Phase Aliasing Correction For Robust Blind Source Separation Using DUET

Yang Wang*†, Özgür Yılmaz‡ and Zhengfang Zhou §

November 4, 2011

Abstract

Degenerate Unmixing Estimation Technique (DUET) is a technique for blind source separation (BSS). Unlike the ICA-based BSS techniques, DUET is a time-frequency scheme that relies on the so-called W-disjoint orthogonality (WDO) property of the source signals, which states that, statistically, the windowed Fourier transforms of different source signals have disjoint supports. In addition to being computationally very efficient, one of the advantages of DUET is its ability, at least in some cases, to separate $n \geq 3$ source signals using only two mixtures. However, DUET is prone to phase wrap-around aliasing, which often leads to incorrectly separated sources with artifacts and distortions. This weakness limits the effectiveness of WDO and DUET as a robust tool for BSS. In this paper we present a method for correcting the phase wrap-around aliasing for WDO-based techniques such as DUET using over-sampled Fourier transform and modulo arithmetic. Experimental results have shown that this phase aliasing correction method is very effective, yielding a highly robust blind source separation for speech mixtures.

1 Introduction

Blind source separation (BSS) is a major area of research in signal processing with a vast literature. It aims to separate source signals from their mixtures without assuming detailed knowledge about the sources and the mixing process. A major focus of BSS is the separation of audio signals. The basic setup for BSS in audio has n audio sources $S_1(t), \ldots, S_n(t)$ and m mixtures $X_1(t), \ldots, X_m(t)$, with the model

$$X_k(t) = \sum_{i=1}^{n} \sum_{j=1}^{N_{k,i}} a_{k,i,j} S_i(t - d_{k,i,j}), \qquad k = 1, \ldots, m \tag{1.1}$$

where $a_{k,i,j}$ are the mixing coefficients and $d_{k,i,j}$ are the delays as a result of reverberations. In BSS these coefficients are unknown. With the presence of reverberations we often refer the mixtures X_k in (1.1) as

*Department of Mathematics, Michigan State University
†Contact Author: ywang@math.msu.edu
‡Department of Mathematics, University of British Columbia, Vancouver
§Department of Mathematics, Michigan State University

convolutive mixtures. The *anechoic model* assumes no reverberation in the mixtures, and thus it yields a much simpler model

$$X_k(t) = \sum_{i=1}^{n} a_{k,i} S_i(t - d_{k,i}), \qquad k = 1, \ldots, m. \tag{1.2}$$

The bulk of the studies in BSS employ the *independent component analysis (ICA)* model, where the source signals S_k, modeled as random variables, are assumed to be independent, see the surveys [4, 9] and the references therein for more details. Under the ICA model, many techniques such as Joint Approximate Diagonalization Eigenmatrices (JADE) [3] and Information Maximization (Infomax) [2, 5] along with their refinements have been developed (e.g. [6, 8]). These techniques use the kurtosis or the information optimizations to find the mixing coefficients and the delays. In many cases they yield excellent results: Clean separation with negligible distortion. They also have some limitations. For example, in the ICA model the sources must be non-Gaussian. When 4th order statistics are used, small perturbations to the tail part of the distribution can yield large errors, which makes kurtosis-based techniques less robust. With information optimization based algorithms there is often no guarantee that the iterative schemes converge to the global optimum. In fact for many schemes there is not even any guarantee that they converge at all. Thus the performance of these schemes can vary significantly from one set of data to another. There have been many attempts to address these problems (see e.g. [9, 14, 11, 12]), with varying degrees of success.

A different approach to BSS uses the time-frequency orthogonality property of the source signals, the so-called *W-Disjoint Orthogonality (WDO)* [10, 16, 17]. One of the advantages of WDO is that it can be used for BSS in the degenerative setting where there are more source signals than mixtures, i.e. $n > m$. The *Degenerate Unmixing Estimation Technique (DUET)* allows for the reconstruction of many sources from only two mixtures under the WDO assumption. DUET is also extremely fast, thus very suitable for both batch and dynamic source separation. There is an extensive literature on BSS using WDO, specifically on DUET as well as some related schemes, e.g., [21, 1, 20, 19, 7, 13, 18].

