Audio Engineering Society

Convention Paper

Presented at the 130th Convention
2011 May 13–16 London, UK

Directional Segmentation of Stereo Audio via Best Basis Search of Complex Wavelet Packets

Jeremy Wells

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

ABSTRACT

A system for dividing time-coincident stereo audio signals into directional segments is presented. The purpose is to give greater flexibility in the presentation of spatial information when two-channel audio is reproduced. For example, different inter-channel time shifts could be introduced for segments depending on their direction. A novel aspect of this work is the use of complex wavelet packet analysis, along with 'best basis' selection, in an attempt to identify time-frequency atoms which belong to only one segment. The system is described, with reference to the relevant underlying theory and the quality of its output for the best bases from complex wavelet packets is compared with methods based on more established analysis and processing methods.

1. INTRODUCTION

For the individual listener, 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 and their own environment and reproduction equipment. As trends in spatial presentation vary over time so there may be a desire to revise the spatial presentation in existing two-channel recordings. Examples such as these of the desirability of flexibility in two-channel audio presentation motivate the work described in this paper. The purpose of the system described is to divide the auditory scene presented by time-coincident (level-panned) audio into directional 'segments'. Having more segments than audio channels offers flexibility in how each segment is presented at the loudspeakers and this is the over-arching aim of this research. As such, this work exists between individual source separation, such as that described in [1], and spatial processing (for example [2-4]). 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.

For a narrow segment, or for a single point-like source within a larger segment, the audio contained within it will have very high inter-channel correlation (as it is time coincident) and the ratio of the channel energies will correspond with the spatial position of the centre of that segment (or the point source within it). To produce more segments than audio channels a time-frequency or time-scale analysis (or a combination of the two) is performed on the audio. This analysis is only perfectly successful when each time-frequency/scale atom belongs to one, and only one, segment. Where sources in different segments are overlapping in time and frequency (as is the case in rhythmic and tonal music which uses harmonic sounds and/or broad-band percussion) then such disjointedness in a fixed time-frequency representation is unlikely to occur. In most situations, the optimal time-frequency/scale representation will be one which adapts to the audio scene.

An adaptive method is investigated for directional segmentation in this paper and compared with other analysis methods used previously for similar applications. Signal analysis is performed using the Dual-Tree Complex Wavelet Transform [5, 10]. This transform is used to obtain a packet decomposition from which a best basis is chosen which is deemed to represent the best adaptation to the signal under analysis. Complex packet analysis offers the possibility of using the phase difference between channels for each packet (atom) and this paper explores whether such information can improve the basis selection for the task of directional segmentation.

In the next section of this paper the nature and purpose of directional segmentation of two channel audio is discussed. Section 3 briefly surveys the different time-frequency\scale analysis methods which are then compared in later sections. 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

Stereo amplitude panning and\or coincident microphone techniques encode spatial information as level differences between the left and right channels. Whilst these level differences do translate into time-of-arrival (TOA) differences at the ears of the listener when replayed over loudspeakers (due to path length differences between each ear and each speaker), spatial information is *not* encoded as time differences between the channels for purely amplitude-panned\coincident recordings. An alternative approach to representing space is to actually introduce a time shift between channels. This is can be achieved with microphones in a spaced configuration or by using inter-channel delays for individual sources. A more detailed discussion and comparison of these techniques can be found in [6].

