SEPARATING HARMONIC AND INHARMONIC NOTE CONTENT FROM REAL MONO RECORDINGS

Mark R. Every
University of York
Department of Electronics

ABSTRACT

Four methods will be described that allow the extraction of both harmonic and inharmonic note content from real mono recordings, with the objective of separating a music recording into individual notes. Harmonics can be separated from the original mixed spectrum by constructing filters in the spectral domain with notches at the harmonic frequencies, and including a modification for overlapping harmonics from multiple instruments[1]. Three different methods will be described for separating overlapping inharmonic content: an autoregressive model based method for separating transient from steady-state audio content, a bandwise noise power interpolation method for separating overlapping impulsive onsets[2], and lastly, a method that separates overlapping noise content using the correlation of the harmonic amplitudes with the shape of the spectral noise envelope.
Keywords – Music note separation, harmonic, inharmonic, transient, noise extraction.

1. INTRODUCTION

The problem of separating pitched music notes from a real mono polyphonic recording is addressed here. Ultimately, this will allow a variety of creative transformations and effects to be applied to individual notes within the recording, and will also be useful for spatialisation and music content description.

A pitched music note, being typically a complex audio event, is viewed as containing a deterministic component of slowly-time varying harmonics, and an inharmonic component, consisting of both transient or impulsive structure and shaped noise. This is a similar characterisation to the sinusoid+transient+residual models used in [4] and [5], although since only pitched notes are being dealt with here, no allowance has been made for partials that are not harmonics.

Harmonic/partial extraction techniques are fairly well established in the fields of speech and music signal processing. Harmonics are often modelled as sinusoids with slowly time-varying frequencies and amplitudes[6] which can be re-synthesised and subtracted from the original signal in the time domain. Alternatively, sinusoids can be subtracted in the spectral domain[7]. In [1], harmonics were removed in the spectral domain using comb-like filters, with allowances made for overlapping harmonics from multiple notes. This last approach will be reviewed in section 2.1.

This leaves us with a difficult and less explored problem: separating overlapping inharmonic content from multiple sources. Transient and noise content, even in pitched notes, is important in timbre perception, and its absence can evoke a sense of unnaturalness and dullness of note attack. It constitutes for example, the 'breathiness' in a flute note, or sharpness of attack in a piano note. At least two problems are encountered when separating overlapping inharmonic content:

1. Transient content arises from highly non-linear excitations, and it is difficult to find a generic parameterised model for all types of transient events. This has in fact led to multiple descriptions or models of transient content.

2. When multiple notes are overlapping, and each note contains a broad-band noise component, there is no direct way of measuring the time and frequency dependent noise content of each individual note, as only the sum of the noise components can be observed.

Three different methods have been designed to tackle the problem. The first separates all transient events from the recording beforehand, and then matches and adds each transient event to a single note. This is described in section 2.2. The second approach, which is aimed at separating multiple impulsive onsets characterised by decaying bandwise noise power envelopes and separated by short time intervals, is reviewed in section 2.3 and described in more detail in [2]. The last approach described in section 2.4, attempts to correlate the shape of the harmonic spectrum with the noise power envelope of each note. Thereby, the noise envelope of each note can be estimated (this can not be measured directly due to point 2 above), by measuring the note's harmonic amplitudes.

2. METHOD & RESULTS

2.1. Harmonic extraction

The harmonics of each note can be extracted from the recording by filtering in the spectral domain[1]. Firstly,

Email: mark.every@gmail.com

the harmonic frequency and amplitude trajectories are estimated by tracking harmonics, where the harmonic frequencies exist at roughly integer multiples of the pitch. Then a set of comb-like filters is constructed in the spectral domain for each note, that when multiplied by the complex spectrum of the mix removes all harmonics belonging to that note. The comb-like filters are of unit amplitude across the width of each harmonic, except when harmonics are overlapping. In the latter case, the filters must be modified to separate the overlapping spectral peak between the interfering notes[1]. Average signal to residual ratios (SRR's) of the separated notes of 21.3, 17.9, 13.8 and 10.8 dB were obtained in [1] for completely random mixes of 2, 3, 4, and 5 synchronous notes respectively, and in some mixes, SRR's exceeding 30 dB were achieved using this method.

