A SPECTRAL-FILTERING APPROACH TO MUSIC SIGNAL SEPARATION

Mark R. Every, John E. Szymanski

Media Engineering Group, Department of Electronics
University of York, York, U.K.
mre104@york.ac.uk, jes1@ohm.york.ac.uk

ABSTRACT

The task of separating a mix of several inter-weaving melodies from a mono recording into multiple tracks is attempted by filtering in the spectral domain. The transcribed score is provided in MIDI format *a priori*. In each time frame a filter is constructed for each instrument in the mix, whose effect is to filter out all harmonics of that instrument from the DFT spectrum. The complication of overlapping harmonics arising from separate notes is discussed and two filter shapes that were found to be fairly successful at separating overlapping harmonics are presented. In comparing the separated audio tracks to the original instrumental parts, signal-to-residual ratios (SRR's) in excess of 20 dB have been achieved. Audio demonstrations are on the internet [1].

1. INTRODUCTION

Music separation, or more specifically, separating a number of instruments playing inter-weaving melodic lines from a mono recording, is nearly impossible to perform perfectly as mixing audio signals almost always results in a loss of information. A more achievable aim is to obtain separation of adequate quality to be useful in a number of applications. These include: audio restoration; de-mixing of old mono recordings before cleaning-up the separated instruments individually and re-mixing; re-mixing mono recordings in stereo/surround sound; structured audio coding; and some creative applications, for example an effects processor that applies an effect to a structured component of a sound rather than the whole.

The task is considered here to be a two-stage process: transcribing the mix into separate instrumental parts for which the pitch and timing of each note are found, and then performing the separation. It is conceivable that the separated results could conversely aid the transcription process, but this is not part of this implementation. As the first stage, automatic music transcription (AMT), is a demanding task in itself and the reader is referred to [2] for an account of some approaches to AMT, this research instead focuses on achieving good separation performance given the score in advance. The score is provided in MIDI format, such that a transcription of each instrumental part is available on a separate MIDI track.

To begin with, the mixed waveform is split into overlapping time frames and the DFT of the signal is computed in each frame. The pitch of a note can vary considerably over its duration, whereas a transcription of a note will most likely assign the note to a constant and discrete pitch. It was also observed that high fidelity separation was only achieved when the variation in pitch over the duration of each note was estimated accurately. Thus, for every time frame, a refinement is made of the MIDI pitches of all notes sounding in this frame. Following this, a filter is designed in the frequency domain for each instrument, whose purpose is to remove the harmonics assigned to that instrument from the spectrum. It is possible that an instrument could be playing more than one note concurrently, and in this case a filter is designed that filters from the spectrum the harmonics of each note played by this instrument. Re-synthesized separated waveforms are produced by calculating the DFT^{-1} of each filtered spectrum, and interpolating between time frames using an overlap-add technique.

Another approach to separating musical instruments [3] also notes the need for an accurate time-varying pitch estimate of each note, but instead takes an additive approach to re-synthesis, whereby the harmonics of each note are synthesized using oscillators whose time-varying frequency, amplitude and phase have been previously estimated in a least-squares sense. Similarly, additive synthesis has been used for the separation of harmonic sounds in [4]. Whilst these approaches may be able to produce fairly realistic synthesized sounds, some difficulty was encountered in the preliminary stages of our research in obtaining a realistic sounding residual using this method. The residual in this respect is the original mix minus the sum of separated sounds produced by additive synthesis. During this time domain subtraction, harmonics are liable to bleed into the residual unless highly accurate phase matching is achieved between the sinusoidal components of the additive synthesis model and corresponding components in the original mix.

Alternatively, one could consider the spectrum of a harmonic sound in a single time frame to consist of a sum of scaled and translated Fourier transforms of the window function centred at the harmonic frequencies, plus a residual component. This type of model is discussed for example in [5]. The separation of harmonic sources could then be achieved by removing the harmonics of each source from the spectrum by subtracting from the mixed spectrum a sum of ideal window shapes, whose amplitude, phase and centre frequencies had all been optimally calculated. On the contrary, in the approach described here, assuming for the moment that a spectral peak we are investigating contains a single harmonic, the harmonic is separated from the mixed spectrum by constructing a filter of unit amplitude across the main lobe of the spectral peak between the troughs in the amplitude spectrum on either side of the peak. Thus if the shape of the spectral peak was indeed the Fourier transform of the window function, this method would not remove the window's side-lobes, but for example in the case of the Hamming window, the largest side-lobes are 43 dB lower than the main lobe, so they may be sufficiently small for this not to be a concern. It is fairly common to observe spectral peaks significantly higher than the noise level that do not closely resemble the shape of the DFT of the window function, even if one takes into account the distortion of their shape due to noise or residual components, the average envelope of which could be interpolated from the surrounding spectrum. A possible explanation is that these distortions arise from frequency and amplitude modulations of a sinusoidal component within the time frame, or alternatively, that the modelling of instrument harmonics as slowly time-varying single sinusoids may not always be very accurate. The filter of unit amplitude removes the majority of the energy attributed to the instrument harmonic, without assuming that the harmonic conforms to a precise shape in the spectral domain. Signal-to-residual ratios of more than 30 dB have been achieved [6] when separating mixes of two simultaneous notes, and this provides some validation for using this filtering approach.