The question of which approach to capturing and encoding spatial information is not closed. It is sufficient here to say that, given the wide range of recording situations and personal tastes for spatial presentation, some flexibility in the representation of space within two-channel recordings is desirable. However, offering such flexibility is far from trivial where the number of source positions exceeds the number of channels (i.e. is greater than two, in this case). To translate level into timing differences requires a separated signal for each point in the transverse plane, since each will require a different inter-channel gain and delay to be applied. The presence of early reflections and reverberation further complicates the problem. Although the reproduced signal will ultimately arrive at an array of only two sensors (the ears of the listener) the introduction of inter-channel delays may require the exposure of individual sounds which have previously been only been heard as a component in a temporally coincident mixture. Where the number of source positions exceeds the number of channels at any point in the time-frequency plane then individual signals for each source cannot be exactly determined. An atomic decomposition of a signal describes it as the energy (and possibly phase) coefficients of a set of atoms: translated and modulated (time-frequency) or translated and scaled (time-scale) functions or sets of functions, or a combination of the two. If a division of the time-frequency\scale plane can be achieved where no more than one source direction contributes to any one atom then perfect directional separation can be achieved, even where there are more source positions than channels. Different combinations of different types of sources will overlap in different regions of time-frequency\scale and therefore adapting the set of functions to match the signal will offer some choice in how atoms map to source directions.

An example of the segmentation of two-channel audio is described in [4]. The main purpose of the segmentation in that work is to generate signals for reproduction over

a larger number of speakers than there are signal channels, a process known as up-mixing, although other applications such as 're-panning' are also described. Processes for extraction of ambience and directional segmentation are detailed. The segmentation begins with the derivation of a 'panogram' from the two-channel audio which is used to determine the likely positions of sources. A narrow Gaussian window is centered on each of the estimated source positions. This window is used to apply a direction dependent gain function to the coefficients of the time-frequency representation, which in this case is the short-time Fourier transform (STFT). From this modified windowed set of coefficients audio representing the directional source is synthesized. For up-mixing this process is preceded by ambience extraction which searches for atoms with low inter-channel coherence and synthesizes an estimate of the ambience from this. This synthesized ambience is then decorrelated from the ambience remaining in the front channels and delayed before being delivered via the surround loudspeakers.

3. SEGMENTATION WITH COMPLEX ADAPTIVE TRANSFORMS

The work presented here investigates alternative atomic decompositions to the STFT used in [4]. Since the output of the STFT is complex the phase can be used to indicate whether more than one component is contributing to an atom in the decomposition. For purely level difference stereo there will be no inter-channel phase difference for an atom containing energy from only one source direction. However, where there are significant contributions to a single atom from more than one direction the phase difference will not be constant. Despite this information being available, the size of STFT atoms is fixed once the window size has been chosen: there is no multiple, packet-like decomposition from which a best energy-preserving basis can be chosen.

The wavelet packet decomposition (WPD) is a generalisation of the discrete wavelet transform (DWT). The DWT yields a fixed, dyadic (hence time-scale) division of the time-frequency plane. An overview of the successive 'low\high pass filtering followed by decimation' operations with which the DWT is calculated are shown in Figure 1. The dyadic division of the time-frequency plane of the DWT is just one of many different divisions of the plane offered by the WPD. Figure 2 gives an overview of the filtering and decimation operations which provide the full tree of the WPD from which a particular basis, offering a particular division of the time-frequency plane, can be chosen. Because of the tree structure of the decomposition fast algorithms exist for searching for the best representation (the 'best basis') for a particular signal [7, 8].

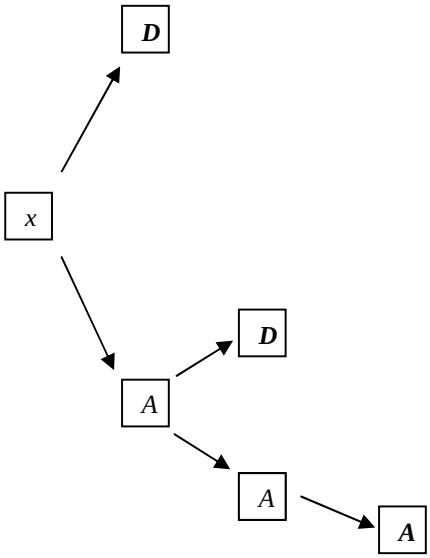

Figure 1: Filtering and decimation operations for a 3 scale DWT. A downward arrow represents a low-pass filter followed by decimation by a factor of 2. An upward arrow represents a high-pass filter followed by decimation by a factor of 2. The signal is successively decomposed into detail (*D*) and approximation (*A*) coefficients at each scale. Those sets of coefficients marked in bold form the dyadic basis of the DWT.

