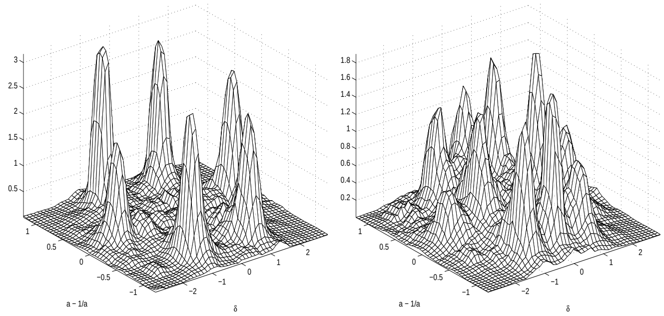


Fig. 14. Six and Ten Source Synthetic Mixing Smoothed Histograms. Each peak corresponds to one source and the peak location corresponds to the associated source’s mixing parameters.

source	SNR in	SNR out	SNR gain	PSR	WDO DUET	WDO 0dB
s_1	-7.29	5.92	13.21	0.76	0.57	0.80
s_2	-7.29	5.24	12.53	0.78	0.55	0.78
s_3	-5.08	6.60	11.67	0.80	0.62	0.81
s_4	-9.29	5.35	14.63	0.79	0.56	0.69
s_5	-5.03	7.06	12.09	0.78	0.63	0.81
s_6	-9.28	5.47	14.75	0.77	0.55	0.66
s_1	-9.74	-0.32	9.42	0.58	-0.04	0.70
s_2	-7.73	3.14	10.87	0.66	0.34	0.77
s_3	-11.64	3.43	15.06	0.68	0.37	0.64
s_4	-9.72	-0.60	9.13	0.58	-0.09	0.67
s_5	-7.73	3.93	11.66	0.66	0.39	0.73
s_6	-11.61	3.14	14.75	0.56	0.29	0.51
s_7	-7.75	2.57	10.31	0.56	0.25	0.74
s_8	-11.62	1.36	12.98	0.61	0.16	0.62
s_9	-9.72	4.70	14.42	0.60	0.39	0.67
s_{10}	-9.74	3.33	13.07	0.60	0.32	0.64

Fig. 15. Six and Ten Source Demixing Performance. Performance of the blind technique is compared against the optimal time-frequency mask, the 0 dB mask.

To show the limits of this technique, a ten source stereo mixture was synthetically mixed. The smoothed histogram for the mixture is shown in Figure 14 and Figure 15 contains the demixing performance. The SNR gains are still high, the average gain above 12 dB, however, the WDO performance has dropped to “very distorted and barely intelligible.” However, as we are trying to demix ten sources from just two mixtures, these results are promising. More promising indeed is the fact that the 0 dB mask’s performance is significantly better showing that there is room for improvement.

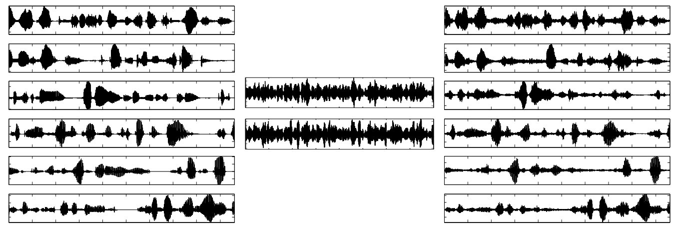


Fig. 16. Six Sources, Stereo Mixture, and Six Demixtures.

