Real-time Pitched/Unpitched Separation of Monophonic Timbre Components

Joseph A. Sarlo

Department of Music, CRCA, Cal-IT^2, University of California, San Diego
jsarlo@ucsd.edu

Abstract

Described here is a new technique for the real-time separation of a monophonic sound source timbre into its pitched and unpitched components (e.g. violin string vibration vs. bow noise.) This technique differs from pre-existing noise extraction and reduction techniques in that it does not rely on a priori knowledge of the signal and does not consider the duration or stability of a component in determining whether that component is pitched, but rather analyzes sinusoidal components based on their harmonic relationship to an estimated fundamental frequency. This has several benefits including independence of time in the determination of un-pitched components and the ability to discern pitched components from unpitched components in situations where the unpitched components may be steady in the time-evolving frequency sense.

1 Introduction

We often consider the timbre of a sound to be a composite phenomenon, being made up of several distinguishable components. As such, it is often desirable to separate a given sound source into some set of constituent components. The components can be categorized in various ways, often based on perceptual differences. One such perceptual category is that of pitched versus unpitched. We can, for example, consider the sound produced by a violin to be the composite of the tones produced by the string vibration (pitched) and the noise produced by the bow (unpitched). We describe here a method for separating a monophonic timbre into these two distinct components.

Many noise reduction techniques for audio applications have been developed. Some rely on the ability to estimate the noise by sampling the signal during "noise-only" portions (Boll 1979). This is clearly an inappropriate technique for extracting the unpitched portion of a monophonic musical instrument timbre since the pitched and unpitched portions are nearly always simultaneous. Others define noise to be the components of a sound that are rapidly changing (Hirsch 1993) and rely on this idea in their determination of which portions of a sound are noise. This, however, requires time to elapse between analyses in order to determine whether a component is steady or not. Also, what is often perceived as unpitched in a timbre is not dependant on steadiness or duration, but on the harmonic relationship that the noise component has with the perceived pitch. For example, in the case where a single piano note is played with the sustain pedal depressed, the non-sympathetic vibrations of the piano strings not struck by the key hammer will not change rapidly over time, yet can be considered unpitched in reference to the perceived pitch.

The technique presented here uses a sinusoidal analysis of the sound source and an estimated fundamental frequency to separate the unpitched portion of a monophonic timbre from the pitched portion. Thus, it has the ability to discern unpitched timbre components from pitched timbre components in cases where the unpitched components are steady in the time-evolving frequency sense. The methods of analysis, separation, and re-synthesis will be presented followed by a brief discussion of possible applications.

2 Analysis, Separation, Re-synthesis

The technique presented here essentially consists of performing both a sinusoidal deconstruction and a pitch estimation of the source sound. The individual sinusoidal components are then considered to be pitched or unpitched based on their harmonic relationship to the estimated pitch. The amplitudes of the sinusoidal components are then adjusted accordingly and a time-domain signal is re-synthesized.

2.1 Analysis

The analysis is accomplished via the phase vocoder with single-sample hop proposed by Puckette and Brown (1998) in which the frequencies of the sinusoidal components are estimated by a rectangular windowed discrete Fourier transform (DFT). This technique is used for its computation speed, its frequency estimation accuracy, and its ability to estimate frequency given a single block of audio data.

Consider the time series input signal $x[n]$ sampled at a rate of s_n Hz. Analysis is done on a sample blocks of length N, where N is a radix-2 integer typically between 512 and 4096. We have that

$$X_\tau[k] = \sum_{n=0}^{N-1} x\left[n + \frac{\tau N}{h}\right] \cdot e^{\frac{-2\pi jkn}{N}}$$

for $\tau = (0,1,2,\cdots)$

$$k = \left(-\frac{N}{2}, -\frac{N}{2}+1, \cdots, -1, 0, 1, \cdots, \frac{N}{2}-2, \frac{N}{2}-1\right)$$

Here, τ is the sample block index, k is the DFT bin index, and h is some radix-2 hop factor between 2 and N, typically 4 or 8. We note that h is only needed for re-synthesis purposes and is not a necessary part of the analysis. We can then estimate the frequency of the k^{th} sinusoidal component of the τ^{th} sample block as

$$f_{\tau,k} = s_n\left(\frac{k}{N} - \text{im}\left(\frac{j}{N} \cdot \frac{X_\tau[k+1] - X_\tau[k-1]}{2X_\tau[k] - X_\tau[k+1] - X_\tau[k-1]}\right)\right)$$

We also estimate the fundamental frequency of the τ^{th} sample block using some pitch estimation technique such as finding $f_{\tau,k}$ for which $|X_\tau[k]|$ is maximum with respect to k for a fixed τ, or using a more sophisticated technique (Puckette, Apel, and Zicarelli 1998). In practice, the method of pitch estimation used is dependant on the sinusoidal structure of the input sound source. In any case, we define F_τ to be the estimated pitch, or fundamental frequency, for the τ^{th} sample block.