However, WDO and DUET are not without flaws. The setup of DUET is in an anechoic model, which raises questions about its effectiveness in a typical echoic setting. Our experiments seem to indicate that it handles echoic environment quite well, although more mathematical analysis is needed to better understand this. Another major flaw of DUET is the phase wrap-around aliasing. As we shall see, in DUET the time-frequency decompositions of the blind source estimates depend on attaining correctly certain phase values from the windowed Fourier transform of the mixtures. But because the phase values are bounded in the interval $(-\pi, \pi]$ correct phase values often cannot be obtained given the phase wrap-around aliasing. This means whenever wrap-around occurs the phase values are incorrect. This failure leads to incorrect time-frequency decompositions in DUET (even though the mixing coefficients and delays may be accurately estimated from low frequency content of the mixtures). Depending on the severity of the problem it can often lead to poor results in the form of uncleanly separated sources with distortion and artifacts. For WDO-based BSS techniques such as DUET to work consistently, the microphones must be placed very close together (say within 2-5 centimeters).

The main contribution of this paper is to provide a phase wrap-around aliasing correction method. This method will allow us to greatly improve the accuracy and performance of WDO-based BSS techniques even with phase wrap-around. The key ingredients for this phase aliasing correction method are the over-sampled Fourier transform and simple modulo arithmetic. Incorporating this phase aliasing correction mechanism to DUET yields high quality separation of source signals in real world audio mixtures, and furthermore, as we shall demonstrate, it is extremely robust.

2 W-Disjoint Orthogonality and DUET

Given a windowing function $W(t)$ and a function $f(t)$, the windowed Fourier transform of f with window W is defined by

$$\widehat{f}^W(\omega,\tau) := \int_{\mathbb{R}} W(t-\tau)f(t)e^{-i\omega t}\,dt. \tag{2.3}$$

We call two functions $S_1(t)$ and $S_2(t)$ *W-disjoint orthogonal* if the supports of the windowed Fourier transforms of $\widehat{S}_1^W(\omega,\tau)$ and $\widehat{S}_2^W(\omega,\tau)$ are disjoint. In other words we have

$$\widehat{S}_1^W(\omega,\tau)\cdot\widehat{S}_2^W(\omega,\tau) = 0 \tag{2.4}$$

for all ω and τ. If any two functions in $S_1(t),\ldots,S_n(t)$ satisfy the WDO property then we say $S_1(t),\ldots,S_n(t)$ satisfy the WDO property.

There is strong evidence that W-disjoint orthogonality is satisfied approximately for a large class of window functions W when the source functions S_i are speeches by different individuals, see e.g. [17]. The success of DUET is also a strong indirect evidence that WDO holds for speech and some other forms of audio signals. To see how WDO leads to DUET consider the anechoic model (1.2) with $m = 2$ and the constant window function $W(t) = 1$. In this anechoic model we have

$$\begin{aligned} X_1 &= \textstyle\sum_{k=1}^n a_{1k}S_k(t-d_{1k}),\\ X_2 &= \textstyle\sum_{i=k}^n a_{2k}S_k(t-d_{2k}). \end{aligned}$$

Note that by normalization we may without loss of generality assume that all $a_{1k} = 1$ and $d_{1k} = 0$. Thus we have

$$\begin{aligned} X_1 &= \textstyle\sum_{k=1}^n S_k(t)\\ X_2 &= \textstyle\sum_{k=1}^n a_k S_k(t-d_k). \end{aligned} \tag{2.5}$$

With $W(t) = 1$ we have

$$\begin{aligned} \widehat{X}_1^W(\omega,\tau) &= \textstyle\sum_{k=1}^n \widehat{S}_k^W(\omega,\tau),\\ \widehat{X}_2^W(\omega,\tau) &= \textstyle\sum_{k=1}^n a_k e^{-id_k\omega}\widehat{S}_k^W(\omega,\tau). \end{aligned}$$

Now assume that the source functions $S_k(t)$ satisfy the WDO property. It follows that for any given ω the function

$$F(\omega,\tau) := \frac{\widehat{X}_2^W(\omega,\tau)}{\widehat{X}_1^W(\omega,\tau)} \tag{2.6}$$

can only take values in the finite set $\{a_k e^{-id_k\omega} : 1 \le k \le n\}$. This observation forms the basis for DUET. More precisely, define the *amplitude-phase function*

$$\Lambda(\omega,\tau) := \big(|F(\omega,\tau)|, -\omega^{-1}\Theta(F(\omega,\tau))\big), \tag{2.7}$$

where $\Theta(z)$ denotes the angle of z, $-\pi < \Theta(z) \le \pi$. Assume there is no phase wrap-around in $\Theta(F(\omega,\tau))$. Then the function Λ only takes values in the finite set $\{(a_k, d_k) : 1 \le k \le n\}$. We may now compute each $\widehat{S}_k^W(\omega,\tau)$ via the following assignment algorithm

$$\widehat{S}_k^W(\omega,\tau) = \begin{cases} \widehat{X}_1^W(\omega,\tau) & \text{if } \Lambda(\omega,\tau) = (a_k, d_k) \\ 0 & \text{otherwise.} \end{cases}$$

The DUET reconstruction of the source signals S_k is now easily achieved from their windowed Fourier transforms $\widehat{S}_k^W(\omega,\tau)$ using standard techniques.

