A COMPARISON OF ANALYSIS AND RESYNTHESIS METHODS FOR DIRECTIONAL SEGMENTATION OF STEREO AUDIO

Jeremy J. Wells,

Audio Lab, Department of Electronics,
University of York, YO10 5DD
York, UK
jjw100@ohm.york.ac.uk

ABSTRACT

A comparison of analysis and resynthesis methods for use with a system for dividing time-coincident stereo audio signals into directional segments is presented. The purpose of such a system is to give greater flexibility in the presentation of spatial information when two-channel audio is reproduced. Example applications include up-mixing and transforming panning from amplitude to time-delay based. Included in the methods are the dual-tree complex wavelet transform and wavelet packet decomposition with best basis search. The directional segmentation system and the analysis and resynthesis methods are briefly described, with reference to the relevant underlying theory, figures of merit are presented for each method applied to three stereo mixtures of contrasting material and the subjective quality of the output (with links to all audio examples) is discussed.

1. INTRODUCTION

Audio recordings represent the capture of an acoustic event, or the rendering of an electronic/digital process, at a particular point in time. If there is more than one discrete channel then spatial information can be included in the recorded information. For two-channel stereo recordings, despite the sparsity of the spatial sampling points, a rich spatial experience can be provided for the headphone listener (particularly if the information is binaurally captured), or for (a) person(s) within a small listening area between two loudspeakers in a good listening environment. That said, there are now increased opportunities for surround sound (i.e. more than two-channel) storage, transmission and playback. Also, for individual listeners, the ideal presentation of the spatial information contained within a two-channel stereo audio recording will depend to a certain extent on their own preferences, listening environment and reproduction equipment. As trends in spatial presentation have varied over time, and continue to vary, so there may be a desire to revise the spatial presentation in existing two-channel recordings. Examples such as these require the 'un-locking' of the spatial information for each source (real and virtual) direction. This represents a considerable challenge where there are more source directions than channels.

The purpose of the target system, for which the analysis and resynthesis methods are compared here, is to divide the auditory scene presented by time-coincident (level-panned) audio into directional 'segments' [1]. Having more segments than audio channels offers flexibility in how each segment is presented at two (or more, if up-mixing is the application) loudspeakers. This is the over-arching aim of this research. As such, this work exists between individual source separation, such as that described in [2], and spatial processing (for example [3-5]). The purpose is not necessarily to provide every single instrument separately for re-mixing, but to provide (distinct or overlapping) zones within a two-channel audio scene.

Previously an adaptive analysis/resynthesis method, based on dual-tree complex wavelets, was investigated and compared for use in this system with other methods traditionally used for this type of application [1]. Whilst the complex wavelet packets demonstrated an ability to adapt to the input, the figures of merit (FoM) used in that study demonstrated that they were always out-performed by another method (albeit not always the same one). However their adaptivity did avoid the transient smearing that was exhibited with short-time Fourier transform (STFT) methods with relatively long window lengths. A version of best basis search of complex wavelet packets, which used the available phase information was also investigated but did not consistently offer an improvement in the FoM and, in one case, caused a significant degradation in performance.

The work in this paper expands the range of analysis/synthesis methods used, introduces a regularised version of the phase-weighted best basis search and includes an additional FoM. Since subjective evaluation is also a crucial part of assessing these methods all audio examples used to generate the FoMs are discussed and made available online, as was done for the previous work.

In the next section of this paper an overview of directional segmentation of stereo audio is given and the segmentation system that all of the methods are tested with is described. Section 3 summarises the different analysis/resynthesis methods used and discusses the necessary theoretical detail. In section 4 the experimental design is explained and section 5 presents results for three different two-channel amplitude-panned mixtures. The final section summarises the paper and presents conclusions based on the results.

2. DIRECTIONAL SEGMENTATION OF TWO-CHANNEL AUDIO

2.1. Application examples

Space is represented in stereo recordings by differences between the signals reproduced at each loudspeaker (or earpiece if headphones are used, although only loudspeaker reproduction is considered in this paper). If there are no differences between the signals then the presentation is monophonic. The inter-channel differences may be amplitude (e.g. coincident microphones, typical panning controls), time (e.g. spaced microphones, time-delay

panning) and/or spectral (e.g. binaural with crosstalk cancellation for loudspeaker reproduction). A detailed discussion of the differences between amplitude- and time-difference presentation of audio via loudspeakers has been given previously [1, 6]. The work in those papers, and that presented here, is motivated by the desirability of reconfiguring spatial audio so that the spatial information can be presented in a different way. This work focuses on processing of amplitude panned (or captured) spatial audio.

If directional segments can be extracted from a two-channel mixture then they could be re-panned using time differences instead, therefore changing the presentation of spatial information. To introduce such position dependent delays for each source direction post-recording/mixing, where there are more source directions than channels, requires a separation system. The work described in this paper tests the effectiveness of different time-frequency analysis and synthesis methods when used in such a system.

Another means of changing the presentation is to change the number, or configuration, of loudspeakers. More than two-channels, delivered via the same number of loudspeakers (or more) can improve localisation, create a greater sense of envelopment and increase the size of the listening 'sweet spot'. For soundfield reconstruction systems (such as high-order ambisonics) increasing the number of loudspeakers reduces spatial aliasing. For panning systems (e.g. so called 'pair-wise' positioning of sources) a greater number of discrete channels concentrates sound energy for a single source into a smaller number of speakers (or a smaller area of the array). This improves localisation over a wider listening area. For example, where there are more loudspeakers but just two discrete audio channels available (such as for the playback of legacy two-channel stereo over 5.1 surround systems) then the listening sweet spot may be enhanced (for example by extracting centre source directions and reproducing the audio via all of the front three speakers) or the spatial presentation may be enhanced by the positioning of source directions into rear speakers (e.g. for improved rendering of reverberation). This process is known as 'up-mixing' (e.g. [5]). Again, this process requires some form of separation algorithm in cases where there are more than two source directions.