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. As a result of this approach the DT-CWT is 100% redundant and approximately shift invariant. 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 (as is the case for the DT-CWT). However 'analyticity' (the extent to which each of the function pairs are Hilbert transforms of each other) is better preserved by an altered scheme where some of the filtering stages of both trees use the same filters [12]. This is the approach used in the experiments for this

paper and is referred to as the Dual-Tree Complex Wavelet Packet Decomposition (DT-CWPD).

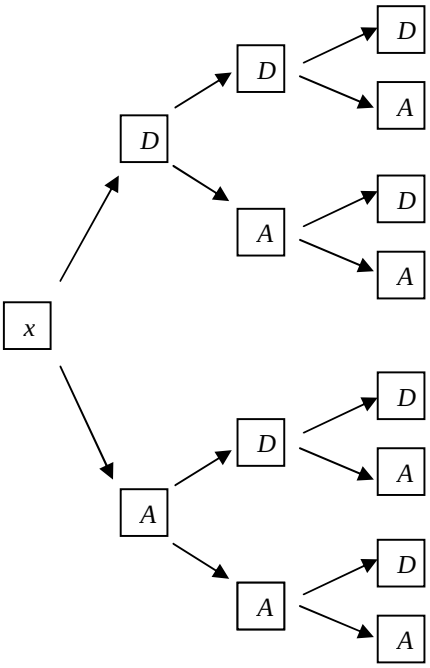

Figure 2: Filtering and decimation operations for a 3 scale WPD. For this decomposition there are 2^3 possible bases.

Local cosine bases given by the Cosine Packet Transform (CPD) are also amenable to fast searching for a best basis [9]. These bases are constructed from windowed cosines of the form (using the notation given in [1]):

$$C[n] = w[n]\sqrt{\frac{2}{2^{-d}N}}\cos\left(\pi\left(k+\frac{1}{2}\right)\frac{n-c_{pd}}{2^{-d}N}\right) \qquad (1)$$

where

$$c_{pd} = 2^{-d}Np - \frac{1}{2} \qquad (2)$$

and N is the length of the input signal (and must be a positive integer power of 2), p is the position of a function, d is the decomposition depth and D is the maximum depth. The length of a basis function at depth d is $N2^{-d}$ and so this basis divides the time axis into partitions belonging to $N2^{-d}\ \ d \in [1,2,..,D]$. Thus the CPD divides the time-frequency plane into time partitions (whose frequency resolution are determined by choice of partition length), whereas the WPD divides the time-frequency plane into frequency partitions (whose length are determined by the choice of bandwidth) [9].

4. COMPARISON OF DECOMPOSITION METHODS FOR DIRECTIONAL SEGMENTATION

The purpose of the experiment described in this paper is to investigate the directional segmentation properties of the three decomposition methods described in the previous section. In order to make a comparison between these different methods they are all tested in the same segmentation framework and on the same two-channel anechoic signal mixtures.

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 [4]. 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:

$$\begin{pmatrix}\text{out}_\text{L}\\ \text{out}_\text{R}\end{pmatrix} = \begin{pmatrix}.8341, .5995, .4005, .1659\\ .1659, .4005, .5995, .8341]\end{pmatrix}\begin{pmatrix}\text{src}_1\\ \text{src}_2\\ \text{src}_3\\ \text{src}_4\end{pmatrix} \qquad (3)$$

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 process described here is designed to divide the audio scene into K+1 segments, where K is the number of segments in the front quadrant and the additional segment contains any residual energy not assigned to the others. The segments are orthogonal to each so that, where no processing is applied individually to segments

prior to recombination, the output is identical to input. This is different to the approach taken in [1], for example, where non-orthogonal segmentations of the soundfield are taken, and different orthogonal bases are used for each source, in order to achieve the best quality for individual sources heard separately. The algorithm proceeds as follows:

1. One of the time-frequency decompositions under test is performed on each channel of the mixture separately.

2. For each left-right atom pair in the decomposition the direction is estimated via:

$$\theta = \mathrm{sgn}\left(a_{\mathrm{R}} - a_{\mathrm{L}}\right)\mathrm{arccot}\left(\frac{\sqrt{a_{\mathrm{R}}} + \sqrt{a_{\mathrm{L}}}}{\sqrt{a_{\mathrm{R}}} - \sqrt{a_{\mathrm{L}}}}\right) \quad (4)$$

where *a* is the atomic energy.