2.2. Transient extraction

The transient component of a note is usually dominant at the note attack, and is characterised by a rapid change in dynamics and non-stationary behaviour. Due to the shortness of duration of the transient event, it was decided that transient events would be extracted from the recording using a time domain model. This is in contrast with other transient extraction methods such as [3], which uses a multi-resolution time-frequency representation. In [3], the phase deviation between frames within each frequency bin is used to distinguish between transient and steady-state information. Here, the recording is initially de-constructed into a transient and non-transient component using an autoregressive (AR) model.

To begin with, the sampled recording $x(n)$ is modelled over short time frames as an AR process of order P ($P = 512$ is appropriate for a sampling rate of 44.1 kHz ≈ 12 ms), which is excited by white noise $e(n)$:

$$x(n) = \sum_{p=1}^{P} a(p)\, x(n-p) + e(n) \tag{1}$$

The set of AR coefficients, $a(p)$, was estimated over a segment of the signal of length $L = 3P$ using the Burg method. The variance of the error signal over a window of length L as a function of time is:

$$\begin{aligned} s(n) &= \frac{1}{L}\sum_{m=0}^{L-1}\left[e(n-m)\right]^2 \\ &= \frac{1}{L}\sum_{m=0}^{L-1}\left[x(n-m)-\hat{x}(n-m)\right]^2 \end{aligned} \tag{2}$$

where

$$\hat{x}(n) = \sum_{p=1}^{P} a(p)\, x(n-p) \tag{3}$$

is the forward predicted AR signal.

The method for transient removal rests on the premise that transient signals, being non-stationary and of very short duration, are not well modelled as stationary AR processes and thus produce sharp increases in the error variance $s(n)$. To satisfy scaling invariance a variable threshold for transient detection is used. If $s_\lambda(n)$ is the output of a median filter of length λ ($\lambda = 400$ ms was used) with input $s(n)$, then a transient event is detected starting at time n_i when:

$$\begin{aligned} s(n_i) &> \alpha\, s_\lambda(n_i) \\ s(n_i-1) &\leq \alpha\, s_\lambda(n_i-1) \end{aligned} \tag{4}$$

where α is a constant threshold height. In practice it is desirable to perform some prior smoothing of $s(n)$ to avoid spurious detections, and so $s(n)$ was convolved with a Hamming window of length 40 ms. The end of the transient event n_f similarly occurs when:

$$\begin{aligned} s(n_f+1) &\leq \alpha\, s_\lambda(n_f+1) \\ s(m) &> \alpha\, s_\lambda(m); \quad m = n_i, \ldots, n_f \end{aligned} \tag{5}$$

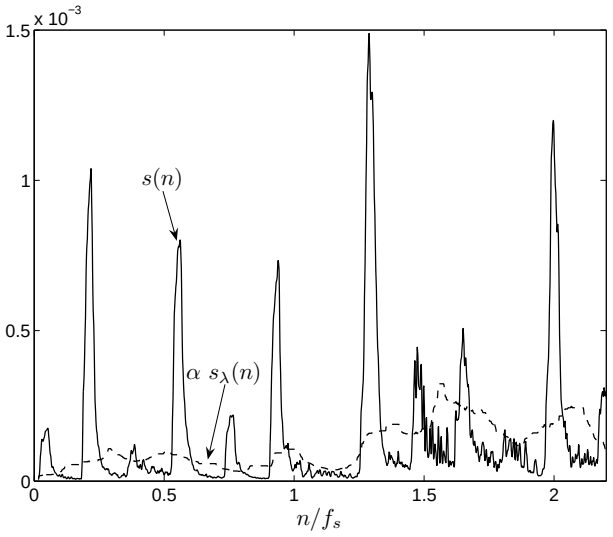

Figure 1. AR model error variance $s(n)$ and transient detection function $\alpha\ s_\lambda(n)$ for a sample recording, with $\alpha = 2$.

