Audio Engineering Society

Convention Paper

Presented at the 117th Convention
2004 October 28–31 San Francisco, CA, USA

Harmonic Sound Source Separation using FIR Comb Filters

Mikel Gainza[1], Bob Lawlor[2], and Eugene Coyle[3]

[1] Dublin Institute of Technology, Dublin, Ireland
mikel.gainza@dit.ie

[2] National University of Ireland, Maynooth, Ireland
rlawlor@eeng.may.ie

[3] Dublin Institute of Technology, Dublin, Ireland
eugene.coyle@dit.ie

ABSTRACT

A technique for separating harmonic sound sources using FIR comb filters is presented. First, a pre-processing task is performed by a multipitch estimator to detect the pitches that the signal is composed of. Then, a method based on the Short Time Fourier Transform (STFT) is utilized to iteratively extract the harmonics belonging to a given source by using FIR comb filters.

The presented approach improves upon existing sinusoidal model approaches in terms of the perceptual quality of the extracted signal.

1. INTRODUCTION

Sound source separation is the process of extracting the audio signals that belong to a given source from a polyphonic audio mixture. The applications are numerous: music enhancement, structured coding of audio signals, music teaching utilities and automatic music transcription are a few examples.

The separation of sound sources is a complex problem. However, if the assumption that the audio mixture be a sum of harmonic musical signals, the audio mixture can be separated into signals that are perceptually close to the original sources before mixing.

There are several techniques that deal with the music separation problem such as Independent Component Analysis (ICA) [1], Degenerative Unmixing Estimating Technique (DUET) [2] or sinusoidal modeling [3, 4]. However, the use of comb filters for music separation remains relatively unexplored, having only been utilized by [5, 6] with synthesized signals for music transcription, with an extension to music separation.

The presented method also uses FIR comb filters, which only requires one channel to perform the music separation, as opposed to ICA and DUET. Also, this approach is applicable to real signals, which is not the case in other FIR comb filtering methods [5, 6]. In addition, the perceptual quality of the extracted musical signals is improved in comparison to sinusoidal modeling approaches.

2. FIR COMB FILTERING

2.1. Theory

By using FIR comb filters, the comb spectral shape can be obtained by summing an input signal $x[n]$ with a delayed version of the same signal $x[n]$. The difference equation is as follows:

$$y[n] = x[n] + g * x[n-D] \qquad (1)$$

where g is a factor which scales the gain of the filter between $1+g$ and $1-g$, and D is the delay in samples.

The FIR comb filter transfer function is represented as follows:

$$H(z) = 1 + g * z^{-D} \qquad (2)$$

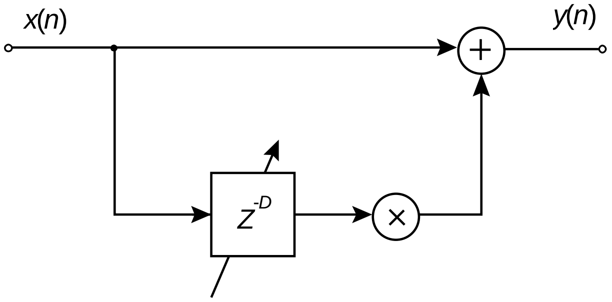

Figure 1: FIR comb filter block diagram

The comb effect results from phase cancellation and reinforcement between the delayed and undelayed signal. This can be appreciated in figure 2 where the magnitude responses of two filters with g = 1 (dashed line) and g = -1 (solid line) respectively are represented.

The sampling rate f_s is 44100 Hz and the delay D is 4 for both filters.

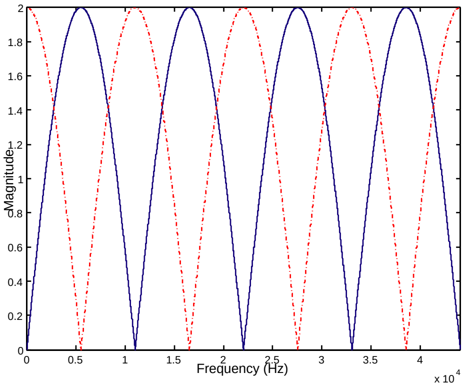

Figure 2: FIR comb filter magnitude response using g=1 (dashed line) and g =-1 (solid line)

Considering the filter with g = 1:

At frequencies $n * \frac{f_s}{D} <= f_s$ (3)

where n is an integer, the delay D causes a 360 degree shift between the original and delayed signal producing reinforcement, which occurs at peaks in the filter magnitude frequency response in the following frequencies: 11025, 22050, 33075 and 44100 Hz.