2.2 Separation

To determine if a component is pitched or unpitched, we compare each sinusoidal component of frequency $f_{\tau,k}$ to the estimated fundamental frequency F_τ. We define the fractional partial index of the k^{th} sinusoidal component of the τ^{th} sample block to be

$$p_{\tau,k} = \begin{cases} \dfrac{f_{\tau,k}}{F_\tau} & \text{for } f_{\tau,k} \geq F_\tau \\ \dfrac{F_T}{f_{\tau,k}} & \text{for } f_{\tau,k} < F_\tau \end{cases}$$

We then calculate the distance of the fractional partial index to the nearest whole partial index as

$$d_{\tau,k} = \left| \lfloor p_{\tau,k} + 0.5 \rfloor - p_{\tau,k} \right|$$

When $f_{\tau,k}$ is a perfect harmonic or sub-harmonic (integer multiple or factor) of F_τ, we have that $d_{\tau,k} = 0$. Otherwise we have that $0 < d_{\tau,k} \leq 0.5$. We could, therefore, consider a sinusoidal component of frequency $f_{\tau,k}$ to be unpitched when $d_{\tau,k} > 0$ and pitched otherwise. In practice, however, inaccuracies in analysis and other physical phenomena make it desirable to define some error bound ε, where $0 \leq \varepsilon \leq 0.5$. We decide that the k^{th} sinusoidal component of the τ^{th} sample block is unpitched when $d_{\tau,k} > \varepsilon$ and pitched when $d_{\tau,k} \leq \varepsilon$. In practice, useful values of ε are dependant on the input sound source.

2.3 Re-synthesis

Using the above method to determine if a sinusoidal component is pitched or unpitched, we alter the amplitude of specific components and then apply phase vocoder analysis/re-synthesis. Specifically, let γ be some gain factor, and let $g_{\tau,k}$ be the gain applied to the k^{th} sinusoidal component of the τ^{th} sample block. We can apply this gain to all of the pitched components by choosing $g_{\tau,k}$ as

$$g_{\tau,k} = \begin{cases} \gamma & \text{for } d_{\tau,k} \leq \varepsilon \\ 1 & \text{for } d_{\tau,k} > \varepsilon \end{cases}$$

Finally, to return to the time-domain, we perform a phase vocoder analysis/re-synthesis utilizing $g_{\tau,k}$ as

$$X'_\tau[k] = \frac{g_{\tau,k}}{\sqrt{N}}\left(\sum_{n=0}^{N-1} w[n] \cdot x\left[n + \frac{\tau N}{h}\right] \cdot e^{\frac{-2\pi jkn}{N}}\right)$$

$$s_\tau[n] = \begin{cases} \dfrac{w[n]}{\sqrt{N}}\left(\displaystyle\sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} X'_\tau[k] \cdot e^{\frac{2\pi jkn}{N}}\right) & \text{for } 0 \leq n < N \\ 0 & \text{for } n < 0, n \geq N \end{cases}$$

$$x'[n] = \sum_\tau s_\tau\left[n - \frac{\tau N}{h}\right]$$

Here, $w[n]$ is a suitable window function, usually a Hanning window. By setting $\gamma = 0$ and choosing some suitable value for ε, we obtain $x'[n]$ that is a representation of only the unpitched components of $x[n]$. We can obtain a representation of the pitched components of $x[n]$ by a similar definition of $g_{\tau,k}$. We can also choose $\gamma \neq 0$ to obtain

a signal that is some combination of pitched and unpitched components or alter γ for various values of τ to obtain a time-evolving combination signal.

3 Applications

Several applications of the technique discussed are evident. For example, it is the case for many timbres that during the attack portion of the amplitude envelope, the unpitched components are much more prominent than during the remainder of the amplitude envelope. One could, therefore, utilize the technique presented here to aid in onset detection. From an electro-acoustic compositional standpoint, applications abound including phasing and spatialization effects, improvements to techniques such as time-stretching and pitch-shifting, and possibilities for convincing timbral morphing.

References

Boll, S.F. (1979). *Suppression of Acoustic Noise in Speech using Spectral Subtraction*. IEEE Transactions on Acoustics, Speech, and Signal Processing (27), pp. 113-120.

Hirsch, H. G. (1993). *Estimation of noise spectrum and its application to SNR-estimation and speech enhancement*. ICSI Technical Report TR-93-012, Intl. Comp. Science Institute, Berkeley, CA.

Puckette, M. S., T. Apel, D. Zicarelli. (1998). "Real-Time Audio Analysis Tools for Pd and MSP." In *Proceedings of the International Computer Music Conference*, pp. 109–112. Ann Arbor: International Computer Music Association.

Puckette, M. S., and J. Brown. (1998). "Accuracy of Frequency Estimates From the Phase Vocoder." *IEEE Transactions on Speech and Audio Processing* 6/2, pp. 166–176.