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 deterministic/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, $E_b(t)$, of a sum of two impulsive sounds with a short delay, with onset time $t_{on}$ and offset time $t_{off}$ 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. A more practical 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
in the noise amplitude occurring at event attacks. As perceptual
studies have shown that critical bands in the human hearing sys-
tem 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
closest 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:

\[
E_b(r) = \sum_{k=k_{\text{min}}^b}^{k_{\text{max}}^b} |F(k, r)|^2
\]

where \( k_{\text{min}}^b \) and \( k_{\text{max}}^b \) are the lower and upper limits in frequency
bands of Bark band \( b \) respectively. The envelope \( E_b(r) \) was con-
volved with a Hamming window of length 8 to smooth the enve-
lope 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 suc-
cessively 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 cri-
teron such as minimum entropy. The dyadic wavelet packet trans-
form (DWPT) can be performed using a similar filter bank imple-
mentation 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 fol-
lowed by pruning according to a minimum (non-normalised) Shan-
non 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 \( d = 6 \).

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 auto-
matic 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 analy-
sis methods. This corresponds to the energy in Bark band \( b \) for
analysis method i, and the squared approximation or detail coeffi-
cients 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, \( r_{on} \), 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 \( E^o_b(r) \) 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, \( E^o_b(r) \) must be limited according to: \( E^o_b(r) = \min(E_b(r), E^o_1(r)) \). 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 \( E^o_b(r) = E^o_{com}^b(r) = E_b(r) - E^o_b(r) \). It is now useful to define two weighting functions: \( w^1_b(r) = \sqrt{E^o_b(r)/E_b(r)} \) and \( w^2_b(r) = \sqrt{E^o_b(r)/E_b(r)} \), 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 \( E^o_b(r) \) as above, and then it is subtracted from \( E_b(r) \) to yield \( E_{com}^b(r) = E_b(r) - E^o_b(r) \). The first event can then be removed from consideration and the power envelope \( E^o_1(r) \) can be estimated from \( E_{rem}^b(r) \). In general, the interpolated envelope of the \( p \)th event, \( E^o_p(r) \) where \( p > 1 \), can be estimated, and then subtracted from \( E_{com}^b(r) \) by assigning \( E_{com}^b(r) = E_{com}^b(r) - E^o_p(r) \). The process is iterated until there is only the last event (the \( p \) remaining). \( E^o_{final}(r) \) is estimated to be the final remainder: \( E^o_{final}(r) = E^o_{com}^b(r) \).

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:

\[
\begin{align*}
& \left[ E^o_p(r) \right]_b = \sqrt{E^o_p(r)/E_b(r)} \\
\end{align*}
\]

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 \( \log_{10}(E_b(r_{fin})) \rightarrow -\infty \).

2.3. Re-synthesis

It was described in the previous section how to estimate the power envelopes 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 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:

\[
F^p_k(r) = \frac{w^p_k(r)}{\sum_{q=1}^{P} w^p_q(r)} \cdot F(k, r)
\]
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 ($\delta T$)

<table>
<thead>
<tr>
<th>mix</th>
<th>$\delta T$ (ms)</th>
<th>MSRR DFT</th>
<th>MSRR DWT</th>
<th>MSRR DWPT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>50</td>
<td>3.1</td>
<td>3.7</td>
<td>3.7</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>17.4</td>
<td>19.8</td>
<td>19.8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>25.6</td>
<td>24.8</td>
<td>24.8</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
<td>11.7</td>
<td>11.2</td>
<td>12.6</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>17.2</td>
<td>23.5</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>41.2</td>
<td>43.0</td>
<td>40.7</td>
</tr>
<tr>
<td>3</td>
<td>50</td>
<td>3.2</td>
<td>10.9</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>12.8</td>
<td>10.9</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>300</td>
<td>21.9</td>
<td>22.2</td>
<td>22.8</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>6.8</td>
<td>9.2</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>11.1</td>
<td>10.8</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>14.5</td>
<td>14.6</td>
<td>14.9</td>
</tr>
</tbody>
</table>

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