SEPARATION OF OVERLAPPING IMPULSIVE SOUNDS BY BANDWISE NOISE INTERPOLATION

Mark R. Every and John E. Szymanski

Media Engineering Group, Department of Electronics, University of York, York, U.K.
mark.every@gmail.com, jes1@ohm.york.ac.uk

ABSTRACT

The task of extracting harmonic content of multiple pitched sources from a mono audio mix has been investigated on several occasions [1, 2, 3, 4]. However, most pitched notes contain an inharmonic component, which is an important perceptual attribute of the sound. This content is usually not dealt with during separation. It would also be interesting in its own right to develop separation techniques for extracting percussive sounds for polyphonic mixes. This paper describes an attempt at separating overlapping impulsive content of multiple sources from a mono mix. The method uses an interpolation within individual frequency bands of the decaying noise envelope of each source across overlapping sections with other sources. Three analysis methods determining the distribution of these bands were tested: the DFT followed by processing in Bark bands, the discrete wavelet transform (DWT), and the dyadic wavelet packet transform (DWPT).

1. INTRODUCTION

Mono audio mixes of pitched notes have been separated into individual sources with some success based upon harmonic models[1, 2, 3, 4]. In [1] these harmonic models take the form of a filter for each source defined in the frequency domain, with notches at the harmonic frequencies, which filters the harmonic content of that source from the mix. In [2, 3, 4] sinusoidal models were used to model harmonics and separation was achieved by re-synthesising each source directly. The residual after the harmonic content is removed from a pitched note often contains an attack transient[5], and also for sustained notes, can contain some sustained noise, for example that perceived as 'breathiness' in a flute note. The attack transient is typically quite prominent for struck or plucked instruments such as piano and guitar, and not as noticeable for sustained notes. The rapidly time-varying and non-linear behaviour at notes attacks precludes the use of sinusoidal or harmonic models as conditions of quasi-stationarity become invalid. Thus, alternative methods would be needed to extract this type of content from mixes of pitched notes. Apart from this application, it would be worthwhile to develop methods for separating overlapping percussive sounds. A potential use of this work could be to re-mix a badly recorded drum track with balanced amounts of each percussive instrument, e.g. snare, hi-hat, cymbals, etc.. As both percussive sounds and attack transients are typically characterised by a sharp increase in broad-band energy followed by a slower decay, they will collectively be referred to hereafter as 'impulsive events' or simply 'events'.

A difficulty in separating overlapping impulsive audio content is that at the start of an excitation, the instrument can exhibit highly non-linear behaviour before stable vibrational modes are established. This behaviour is difficult to predict and is specific to the particular instrument. Also, whereas the primary sound production mechanisms in pitched instruments, e.g. the string and tube, can be roughly modelled at a short time after the attack using simple 1-dimensional physical equations leading to equally spaced harmonics, the primary sound production mechanisms in percussive instruments are inherently 2 or 3-dimensional. This means that if any stable vibrational modes actually exist, their distribution will be intimately linked with the instrument type, and so it is difficult to design a generic signal model suited to all percussive instruments. Instead we borrow an idea from the deterministic plus stochastic decomposition model known as spectral modelling synthesis (SMS)[6]. In SMS, a completely impulsive sound, i.e. without any deteministic/sinusoidal content, is modelled as white-noise filtered by a frequency-dependent envelope that is allowed to change in shape slowly over time.

2. METHOD

This paper uses the concept of a time-varying noise envelope, and makes the assumption that during the decay of an impulsive sound, the power in any particular frequency band decays uniformly over the duration of the excitation. The reason for bandwise processing is that the energy distribution of an impulsive event is frequency dependent, and also we must allow for different frequency regions of an impulsive event having different rates of decay. The energy in a sum of random zero-mean distributed noise sources is equal to the sum of the energies of the unmixed sources. Thus, it is argued that within a particular frequency band, the sum of the noise power envelopes of a set of unmixed sources should sum to the measured noise power envelope of the mix of these sources. Fig. 1 depicts the noise power envelope in band b, $E_b(r)$, of a sum of two impulsive sounds with a short delay, with onset time r^1_{on} and offset time r^1_{off} for the first source, and similarly for the second source. If the envelope of the first source is interpolated across the duration of the overlap with the second source as shown, then the envelope of the second source can be estimated by simply subtracting the first envelope from the mixed envelope.