At frequencies $(2 * n - 1) * \frac{1}{2} \frac{f_s}{D} <= f_s$ (4)

the same delay D causes a 180 degree shift between the original and delayed signal causing cancellation in the sum, which occurs at nulls in the filter magnitude frequency response in frequencies 5512.5, 16537.5, 27562.5 and 38587.5 Hz.

2.2. Other FIR Comb Filters Systems

The use of comb filters for music separation remains relatively unexplored. To separate the sound sources, Miwa [5, 6] removes other sources from the polyphony. E.g. if C_3 and G_3 are playing together, a filter extracting all the harmonics of G_3 will separate C_3. To achieve the latter, he utilized FIR comb filters with g=-1, where the first null (dip) of each filter matches the frequency of the note to be extracted.

However, the overlapped harmonics in a separated source with other sources will have a zero amplitude value, since they were extracted by other source filters. These methods [5, 6], were tested using synthesized signals. If real signals are used instead, to obtain zero outputs would be much more complicated due to the frequency modulations, inharmonicities or vibrato effects.

3. SYSTEM DESCRIPTION

3.1. Introduction

A technique for separating harmonic sound sources using FIR comb filters is presented. First, a pre-processing task is performed by a multipitch estimator to detect the pitches that the signal is composed of. Then, the pitches evolution over time (frame) is analyzed, and a first grouping of frames with similar pitches between consecutive frames is formed. Next, a method based on the Short Time Fourier Transform (STFT) [7] is utilized to iteratively extract the harmonics belonging to a given source by using FIR comb filters. Then, the amplitude value of the overlapped harmonics of the separate sources will be corrected, and finally the extracted groups will be associated with a source of production.

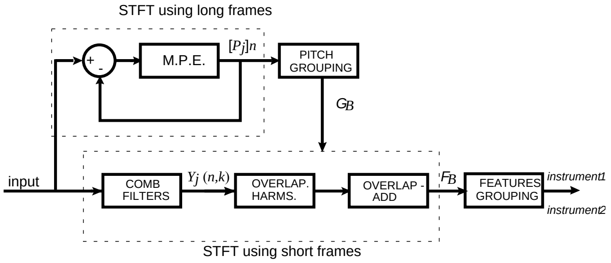

Figure 3: S*ystem block diagram*

3.2. Pre-processing Task: Multipitch Estimator

A pre-processing task is first performed to detect the pitches that the signal is composed of. The multipitch estimator (MPE) is based on the Short Time Fourier Transform (STFT) for analyzing the signal over time.

$$X(n,k)=\sum_{m=0}^{L-1} x(m+nH)w(m)*e^{-j(2\pi/N)k.m} \qquad (5)$$

where *w(m)* is the window that selects a *L* length block from the input signal *x(m)*, *n* is the frame number, *N* is the FFT length and *H* is the hop length in samples.

A recursive estimation is performed in every frame, which in each iteration detects the frequency bin number *p* of the most prominent peak. However, in some instruments (e.g.: the oboe), the most prominent harmonic is not the fundamental. Also, the fundamental in the onset part of the signal in some instruments is delayed relative to other harmonics, which could result in having less energy in the delayed fundamental than in other harmonics, if a frame is analyzing that part of the signal. Thus, if there is a peak located in the frequency bin 90, and there are also peaks in frequency bins 60 and 30, then bin 30 is considered as the bin where the pitch is located.