B. Anechoic and Echoic Mixing Results

We also tested DUET on speech mixtures recorded in an anechoic room. For the tests, each speech signal was recorded separately and then the signals were mixed additively to generate the mixtures for the tests. Knowledge of the actual signals present in each mixture allows us to calculate the performance measures exactly. For the recordings, the microphones were separated by 1.75 cm and the speech signals were played from various positions on a semicircle around the microphones with the microphone axis along the line from the 0° position to the 180° position. Two female (F1 and F2) and one male (M1) TIMIT sound files were used for the tests. Pairwise mixing results for female-female and male-female mixtures are shown in Figure IV-B. Again, for comparison purposes, the WDO obtained by the DUET algorithm is compared to the optimal WDO which is obtained using the 0 dB mask. The separation obtained by DUET is nearly perfect and in all but the 30° case: the DUET mask’s performance is essentially the same as the performance of the optimal mask.

test	SNR in	SNR out	SNR gain	PSR	WDO DUET	WDO 0dB
F1 0°	-0.58	12.69	13.26	0.92	0.87	0.96
F2 30°	0.58	11.25	10.68	0.96	0.89	0.96
F1 0°	-0.54	15.97	16.51	0.98	0.95	0.96
F2 60°	0.54	17.21	16.68	0.98	0.96	0.96
F1 0°	-0.62	15.29	15.91	0.97	0.94	0.94
F2 90°	0.62	15.69	15.07	0.98	0.95	0.95
F1 0°	-0.49	17.50	17.99	0.98	0.96	0.96
F2 120°	0.49	17.36	16.87	0.98	0.97	0.97
F1 0°	-0.50	15.79	16.29	0.97	0.94	0.94
F2 150°	0.50	15.51	15.01	0.98	0.95	0.95
F1 0°	-0.44	16.29	16.73	0.96	0.94	0.94
F2 180°	0.44	14.49	14.05	0.98	0.94	0.95
F1 0°	3.54	13.99	10.46	0.96	0.92	0.97
M1 30°	-3.54	10.35	13.88	0.91	0.83	0.94
F1 0°	3.60	18.42	14.81	0.99	0.97	0.98
M1 60°	-3.60	15.41	19.01	0.97	0.94	0.95
F1 0°	3.63	18.92	15.29	0.99	0.98	0.98
M1 90°	-3.63	15.91	19.54	0.97	0.95	0.95
F1 0°	3.69	19.91	16.22	0.99	0.98	0.98
M1 120°	-3.69	15.79	19.48	0.98	0.95	0.95
F1 0°	3.75	19.57	15.82	0.99	0.98	0.98
M1 150°	-3.75	16.37	20.12	0.97	0.95	0.95
F1 0°	3.90	18.47	14.57	0.99	0.97	0.98
M1 180°	-3.90	15.51	19.41	0.97	0.94	0.94

Fig. 17. Pairwise Anechoic Demixing Performance.

Higher order mixing results are listed in Figure 18. In addition to the three, four, and five source anechoic mixtures tested, a three source echoic mixture was tested. All of the speech signals, three female (F1, F2, and F3) and two male (M1 and

M2), were taken from the TIMIT database. The echoic recording was made in an echoic office environment. As the number of sources increases, the demixing performance decreases, although the performance is still acceptable in the five source mixture. As expected, the performance drops off significantly when switching from the anechoic to the echoic environment as the method is based on an anechoic mixing model. However, some separation is still achieved. Figure 19 compares the one source histograms for anechoic and echoic recordings for sources at three different angles. The histograms corresponding to the summation of the three sources are also shown. The anechoic histograms are well localized and the peak regions are clearly distinct, even in the histogram corresponding to the summation of the sources. The echoic histograms peak regions are spread out and overlap with one another. This overlap results in reduced demixing performance. Note, however, that the 0 dB mask still performs well in the echoic case, so there remains a gap between what we can separate blindly and what we can separate with knowledge of the instantaneous time-frequency amplitudes when using time-frequency masking to demix.