2.2. Directional segmentation via time-frequency analysis and resynthesis

Time-frequency analysis, and resynthesis, is concerned with the decomposition, and construction, of signals as combinations of individual components that have certain positions and distributions in time and frequency [7]. The time-frequency plane for a signal is the distribution of these components across these two dimensions. An overview of the use of time-frequency analysis and resynthesis for directional segmentation of audio, along with a discussion of important prior work, is given in [1] and the reader is directed there for further information.

The context for the comparison of time-frequency analysis and resynthesis methods which is reported in this paper is a system that is described in detail in [1] and, again, the reader can find more information there. In that paper the possibility of using a phase-weighted entropy measure, in cases where the analysis-resynthesis method was both adaptive and complex, was examined. This phase-weighted entropy measure was given by:

$$H = -\sum_{p=1}^{P} \frac{\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right) \log_2\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right)}{\left|\phi_{\mathrm{R}}(p) - \phi_{\mathrm{L}}(p)\right|} \quad (1)$$

where a and φ are the energy and phase of an individual atom of the decomposition, (L and R designate which spatial channel the atom belongs to) and H is the entropy for a particular basis of P atoms. It was found that this measure did not consistently improve performance. For this paper a regularised version of (1) is employed to investigate whether this improves consistency and/or performance, where r is the regularisation constant:

$$H = -\sum_{p=1}^{P} \frac{\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right) \log_2\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right)}{\left|\phi_{\mathrm{R}}(p) - \phi_{\mathrm{L}}(p)\right| + r} \quad (2)$$

For real packet decompositions this version of the cost function cannot be used since no phase information is available. No best basis search is performed where the analysis-synthesis basis is fixed (i.e. the method is non-adaptive).

As described in [1] overlapping directional windows are used rather than the binary functions that have been commonly used in other studies (e.g. [2]). These directional windowing functions are shown in Figure 1 four equally spaced segments (which is the scenario tested in this paper). In most situations it will be desirable for a segment to be centred on a single source, and encompass that source only. In the case where sources are not regularly spaced, a modified windowing function would be required to ensure that segments are source-centred and preserve energy when combined. This could be achieved by using Hann-like tapering at the ends of constant functions as described in [26].

The windowing functions shown in Figure 1 only fully cover the front and rear quadrants (not the sides) of the recorded space. In anechoic situations where sources are only placed within the front quadrant (as is tested here) then the presence of energy outside of these regions (the residual after separation) indicates that separation has not been completely successful - the lower the energy level in the residual, the more successful the capture of sources within directional segments has been. Therefore the relative amount of energy in this residual is used as an FoM in the results presented in this paper. In echoic situations then this residual may also (correctly) contain reverberation/reflections from the side.

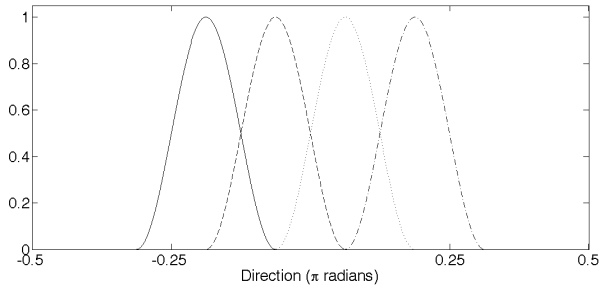


Figure 1: *Directional segmentation windows applied to an audio scene containing four equidistantly and symmetrically spaced sources.*

3. TIME-FREQUENCY ANALYSIS/SYNTHESIS METHODS

This section surveys the different analysis/resynthesis methods which are tested within the system discussed in sub-section 2.2. They can be grouped in two different ways: real and non-redundant versus complex and redundant, or adaptive versus non-adaptive. Since extensive coverage of many of the methods has been provided previously, what is presented here is a short summary of the information in [1], with additional detail on methods which have been used here for the first time.

3.1. Discrete wavelet transform (DWT)

This transform, which is exhaustively covered in the existing literature (e.g. [8]) is characterised by successive high and low pass filtering operations followed by decimation by a factor of two which yields a dyadic division of the time-frequency plane (fixed basis). The DWT is non-redundant, shift-variant and is sometimes referred to as the 'fast' or 'decimated' wavelet transform, to differentiate it from undecimated wavelet transforms (which are redundant). The nature of the wavelet (e.g. its distribution in time-frequency) is determined by the coefficients used in the filters. Four different sets of filter coefficients are used here. The first set are those of Daubechies with six vanishing moments ('db6', 12 tap filters), the second are Daubechies with fourteen vanishing moments ('db14', 28 tap) and the third are those of Vaidyanathan, designed for narrow transition from pass- to stop-band ('vaid', 24 tap) [8]. These filter sets are available either in the Mathworks Wavelet Toolbox, the Wavelab Toolbox or the Dual-Tree Wavelet Packet Toolbox [9-11]. The fourth set has been generated using the Filter Design Toolbox in Matlab (`firpr2chfb` function). These are 48 tap power-symmetric filters. The magnitude response of the low-pass filter is shown in Figure 2 (since the filters are power-symmetric the high-pass response is the exact reverse of that shown in the figure). In the experiments conducted for this paper, the DWT is carried out over eleven stages, yielding an eleven-scale decomposition. All four filter sets are orthogonal (i.e. the synthesis filters are the time reverse of the analysis filters).