Although in practice the window function W is often taken to be a bell-shaped function such as the Hamming window instead of a constant function, the above method for reconstructing the source signals S_k as an approximation is still quite effective and often yields good results. Instead of taking exactly n values $\{(a_k, d_k) : 1 \le k \le n\}$, the range of the amplitude-phase function $\Lambda(\omega,\tau)$ is now concentrated at these values. The histogram of Λ should shows n peaks at these points on the amplitude-phase plane. A clustering algorithm is used for assigning $\widehat{X}_1^W(\omega,\tau)$ to each $\widehat{S}_k^W(\omega,\tau)$ based on WDO. A simple yet effective clustering algorithm is to first identify the n peaks in the histogram. They are assumed to be the points $\{(a_k, d_k) : 1 \le k \le n\}$. Then for each (ω,τ) we assign $\widehat{X}_1^W(\omega,\tau)$ to $\widehat{S}_k^W(\omega,\tau)$ if (a_k, d_k) is the closest to $\Lambda(\omega,\tau)$.

The analysis above assumes that $\Theta(F(\omega,\tau)) = \Theta(e^{-id_k\omega}) = -d_k\omega$. However, this is no longer true if $|d_k\omega| > \pi$ as a result of phase wrap-around. Once this happens we may not obtain the correct assignment and hence the correct reconstruction of the source signals.

3 Phase Wrap-Around Aliasing Correction for DUET

As we have mentioned, one of the major problems of DUET is the phase wrap-around, which results in faulty assignments in computing the functions $\widehat{S}_k^W(\omega,\tau)$. This may lead to poor separation of the source signals as well as distortions and artifacts. We illustrate this problem with a simple example. Assume that

$$X_1 = S_1(t) + S_2(t), \qquad X_2 = S_1(t-5) + S_2(t+3).$$

The amplitude-phase function $\Lambda(\omega,\tau)$ defined in (2.7) is now either $(1, -\omega^{-1}\Theta(e^{-i5\omega}))$ or $(1, -\omega^{-1}\Theta(e^{i3\omega}))$. However, if $|\omega| > \frac{\pi}{5}$ then $\Theta(e^{-i5\omega}) \ne -5\omega$ because of the phase wrap-around and $-\pi < \Theta(z) \le \pi$ for all z. Similarly, if $|\omega| > \frac{\pi}{3}$ then $\Theta(e^{i3\omega}) \ne 3\omega$. Thus $\Lambda(\omega,\tau)$ no longer takes only two values. Instead the phase component of $\Lambda(\omega,\tau)$ is virtually arbitrary when $|\omega| > \frac{\pi}{3}$, leading to faulty assignments in the reconstruction.

As we can see, the larger the d_i is, the fewer percentage of assignments will be correct, which leads to poorer separation. In the above example, only when $|\omega| < \frac{\pi}{5}$ we can be sure that the correct assignments are made. In other cases we might just as well toss a coin for each assignment. One solution to avoid the wrap-around problem is to place the microphones very close so that the delays d_i are small. There is frequently technical restrictions for doing so. For example, with 16kHz sampling frequency one would need the microphones to be within 2.2 centimeters to ensure $|d_i| \leq 1$. Depending on the models, microphones often cannot be placed this close. For speech separation, since human speech has a rather narrow frequency range, the 16kHz sampling frequency represents a considerable over-sampling. Thus even with $|d_j|$ slightly greater than 1 the separation using DUET can still be quite good. However, as we shall demonstrate in the next section, the performance of DUET deteriorates rather quickly as we move the microphones further apart.