3. Where a packet decomposition has been performed the best basis is searched using the Shannon entropy as the cost function:

$$H = -\sum_{p}^{P}\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right)\log_2\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right) \quad (5)$$

where *H* is the entropy for a particular basis of *P* atoms. The DT-CWPD is also tested using a cost function which is weighted by the phase difference between atoms within a pair. Here the phase-weighted entropy for a basis is given by:

$$H = -\sum_{p}^{P}\frac{\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right)\log_2\left(a_{\mathrm{L}}(p) + a_{\mathrm{R}}(p)\right)}{\left|\phi_{\mathrm{R}}(p) - \phi_{\mathrm{L}}(p)\right|} \quad (6)$$

For the STFT the basis is fixed and cannot adapt to the signal.

4. Once the basis has been determined then the atoms for that basis are segmented into *K* two-channel segments. For each segment centered at a particular source angle, segmentation is performed by a Hann window centered at that source position:

$$a_{\mathrm{L},k} = \begin{cases} a_{\mathrm{L}}\left(1 + \cos\left(2K\theta - \Theta_k\right)\right), \left|\theta - \Theta_k\right| < \dfrac{\pi}{2K} \\ 0, \text{ otherwise} \end{cases} \quad (7)$$

$$a_{\mathrm{R},k} = \begin{cases} a_{\mathrm{R}}\left(1 + \cos\left(2K\theta - \Theta_k\right)\right), \left|\theta - \Theta_k\right| < \dfrac{\pi}{2K} \\ 0, \text{ otherwise} \end{cases} \quad (8)$$

where *k* is the source index and Θ_k is the position of that source (known *a priori*). These directional windowing functions are shown in Figure 3 for $K = 4$ (the situation tested in this experiment). Clearly, where sources are not equally spaced, a modified windowing function would be required to ensure segments are source-centered and preserve energy when combined, possibly using Hann-like tapering at the ends of constant functions.

5. The left and right channels for each separated source are then resynthesized.

6. The windowing functions in equations (7-8) and shown in Figure 3 only fully cover the front quadrant. Where the source separation has not been entirely successful then there will be energy outside of these regions which is the 'residual after separation'. This residual can be obtained by subtracting the sum of the separated sources from the input signal. The lower the energy level in the residual, the more successful the capture of sources within directional segments has been.

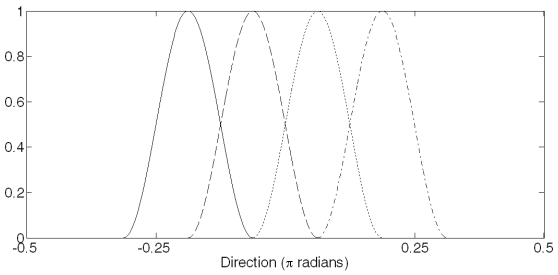

Figure 3: Directional windows applied to atoms for source 1 (solid line), source 2 (dashed), source 3 (dotted) and source 4 (dot dashed).

5. RESULTS

5.1. Decompositions

The segmentation algorithm described in the previous section has been tested with the following time-frequency decompositions:

5.1.1. DT-CWT

This is performed using the filters and algorithms provided in the toolbox for the DT-CWPD [13]. Two sets of filters are provided based around the Daubechies wavelets with 6 and 14 vanishing moments respectively (db6 and db14) [9]. The algorithm is tested with both of these filters (DT-CDWT short, using db6, and DT-CDWT, using db14). The decomposition is performed up to and including scale 11. Figure 4 shows the real and imaginary wavelets for DT-CDWT short at scale 11. No best basis search is possible with the DT-DWT although its basis functions are included in the DT-DWPD.