Fig. 1 shows $s(n)$ and the transient detection function $\alpha\ s_\lambda(n)$ for the a sample recording containing some percussive onsets. Once the boundaries of each transient event n_i and n_f have been estimated, the non-transient component between the boundaries is estimated as a weighted sum of the forward and backward AR predictions from either side of the transient event. We define the forward and backward predictions for $n_i \leq n \leq n_f$ as:

$$\hat{x}_f(n) = \sum_{p=1}^{n-n_i} a_f(p)\, \hat{x}_f(n-p) + \sum_{p=n-n_i+1}^{P} a_f(p)\, x(n-p) \tag{6}$$

$$\hat{x}_b(n) = \sum_{p=1}^{n_f-n} a_b(p)\, \hat{x}_b(n+p) + \sum_{p=n_f-n+1}^{P} a_b(p)\, x(n+p) \tag{7}$$

where $a_f(p)$ are the AR coefficients computed using the Burg method over a segment of the signal immediately

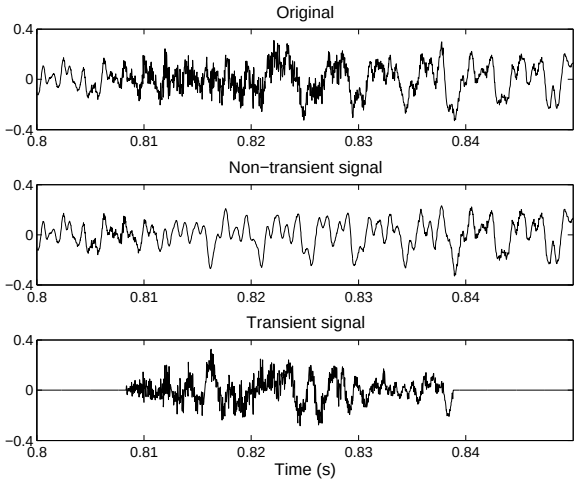

Figure 2. The separation of a transient event from a short segment of audio.

preceding the event boundary: $n_i - L \leq n \leq n_i - 1$, and $a_b(p)$ are the AR coefficients computed from the time reversed signal immediately following the event boundary: $n_f + 1 \leq n \leq n_f + L$. The weighted sum of the forward and backward predictions determines the non-transient estimation:

$$x_{nt}(n) = \frac{\beta^{-1}(n_f - n)\,\hat{x}_f(n) + \beta(n - n_i)\,\hat{x}_b(n)}{\beta^{-1}(n_f - n) + \beta(n - n_i)} \quad (8)$$

where $\beta > 1$ favours the backwards prediction over the forward prediction, and vice versa for $\beta < 1$. Eqn. 8 is simply a sum of $\hat{x}_f(n)$ and $\hat{x}_b(n)$ after each is multiplied by a scalable triangular window. Since transient events typically have a sharp attack followed by a slower decay, the backwards AR prediction is generally more accurate, and so $\beta = 2$ has been used. Finally, the transient event $x_t(n)$ is obtained by subtracting the non-transient prediction $x_{nt}(n)$ from $x(n)$ for $n_i \leq n \leq n_f$. Fig. 2 shows the separated transient and non-transient components of a short segment of audio during which a transient event occurs.

So far it has been shown how to separate a set of transient events which are localised in time from the original mix. As the objective is actually to separate the transient content of each note from the recording, the events must now be identified with individual notes. As transient content is usually dominant at the note attack, each transient event can simply be identified with the nearest note onset, which is valid unless the transient event is actually the result of several notes whose transient events overlap in time. There is of course potential for a more sophisticated matching process and the case of overlapping transients is partly dealt with in section 2.3. That aside, the transient event is simply added to the extracted harmonic and/or noise content of the note which it has been matched to.