Anechoic

test	SNR in	SNR out	SNR gain	PSR	WDO DUET	WDO 0dB
M1 0°	-2.72	13.67	16.39	0.92	0.88	0.90
F1 90°	-2.05	7.96	10.00	0.96	0.80	0.93
M2 180°	-4.37	13.32	17.70	0.88	0.84	0.87
M1 0°	-6.93	9.89	16.83	0.78	0.70	0.80
F1 60°	-3.19	7.11	10.30	0.92	0.74	0.91
M2 120°	-4.37	6.98	11.35	0.85	0.68	0.89
F2 180°	-5.05	10.08	15.12	0.86	0.78	0.90
F1 0°	-9.77	7.97	17.74	0.73	0.62	0.76
M1 60°	-4.30	7.16	11.46	0.83	0.67	0.86
F2 90°	-3.77	5.99	9.76	0.91	0.68	0.91
M2 120°	-5.60	7.05	12.65	0.80	0.65	0.85
F3 180°	-8.59	8.53	17.11	0.76	0.65	0.82

Echoic

test	SNR in	SNR out	SNR gain	PSR	WDO DUET	WDO 0dB
M1 0°	-5.20	5.38	10.58	0.56	0.40	0.81
M2 90°	0.07	4.33	4.26	0.89	0.56	0.91
F1 180°	-4.48	6.03	10.51	0.65	0.49	0.87

Fig. 18. Higher Order Demixing Performance. Results for three source, four source, and five source anechoic mixtures, as well as three source echoic mixing.

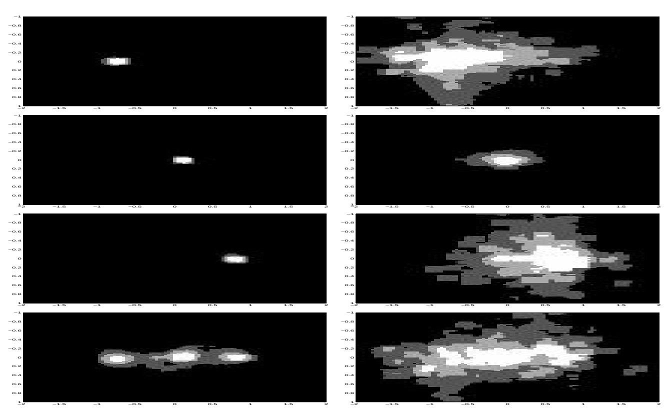


Fig. 19. Anechoic vs. Echoic Histogram Comparison. The left column images are of the histograms for three anechoic sources at 0°, 90°, 180°, and their mixture. The histogram of the mixture is essentially the summation of the individual histograms and the peak regions in the histogram are clearly separated. The right column images are of the histograms for three echoic sources 0°, 90°, 180°, and their mixture. While the individual histograms show some level of localization (left, center, right), peak regions in each histogram overlap and the peaks are difficult to identify in the summation image. Thus, the algorithm performs worse on echoic mixtures.

Figure 20 shows the histogram for the Te-Won Lee real office room recording consisting of two speakers[22]. The histogram shows a number of peaks, the peaks with $a - 1/a > -0.5$ are all associated with the Spanish speaker, and those along the $a - 1/a = -1.0$ line correspond to the English speaker. Note that for this recording, it is the attenuation direction in the histogram that allows for the separation and that a method that only relied on delays would not be able to separate the sources. Demixtures generated from this recording using the DUET algorithm are compared to several other BSS techniques here [23].

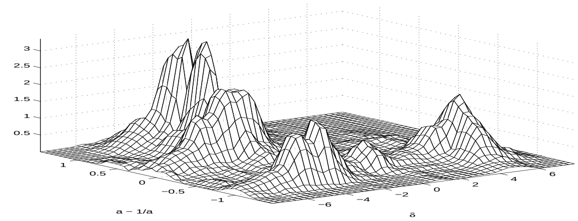


Fig. 20. Histogram for Te-Won Lee's "A real Cocktail Party Effect" Echoic Mixing Example.

V. Conclusions

In this paper, we presented a method to blindly separate mixtures of speech signals. We first illustrated experimentally that binary time-frequency masks exist that can separate as many as 10 different speech signals from one mixture. This relies upon a property of the Gabor expansions of speech signals, which we refer to as W-disjoint orthogonality. W-disjoint orthogonality in the strict sense is satisfied by signals which have disjoint time-frequency supports. Speech signals, as a result of the sparsity of their Gabor expansions, satisfy an approximate version of the W-disjoint orthogonality property. In Section II-A, we introduced a means of measuring the degree of W-disjoint orthogonality of a signal in a given mixture with respect to a windowing function W. Listening experiments showed that there is a fairly accurate relationship between the WDO value of a particular signal in a mixture for a given time-frequency mask and the subjective performance of the time-frequency mask to separate the signal from the mixture.