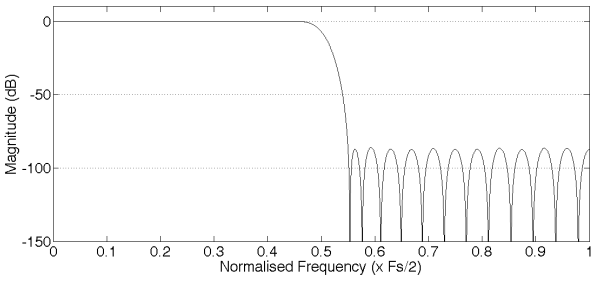


Figure 2: *Magnitude frequency response of the 48 tap filter low-pass filter.*

3.2. Wavelet packet decomposition (WPD)

The wavelet packet decomposition (WPD) is a generalisation of the DWT. Dyadic is just one of many different divisions of the time-frequency plane which are achieved when both low and high pass filtering operations are carried out on each set of coefficients at each decomposition level. A number of different decompositions can be achieved by different combinations of high- and low-pass filtering operations and from these a single decomposition, offering a particular division of the time-frequency plane, can be chosen. Because of the binary tree structure of the decomposition, fast algorithms exist for searching for the best representation (the 'best basis') for a particular signal [12, 13].

The same four sets of filters that are used to implement the DWT are used for the WPD. Although the WPD can be considered to include the DWT, results for the DWT are presented separately in the next section since deriving a DWT only is a much cheaper operation computationally (but the basis is fixed). As for the DWT, the WPD is carried out over eleven scales, dividing the frequency axis into 2048 components for a full packet decomposition at this scale.

3.3. Cosine Packet Decomposition

Local cosine bases given by the Cosine Packet Decomposition (CPD) are also amenable to fast searching for a best basis [8]. The reader is directed to [1] for details of the implementation used in these experiments. The CPD divides the time-frequency plane into time partitions (whose frequency resolution are determined by choice of partition length), as opposed to the WPD, which divides the time-frequency plane into frequency partitions (whose length are determined by the choice of bandwidth) [8]. In both cases, many different combinations of different length (or bandwidth) segments can be chosen to form a number of orthogonal transforms (bases) from which a best basis can be chosen. As for the WPD with best basis, the CPD with best basis gives real coefficients of a non-redundant transform. The CPD is implemented here with the Wavelab toolbox [10]. A 'sine' taper is used and D is chosen so that the shortest packet is 512 samples long, given N. Where necessary the input signal is appended with zeros so that its length N is a power of two.

3.4. Short-time Fourier transform (STFT)

The STFT is perhaps the most widely known and well understood time-frequency analysis-resynthesis method for audio signals. A detailed discussion and description can be found in many sources (e.g. [14]). This method decomposes signals into equal length frames, which can be overlapping and tapered. A discrete Fourier transform (DFT) is applied to each frame and this gives a set of complex coefficients for sinusoids which are harmonics of the frame period, at the centre of each frame. The amount of overlap, and hence redundancy, can be arbitrarily set but is constrained by the shape of the tapering window applied to the frame (e.g. a minimum 50% overlap is required for the Hann window) and the distance from the centre of one frame to the next cannot be more than the frame length itself. Although the STFT can be non-redundant, tapering is usually applied to prevent energy spreading due to discontinuities at frame boundaries, and this renders the STFT redundant. For example, an overlap of 50% yields an STFT with 100% redundancy (providing zero-padding is not used). For the work described in this paper two sets of five STFT types are employed: one set applies a Hann window with 50% overlap prior to the DFT but no windowing of the output of the inverse DFT (IDFT), the second set has a 75% overlap and a Hann window is applied prior to DFT and after IDFT (where a Hann window is applied twice, the minimum overlap is 75%). Within each STFT set five frame lengths are used: 512, 1024, 2048, 4096 and 8192 samples. The frames are not zero-padded prior to analysis.

3.5. Dual-Tree Complex Wavelet Transform (DT-CWT)

The Dual-Tree Complex Wavelet Transform (DT-CWT) of Kingsbury is an extension of the DWT whereby a signal is decomposed by two sets of basis functions for which each corresponding pair of functions are approximately Hilbert transforms of each other [15]. As a result of this approach the DT-CWT is 100% redundant and approximately shift invariant. The Q-shift method of achieving approximate analyticity is used to determine the filter coefficients for level two of the decomposition onwards [16]. A different set of filter coefficients is used for the first stage of the transform: this filter set is used for both 'trees' with a one sample relative delay. At subsequent stages the Q-shift (quarter sample delay) filter set is used in both trees. In the second tree these filter coefficients are used in reverse order, giving a three-quarter delay and, therefore, the half sample relative delay between trees needed for analyticity. The longer the Q-shift filters are, the closer the two sets of basis functions are to being Hilbert transform pairs. Four sets of filter coefficients are used here to implement the DT-CWT: 'db5' (first stage) followed by the 14-tap Q-shift filter coefficients given in Table 2 of [15], 'db14' followed by the same 14-tap Q-shift filter coefficients, 24 tap Vaidyanthan followed by 24 tap Q-shift filters and, finally, the 48 tap filters described at the end of Section 3.1 followed by 48 tap Q-shift filters. The last two sets of Q-shift filters were designed using the Q-shift filter design toolbox [17]. For comparison the magnitude response of the 14, 24 and 48 tap Q-shift filters are shown in Figure 3. As for the real DWT, the number of scales in the following experiments is eleven.