We propose a method for correcting the phase wrap-around problem. The key ingredient is a revised amplitude-phase function that makes effective use of the continuity of the Fourier transform. Using the continuity of Fourier transform to solve the aliasing problem in DUET was first proposed by Rickard in [15], although no experimental details were given there. Our approach differs slightly from that of [15] as we use modulo arithmetic and over-sampled Fourier frame expansions. Furthermore, we shall demonstrate the effectiveness of our approach using extensive tests on synthetic mixtures as well as real life speech recordings. The revised amplitude-phase function $\overline{\Lambda}(\omega, \tau)$, replacing the original amplitude-phase function $\Lambda(\omega, \tau)$, yields a much more accurate assignment scheme for obtaining $\widehat{S}_k^W(\omega, \tau)$ from $\widehat{X}_1^W(\omega, \tau)$, leading to superior performance in DUET.

Assume that WDO holds for the source signals $S_1, \ldots, S_n$. Fix a point $\omega = \omega_0$. Our approach relies on an *assumption of continuity*: if we have $\widehat{X}_1^W(\omega_0, \tau) = \widehat{S}_k^W(\omega_0, \tau)$ then more likely we also have $\widehat{X}_1^W(\omega_0 + \varepsilon, \tau) = \widehat{S}_k^W(\omega_0 + \varepsilon, \tau)$ when $\varepsilon > 0$ is small. Like the WDO, the validity of this assumption needs to be further analyzed mathematically. Indirect evidence for its validity can be seen from the improvement that our phase aliasing correction provides over DUET, which we present in the next section. Let us set $\varepsilon = 2\pi/M$ where M is a large positive integer. Hence the function $F(\omega, \tau)$ defined by (2.6) satisfies

$$F(\omega_0, \tau) = a_k e^{-id_k\omega_0}, \qquad F(\omega_0 + \varepsilon, \tau) = a_k e^{-id_k(\omega_0+\varepsilon)}.$$

Now

$$\begin{aligned} -d_k\omega_0 - \Theta(F(\omega_0, \tau)) &\equiv 0 \pmod{2\pi}, \\ -d_k(\omega_0 + \varepsilon) - \Theta(F(\omega_0 + \varepsilon, \tau)) &\equiv 0 \pmod{2\pi}. \end{aligned}$$

It follows that

$$d_k\varepsilon \equiv \Theta(F(\omega_0, \tau)) - \Theta(F(\omega_0 + \varepsilon, \tau)) \pmod{2\pi}.$$

With $\varepsilon = \frac{2\pi}{M}$ we have

$$d_k \equiv \frac{M}{2\pi}\Big(\Theta(F(\omega_0, \tau)) - \Theta(F(\omega_0 + \frac{2\pi}{M}, \tau))\Big) \pmod{M}.$$

Note that once we know the bound for d_k this uniquely determines d_k when M is sufficiently large. An alternative but equivalent perspective is that under the assumption of continuity, $-d_k$ is the derivative of

$\Theta(F(\omega,\tau))$ (after the obvious correction of aliasing discontinuity) with respect to ω. Finding d_k via (3.8) amounts to computing $-\partial\Theta(F(\omega_0,\tau))/\partial\omega$ by first order forward difference, i.e.

$$d_k \equiv D_1(\omega,\tau,\varepsilon) := \frac{1}{\varepsilon}\Big((\Theta(F(\omega_0,\tau)) - \Theta(F(\omega_0+\varepsilon,\tau))\Big) \pmod M \tag{3.8}$$

with $\varepsilon = \frac{2\pi}{M}$. This forward difference yields only a first order approximation of the derivative. So naturally we may also consider higher order differences to approximate the derivative for potentially better performace. One candidate is the second order difference

$$d_k \equiv D_2(\omega,\tau,\varepsilon) := \frac{1}{2\varepsilon}\Big(\Theta(F(\omega_0+\varepsilon,\tau)) - \Theta(F(\omega_0-\varepsilon,\tau)\Big) \pmod M \tag{3.9}$$

and the five-point approximation

$$d_k \equiv D_3(\omega,\tau,\varepsilon) := \frac{1}{12\varepsilon}\Big(-\Theta(F(\omega_0+2\varepsilon,\tau))+8\Theta(F(\omega_0+\varepsilon,\tau))-8\Theta(F(\omega_0-\varepsilon,\tau))+\Theta(F(\omega_0-2\varepsilon,\tau)\Big) \pmod M. \tag{3.10}$$

The modified amplitude-phase function $\overline{\Lambda}(\omega,\tau)$ is now defined by

$$\overline{\Lambda}(\omega_0,\tau) := \Big(|F(\omega_0,\tau)|, D(\omega_0,\tau,\varepsilon)\Big), \tag{3.11}$$

with $\varepsilon = \frac{2\pi}{M}$ and D can be any of the derivative approximation functions such as D_1, D_2, D_3 defined above.