2. PRE-PROCESSING

All instrumental note samples used were in `.wav` format, mono, sampled at 44.1 kHz, 16 bit resolution, and all except the piano samples were recorded in an anechoic chamber. Mixed samples of 5-20 seconds in length consisting of multiple inter-weaving melodies were produced within a software sequencer such that audio and MIDI tracks for each instrument were recorded in parallel. MIDI note messages were used to trigger real audio samples, and it was possible for each instrument to be playing more than one note concurrently. The original mix was then split into overlapping time frames of length N_{win} = 8192 samples (186 ms), with an overlap of 87.5%. In each frame, after time weighting the signal with a Hamming window, an FFT was used to transform to the spectral domain.

In each time frame, the number of simultaneously sounding notes in the original mix was found from the MIDI data. As the transcribed pitches of these notes in the MIDI data were restricted to the notes of a keyboard, and considerable pitch variations over the duration of a note are not uncommon, a pitch-refinement process was used to accurately estimate all pitches present in each frame. Each refined pitch estimate was taken to be the mean of $\{f_j^p/j\ ;\ j = 1 \ldots J\}$, where f_j^p is the frequency of the j^{th} harmonic of pitch p. The harmonic frequencies $\{f_j^p\}$ were found using an iterative process starting with the identification of the fundamental frequency spectral component and then searching for spectral peaks at successively higher harmonics.

An effective method for detecting prominent spectral peaks was necessary both during pitch refinement and later in the filter design. The aim was to detect all local peaks in the amplitude spectrum significantly higher than the noise floor. A frequency-dependent threshold is usually necessary to detect all harmonics, whilst keeping the number of spurious spectral peaks or noise components above the threshold to a minimum. This frequency-dependent thresholding was implemented by dividing the spectrum by $Env(f)^c$ where c was chosen to be between 0 and 1, and $Env(f)$ is the convolution of the amplitude spectrum with a Hamming window of length $1 + N_{win}/64$. Local peaks were found above the threshold using a neighbourhood search. Harmonics right up to the Nyquist frequency were detected effectively using this method.

Finally, the baricentric interpolator [7] was used to interpolate the spectral peak centre frequencies to sub-bin frequency resolution. This interpolator was compared with others such as Grandke's, Quinn's and the parabolic interpolator, and found to be quite effective for Hamming windowed data.

3. FILTER DESIGN

The basic idea in this spectral-filtering approach to separation is that if the pitches are known of all notes present during a particular time frame, and the number of notes is not too large, then it is possible to identify most of the prominent spectral peaks uniquely with single harmonics, and to construct filters notches of unit amplitude across the width of each peak to remove the corresponding harmonic from the spectrum. A separate filter is designed for each pitch whose effect is to remove all the harmonics of this pitch from the spectrum, and the width of the notches are taken to be between the troughs in the amplitude spectrum on either side of the peak maxima. A difficulty arises when harmonics of more than one pitch are overlapping in the spectrum. This problem was resolved in [8] for combinations of two overlapping partials in a stereo mix. In our case, the sum of the filter amplitudes for all pitches, is set to unity across the width of this peak, and the shape of each filter notch is designed so that a suitable division of the energy in the spectral peak is achieved. This will be discussed in more detail below.

To begin with, it was necessary to ascertain whether each prominent spectral peak was attributable to a single harmonic or multiple harmonics. For the former case, we will refer to the spectral peak as a single-component peak and in the latter, a multi-component peak. A peak was matched to the j^{th} harmonic of note p if its centre frequency f_k was within a fixed range δ of the predicted harmonic frequency f_j^p , where $f_j^p \approx j \cdot f_0^p$, and f_0^p is the pitch of note p. The values of the f_j^p were allowed to deviate from exact harmonicity ($f_j^p = j \cdot f_0^p$), such that if a single-component peak at f_k was found to be very close to a predicted harmonic f_j^p, then f_j^p would be set equal to f_k. In either case, the next predicted harmonic would be at $f_{j+1}^p = f_j^p + f_0^p$. This modification improved separation performance, probably due to the fact that instruments whose harmonics are slightly de-tuned are treated more appropriately, and also that any slight pitch errors would not necessarily be compounded when multiplying by j to find the j^{th} harmonic.