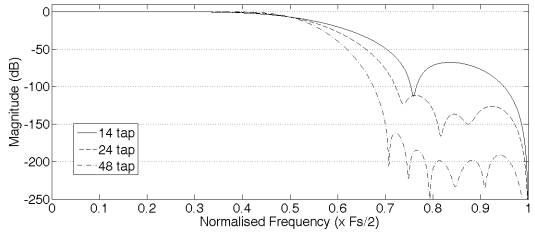


Figure 3: *Magnitude frequency responses of each low-pass Q-shift filter.*

3.6. Dual-Tree Complex Wavelet Packet (DT-CWPD)

The Dual-Tree Complex Wavelet Packet Decomposition (DT-CWPD) is the complex equivalent of the WPD, in the same way that the DT-CWT is the complex equivalent of the DWT. It yields bases with 100% redundancy. Since the DT-CWT consists of two orthogonal decompositions of the same signal, a straightforward approach to deriving a wavelet packet decomposition is to treat the two 'trees' as completely independent with their own sets of filters, where, after the first decomposition stage, the set used in one tree is the time-reverse of the set used in the second tree (as is the case for the DT-CWT). However 'analyticity' is better preserved by an altered scheme where some of the filtering stages of both trees use the same filters [18]. This scheme is employed here for the DT-WPD and it is implemented using (with some modifications) the toolbox provided at [19]. The same filter sets are used as for the DT-CWT (except that the first stage filter is 'db5' rather than 'db6', although it is replaced with 'db6' when non Q-shift filters are used in subsequent stages, see [19]). In fact, the first two filter sets are the same as those provided as examples at [19]. The maximum decomposition level is, again, eleven.

4. COMPARISON OF ANALYSIS/RESYNTHESIS METHODS

In order to compare the methods described in Section 3, they are tested using the system discussed at the end of Section 2. They are tested with three different anechoic audio mixtures, ranging from two to seven seconds in length, each containing four equally spaced point sources. The use of mixtures of anechoic sources allows the Signal to Residual Ratio (SRR, the ratio of energy in the residual segment to the energy contained in all of the other segments) to be used as an FoM. As for the experiments described in [1], for the purposes of this test the source positions for each mixture are the same and are known *a priori*. Whilst *a priori* knowledge of source positions is unlikely to be available in real-world applications it is the ability of the decomposition methods for segmentation which is specifically being tested here. In practice, *a posteriori* knowledge of source positions could be gained from global statistics for the mixture, such as the 'panogram' described in [5]. Each mixture contains four sources ($src_{1\text{-}4}$ and each of these are panned to the left and right outputs (out_L, out_R) of the mixture via:

$$\begin{pmatrix} out_L \\ out_R \end{pmatrix} = \begin{pmatrix} .8341, .5995, .4005, .1659 \\ .1659, .4005, .5995, .8341] \end{pmatrix} \begin{pmatrix} src_1 \\ src_2 \\ src_3 \\ src_4 \end{pmatrix} \quad (3)$$

This mixing matrix gives the same ratio between left and right energy that would occur for four sources spaced equidistantly in an arc within the front quadrant of a coincident pair of dipole microphones at 90 degrees to each other: sources positioned at –33.75 degrees (-3π/16 radians), -11.25 (-π/16), 11.25 (π/16) and 33.75 (3π/16) from the centre of the front quadrant. The centres of the windows shown in Figure 1 are at these positions and each position is covered by that one window only (at the centre of one window, the other three windows are at zero).

4.1. Mixture 1: pitched instruments

The individual sources for this mixture are clarinet, violin, soprano singer and viola performing an excerpt from a Mozart opera. The sources are obtained from [20].

4.2. Mixture 2: speech babble

This is a combination of four speakers talking simultaneously. The mixture comprises two male adults, one female adult and one male child. The sources are obtained from [21].

4.3. Mixture 3: percussion with single pitched instrument

This mixture consists of three hand percussion instruments and a single note with swept pitch from a Shakuhachi-like instrument. The sources are obtained from [22].

4.4. Figures of merit (FoM)

The quality of the segmentations is objectively measured by four quantities for each separated source: the energy weighted inter-channel correlation, the signal to residual energy ratio (SRR), the azimuth error and the signal to distortion ratio (SDR). The SDR is described in [23] and can be evaluated using the BSS_Eval Toolbox [24]. It compares the separated sources with the original

un-mixed sources and attempts to measure the ratio of the actual source energy to the energy due to artefacts of the separation algorithm and interference from other sources. It requires prior knowledge of the individual sources, which is available here. The SDR is designed for monophonic separated sources so it is applied here to the sum of each channel of the stereo separated outputs.