Of course, in practice one uses discrete windowed Fourier transforms, which are computed via FFT. Assume that the size of window (the support of the window function W in this case) is N. Then the variable ω takes on values $\frac{2\pi m}{N}$ where $-\frac{N}{2} < m \le \frac{N}{2}$ in DUET. With our method, the FFT is being replaced with an *over-sampled* FFT. Instead of computing FFT on the size N data $g_\tau(t) := W(t-\tau)X_i(t)$ we compute the M-point FFT of $\tilde{g}_\tau(t)$ that we obtain by padding $M-N$ zeros to $g_\tau(t)$. Here the integer M can be substantially larger than N. We choose $M = pN$ for some integer $p \ge 1$. A good balance between performance and computational demand (mainly memory demand) is $3 \le p \le 5$. The over-sampled discrete Fourier transform is equivalent to the *harmonic frame transform* familiar to the study of tight frames. With the over-sampled FFT the variable ω now takes on values $\frac{2\pi m}{pN}$ where $-\frac{pN}{2} < m \le \frac{pN}{2}$. Since these data are redundant, for the reconstruction of the source signals in DUET only a portion of the data where $\omega = \frac{2\pi pm}{pN} = \frac{2\pi m}{N}$ for $-\frac{N}{2} < m \le \frac{N}{2}$ are needed. The modified amplitude-phase function $\overline{\Lambda}(\omega,\tau)$ defined in (3.11) now utilizes the over-sampled FFT in place of the Fourier transform.

4 Experimental Results

We present several experimental examples using both synthetic mixtures and real world recordings. All our real recordings were done in ordinary rooms with moderate reverberations. Although in theory DUET is based on anechoic model only, it actually works in practice for echoic settings. In fact, after addressing the phase wrap-around aliasing using our method it is an extremely robust technique for audio BSS. For an online demo of our proposed algorithm, see [22].

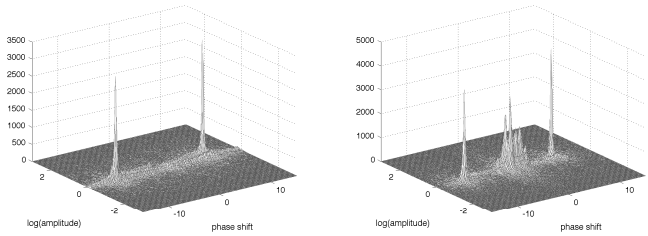

Figure 1: Histograms of $\overline{\Lambda}(\omega,\tau)$ (left) and $\Lambda(\omega,\tau)$ (right) from artificial mixtures of two speech sources. With phase aliasing correction the histogram on the left shows two sharp peaks at the desired locations. However, without the correction the histogram on the right has at least three peaks, with the peak in the middle being an artifact of the wrap-around aliasing.

Example 1. We first show the result using synthetic mixtures. The two source signals S_1, S_2 are two downloaded radio recordings at 16kHz sampling rate. They are artificially mixed using

$$X_1(t) = S_1(t) + S_2(t), \quad X_2(t) = 0.9S_1(t-8) + 1.1S_2(t+10).$$

Figure 1 shows that using the phase aliasing correction the histogram of the amplitude-phase function $\overline{\Lambda}(\omega,\tau)$ has two very clear peaks at the desired locations, while the histogram of the original amplitude-phase function $\Lambda(\omega,\tau)$ without phase aliasing correction looks rather ambiguous, with a large "mountain" in between the two correct peaks.

Example 2. Here we have a real recording of mixed speeches by two speakers. The two mixtures are recorded in a medium sized room with moderate amount of reverberation. A pair of inexpensive omni-directional microphones (about \$13 apiece) are used for the recording, and they are placed 12cm apart. Using the phase aliasing correction the histogram of the amplitude-phase function $\overline{\Lambda}(\omega,\tau)$ (Figure 2 left) shows two very clear peaks. The separated signals have very little distortion and artifacts. One of the mixtures and the separated signals are plotted in Figure 3. On the other hand, without the phase aliasing correction the histogram of the amplitude-phase function $\Lambda(\omega,\tau)$ (Figure 2 right) shows three peaks, in which the largest peak in the middle is a false one. The separated signals show distortion and artifacts.