2.1. Distribution of frequency bands

The motivation for bandwise processing has been given, so now the precise distribution of the frequency bands should be decided. If the bands are too narrow, the noise envelopes are likely to be too noisy, but if they are too large there could be blurring of the frequency dependent nature of the sound. Another consideration is that any division of the signal into time frames should be of sufficiently high time resolution so as not to blur rapid changes

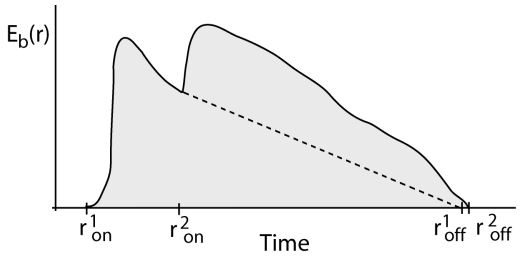

Figure 1: *The noise power envelope of a sum of two impulsive sounds with event onset and offset times indicated.*

in the noise amplitude occurring at event attacks. As perceptual studies have shown that critical bands in the human hearing system are roughly equally spaced on a logarithmic scale, a variable resolution analysis method such as wavelet analysis or a constant-Q representation is a natural choice. It was decided to compare three different analysis methods: the DFT followed by grouping of frequency bins into bands equally spaced on a Bark scale, the discrete wavelet transform (DWT) and the dyadic wavelet packet transform (DWPT). All samples used were sampled at 44.1 kHz and 16-bit resolution.

*2.1.1. Analysis method **(i)**: Processing in Bark bands*

A 1024 sample (23 ms) DFT was calculated with a hop size of 256 samples (6 ms) between frames. The positive frequency axis was split into $B = 24$ non-overlapping bands equally distributed on the perceptually motivated Bark scale[7]. $F(k, r)$ denotes the complex value of the k^{th} frequency bin of the DFT computed in the r^{th} time frame. The power in the b^{th} Bark band as a function of time is:

$$E_b(r) = \sum_{k=k^b_{min}}^{k^b_{max}} |F(k, r)|^2 \quad (1)$$

where k^b_{min}, k^b_{max} are the lower and upper limits in frequency bins of Bark band b respectively. The envelope $E_b(r)$ was convolved with a Hamming window of length 8 to smooth the envelope slightly.

*2.1.2. Analysis method **(ii)**: Discrete Wavelet Transform*

The discrete wavelet transform (DWT) of the signal $x(n)$ on a dyadic time-frequency lattice was calculated using the two-channel subband coder implementation using quadrature mirror filters as shown in fig. 2a. A depth $d = 6$ was used with the Daubechies-6 ('db–6') wavelet, where the depth is the number of low-pass filtering plus downsampling operations needed to go from the signal to the deepest level of the tree structure. The output of the DWT is a set of approximation coefficients which encode the largest scale signal features, and d sets of detail coefficients which encode successively smaller and smaller scale signal features. As scale is inversely proportional to frequency, effectively the signal is split into $d+1$ overlapping bands, with d of these bands equally spaced on a logarithmic frequency axis and the last band is the low-pass filtered signal.

*2.1.3. Analysis method **(iii)**: Dyadic Wavelet Packet Transform*

Wavelet packet analysis seeks to determine a best basis for which to encode the signal features that is optimal according to some criterion such as minimum entropy. The dyadic wavelet packet transform (DWPT) can be performed using a similar filter bank implementation to the DWT, as shown in fig. 2b. The difference is that low-pass and high-pass filtering is performed on both the detail and approximation coefficients at each level. The DWPT was followed by pruning according to a minimum (non-normalised) Shannon entropy criterion to reduce the full tree structure to a minimal set of approximation and detail levels. The resulting frequency distribution of bands is dependent on the actual signal. Again the 'db–6' wavelet was used up to a depth of $d = 6$.