2.3. Note attacks separated by a short time interval

The second method is designed for separating multiple impulsive attacks or transient events that are overlapping in time, but whose onsets are separated by a short delay. The method is described in detail in [2]. It interprets each impulsive event as a random signal with a frequency and time dependent noise power envelope. The original mix of impulsive sounds is initially split into frequency bands. 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). It is assumed that in each band, the power of each impulsive event in the mix decays uniformly after the initial attack. As the events are expected to be random zero-mean distributed noise sources, the sum of their power envelopes should ideally be roughly equal to the power envelope of the mix. Thus, by interpolating the uniformly decaying power envelope of a first impulsive event across an overlapping section with a second event, the power envelope of the second event can be estimated simply by subtracting the envelope of the first event from the envelope of the mix. This method relates to point 2 in section 1, in that it is an indirect way of estimating the noise envelope of each source. Once the envelopes of all events have been estimated in this manner, weighting functions are designed that split the original DFT or wavelet coefficients in each band of the mix between the multiple events. The division of the energy in the DFT or wavelet coefficients is made in proportion to the contributions in each band of the individual event power envelopes to the total power envelope. This allows re-synthesis of the individual events whilst still retaining the phase information of the original mix.

Mean SRR's using this method when separating mixes of 2-3 percussive sounds with a delay of 100 ms between consecutive attacks were around $10-24$ dB. Audio demonstrations are available on the internet[8]. There are some artifacts in the re-synthesised impulsive sounds, which is not surprising given the simplicity of the model. However, the non-stationary behaviour of impulsive events make introducing more detailed modelling assumptions difficult, and risk making the model restrictive to only particular types of impulsive events.

2.4. Connection of harmonic & inharmonic content

The last technique for separating the inharmonic content of a note from a mix of notes is aimed specifically at splitting the noise content of the mix between the constituent notes. Similarly to section 2.3, an indirect method of estimating the individual time and frequency dependent noise envelopes of each note is used to balance the noise component of the mix between the notes. In this case however, a single short-time Fourier transform (STFT) representation has been used.

The method assumes a correlation between the harmonic and noise content of each note. In an acoustic

instrument, the two components generally arise from the same excitation, supporting this view. The method uses measurements of the harmonic trajectories to provide an estimate of the noise envelope, which can not be directly measured when the note is overlapping with other notes (point 2 in section 1). As a fairly simple example, suppose that on average both the set of harmonic amplitudes and the shape of the noise power envelope are roughly linear on a logarithmic amplitude scale. This is motivated by the general observation of a roughly $1/f$ decay in the average amplitude spectra of musical signals. Let $a_h^p(r) = m_h^p(r) \cdot k + c_h^p(r)$ be the least-squares error fit to the logarithm of the squared harmonic amplitudes of note p in frame r as a function of frequency bin k. Suppose too that the gradient m_h^p and y-intercept c_h^p are related to the corresponding parameters of the linear fit to the noise power envelope, m_n^p and c_n^p, through two constants μ^p and ν^p:

$$\begin{aligned} m_n^p(r) &= \mu^p \, m_h^p(r) \\ c_n^p(r) &= \nu^p + c_h^p(r) \end{aligned} \tag{9}$$