One may have noticed that the two peaks in this example are no longer as sharp as the two peaks in the previous example. Actually, a huge difference between real recordings and artificial mixtures is that in real recordings the peaks are not nearly as sharp, even when the recording is done in an anechoic environment. Reverberations will further reduce the sharpness of the peaks. However, our tests have shown that for two-speaker mixtures, when the phase wrap-around aliasing correction is used, even with strong reverberation and inexpensive microphones the algorithm works very well, and it is surprisingly robust. In fact, when the microphones are placed between 5cm to 25cm, we have not had any failures.

Example 3. Here we show the result from a degenerative case, where we have two mixtures of three speech

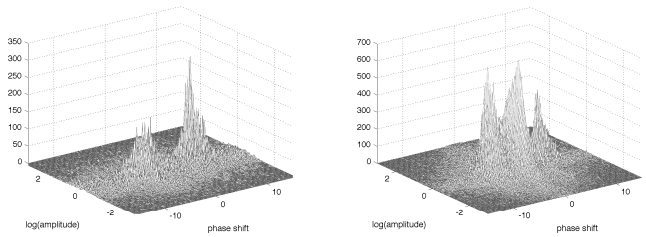

Figure 2: Histograms of $\overline{\Lambda}(\omega, \tau)$ (left) and $\Lambda(\omega, \tau)$ (right) from a real recording.

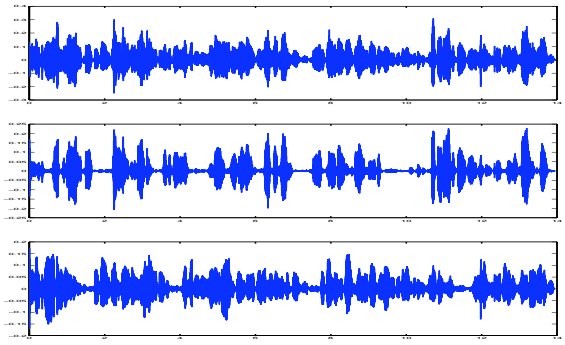

Figure 3: Top plot: one of the two real recordings of two-speaker mixtures. Bottom two plots: separated sources using DUET with phase aliasing correction.

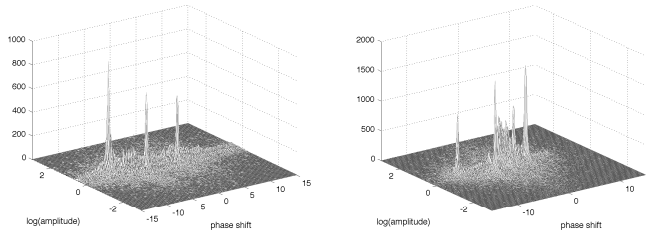

Figure 4: Histograms of $\overline{\Lambda}(\omega,\tau)$ (left) and $\Lambda(\omega,\tau)$ (right) from artificial mixtures of three speech sources. With phase aliasing correction the histogram on the left shows three sharp peaks at the desired locations. Without the correction the histogram on the right is much more ambiguous.

sources. The mixtures are created artificially via the following formula:

$$X_1(t) = S_1(t) + S_2(t) + S_3(t), \quad X_2(t) = 1.1S_1(t+4) + S_2(t) + 0.9S_3(t+9).$$

Again, using phase aliasing correction the histogram of the amplitude-phase function $\overline{\Lambda}(\omega,\tau)$ has three very sharp peaks at the desired locations (Figure 4 left), while the histogram of the original amplitude-phase function $\Lambda(\omega,\tau)$ without phase aliasing correction looks more ambiguous (Figure 4 right). Indeed, the ability to separate more than two sources from only two mixtures is one of the great strengths of DUET. Note that if there are more than two mixtures at our disposal, such additional information can be utilized to improve the separation performance. Such an algorithm was proposed in [18], and it was demonstrated that having more mixtures indeed improves the quality of separated sources significantly.