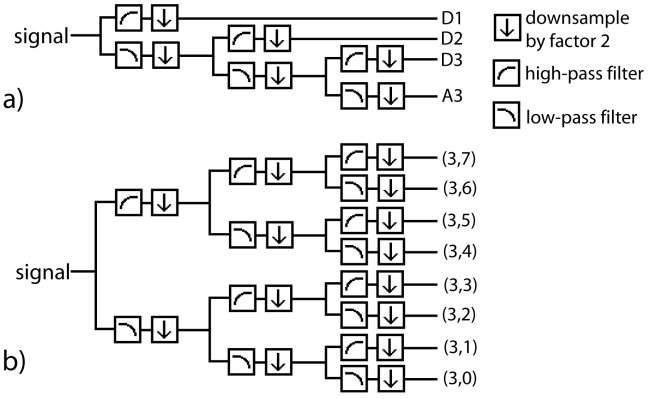

Figure 2: *Filter bank implementation of the **(a)** discrete wavelet transform (DWT), **(b)** dyadic wavelet packet transform (DWPT).*

For analysis methods ii and iii, a similar derivation of the power envelope $E_b(r)$ to the DFT method is required. The levels of approximation and detail coefficients in the filter bank tree were arranged with increasing centre frequency corresponding to bands $b = 1 \ldots B$. Then the power in the b^{th} frequency band $E_b(r)$ was taken to be the square of the approximation or detail coefficients at this level. In this context, the time index r has a time resolution that will be a power of 2^d larger than the original sampling period, where d is the depth within the tree structure corresponding to the particular frequency band. Once again a Hamming window was convolved with E_b to obtain a smoother envelope for improving interpolation.

2.2. Envelope Interpolation

The onset times of all impulsive events in the mix were assumed to be known beforehand. Although event offset times are useful, they are not essential in the proposed interpolation scheme. In an automatic system, event onset and offset times could be estimated from the signal. The noise power of the mix in a particular frequency band b has been generically denoted by $E_b(r)$ for the three analysis methods. This corresponds to the energy in Bark band b for analysis method i, and the squared approximation or detail coefficients at a node of the DWT (ii) or DWPT (iii). $E_b(r)$ is illustrated in fig. 1 for a sum of two impulsive sounds with event onset and offset times shown. One can see in the figure how the envelope of the first event can be interpolated across the overlapping section.

This is done by setting the starting point of the interpolation to immediately precede the onset of the second event, r^2_{on}, and setting the ending point as the offset of the first event. If the offset times are not provided, they could be set to a position at which the mixed envelope falls beneath some predefined threshold. Let $E^1_b(r)$ be the interpolated power envelope of the first event in band b. Since we assume that the powers of multiple events are additive in a particular band, $E^1_b(r)$ must be limited according to: $E^1_b(r) = \min[E_b(r), E^1_b(r)]$. It then follows that the estimate of the remaining power in band b, $E^{rem}_b(r)$, which is assigned to the second event, is $E^2_b(r) = E^{rem}_b(r) = E_b(r) - E^1_b(r)$. It is now useful to define two weighting functions: $w^1_b(r) = \sqrt{E^1_b(r)/E_b(r)}$ and $w^2_b(r) = \sqrt{E^2_b(r)/E_b(r)}$, which when squared, are estimates of the proportion of energy in band b as a function of time, contributed by each of the two events. The weighting functions have a range $[0, 1]$.

In a mix of several impulsive events, it is always possible to find at least one event whose power envelope can be interpolated across overlapping sections, unless two or more events begin at exactly the same instant. We begin by estimating $E^1_b(r)$ as above, and then it is subtracted from $E_b(r)$ to yield $E^{rem}_b(r) = E_b(r) - E^1_b(r)$. The first event can then be removed from consideration and the power envelope $E^2_b(r)$ can be estimated from $E^{rem}_b(r)$. In general, the interpolated envelope of the p^{th} event, $E^p_b(r)$ where $p > 1$, can be estimated, and then subtracted from $E^{rem}_b(r)$ by assigning $E^{rem}_b(r) = E^{rem}_b(r) - E^p_b(r)$. The process is iterated until there is only the last event ($p = P$) remaining. $E^P_b(r)$ is estimated to be the final remainder: $E^P_b(r) = E^{rem}_b(r)$. A disadvantage of the procedure is that the envelopes determined for each event are dependent on the order in which successive events were extracted from the mix. The p^{th} weighting function in band b is finally:

$$w^p_b(r) = \sqrt{\frac{E^p_b(r)}{E_b(r)}} \tag{2}$$

There are a number of possible interpolation methods, such as linear, exponential and logarithmic interpolation, that can be used once the starting and end points for interpolation have been established. It was found that on the whole, linear interpolation of the logarithm of the envelope, $\log_{10}[E_b(r)]$ performed fairly well. The amplitude of the end point of the interpolation needs to be set to a small positive value to avoid $\log_{10}[E_b(r^p_{off})] \rightarrow -\infty$.