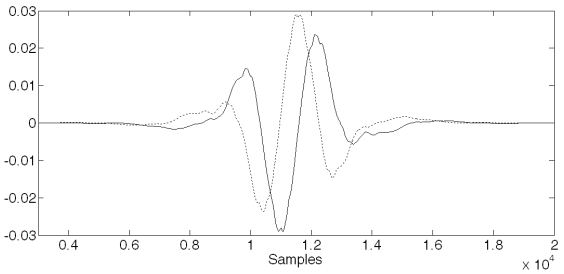

Figure 3: Real (solid line) and imaginary (dotted) wavelets for DT-CDWT short at scale 11.

5.1.2. DT-CWPD

Again, this is performed using both sets of filters provided in [13] and the decomposition is performed up to and including scale 11, dividing the frequency axis into 2048 partitions for wavelet packets at this scale. Following this decomposition a best basis search is performed. As described in the previous section, this decomposition method is tested with both a standard and a phase-weighted entropy measure (equations 5 and 6). Where the latter entropy measure has been used this is indicated by the symbol φ.

5.1.3. CPD

This is implemented with the WaveLab toolbox for Matlab [14]. The maximum decomposition level D is chosen such that the shortest packet length is 512 samples. A best basis search is performed, but since the output of the CPD is not complex only the standard entropy measure (equation 5) is used.

5.1.4. STFT

This is tested with 50% overlapped Hann windowed frames of length 512, 1024, 2048, 4096 and 8192 samples. No best basis search is possible with the STFT.

5.2. Mixtures

All of the decomposition methods described in the previous sub-sections have been tested on three short audio mixtures ranging between 2 and 6 seconds duration, each combined according to equation (3).

5.2.1. Pitched instruments

The individual sources for this mixture are clarinet, violin, soprano singer and viola performing an excerpt from a Mozart opera. The sources were obtained from [15].

5.2.2. 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 were obtained from [16].

5.2.3. Percussion

This mixture consists of three hand percussion instruments and a single note with swept pitch from a Shakuhachi-like instrument. The sources were obtained from [18].

5.3. Data

The quality of the segmentations is objectively measured by three quantities for each separated source: the energy weighted inter-channel correlation, the signal to residual energy ratio (SRR) and the azimuth error.

For perfect segmentation of an anechoic mixture each individual segment should represent a single point source. For sources which are amplitude-panned there should be no phase difference between left and right channels – the signals should be identical apart from amplitude scaling. In this situation the zero-lag cross-correlation X between channels will be 1.0, where X is given by:

$$X = \frac{\mathbf{src'}_L \cdot \mathbf{src'}_R}{|\mathbf{src'}_L||\mathbf{src'}_R|} \quad (8)$$

and $\mathbf{src'}_L$ and $\mathbf{src'}_R$ are vectors containing the samples of the left and right channels of the segmented source. The higher the correlation value, the better the segmentation process has isolated components arriving from a single direction. For each decomposition method the mean of this correlation for each source (weighted according to the energy in the separated signals) is given in the next sub-section.

There will be no energy in the residual signal where there has been perfect separation, since the value of one of the Hann windows is 1 (and the value of the others is 0) at each of the source positions. Where there is residual energy this is an indication that atoms in the basis are contributed to by more than one source, giving rise to a spurious source direction outside of the front quadrant. The higher the SRR ratio the better the segmentation process is at placing energy for sources within the front quadrant.

Finally the ratio of left to right energy is an indicator of how much energy in a separated segment is due to the correct source. From this ratio the azimuth of the separated source can be found, using equation (4). The absolute azimuth error can then be found since the actual source azimuth is known. The energy-weighted mean azimuth error is presented for each method in the next sub-section.