Denote the DFT of the mix in frame r with partial content already removed, by $F(k,r)$. The noise power envelope, $R(k,r)$, which is measured on a logarithmic amplitude scale, is obtained by smoothing $|F(k,r)|^2$. As stated in section 2.3, it is expected that the power in a sum of multiple overlapping noise signals is equal to the sum of the powers of the individual signals. Thus, $R(k,r)$ should be roughly equal to the sum of the noise power envelopes of the P separate notes $R^p(k,r)$ (defined suitably for a logarithmic scale):

$$R(k,r) \approx \log_{10} \left[\sum_{p=1}^{P} 10^{R^p(k,r)} \right] \tag{10}$$

Now, if $\hat{R}^p(k,r)$ is the straight line approximation to the noise envelope of note p in frame r:

$$\begin{aligned} \hat{R}^p(k,r) &= m_n^p(r) \cdot k + c_n^p(r) \\ &= \mu^p \, m_h^p(r) \cdot k + \nu^p + c_h^p(r) \end{aligned} \tag{11}$$

then the set of parameters $\{\mu^1, \nu^1, \ldots, \mu^P, \nu^P\}$ can be determined by finding the global minimum of the least-squares error function:

$$E(\{\mu^p, \nu^p\}) = \sum_{k,r} |R(k,r) - \hat{R}(k,r)|^2 \tag{12}$$

where

$$\hat{R}(k,r) = \log_{10} \left[\sum_{p=1}^{P} 10^{\hat{R}^p(k,r)} \right] \tag{13}$$

is the sum of the predicted noise power envelopes. Table 1 gives the values of μ and ν calculated for 2 s samples of ten different instrument types all playing the pitch C4 (262 Hz). In an informal listening test, the range of ν values correlated fairly well with perceptual judgements of the degree of noisiness.

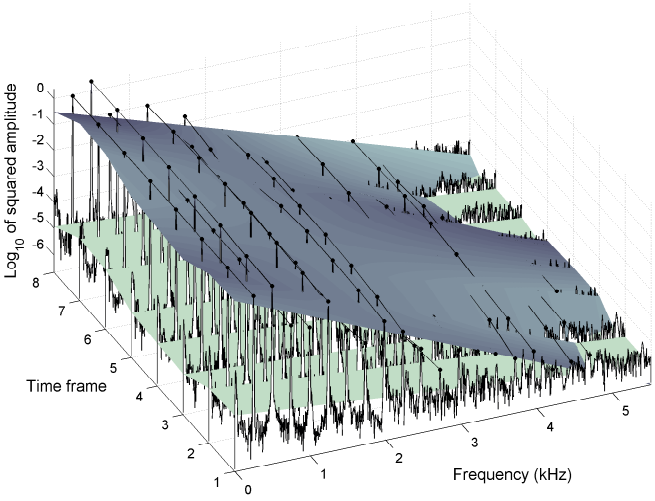

Figure 3. The logarithm of the magnitude-squared STFT of a single note, with harmonic trajectories also shown. The upper plane is a straight line fit to the harmonic amplitudes, and the lower plane is a straight line fit to the noise power envelope.

A weighting function is defined as:

$$w^p(k,r) = \left[10^{\hat{R}^p(k,r) - \hat{R}(k,r)} \right]^{\frac{1}{2}} \tag{14}$$

such that $[w^p(k,r)]^2$ is an estimate of the proportion of energy at frequency bin k and time frame r of the STFT of the original mix that is contributed by note p. Consequently,

$$\hat{F}^p(k,r) = \frac{w^p(k,r)}{\sum_{q=1}^{P} w^q(k,r)} \, F(k,r) \tag{15}$$

is an estimate of the DFT of the noise component of note p. The normalisation in eqn. 15 ensures that $F(k,r)$ is split entirely between the p notes.

Fig. 3 shows the logarithm of the magnitude-squared STFT of a segment of a single note, and both the straight line fit to the harmonic amplitudes, and straight line fit to the noise power envelope $R(k,r)$ in each time frame. It should also be mentioned that if the parameters $\{\mu^p, \nu^p\}$ are chosen to be identical for a particular instrument or source, then eqn. 12 can be optimised over a smaller number of parameters relating to the separate instrument types rather than the entire set of notes in the mix.