2.3. Re-synthesis

It was described in the previous section how to estimate the power envelope for each source in any particular frequency band. The original mixed envelope is derived from the modulus squared of the DFT coefficients for analysis method i, and the squared approximation or detail coefficients for methods ii and iii. As the phase of the original DFT coefficients and sign of the approximation or detail coefficients are therefore lost due to the squaring operation, it is undesirable to re-synthesize the separate sources directly from the interpolated power envelopes. Instead the weighting functions, $w^p_b(r)$, are normalised and multiplied by the original (complex) DFT coefficients or (signed) approximation/detail coefficients of the mix. For analysis method i, the extracted DFT coefficients of event p in time frame r are obtained using:

$$F^p(k, r) = \frac{w^p_b(r)}{\sum_{q=1}^{P} w^q_b(r)} \cdot F(k, r) \tag{3}$$

where $b \equiv b(k)$ is the frequency band in which bin k is situated. It is clear that $F^p(k, r)$ has the same phase as the original complex spectrum $F(k, r)$. The set of weighting functions effectively share $F(k, r)$ between the P sources, in a way that reflects the estimated distribution of energy amongst the sources. The normalisation also ensures that:

$$\sum_{p=1}^{P} F^p(k, r) = F(k, r) \tag{4}$$

Thus $F(k, r)$ is split entirely between the P sources without any residual. The waveform of event p is then re-synthesized from $F^p(k, r)$ using the DFT^{-1} and an overlap-add method to interpolate between time frames.

Similarly for analysis methods ii and iii, if $c_b(r)$ are the set of approximation/detail coefficients in band b computed from the original mix, then the corresponding coefficients for source p are:

$$c^p_b(r) = \frac{w^p_b(r)}{\sum_{q=1}^{P} w^q_b(r)} \cdot c_b(r) \tag{5}$$

The normalisation once again ensures that:

$$\sum_{p=1}^{P} c^p_b(r) = c_b(r) \tag{6}$$

Re-synthesis of source p from its approximation and detail coefficients is then performed using an inverse filter bank implementation of the inverse DWT (for method ii) or inverse DWPT (for method iii). Analysis and re-synthesis using the filter bank implementations of the wavelet and wavelet packet transforms in fig. 2 followed immediately by their inverse processes, are capable of perfect reconstruction of the original signal.

3. RESULTS

The proposed source separation method was applied to three pairs and one mix of three percussive sounds, with different time delays between the onsets. The audio results are available on the internet[8]. As the method is designed for sounds whose partial/sinusoidal content has already been subtracted, any stable partial content in the original percussive sounds was removed beforehand using an SMS decomposition[6] into sinusoids plus residual. For a quantifiable measure of separation performance, we define the mean signal to residual ratio (MSRR) in decibels for a sum of P sources:

$$MSRR = \frac{10}{P} \sum_{p=1}^{P} \log_{10} \left\{ \sum_{n=1}^{N} \frac{[(x_p(n)]^2}{[x_p(n) - x'_p(n)]^2} \right\} \tag{7}$$

where $x_p(n)$ and $x'_p(n)$ are the original and separated waveforms respectively of source p, and the waveform of the original mix is $x(n) = \sum_{p=1}^{P} x_p(n)$. The MSRR's for the above sample mixes are given in table 1 for the three analysis methods, with the delay between consecutive event onsets varying between 50 ms and 300 ms.