Example 4. In the above examples, we observed that our phase aliasing correction technique results in significantly improved separation in isolated examples. Next, we run an extensive test on a large number of synthetic mixtures. We generated mixtures $X_1(t) = S_1(t) + S_2(t), \quad X_2(t) = S_1(t+d_1) + 0.98S_2(t+d_2)$ where S_1 and S_2 are 10-second speech signals obtained from downloaded radio recordings, sampled at 16 kHz. In these mixtures, we fixed $d_2 = -2$ and varied d_1 from 1 to 60. From each mixture, we computed the estimated sources $\overline{S}_j$ (i) using DUET (no phase aliasing correction), and (ii) using the proposed phase aliasing correction method with the derivative estimators D_1, D_2, D_3 as shown in (3.8),(3.9), and (3.10) respectively, each with over-sampling rates $p = 1$ (i.e., no oversampling), $p = 3$, and $p = 5$. In each case we repeated the experiment with 10 different source pairs S_1 and S_2, and computed the resulting signal-to-noise-ratio via

$$\mathrm{SNR}_j = 20\log_{10}\big(\|S_j\|_2/\|S_j - \overline{S}_j\|_2\big), \ j = 1, 2.$$

The average SNR corresponding to each method is reported in Figure 5 and Figure 6.

Assuming that the speed of sound is $c = 343$ m/s, one sample delay corresponds to a microphone distance no less than 2.14 cm. Furthermore, note that, assuming the speech signals have a bandwidth of 8 kHz, their Fourier transforms will be full band when sampled at 16 KHz. In this case, phase aliasing will occur whenever the time delays d_j are greater than 1 sample. Indeed, as seen, e.g., in Figure 5, the separation performance

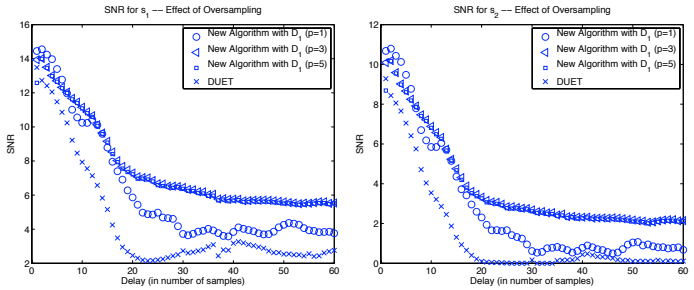

Figure 5: Average SNR for the estimated source $\overline{S}_1$ over 10 experiments per data point. We implemented the phase aliasing correction method with over-sampling rates $p = 1, 3, 5$ using the derivative estimator D_1 defined in (3.8).

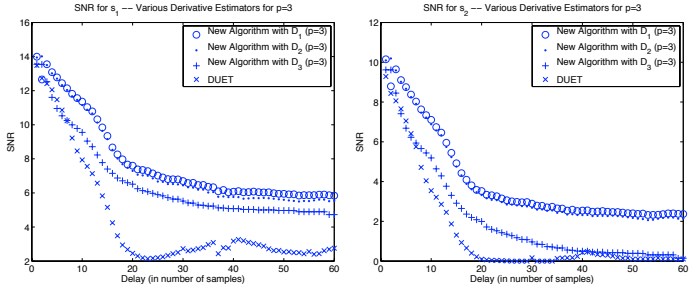

Figure 6: Average SNR for the estimated source $\overline{S}_1$ over 10 experiments per data point. We fix the over-sampling to $p = 3$ and compare the derivative estimators D_1, D_2, and D_3 (as defined in (3.8),(3.9), and (3.10) respectively).

of DUET (where no phase wrap aliasing correction is done) deteriorates rapidly for both sources when d_1 is increased from 1 to 60. In fact, for $d_1 > 7$, the estimated mixtures have very poor quality, and are sometimes unintelligible. With phase aliasing correction, on the other hand, the performance improves significantly (almost by 6 dB) , and even when $d_1 = 60$, which corresponds to a microphone distance of more than 128 cm, both estimated sources are intelligible and still without any major artifacts.

We finish this section by commenting on the effect of oversampling and the use of higher order derivative estimates. As seen in Figure 5, for relatively small values of d_1, e.g., $d_1 < 10$, oversampling bears no benefit (in fact the SNR values corresponding to $p > 1$ are slightly lower when d_1 is in this range). However, if the microphone distance is large, i.e., when $d_1 > 20$, the separation performance improves by almost 2 dB when $p = 3$ or $p = 5$ (no major difference in performance for these two values of p, so we propose to use $p = 3$ as this is computationally less expensive). Finally, the use of higher-order derivative estimates does not seem to make a significant difference in separation performance, see Figure 6.

5 Conclusion

DUET is a technique for blind source separation based on the W-disjoint orthogonality. It is a simple and fast algorithm that can easily be used for dynamic separation of source signals from mixtures. A major problem with DUET is the phase wrap-around aliasing that may lead to inaccurately separated sources with distortions and artifacts. In this paper we have presented a method based on over-sampled FFT and simple modulo arithmetic to effectively overcome the phase wrap-around aliasing. This substantially improves the performance of DUET for blind source separation. Tests on synthetic mixtures as well as real recordings of audio mixtures in moderately echoic environments have shown extremely robust performance and excellent source separation.