Tables 1-3 present results for each of the three sound mixtures: the pitched instrument mix, speech mix and percussion mix. For each mixture the energy weighted correlation and the signal to residual ratio is presented for each of the decomposition methods. For each of the three mixtures the distribution of functions for the best basis across each scale is shown in Figures 4-6 for the standard and phase-weighted entropy. Either the DT-CWPD or short DT-CWPD is shown depending on which of these is the most successful at segmenting a particular mixture. Whilst the horizontal axes of these plots do show the relative bandwidths of the basis functions they are 'natural' (or Paley), rather than frequency, ordered and so there is not a sequential mapping between their position on the horizontal axes and their centre frequency (See the discussion of frequency\sequency and natural\Paley ordering in [7]). Audio examples of the segmentations are provided online, along with Matlab code for producing them [17].

Decomposition method	Corr.	SRR(dB)	Azimuth error
DT-CDWT	0.9161	19.3620	3.7838
DT-CDWT short	0.9159	19.3680	3.7838
DT-CWPD	0.9666	24.9131	1.2617
DT-CWPD φ	0.9665	25.1797	1.2588
DT-CWPD short	0.9648	24.7197	1.3293
DT-CWPD short φ	0.9641	25.0425	1.3419
CPD	0.9880	21.1404	0.7867
STFT 512	0.9425	19.0438	2.1394
STFT 1024	0.9588	22.0967	1.5092
STFT 2048	0.9679	24.7309	1.1700
STFT 4096	0.9713	25.7743	1.0835
STFT 8192	0.9684	26.3274	1.0863

Table 1: Segmentation results for the pitched instrument mixture.

Decomposition method	Corr.	SRR(dB)	Azimuth error
DT-CDWT	0.9065	18.8201	2.7823
DT-CDWT short	0.9066	18.8303	2.7800
DT-CWPD	0.9327	20.1290	1.7641
DT-CWPD φ	0.9317	20.2468	1.7859
DT-CWPD short	0.9339	20.4221	1.6639
DT-CWPD short φ	0.9353	20.0872	1.6341
CPD	0.9788	17.5386	0.8371
STFT 512	0.9117	18.6367	2.4099
STFT 1024	0.9303	19.9590	1.6564
STFT 2048	0.9463	21.2447	1.2576
STFT 4096	0.9464	20.7891	1.3281
STFT 8192	0.9334	19.9363	1.7269

Table 2: Segmentation results for the speech babble mixture.

Decomposition method	Corr.	SRR(dB)	Azimuth error
DT-CDWT	0.8885	17.3829	2.1194
DT-CDWT short	0.8885	17.3833	2.1194
DT-CWPD	0.9547	23.2966	0.9156
DT-CWPD φ	0.9483	19.0078	1.2382
DT-CWPD short	0.9495	22.6258	0.9769
DT-CWPD short φ	0.9491	22.6843	0.9872
CPD	0.9808	19.4562	0.5311
STFT 512	0.9381	21.4797	1.0892
STFT 1024	0.9527	23.0793	0.8577
STFT 2048	0.9593	24.7444	0.7792
STFT 4096	0.9594	24.6397	0.8216
STFT 8192	0.9584	23.7691	0.9116

Table 3: Segmentation results for the percussion mixture.

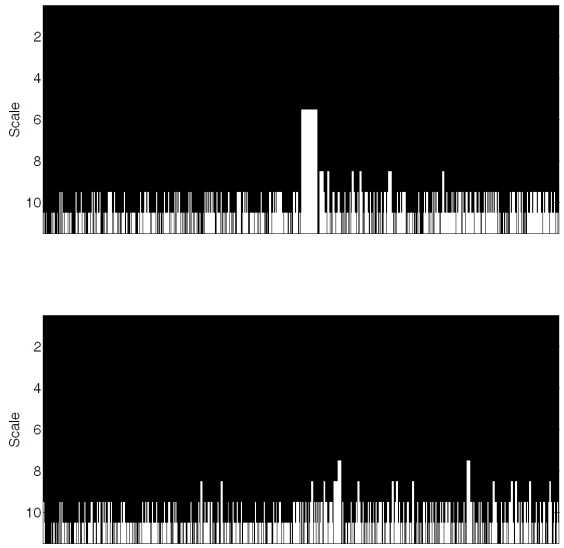

Figure 4: Function distribution for standard (top) and phase-weighted (bottom) best basis of the DT-CWPD for the instrument mixture