Next, we addressed the problem of blindly constructing a binary time-frequency mask that demixes. To this end, we considered the two mixture case. For strictly W-disjoint orthogonal signals, we showed that the DUET attenuation and delay estimators are the anechoic mixing parameters and, using this fact, described a simple algorithm to construct a binary time-frequency mask that demixes perfectly. Next we showed, by modeling the contributions of the interfering sources as independent Gaussian white noise, that the ML estimators for the mixing parameters are given by weighted averages of the instantaneous DUET estimates. Motivated by this, we constructed a weighted histogram that is used to enumerate the sources and partition the time-frequency representation of one of the mixtures to demix.

In Section IV we presented experimental results showing that the algorithm works extremely well for synthetic mixtures of speech as well as for speech mixtures recorded in an anechoic room. That is, the performance of the mask generated using the algorithm was close to that of the ideal mask. In an echoic room the anechoic model is violated and the quality of the demixing is reduced. In the echoic case, the demixtures contain some crosstalk and distortion, but are intelligible. There is, in the echoic case, a performance gap between the ideal binary time-frequency mask and the mask generated using the DUET algorithm. Closing this gap is the one goal of our future research.

The algorithm presented in this paper can certainly be improved by using more flexible time-frequency methods. One possible direction is to choose window functions adaptively instead of using a fixed window. An algorithm of this type is presented in [24]. It would be also interesting to investigate whether one could obtain a similar separation algorithm using wavelet expansions of speech.

In Section III-B.1, we derived the ML estimators for the mixing parameters by using a naive stochastic model for the contributions of interfering sources. A better estimator can perhaps be obtained by using a more realistic stochastic model for the Gabor expansions of speech signals and considering the whole problem in a stochastic setting. [25] used dependent Bernoulli random variables statistically enforce strict W-disjoint orthogonality. [26] eliminated the dependence of the Bernoulli random variables and the signal class for which the algorithm is intended includes signals which occasionally contain hits, time-frequency points where more than one source is active. For actual speech signals, [8] suggests a distribution with a sparsity factor, which may be well suited for further analysis.

As we mentioned in Section IV, the enumeration and identification of the histogram peaks is an issue that should be studied. One technique for identifying and tracking the peak locations through time is presented in [25]. Rather than constructing a histogram, [25] tracks the mixing parameters using a gradient method based on the instantaneous estimates. This allows the speakers to move in the environment without affecting the separation results. Source enumeration, however, is not addressed and remains an topic of future research for approximately W-disjoint orthogonal signals.

References

[1] J.-K. Lin, D. G. Grier, and J. D. Cowan. Feature extraction approach to blind source separation. In *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 398–405, Amelia Island Plantation, Florida, September 24–26 1997.

[2] T-W. Lee, M. Lewicki, M. Girolami, and T. Sejnowski. Blind source separation of more sources than mixtures using overcomplete representations. *IEEE Signal Processing Letters*, 6(4):87–90, April 1999.

[3] M. Van Hulle. Clustering approach to square and non-square blind source separation. In *IEEE Workshop on Neural Networks for Signal Processing (NNSP)*, pages 315–323, Madison, Wisconsin, August 23–25 1999.

[4] A. Jourjine, S. Rickard, and Ö. Yılmaz. Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 5, pages 2985–2988, Istanbul, Turkey, June 5–9 2000.

[5] P. Bofill and M. Zibulevsky. Blind separation of more sources than mixtures using sparsity of their short-time Fourier transform. In *International Workshop on Independent Component Analysis and Blind Signal Separation (ICA)*, pages 87–92, Helsinki, Finland, June 19–22 2000.