The other FoMs were introduced and used in [1] and so are only briefly summarised here. The zero-lag inter-channel cross-correlation between two channels for a single point source will be 1.0 since there are identical signals at each channel (albeit with different gains, if not positioned centrally) and there is no relative delay between then. Therefore, the closer this FoM is to 1.0, the better this segment has captured audio from one source only. The zero-lag cross correlation is given by:

$$X = \frac{\mathbf{src'}_\mathbf{L} \cdot \mathbf{src'}_\mathbf{R}}{|\mathbf{src'}_\mathbf{L}||\mathbf{src'}_\mathbf{R}|} \tag{4}$$

where $\mathbf{src'}_\mathbf{L}$ and $\mathbf{src'}_\mathbf{R}$ are vectors containing the samples of the left and right channels of the segmented source. An overall FoM for all of the separated sources is given by the energy weighted mean of X of the sources. Whilst the SDR and the cross-correlation give an indication of the quality of the segmentation, the SDR does not take account of gain errors and the cross-correlation does not take account of gain or frequency response errors (it just measures the localisation of energy for a source – not how it is distributed in frequency). For anechoic sources the relative level of energy in the residual segment is an indicator of how successful the segmentation is in capturing the elements of the signal. The Signal to Residual ratio (SRR, measured in dB) is the ratio of the residual energy to the energy in the input mixture. The azimuths of individual separated sources can be calculated using

$$\theta' = \operatorname{sgn}\left(|\mathbf{src'}_\mathbf{R}| - |\mathbf{src'}_\mathbf{L}|\right) \operatorname{arccot}\left(\frac{|\mathbf{src'}_\mathbf{R}| + |\mathbf{src'}_\mathbf{L}|}{|\mathbf{src'}_\mathbf{R}| - |\mathbf{src'}_\mathbf{L}|}\right) \tag{5}$$

and from this the azimuth error can be found, since actual the source directions are known. The energy-weighted mean azimuth error for all sources is an indicator of the extent to which segments are contaminated by each other, since azimuths will be biased by the presence of energy from other sources.

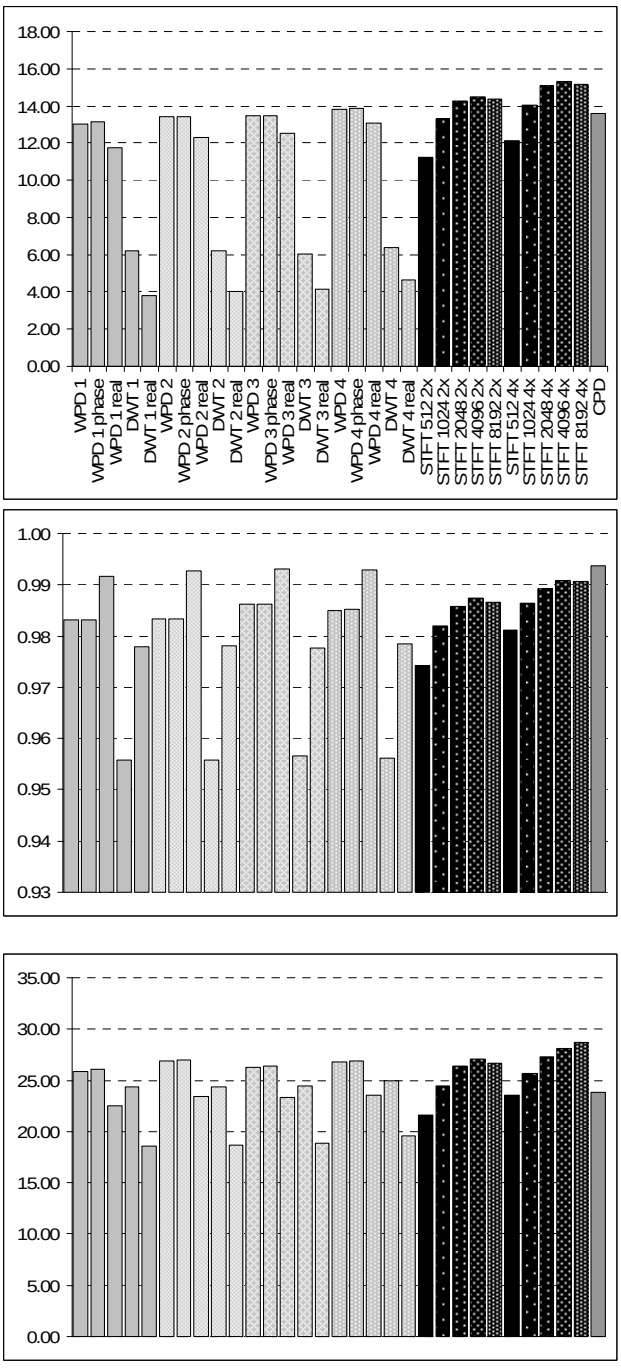
5. RESULTS

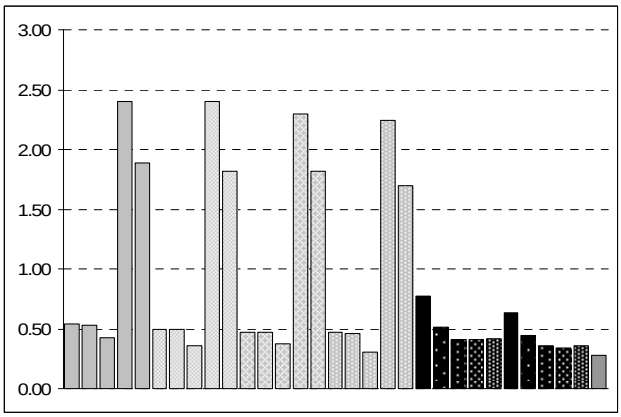
Three sets of plots are presented, one for each mixture. Within each set there are four plots which compare the performance of the different analysis and resynthesis methods for each FoM. The following abbreviations are used:

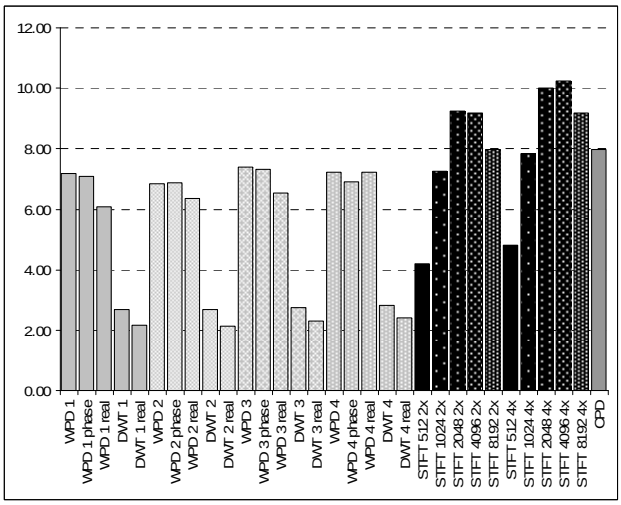
DT-CWPD 1, DT-CWT 1: 14 tap Q-shift filters, non Q-shift filters are 'db5' at the first stage, 'db6' thereafter.
DT-CWPD 2, DT-CWT 2: 14 tap Q-shift filters, non Q-shift filters are 'db14' at all stages.
DT-CWPD 3, DT-CWT 3: 24 tap Q-shift filters, non Q-shift filters are 24 tap Vaidyanathan filters.
DT-CWPD 4, DT-CWT 4: 48 tap Q-shift filters, non Q-shift filters are 48 tap power-symmetric filters.
WPD 1, DWT 1: 'db6' filters.
WPD 2, DWT 2: 'db14' filters.
WPD 3, DWT 3: Vaidyanathan 24 tap filters.
WPD 4, DWT 4: 48 tap Q-shift filters, non Q-shift filters are 48 tap power-symmetric filters.
2x: STFT with 50% overlapping windows
4x: STFT with 75% overlapping windows

'Phase' indicates that the best basis has been determined using equation (2), rather than (1). The value of r, heuristically determined, is set at 0.01 for all mixtures. For each set of figures, the x-axis labels, which indicate the type of analysis/synthesis method under test, are provided in the first of the four plots.

5.1. Figures of merit







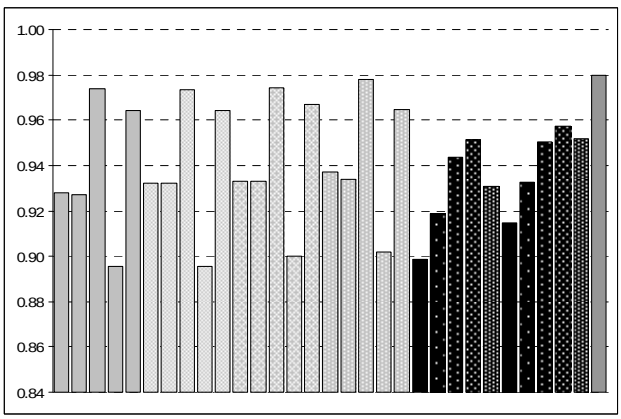
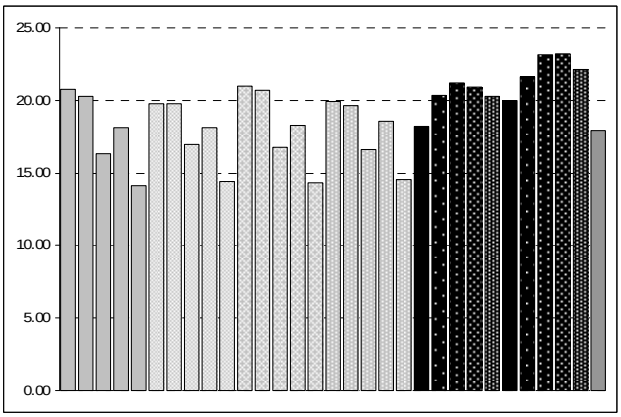


Figure 4: *FoM for the instrument mixture: SDR (dB, top of previous page), correlation (middle of previous page), SRR (dB, bottom of previous page), azimuth error (radians, above)*



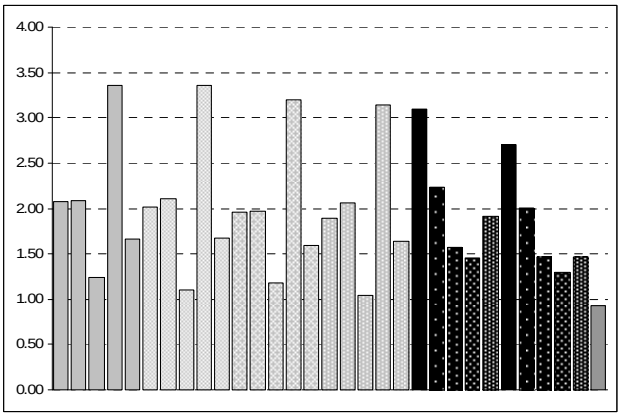
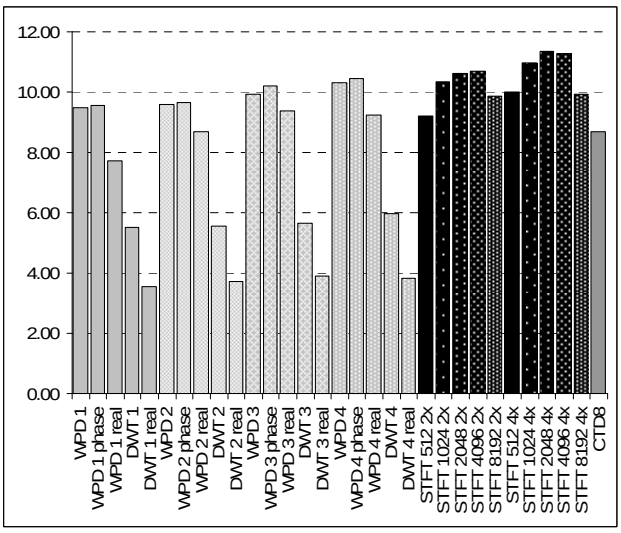