A peak is a local maximum in the magnitude spectrum, however each peak has an estimation error of a maximum of half of a bin, which is equal to *fs*/*N*. To determine a more precise frequency estimation, parabolic interpolation is used [8], which only uses 3 points of the magnitude spectrum around the peak. Then, the existing harmonic peaks *p*n* are disabled as candidate pitches, and the same process is repeated a number of iterations *it*. $[P_j]n$ denotes the *j* detected pitch in a frame *n*, where *j* € [1..*it*]

3.3. First Grouping: Pitch Evolution over Time

Considering a frame length *L* equal to 4094 samples and a sampling rate of 44100 Hz, the duration in time is approximately equal to 90 ms. Thus, if each two consecutive frames between frames *n* and *n*+*Z* have a similar pitch (a certain pitch deviation P_d is allowed), a group of pitches G_B with length *Z* will be formed as follows:

$$G_b=\{P_n,P_{n+1}...P_{n+Z}\} \text{ if } P_{n+o}-P_{n+o-1}<P_d \qquad (6)$$

where $o\in\{1...Z\}$, $B\in\{1...n_t\}$ and n_t is the number of tracks formed.

This process is similar to the tracking stage in sinusoidal modeling [8,9]. However, in the presented system only the pitch is tracked, as opposed to tracking all the peaks as in sinusoidal modeling.

3.4. FIR Comb Filters

3.4.1. Proposed Filter

For sound source separation, the proposed approach uses FIR comb filters equal to one at *g*=1 (see equation 1), with the magnitude response of the filter equal to one at the frequency of the pitch and its corresponding harmonics of the source to be separated. As illustrated in figure 2, a filter with *g*=1 and *D*=4, matches exactly the frequencies 11025, 22050, 33075 and 44100 Hz. Then, by using the information of the frequency of the pitches provided by the MPE, the delay *D* required to develop an adequate FIR comb filter for a given note can be calculated as follows:

$$D = \frac{f_s}{f_p} \tag{7}$$

where f_s is the sampling rate and f_p is the frequency of the pitch.

However, the sampling rate rarely is a multiple of the note frequency. Thus, to improve the accuracy of the separation, resampling techniques are used to convert the sampling rate to a multiple of the frequency note pitch.

The order of the filters varies depending on the sampling rate and the note that the instrument is playing. Using a sampling rate equal to 44100 Hz, a filter for the note C_4 (frequency 261.6 Hz), will require 169 coefficients, and a filter for the note C_7 (784.8 Hz) will require 56 coefficients. Thus, the higher the pitch of a note, the fewer the coefficients required.

The magnitude filter response of the basic comb filter has very wide peaks (see figure 2) and will also filter energy belonging to other sources. To avoid this, only a configurable number of frequency bins, *nb*, is kept on each side of every peak. This is performed by using a magnitude threshold *T*, which sets to zero all the frequency bins that don't reach the threshold value.

Considering the magnitude response *H*(*z*) in figure 2 where *g*=1. It can be appreciated that *H*(*z*) at frequency *f*=0 Hz, which corresponds to the bin number *k* equal to 1, is at a peak maximum. Thus, the magnitude value at bin number equal to *nb* corresponds to the threshold value as follows:

$$\begin{aligned} &T = \left|(H(n_b+1))\right| \\ &\text{if } H(z) \geq T \longrightarrow H(z) = H(z) \\ &\text{if } H(z) < T \longrightarrow H(z) = 0 \end{aligned} \tag{8}$$

Considering figure 4, a filter capable of separating 5 bins (*nb*=2) of each harmonic belonging to a piano playing C_4 is depicted in dashed line in the middle plot.

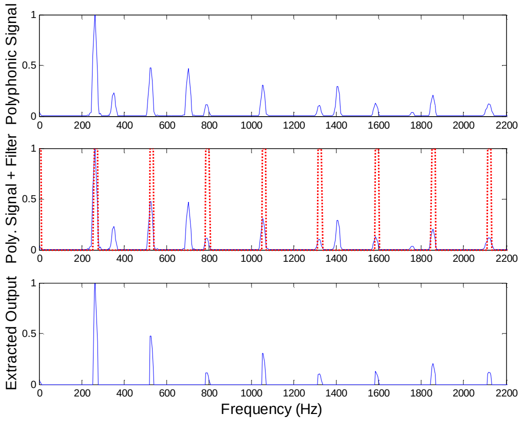

Figure 4: A piano playing C_4 (bottom plot) is extracted from a polyphonic signal (top plot) by using a comb filter (dashed line in middle plot)

3.4.2. Filtering Over Time

A different STFT analysis is performed using a window length less than the one used in the MPE analysis. Thus, frequency resolution is favored in the MPE analysis, and time resolution is favored at the sound separation stage.