[6] M. Aoki, M. Okamoto, S. Aoki, H. Matsui, T. Sakurai, and Y. Kaneda. Sound source segregation based on estimating incident angle of each frequency component of input signals acquired by multiple microphones. *Acoustical Science & Technology*, 22(2):149–157, February 2001.

[7] L.-T. Nguyen, A. Belouchrani, K. Abed-Meraim, and B. Boashash. Separating more sources than sensors using time-frequency distributions. In *International Symposium on Signal Processing and its Applications (ISSPA)*, pages 583–586, Kuala Lumpur, Malaysia, August 13–16 2001.

[8] L. Vielva, D. Erdogmus, C. Pantaleon, I. Santamaria, J. Pereda, and J. C. Principe. Underdetermined blind source separation in a time-varying environment. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 3049–3052, Orlando, Florida, May 13–17 2002.

[9] P. Bofill. Underdetermined blind separation of delayed sound sources in the frequency domain. *preprint*, 2002.

[10] S. Rickard and Ö. Yılmaz. On the approximate W-disjoint orthogonality of speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 529–532, Orlando, Florida, May 13–17 2002.

[11] S. T. Roweis. One microphone source separation. In *Neural Information Processing Systems 13 (NIPS)*, pages 793–799, 2000.

[12] K. Suyama, K. Takahashi, and R. Hirabayashi. A robust technique for sound source localization in consideration of room capacity. In *IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, pages 63–66, New Paltz, New York, October 21–24 2001.

[13] R. Balan, A. Jourjine, and J. Rosca. AR processes and sources can be reconstructed from degenerate mixtures. In *International Workshop on Independent Component Analysis and Blind Source Separation (ICA)*, pages 467–472, Aussois, France, January 11–15 1999.

[14] P. Comon and O. Grellier. Non linear inversion of underdetermined mixtures. In *International Workshop on Independent Component Analysis and Blind Source Separation (ICA)*, pages 461–465, Aussois, France, January 11–15 1999.

[15] Michael Zibulevsky and Barak A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computation*, 13(4):863–882, 2001.

[16] I. Daubechies. *Ten Lectures on Wavelets*, chapter 3. SIAM, Philadelphia, PA, 1992.

[17] W. Kozek. Time-frequency signal processing based on the Wigner-Weyl framework. *Signal Processing*, 29:77–92, 1992.

[18] B. Berdugo, J. Rosenhouse, and H. Azhari. Speakers' direction finding using estimated time delays in the frequency domain. *Signal Processing*, 82:19–30, 2002.

[19] A. Zeira and B. Friedlander. *Gabor representation and signal detection*, chapter 11. Gabor Analysis and Algorithms: Theory and Applications, H. G. Feichtinger and T. Strohmer, Eds., Birkhäuser, Boston, 1998.

[20] R. Balan, J. Rosca, S. Rickard, and J. O'Ruanaidh. The influence of windowing on time delay estimates. In *Conference on Information Sciences and Systems (CISS)*, volume 1, pages WP1–(15–17), Princeton, NJ, March 15–17 2000.

[21] J. Huang, N. Ohnishi, and N. Sugie. A biomimetic system for localization and separation of multiple sound sources. *IEEE Transactions on Instrumentation and Measurement*, 44(3):733–738, June 1995.

[22] `http://www.cnl.salk.edu/~tewon/Blind/blind_audio.html`.

[23] `http://www.princeton.edu/~srickard/bss.html`.

[24] R. Gribonval. Sparse decomposition of stereo signals with matching pursuit and application to blind separation of more than two sources from a stereo mixture. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 3057–3060, Orlando, Florida, May 13–17 2002.

[25] S. Rickard, R. Balan, and J. Rosca. Real-time time-frequency based blind source separation. In *International Workshop on Independent Component Analysis and Blind Source Separation (ICA)*, pages 651–656, San Diego, CA, December 9–12 2001.

[26] R. Balan and J. Rosca. Statistical properties of STFT ratios for two channel systems and applications to blind source separation. In *International Workshop on Independent Component Analysis and Blind Source Separation (ICA)*, pages 429–434, Helsinki, Finland, June 19–22 2000.