Figure 5: *FoM for the speech mixture: SDR (dB), correlation, SRR (dB), azimuth error (radians)*



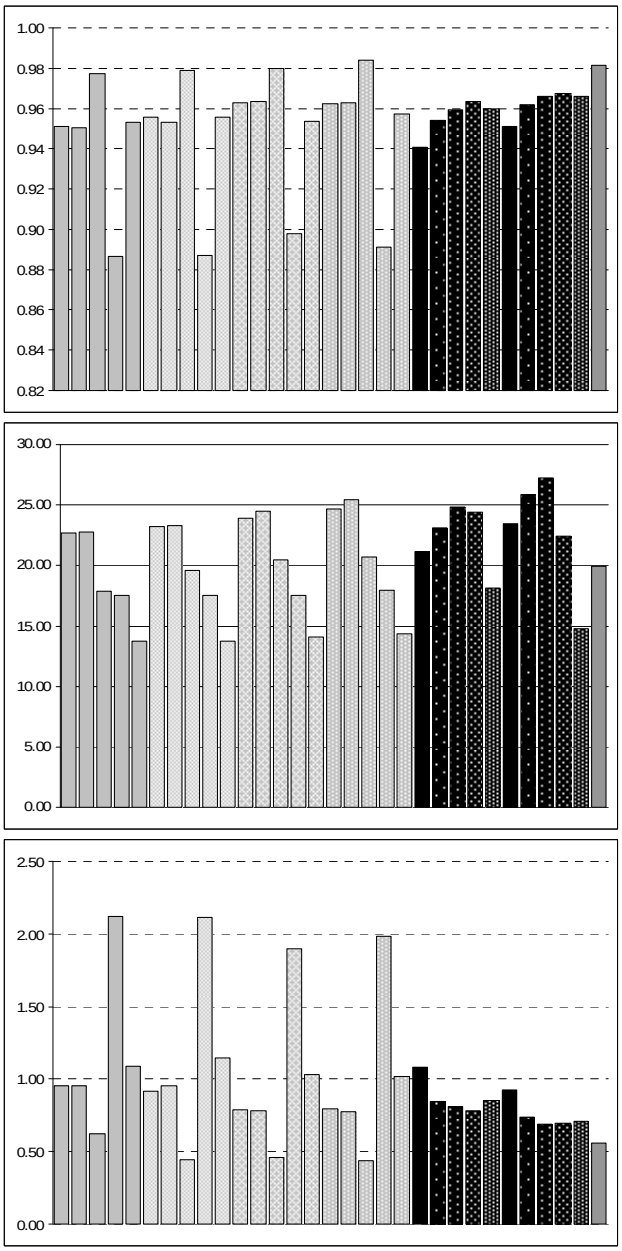


Figure 6: *FoM for the percussion mixture: SDR (top), correlation, SRR, azimuth error (bottom)*

5.2. Online audio examples

Audio files of the original sources, mixtures and separated sources for each method are provided in an online archive so that they can be auditioned [25].

5.3. Discussion

Some clear trends can be seen in Figures 4-6. Redundancy, whether achieved through introducing a second orthonormal transform whose basis functions are an approximate Hilbert transform pair with the first, or by increasing the overlap of basis functions improves the SDR performance of these methods for directional segmentation: STFTs using 75% overlapping windows achieve better results than those using 50% overlap and the complex DWT or WPD always outperforms its real counterpart. Whilst 'real' methods do relatively well in terms of cross-channel correlation and azimuth error, they perform poorly in terms of SRR and their SDR performance is markedly worse than complex versions of the same methods in many cases. This shows that real analysis methods produce individual sources which have close to the correct azimuth and have narrow width, but this is at the cost of additional energy appearing in the residual.

The STFT with 75% overlap achieves the best FoMs for all mixtures. The 4096 frame-length STFT is best for the pitched instrument and speech mixtures, the 2048 frame-length version doing slightly better for the percussion mixture. The DT-CWPD using the fourth filter set performs best in terms of SDR out of the wavelet methods for all except the speech mixture. However it is out-performed by the CPD for all but the percussion mixture. As was found in [1], the use of phase-weighting in the entropy measurement for the best basis search does not have a dramatic positive impact on the FoMs. However the incorporation of a regularisation constant (not employed in [1]) does improve the consistency of phase-weighting overall (preventing serious anomalous degradations as occurred in [1]). Overall it is also more effective than the non-phase weighted measure, but the difference in performance is insufficient to be conclusive.

Listening to the audio outputs for the percussion mixture the drawback of long frame-length STFT analysis and resynthesis is clearly audible: transient smearing is much worse (although the separation is audibly better) than it is, for example, for the DT-CWPD with the filter set 3. The CPD performs well in the first half of the separation but then time definition is lost completely. Although transient smearing is both time-varying gain and spectral change, both of which the SDR should penalise, it does not have much impact on this FoM. It is worth noting here that in [2] the maximum STFT size was limited to 1024 because of the damage that longer frame sizes did to note onsets.

The longer-frame STFT methods audibly perform very well on speech and the pitched instrument mixture, although occasionally consonants and note onsets are degraded. Applying a window to the output of the IDFT, as well as the input to the DFT, is helpful in removing annoying ticks that are due to end-of-frame discontinuities introduced by the segmentation process.