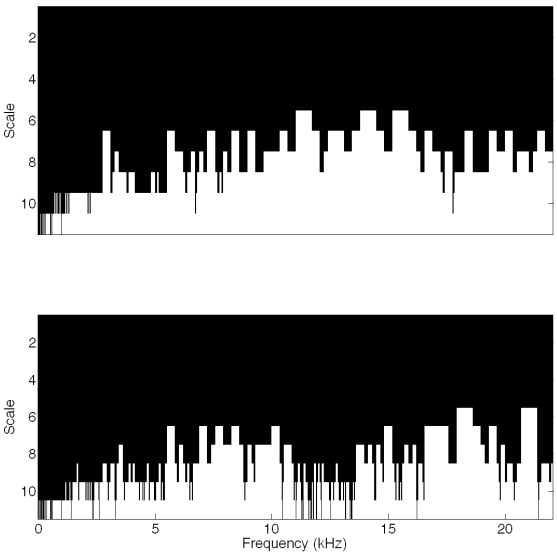

Figure 5: Function distribution for standard (top) and phase-weighted (bottom) best basis of the short DT-CWPD for the speech mixture.

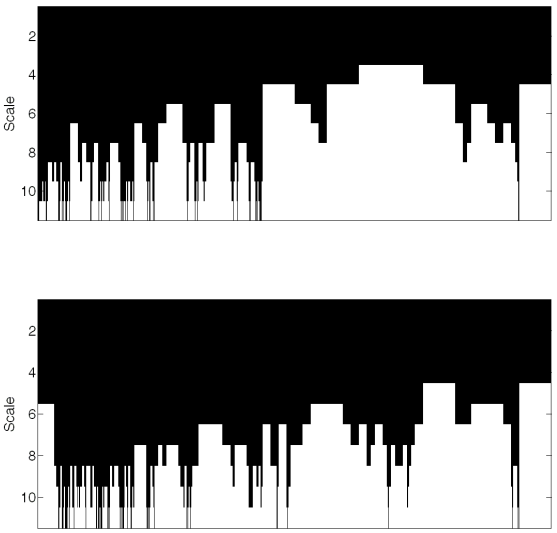

Figure 6: Function distribution for standard (top) and phase-weighted (bottom) best basis of the DT-CWPD for the percussion mixture.

5.4. Discussion

For all three mixtures the DT-CWT performs the worst of all the decompositions in all three performance

measures. This indicates that the fixed, dyadic basis of this transform is not well suited to separating the different sources in the audio mixtures. The results for this transform are almost identical for the db6 and db14 filters. This decomposition method is clearly not suited to this task.

For all of the mixtures the best basis of the CPD produces the lowest azimuth error and highest inter-channel correlation. This demonstrates that this method is best able to produce segmentations containing energy which is centered closest to the original source positions and with the best localisation (i.e. the narrowest energy distribution). However the best basis of the CPD performs relatively badly in terms of SRR, indicating that its azimuth accuracy and localisation comes at a cost of mis-placing a significant amount of the mixture energy outside of the front quadrant. In terms of SRR the STFT performs best for all mixtures. However it is not the same frame length STFT for all mixtures. The 8192 frame length STFT performs best for the instrument mixture and the 2048 performs best for the other two.

The DT-CWPD consistently out-performs the CPD in terms of SRR, but is always inferior to the CPD, and one or more the STFTs, in terms of azimuth error and correlation. Overall the DT-CPWD employing the db14 filters fares best. Although, as can be seen from Figures 4-6 the entropy measure used does have some effect on the chosen basis, the phase-weighted measure does not perform better than the standard measure and, for these example mixtures, this use of the phase information available is not of benefit. As perhaps would be expected the frequency localisation of the basis decreases (and therefore, due to time-frequency duality, the time resolution increases) from the pitched instrument mixture (slowly changing, harmonic components) through the speech mixture (faster changing, largely harmonic components) to the percussion mixture (many transient components).