The filtering process is performed in the frequency domain as follows:

$$Y_j(n,k) = X(n,k) \times H_j(n,k) \tag{9}$$

where $H_j(n,k)$ is an adequate FIR comb filter that extracts a note with pitch $[P_j]$ in the frame *n*.

3.5. Overlapped Harmonics

Notes played in Western music are separated by consonant intervals. Thus, overlapping harmonics between the different sources that compose the polyphonic tune are likely to occur.

In [10], Klapuri proposed a solution for dealing with overlapping harmonics by utilizing another auditory organization cue: the spectral smoothness [11], which refers to the expectation that the spectral envelopes tend to be continuous.

The proposed approach is also based on the spectral smoothness principle [10], but using a different algorithm: the MPE provides information on the harmonics that overlap, and the amplitude value of every overlapped harmonic, *ah,* is replaced by the mean between the previous harmonic, *ah*-1, and the next harmonic, *ah*+1, which we assume are non-overlapped with other sources. The equation is as follows:

$$a_h = \frac{a_{h-1} + a_{h+1}}{2} \qquad (10)$$

The modified output is denoted as $Y'_j(n,k)$

Using information of expected spectral information of different signals, could also contribute to a good approximation of the amplitude value of the overlapped harmonic. However, at this analysis stage, we still don't know what instrument the signal corresponds to.

3.6. Overlap-Add

The IFFT of $Y'_j(n,k)$ is calculated, and the result is overlap-added with the previously filtered signals of the group G_B that the pitch $[P_j]n$ belongs to.

$$F_B = \sum_{m+n}^{n+Z} y(m + nH) \qquad (11)$$

where $y(m)$ is the IFFT of $Y'_j(n,k)$.

3.7. Musical Features Grouping

The next step is to associate the groups F_B with a source of production, based on features of the signal such as the spectral centroid or the signal frequency range.

Signal frequency range: each instrument has a frequency range made by the notes that the instrument can play. Thus, a group F_B will be a candidate to be associated with a given instrument, if the pitch falls within the frequency range of that instrument. E.g.: the *D* key tin whistle can play notes within the $D_5 - B_6$ range, which corresponds to the 587.33-1975.5 Hz frequency range. Therefore, if the *D* key tin whistle is playing, only note pitches within that range will be considered.

Spectral centroid: the spectral centroid is an audio feature that measures how rich a signal is in harmonics, by calculating the centre of gravity of a sound related to a pitch using the following equation: [12, 13]

$$centroid = \frac{\sum_{k=0}^{N/2} k|F(k)|^2}{\sum_{k=0}^{N/2} |F(k)|^2} \qquad (12)$$

where $F_B(k)$ is the FFT of the filtered group F_B.

E.g.: the tin whistle playing C_5 (frequency = 525 Hz) has the spectral centroid very close to the pitch frequency, since the fundamental is the most prominent harmonic. However, the banjo's spectral centroid for the same C_5 note, will be in a frequency greater than 2000 Hz, due to the harmonic richness of the banjo.

Thus, the spectral centroid of the F_B groups will be calculated, to utilize the result as a feature to associate the groups to a given source of production.

4. SEPARATION EXAMPLE

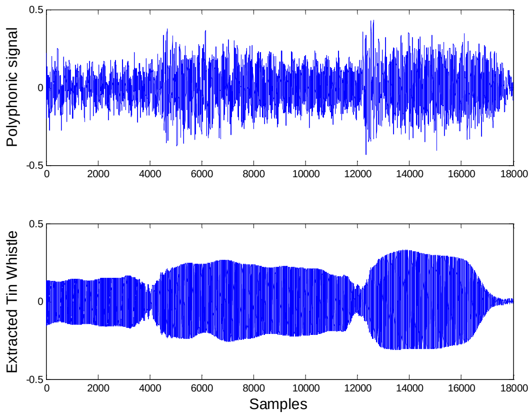

Figure 5: A tin whistle playing C_4 (bottom plot) is extracted from a polyphonic signal (top plot)

Considering figure 5, it can be appreciated how a tin whistle was separated from a polyphonic signal composed of a banjo, a guitar and the separated tin whistle. The spectral centroid value of the extracted signal is considerably lower than in the rest of instruments present in the polyphonic signal, which facilitated the grouping.