When a spectral peak was matched to more than one harmonic from separate notes, then corresponding to each note p contributing to that peak, a filter notch was designed that depended on the predicted frequency and predicted amplitude of its harmonic within the peak: $f_j^p,\ A_j^p$, where it is implicit that $j \equiv j(p)$. The prediction of harmonic frequencies was discussed previously, and the predicted harmonic amplitudes were obtained by linear interpolation between the amplitudes of the nearest harmonics of this pitch, above and below f_j^p, that were matched to single-component peaks. Two similar filter notch designs were tested, both achieving comparable performance after fine-tuning their parameters. The filter notches $H^p(f)$ were defined for frequencies f between the troughs on either side of the peak: f_k^l and f_k^r. For the first design, the filters obeyed equation (1a):

$$\hat{H}^p(f) = A_j^p \cdot \exp\left[-\frac{|f - f_j^p|}{\sigma}\right],\ \forall\, p \in Q \tag{1a}$$

followed by a normalisation:

$$H^p(f) = \frac{\hat{H}^p(f)}{\sum_{q \in Q} \hat{H}^q(f)} \tag{1b}$$

where

$$Q = \{p\ ;\ \exists\, j(p) \text{ s.t. } |f_k - f_j^p| < \delta,\ p = 1 \ldots P\} \tag{1c}$$

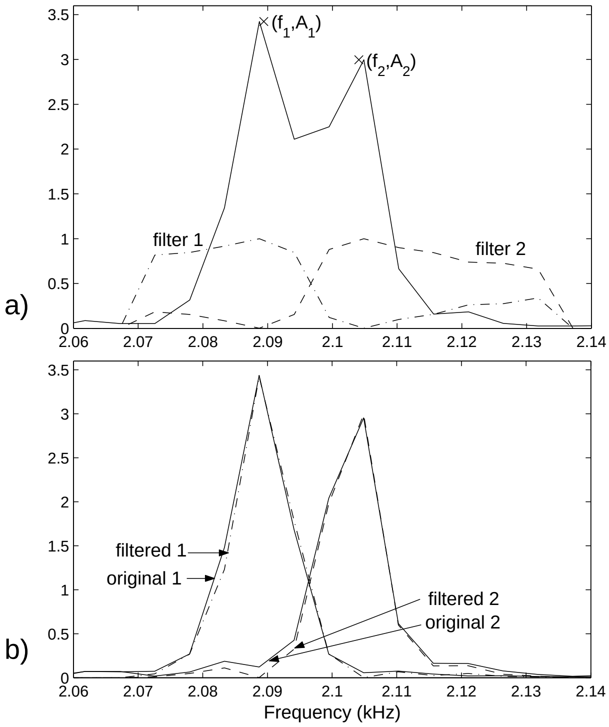

Figure 1: *Filtering of a spectral peak arising from two overlapping harmonics: (a) Construction of the filters using equation (2) is determined by the predicted harmonic frequencies f_1, and f_2, and predicted harmonic amplitudes A_1 and A_2; (b) Comparison of the filtered spectra and original spectra of the individual notes.*

and a suitable value for σ was found to be about $0.02 \cdot (f_k^r - f_k^l)^2$.

For the second filter notch design, if $\mathcal{F}_{win}(f)$ is the DFT of the window function truncated to frequencies between zero and the Nyquist limit, then the filters notches were designed according to:

$$\hat{H}^p(f) = A_j^p \cdot |\mathcal{F}_{win}(\epsilon \cdot |f - f_j^p|)| \qquad (2)$$

where $0.5 < \epsilon < 1$, and again normalised using equation (1b) to obtain $H^p(f)$.

The shape of the filters designed using equation (2) is illustrated in Figure 1a for a peak composed of two overlapping harmonics, and the two resulting filtered peaks are compared with the original spectra of the individual harmonics in Figure 1b.