Listening to the segmentations of the instrument mixture the 8192 length STFT seems to keep each instrument the most ‘intact’ but at a cost of slightly more interference from other directions within each segment than for the CPD. Whilst the CT-WPD performs reasonably well in both regards, each segment is accompanied by a ‘wheezing’ sound, apparently from amplitude modulated upper harmonics from other directions. The performance of the DT-CWT is audibly much worse than any of the other decompositions. For the speech mixture none of the decomposition methods are able to perform as well as for the instrument mixture and this can be heard clearly in the audio examples. In terms of audible quality each method introduces clearly audible degradations but each with a different character. The 2048 length STFT performs better in terms of audible interference from other directions but there is an annoying clicking sound, due to the Hann window shape not being preserved upon resynthesis and discontinuities therefore appearing in the audio. Listening to the outputs for the percussion mixture the CT-WPD and 2048 length STFT both produce relatively good segmentations. The CPD performance varies during the excerpt. For example, percussive onsets are better preserved at the start of the excerpt than at the end, illustrating the effect of the shift variance of this decomposition method – as the note onset positions move relative to the position of the time axis partitions so there is significant variation in the smearing of the onset.

6. CONCLUSIONS

This paper has described an approach to directional segmentation of audio and it has been tested on three different sound mixtures using different time-frequency\scale decompositions. Some of these decomposition methods are highly redundant and allow for a best basis, which is adapted to the mixture to be separated, to be determined. Most of the decomposition methods (STFT and dual tree wavelet based) are complex (quasi complex in the wavelet case) and 100% redundant. These complex methods also exhibit shift invariance (this is approximate in the wavelet case). The overall segmentation algorithm is designed to allow for perfect reconstruction of segmented signals where no additional processing, such as the introduction of inter-channel delays for time-shift panning, is performed.

For all mixtures the critically sampled CPD best basis performed best in terms of azimuth error and inter-channel correlation for segments but it has a higher SRR than all of the other methods, apart from the DT-CWT which performs poorly in all respects. The 8192 STFT performs best in terms of SRR for the instrument mixture but worse than the DT-CWPD best basis for the speech babble. The DT-CWPD is out-performed by the CPD best basis and at least one of the STFT decompositions for each mixture in terms of correlation and azimuth error, but always performs better than the CPD in terms of SRR. The effect of shift variance on the CPD can be clearly heard in segments from the

percussion mixture. The results do not demonstrate the benefit of using the inter-channel phase difference to inform the basis selection.

Further work is needed to provide a clearer picture of which method is most widely suited to the directional segmentation process. The complex wavelet decompositions considered have only used two different filter sets and these are of the Daubechies type, originally designed for signal compaction and may well not be suited to this kind of audio processing. Future experiments will examine other filter sets for complex wavelets and investigate a complex version of cosine packets to determine whether the residual energy can be reduced for this decomposition type, whilst retaining the good energy localisation for the segments. More complicated scenes with more sources, which are non-uniformly spaced, also with early reflections and reverberation due to room geometry, must also be considered.

7. REFERENCES

[1] Nesbit, A. et al., "Audio source separation with a signal-adaptive local cosine transform", *Signal Processing*, Vol. 87, No. 8, August 2007, pp. 1848-1858.

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

[3] Faller, C., "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.

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

[5] Kingsbury, N., "A Dual Tree Complex Wavelet Transform with Improved Orthogonality and Symmetry Properties", *Proceedings of the IEEE Conference on Image Processing*, 2000.

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

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

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

[9] *of Signal Processing*, 2nd edition, Academic Press, San Diego, 1999.

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

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

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

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

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

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

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

[17] Audio examples for this paper available at: www.jezwells.org/directional_segmentation.

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