6. CONCLUSIONS AND FUTURE WORK

This paper, along with its accompanying online resource of audio examples, has presented a comparison of a number of different time-frequency analysis/resynthesis methods for use in directional segmentation. The FoMs used clearly indicate that long-frame STFT methods with relatively high redundancy work best, although audition of the segmentations, particularly for percussion, provide a caution about using such objective measures as a sole indicator of quality. Whilst the dual-tree versions of the wavelet methods perform better than their real counterparts, and complex packets with long filters (including Q-shift) generally perform best, they do not begin to compete (numerically at least) with the STFT (or the CPD, considering just the speech mixture).

It is highly desirable to have an adaptive method that can perform as well as the STFT and there are many parameters and possibilities of the DT-CWPD that have yet to be fully investigated. Filters of 48 taps may still be too short for general audio applications and the benefit of phase-weighting may become

more apparent with longer Q-shift filters. The development of an adaptive method which can match the STFT's performance within the system, and on the example mixtures, tested here, remains a challenge. However the challenge is a worthwhile one, given the potential benefits of high-quality directional segmentation. Of course, some consideration should also be given to computational cost, and more redundant methods are usually more expensive. But, for this application, redundant time-frequency representations seem to perform best overall.

7. REFERENCES

[1] J. Wells, "Directional Segmentation of Stereo Audio via Best Basis Search of Complex Wavelet Packets", presented at the 130th Audio Engineering Society International Convention, London, UK, 2011Convention, Preprint No. 8436.

[2] A. Nesbit et al., "Audio Source Separation with a Signal-Adaptive Local Cosine Transform", *Signal Processing*, Vol. 87, No. 8, August 2007, pp. 1848-1858.

[3] V. Pulkki, "Spatial Sound Reproduction with Directional Audio Coding", *Journal of the Audio Engineering Society*, Vol. 55, No. 6, June 2007, pp. 503-516.

[4] C. Faller, "Modifying the Directional Responses of a Coincident Pair of Microphones by Postprocessing", *Journal of the Audio Engineering Society*, Vol. 56, No. 10, October 2008, pp. 810-822.

[5] C. Avendano and J. Jot, "A Frequency Domain Approach to Multichannel Upmix", *Journal of the Audio Engineering Society*, vol. 52, No. 7/8, July/August 2004, pp. 743-749.

[6] J. Wells, "Modification of Spatial Information in Coincident Pair Recordings", presented at the 128th Audio Engineering Society International Convention, London, UK, 2010, Preprint No. 7983.

[7] L. Cohen, Time-Frequency Analysis, Prentice Hall, 1994.

[8] S. Mallat, A Wavelet Tour of Signal Processing, 2nd edition, Academic Press, 1999.

[9] M. Misiti et al., "Wavelet Toolbox User's Guide", available from http://www.mathworks.com

[10] Donoho, D. et al, "The WaveLab Matlab Toolbox", available from: http://www-stat.stanford.edu/~wavelab/

[11] Bayram, I., "The Dual-Tree Complex Wavelet Packet Transform Matlab Toolbox", available from http://web.itu.edu.tr/~ibayram/dtcwpt/

[12] M. Wickerhauser, *Adapted Wavelet Analysis from Theory to Software*, A K Peters, 1994.

[13] R. Coifman and M. Wickerhauser, "Entropy Based Algorithms for Best Basis Selection", *IEEE Transactions on Information Theory*, Vol. 38, No.2, pp. 713-718, 1992.

[14] U. Zölzer, Ed., *DAFX – Digital Audio Effects*, J. Wiley & Sons, 2002.

[15] I. Selesnick et al, "The Dual-Tree Complex Wavelet Transform", *IEEE Signal Processing Magazine*, pp. 123-151, 2005.

[16] N. Kingsbury, "Complex Wavelets for Shift Invariant Analysis and Filtering of Signals", *Journal of Applied and Computational Harmonic Analysis*, Vol. 10, No. 3, pp. 234-253, 2001.

[17] N. Kingsbury, "Complex Wavelet Design Package", available from http://www-sigproc.eng.cam.ac.uk/~ngk/

[18] I. Bayram and I. Selesnick, "On the Dual-Tree Complex Wavelet Packet and M-Band Transforms", *IEEE Transactions on Signal Processing*, Vol. 56, No. 6, pp. 2298-2310, 2008.

[19] I. Bayram, "The Dual-Tree Complex Wavelet Packet Transform Matlab Toolbox", available from http://web.itu.edu.tr/~ibayram/dtcwpt/

[20] T. Lokki et al., "Anechoic Recordings of Symphonic Music", available at http://auralization.tkk.fi/

[21] Howard, D. et al., CD of audio examples accompanying Howard, D. and Angus, J., *Acoustics and Psychoacoustics* (3rd Edition), Focal Press, London, 2006.

[22] Various, *Roland Sampling Showcase*, Time and Space Audio CD, 1994.

[23] E. Vincent et al., "Performance measurement in blind audio source separation", *IEEE Transactions on Audio, Speech and Language Processing*, Vol. 14, No. 4, pp. 1462 – 1469, 2006.

[24] C. Févotte et al., "BSS_Eval Toolbox User Guide Revision 2.0", available at http://bass-db.gforge.inria.fr/bss_eval/user_guide.pdf

[25] Audio examples and Matlab code for this paper is available at: Audio examples for this paper available at: www.jezwells.org/directional_segmentation.

[26] A. Master, "Stereo Music Source Separation via Bayesian Modeling", PhD Thesis, Stanford University, USA, 2006.