4. RESULTS

The signal-to-residual ratio (SRR) has been used as a quantifiable measure of separation performance. The residual in this case is the difference between the original **x** and separated **x'** waveforms of each instrumental part. Explicitly,

$$SRR_{\mathbf{x}}(\mathbf{x}')\ [dB] = 10 \log \frac{\sum_n x_n^2}{\sum_n (x_n - x_n')^2} \qquad (3)$$

Another measure of separation performance is the average increase in the sum of SRR's for the M instrumental parts:

$$\frac{\pi(\mathbf{x_i}, \mathbf{x_i'}, \mathbf{y})}{M} = \frac{1}{M} \cdot \sum_{m=1}^{M} (SRR_{\mathbf{x_m}}(\mathbf{x_m'}) - SRR_{\mathbf{x_m}}(\mathbf{y})) \qquad (4)$$

Table 1: *Mean signal-to-residual (SRR) ratios and π/M, for sample mixes of 2-4 instrumental parts*

polyphony	2	3	4
mean $SRR_{x_i}(x_i')$	23.2	11.4	10.4
$\pi(\mathbf{x_i}, \mathbf{x_i'}, \mathbf{y})/M$	23.2	14.4	15.3

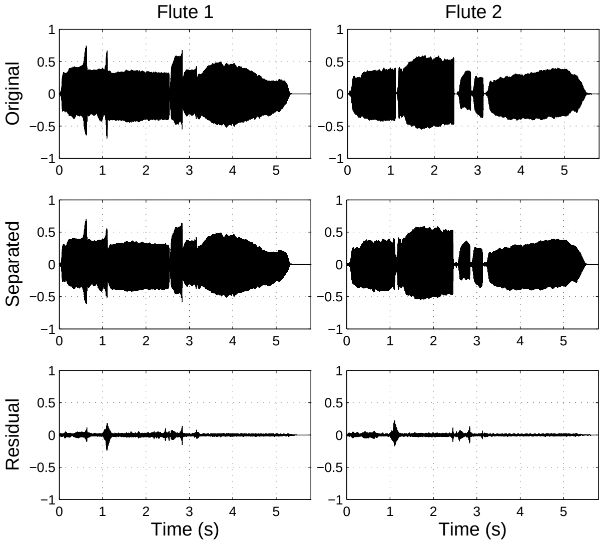

Figure 2: *Original, separated and residual waveforms for a mix of two flute melodies.*

where $\mathbf{y} = \sum_m \mathbf{x_m}$ is the mixed original signal, and larger values of π/M correspond to better separation performance.

The average SRR's and π/M are presented in Table 1 for some samples mixes of two to four instrumental parts. The waveforms of the original mixes were between 5 and 20 seconds in length. The dual polyphony sample was a mix of two harmonising flute melodies, the polyphony of three corresponded to a few upbeat bars in a major key played by a mix of flute, clarinet and French horn, and the example with a polyphony of four was a rough rendition of a few bars of Barber's '*Adagio For Strings*' played on flute, French horn and two soprano saxophones. The audio files corresponding to these test cases have been put on the internet [1] for comparison.

A visual representation of the original, separated and residual time waveforms of each instrumental part in the mix, for the sample consisting of a mix of two flute melodies in Table 1, is given in Figure 2. For the same sample mix, the spectrograms of the original mixed sound and the separated flute parts after filtering are shown in Figure 3. One can see from this last figure a clear separation of the set of harmonics belonging to each instrument, and also note that the noise level in the separated spectrograms is of lower amplitude than that of the original mix, i.e. the noise components of the original mix have mostly gone into the residual waveform.

5. DISCUSSION

Although the results describe only a small selection of test cases, both the quantitative results given in Table 1 and direct comparison by listening to the original and separated audio files, show that this

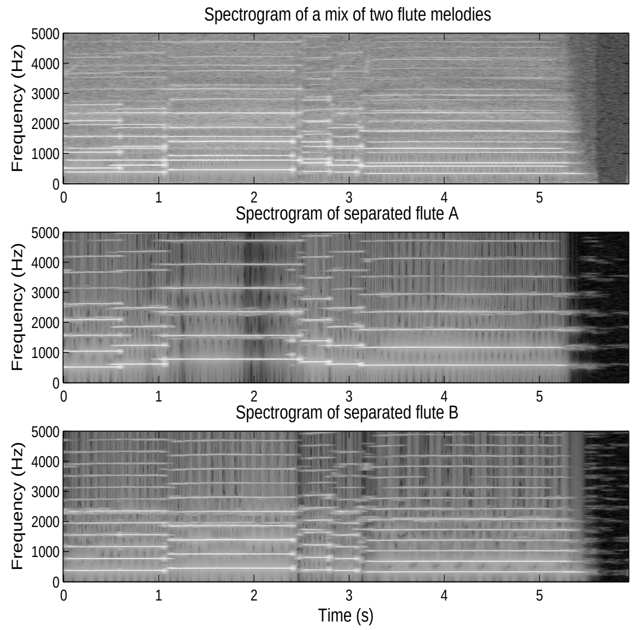

Figure 3: *Spectrograms of a sample mix of two flute melodies, and the separated individual flute parts after filtering. (The gray-scale is equal for each spectrogram and the spectrogram specifics are: sample rate $f_s = 44.1\ kHz$, FFT length 4096 samples, 87.5 % overlap, Hamming windowed).*