The system provided reliable results when two or three instruments were playing together. However, the quality of the separated signals degrades considerably with an increasing number of instruments in the polyphony and the consonance of the frequency interval used. Also, signals with similar spectral centroid and frequency range result in an inadequate signal grouping.

5. CONCLUSIONS AND FUTURE WORK

A system that separates harmonic sound sources using FIR comb filters is presented. The system provides several advantages over Sinusoidal Modeling. Once the pitch is detected, the system extracts all the harmonics that the signal is composed of by using only one operation, thus, reducing the computational cost slightly. The presented approach maintains a configurable number of bins on the sides of spectral peaks, which, as noticed in preliminary listening test, improves the sound fidelity in situations where frequency modulations, such as vibrato, occur within a frame; thus overcoming a recognized problem within sinusoidal modeling implementations.

However, the system has several limitations, the FIR comb filters assume that the pitches were estimated correctly by the M.P.E., which is not always the case. The musical feature grouping remains the most undeveloped part of the system. Another feature to consider is the onset of the signal, which has been widely used in timbre recognition. Connecting an onset detector to the music separation system should improve the quality of the extracted signals. In some cases, the FIR Comb filters fail to extract the onset of the signal, such as the breath of the tin whistle player, which could be very noisy. By using an onset detector, the onset of the extracted signal could be manipulated to improve the perceptual quality of the signal. Also, unexpected frequency components that occur during the onset of a musical signal, can affect the accuracy of the multi pitch estimator (MPE). However, by using the onset detector information, the MPE could be notified whether the part of the signal that it being analyzed contains an onset. Thus, the multi pitch estimations will only arise during the harmonic part of the signal.

6. REFERENCES

[1] A. Hyvärinen, J. Karhunen, et al., Independent Component Analysis: John Wiley & Sons, 2001.

[2] A. Jourjine, S. Rickard, et al., "Blind separation of disjoint orthogonal signals: demixing N sources from 2 mixtures," in Proc. Acoustics, Speech, and Signal Processing, 2000. ICASSP '00, pp. 2985-2988 vol.5.

[3] T. Virtanen and A. Klapuri, "Separation of Harmonic Sound Sources Using Sinusoidal Modeling," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2000.

[4] T. Virtanen and A. Klapuri, "Separation of harmonic sounds using linear models for the overtone series," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, 2002.

[5] T. Miwa, Y. Tadokoro, et al., "Musical pitch estimation and discrimination of musical instruments using comb filters for transcription," in Proc. Circuits and Systems, 1999. 42nd Midwest Symposium on, 1999, pp. 105-108 vol. 1.

[6] T. Miwa, Y. Tadakoro, et al., "The Problems of Transcription using Comb Filters for Musical Instrument Sounds and Their Solutions," TECHNICAL REPORT OF IEICE. SP2000-59, 2000.

[7] J. B. Allen, "Short Term Spectral Analysis, Synthesis, and Modification by Discrete Fourier Transform," IEEE Trans. on Acoust., Speech, and Sig. Proc, vol. ASSP-25(3), pp. 235-238, 1977.

[8] J. O. Smith and X. Serra, "PARSHL: An Analysis/Synthesis Program for Non-Harmonic Sounds Based on a Sinusoidal Representation," Standford University, Standford, California, Technical CCRMA STAN-M-43, 1987.

[9] R. J. Mc Aulay and T. F. Quatieri, "Speech analysis/Synthesis based on a sinusoidal modelling representation," IEEE Transactions on Acoustics, Speech, Signal Processing, vol. 26, pp. 45-45, 1986.

[10] A. Klapuri, "Multipitch estimation and sound separation by the spectral smoothness principle," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2001.

[11] Bregman, Auditory Scene Analysis: MIT Press, 1990.

[12] G. Agostini, M. Longari, et al., "Musical instrument timbres classification with spectral features," in Proc. Multimedia Signal Processing, 2001 IEEE Fourth Workshop on, 2001, pp. 97-102.

[13] A. Eronen, "Comparison of features for musical instrument recognition," in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2001.