The results in table 1 and previous tests indicate comparable performance between the three analysis methods. Phase artifacts typical of the DFT can be heard in the separated sources when using the first analysis method, but on the other hand, for analysis methods ii and iii some artifacts typical of processing with wavelets can be heard. As expected the MSRR results decrease as

Table 1: Mean signal-to-residual ratios (MSRR's) in dB for 4 percussive sample mixes as a function of the analysis method and time between consecutive onsets (δT)

mix	δT (ms)	MSRR		
		DFT	DWT	DWPT
1	50	3.1	3.7	3.7
	100	17.4	19.8	19.8
	200	25.6	24.8	24.8
2	50	11.7	11.2	12.6
	100	17.2	23.5	23.8
	200	41.2	43.0	40.7
3	50	3.2	10.9	9.0
	100	12.8	10.9	10.3
	300	21.9	22.2	22.8
4	50	6.8	9.2	9.0
	100	11.1	10.8	10.7
	200	14.5	14.6	14.9

the time between onsets becomes smaller. This is partly due to the fact that relatively more content becomes overlapping, but also the assumption that each event is in a state of uniform decay becomes less and less likely. If an event has not yet reached a decay state before the onset of the next event, then the interpolated amplitude of its envelope across the overlapping region could be unrealistically small. However, MSRR's of around $10 - 24$ dB were achieved for some mixes of two impulsive events when the distance between onsets is 100 ms. These are comparable to average MSRR's previously achieved for separating harmonic content from mixes of pitched notes[1].

4. CONCLUSION

A method for separating multiple overlapping impulsive sounds has been developed. It makes use of one of three analysis methods: the DFT followed by processing in Bark bands, the discrete wavelet transform (DWT) and the dyadic wavelet packet transform (DWPT). The power envelope of the mix is computed in frequency bands which are distributed according to the analysis method. From this, the power envelopes of the component sources are estimated by interpolating each source envelope across any sections overlapping in time with other sources. A set of weighting functions are then computed for each band that when normalised and multiplied by the DFT or wavelet approximation/detail coefficients of the mix, produce a set of DFT or approximation/detail coefficients from which the component sources can be re-synthesised. The phase information of the DFT of the original mix or sign of the approximation/detail coefficients are retained during re-synthesis.

A couple shortcomings of the method are: firstly, it assumes each impulsive event is in a state of uniform decay by the time the next event begins, which means it is unable to separate impulsive events that begin simultaneously. Secondly, as an iterative rather than a joint estimation of the source envelopes is made, then the shapes of the source envelopes depend on the order in which they were subtracted from the original mix envelope. The method assumes that each impulsive event can be modelled as a frequency dependent noise envelope whose shape changes over time. Whilst this is somewhat of a simplification, it is this which makes the method applicable generically to all percussive instrument types and attack transients for pitched instruments.

Future work will involve integrating this algorithm into an existing system for separating overlapping harmonic content in real recordings[9]. In this system, a MIDI score aligned to the recording provides note timing information.

5. REFERENCES

[1] M. Every and J. Szymanski, "Separation of synchronous pitched notes by spectral filtering of harmonics," *accepted to IEEE Transactions on Speech and Audio Processing*, 2005.

[2] A. Klapuri, T. Virtanen, and J.-M. Holm, "Robust multipitch estimation for the analysis and manipulation of polyphonic musical signals," in *Proc. COST-G6 Conference on Digital Audio Effects (DAFx-00)*, Verona, Italy, Dec. 2000.

[3] T. Tolonen, "Methods for separation of harmonic sound sources using sinusoidal modeling," *presented at AES 106th Convention*, Munich, Germany, May 1999.

[4] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using multipitch analysis and iterative parameter estimation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA-01)*, New Paltz, New York, 2001.

[5] P. Masri and A. Bateman, "Improved modelling of attack transients in music analysis-resynthesis," in *Proc. Int. Computer Music Conference (ICMC-96)*, pp. 100–103, 1996.

[6] X. Serra and J. Smith, "Spectral modelling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–24, 1990.

[7] E. Zwicker and E. Terhardt, "Analytical expression for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Am.*, vol. 68, pp. 1523–1525, 1980.

[8] "Percussive Separation Demonstrations - DAFx'05." http://www-users.york.ac.uk/∼jes1/Separation3.html.

[9] M. Every and J. Szymanski, "A spectral-filtering approach to music signal separation," in *Proc. 7th Int. Conference on Digital Audio Effects (DAFx'04)*, Naples, Italy, pp. 197–200, Oct. 5–8 1994.