Acknowledgment

The authors would like to thank Radu Balan for very stimulating discussions on DUET and its recent developments. The authors would also like to thank Ser Wee, Sean Wu and Na Zhu for their helpful discussions. Yang Wang's research is supported in part by the National Science Foundation, grants DMS-0813750 and DMS-0936830. Yılmaz's work was supported in part by a Natural Sciences and Engineering Research Council of Canada Discovery Grant.

References

[1] R. Balan and J. Rosca. Sparse source separation using discrete prior models. In *Proceedings of Signal Processing with Adaptative Sparse Structured Representations Workshop (SPARS05)*.

[2] A.J. Bell and T.J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7(6):1129–1159, 1995.

[3] J.F. Cardoso and A. Souloumiac. Blind beamforming for non Gaussian signals. *IEE Proceedings F*, 1993.

[4] S. Choi, A. Cichocki, H.M. Park, and S.Y. Lee. Blind source separation and independent component analysis: A review. *Neural Information Processing-Letters and Reviews*, 6(1):1–57, 2005.

[5] A. Cichocki and S. Amari. *Adaptive blind signal and image processing: learning algorithms and applications*. Wiley, 2002.

[6] S.C. Douglas and M. Gupta. Scaled natural gradient algorithms for instantaneous and convolutive blind source separation. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Apr 2007, pp. II-637–II-640.

[7] E. Hoffmann, D. Kolossa, and R. Orglmeister. A Batch Algorithm for Blind Source Separation of Acoustic Signals Using ICA and Time-Frequency Masking. *Lecture Notes in Computer Science*, 4666:480–487, 2007.

[8] A. Hyvärinen and E. Oja. A fast fixed-point algorithm for independent component analysis. *Neural Computation*, 9(7):1483–1492, 1997.

[9] A. Hyvärinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

[10] A. Jourjine, S. Rickard, and Ö. Yılmaz. Blind separation of disjoint orthogonal signals: Demixing n sources from 2 mixtures. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, Jun 2000, pp. 2985–2988.

[11] J. Liu, J. Xin, and Y. Qi. A dynamic algorithm for blind separation of convolutive sound mixtures. *Neurocomputing*, 72(1-3):521–532, 2008.

[12] J. Liu, J. Xin, and Y. Qi. A Soft-Constrained Dynamic Iterative Method of Blind Source Separation. *Multiscale Modeling & Simulation*, 7:1795, 2009.

[13] T. Melia and S. Rickard. Underdetermined blind source separation in echoic environments using DESPRIT. *EURASIP Journal on Advances in Signal Processing*, Article ID 86484, 2007.

[14] N. Murata, S. Ikeda, and A. Ziehe. An approach to blind source separation based on temporal structure of speech signals. *Neurocomputing*, 41(1-4):1–24, 2001.

[15] S. Rickard. The DUET blind source separation algorithm. In *Blind Speech Separation*, Springer, 2007.

[16] S. Rickard, R. Balan, and J. Rosca. Real-time time-frequency based blind source separation. In *Proceedings of ICA-2001*, Dec 2001, pp. 651–656.

[17] S. Rickard and Ö. Yılmaz. On the approximate W-disjoint orthogonality of speech. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2002, pp. 529–532.

[18] R. Saab, Ö. Yılmaz, M. McKeown, and R. Abugharbieh. Underdetermined Anechoic Blind Source Separation via ℓ^q-Basis-Pursuit with $q < 1$. IEEE Transactions on Signal Processing, 55(8):4004–4017, 2007.

[19] R. Sukegawa, S. Uchida, T. Nagai, and M. Ikehara. Blind source separation using correlation at neighboring frequencies. In *Proceedings of International Symposium on Intelligent Signal Processing and Communications*, Dec 2006, pp. 451–454.

[20] M. Swartling, N. Grbic, and I. Claesson. Direction of arrival estimation for multiple speakers using time-frequency orthogonal signal separation. In *Proceedings of IEEE International Conference on Acoustic, Speech and Signal Processing*, May 2006, pp. IV-833–IV-836.

[21] Ö. Yılmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, 2004.

[22] Online demo. `http://www.math.msu.edu/~ywang/DUET.html`