fairly straight-forward approach to music signal separation is quite successful. Mean SRR's of between 10.4 and 23.2 were obtained in Table 1 which represents a factor of about 11 to 210 times more energy in the original un-mixed sounds $\mathbf{x_m}$, than in the residuals $\mathbf{x_m} - \mathbf{x'_m}$. These can be compared with the mean SRR's achieved for separating mixtures of single notes in [6]. In [6], mean SRR's of 26.0 and 18.8 dB were obtained for polyphonies of 2 and 4 respectively, as an average over many almost random sample mixes. These samples mixes were chosen by randomly selecting an instrument out of a group of 10 orchestral instrument types and then randomly choosing a pitch out of each instrument's pitch range. In this paper, the sound examples studied consisted of instrumental parts that harmonised with each other, i.e. notes intervals such as octaves, fifths and thirds were common, making separation considerably more difficult than in random note mixtures due to the fact that many more harmonics would be overlapping in the spectral domain. This is believed to be one of the main reasons that higher SRR's where achieved in [6]. Hence, the issue of how to separate overlapping harmonics is relevant to separating typical musical signals. It is also worth considering that notes usually contain a significant noise or inharmonic component, and given that these algorithms only attempt to remove prominent spectral peaks from the mixed spectrum, even if the harmonics of each note were perfectly subtracted from the spectrum, the maximum SRR's achievable using this approach would be limited by the amount of inharmonic content produced by each instrument.

During listening, as expected the most noticeable differences between the original and separated sounds occur at note onsets. This is partly due to the fact that there is usually a larger inharmonic component of a note at its onset than during the sustained section. Also, the accuracy of the note timing information is an important factor in separation performance. If for example, a note actually starts sounding slightly later than the note onset time provided in the MIDI data, then it is possible that the filter corresponding to this instrument will be filtering content from the mixed spectrum in the few time frames preceding the first time frame that the note is actually present.

Lastly, we have found that the separation algorithms tend to produce interesting sounding residuals that seem to preserve the inharmonic characteristics of each instrument, for example the 'breathiness' of a flute or percussiveness of a piano note. There is potential for further research in finding ways of separating the mixed residual into instrumental parts and recombining these with the separated harmonic components in such as way as to produce more natural sounding results. Furthermore, these residuals may be useful in creative applications such as adding natural sounding, inharmonic instrument characteristics to synthesized sounds.

6. ACKNOWLEDGEMENTS

We would like to show our appreciation to the University of Iowa, Electronic Music Studios for providing all the instrumental samples that were used in this research.

7. REFERENCES

[1] M. R. Every and J. E. Szymanski. (2004, July) "Melody Separation Demonstrations," [Online]. Available: `http://www-users.york.ac.uk/~jes1/Separation2.html`

[2] A. Klapuri, "Automatic transcription of music," *Proc. Stockholm Music Acoustics Conf., SMAC-03*, Stockholm, Sweden, Aug. 6-9, 2003.

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

[4] T. Virtanen and A. Klapuri, "Separation of harmonic sound sources using sinusoidal modeling," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc., ICASSP-00*, Istanbul, Turkey, 2000.

[5] X. Rodet, "Musical sound signal analysis/synthesis: Sinusoidal+residual and elementary waveform models," *Proc. IEEE Conf. on Time-Frequency and Time-Scale Analysis, TFTS-97*, 1997.

[6] M. R. Every and J. E. Szymanski, "Separation of synchronous pitched notes in the spectral domain," submitted to *IEEE Trans. on Speech and Audio Proc.*.

[7] M. Donadio. (1999, May). "How to Interpolate Frequency Peaks," dspGuru, Iowegian Intern. Corp. [Online]. Available: `http://www.dspguru.com/howto/tech/peakfft2.htm`

[8] H. Viste and G. Evangelista, "An extension for source separation techniques avoiding beats," *Proc. 5th Int. Conf. on Digital Audio Effects, DAFx-02*, Hamburg, Germany, 2002.