To summarise, the method shows how to estimate the noise spectrum of the p^{th} note, $\hat{F}^p(k,r)$, given the measured noise spectrum $F(k,r)$. $\hat{F}^p(k,r)$ has the same phase as $F(k,r)$ due to eqn. 15, and from it the noise component of note p can be re-synthesised using the DFT^{-1} combined with an overlap-add method. The logarithm of the squared harmonic amplitudes and noise power envelopes were modelled as linear in frequency, with the parameters of each linear fit connected according to eqn. 9. This is a somewhat simplistic spectral model, and the choice of harmonic/noise envelope model and connection between the two models requires further investigation.

Table 1. μ and ν for a range of instruments playing the pitch C4, sorted in decreasing degree of perceived noisiness

	$B^\flat$ Clarinet	Flute	Soprano Sax	Bassoon	Trombone	French horn	Oboe	Violin	Cello	Piano
μ	0.30	0.25	0.35	0.25	0.37	0.35	0.36	0.46	0.17	0.22
ν	-4.3	-4.5	-4.8	-5.0	-5.1	-5.4	-5.4	-4.6	-5.6	-6.0

3. DISCUSSION

A number of techniques have been described for separating the harmonic and inharmonic content of a note from a mix of notes in a recording. A method for spectral filtering of harmonic content was mentioned as an alternative to sinusoidal estimation and subtraction. As a result of its random and non-stationary behaviour, it is difficult to construct a model of inharmonic content that is separable into the individual contributions of each note in the mix. An approach has been taken in sections 2.3 and 2.4 with regards to noisy content, that attempts to split or balance the noise content of the mix between the individual notes, in a way that reflects an expectation of the time and frequency dependent noise power envelope of each note. In relation to time-localised transient events, section 2.2 alternatively proposes that these events be removed from the mix beforehand, and afterwards assigned and added to individual separated notes from the residual based upon temporal location. Section 2.4 suggests that the correlation of harmonic and inharmonic content can be used as a method for separating overlapping inharmonic content. This is similar in principle to the way that the correlation of partial amplitude and frequency trajectories have been used to separate harmonic content from note mixes[9]. A final consideration which is especially pertinent to the separation of noise content, is the difficulty in distinguishing between real noise content and artifacts of the signal representation such as spectral leakage. It has been attempted to minimise this ambiguity by using a separate signal representation suited to the analysis of each of the different types of signal content. These include the time domain waveform, STFT, discrete wavelet transform and dyadic wavelet packet transform.

4. REFERENCES

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

[2] Every M.R. and Szymanski J.E. Separation of overlapping impulsive sounds by bandwise noise interpolation, *to appear in Proc. 8th Int. Conf. on Digital Audio Effects (DAFx'05)*, Sep. 20-22, 2005.

[3] Duxbury C., Davies M. and Sandler M. Separation of transient information in musical audio using multiresolution analysis techniques, *Proc. Digital Audio Effects Conference (DAFx'01)*, Limerick, Ireland, Dec. 6-8 (2001).

[4] Verma T.S. and Meng T.H.Y., Extending spectral modeling synthesis with transient modeling synthesis, *Computer Music Journal*, Vol. 24, No. 2, pp. 47–59, 2000.

[5] Hamdy K.N., Ali M. and Tewfik A.H., Low bit rate high quality audio coding with combined harmonic and wavelet representations, *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Atlanta, Georgia, Vol. 2, pp. 1045–1048, May 1996.

[6] McAulay R.J. and Quatieri T.F., Speech analysis/synthesis based on a sinusoidal representation, IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 34, No. 4, pp. 744–754, Aug. 1986.

[7] Boll S., Suppression of acoustic noise in speech using spectral subtraction, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 2, Apr. 1979.

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

[9] Virtanen T. and Klapuri A., Separation of harmonic sound sources using sinusoidal modeling, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Istanbul, Turkey, pp. 765–769, June 2000.