Tracking Tone Complexes in Audio Signals Using Structures across Time and Frequency

Bastian Bechtold[1], Joachim Thiemann[2], Steven van de Par[2]

[1] *Carl von Ossietzky Universität / Jade Hochschule, 26121 Oldenburg, Deutschland, Email: bastian.bechtold@jade-hs.de*

[2] *Carl von Ossietzky Universität, 26129 Oldenburg, Deutschland*

Abstract

Most natural sounds, voices, and musical instruments produce modulated tone complexes. The frequency modulation of these tone complexes is of vital interest for topics like melody extraction, speech recognition, and computational auditory scene analysis. In this work, we introduce a new approach to tracking the modulation of tone complexes. This algorithm, called Stretch-Correlation, tracks the modulation of tone complexes by comparing successive short-time spectra using resampling and spectral correlation. This algorithm is compared with two well-known base frequency estimators YIN and PEFAC, and is shown to outperform both at positive signal-to-noise ratios for both synthetic tone complexes and real instrument recordings.

Introduction

A key characteristic of many sounds in nature, music, and speech is the sound's fundamental frequency and frequency modulation. If these are accurately estimated, they form the basis of melody extraction, music transcription, and play an important role in speech recognition and computational auditory scene analysis. Most of these sounds are modulated tone complexes, and consist of a fundamental frequency and a number of partial frequencies at fixed multiples of the fundamental frequency.

Many early fundamental frequency estimators assumed the tone complexes to be harmonic, and thus that partial frequencies only occur at integer multiples of the base frequency [5]. Adding a pre-whitening stage to this process [4] or interpreting the tone complex spectra in the logarithmic frequency domain [1] was later shown to improve the performance of these algorithms. More recent algorithms incorporated specialized partial patterns for music [3] or speech [2]. Still, all of these algorithms are fundamentally limited by their assumption of a strictly harmonic tone complex.

Stretch-Correlation

The present algorithm tracks the frequency modulation of tone complexes by correlating differently-stretched versions of short-time spectra with one another. A tone complex spectrum $S[f]$ consists of a spectral peak at a base frequency f_0 and P partials at arbitrary but unchanging factors m_p of that base frequency.

$$S[f] = \sum_{p=0}^{P} a_p \cdot \Lambda\left(f - f_0 m_p\right) \tag{1}$$

where Λ is a peak function, a_p is the amplitude of the pth peak and p is the partial index. When this tone complex is frequency modulated by a factor σ, both the base frequency and all partial frequencies change by that same factor. For sufficiently peaky and narrow peak functions, this is equivalent to stretching the whole spectrum along the frequency axis by σ^{-1}:

$$S[\sigma^{-1} \cdot f] \approx \sum_{p=0}^{P} a_p \cdot \Lambda\left(f - \sigma \cdot f_0 m_p\right) \tag{2}$$

For computed short-time spectra, spectral stretching can be implemented as resampling of the spectrum.

With that, the modulation difference σ between two short-time spectra S_k and S_l of a modulated tone complex can be estimated as

$$\sigma_{k,l} = \underset{\sigma}{\operatorname{argmax}}\, S_k[\sigma^{-1} \cdot f] \star S_l[f] \tag{3}$$

where $\star$ denotes correlation.

This stretch factor can be calculated for every successive pair of short-time spectra to form a frequency track

$$T_m = \prod_{k=1}^{m} \sigma_{k,k-1} \tag{4}$$

The robustness of the frequency track is further improved by comparing each spectrum against a rolling mean of past spectra, where the mean spectrum is stretched to match each spectrum before averaging.

For tone complexes and white noise, spectral stretching is equivalent to modulation. However, this assumption does not hold for non-white background noise, where stretching would alter the spectral shape of the noise. To compensate for this, the background noise has to be whitened before stretch-correlation is applied.

For this purpose, a simple smoothing algorithm smooths the spectra using a brick-wall filter $w[f'] = 1$ if $f' > f'_w$ else 1 with a very low cut-off frequency-frequency $f'_w = 4$ in the spectrum-of-spectrum f' domain:

$$\overline{S} = S - \left|\mathrm{IFFT}\left(w[f'] \cdot \mathrm{FFT}\left(S[f]\right)\right)\right| \tag{5}$$

where $\overline{S}$ is the smoothed spectrum.

Evaluation

The performance of Stretch-Correlation was evaluated with a large number of varying conditions, and compared to the performance of two well-known base frequency estimators YIN[6] and PEFAC[2]. The evaluation was completed with one set of 450 synthetic signals with various partial patterns and background noises and 5 single instrument recordings in different background noises.

The synthetic signals contained a base frequency and ten partials at different multiples of the base frequency at different amplitudes. Amongst the partial distributions were harmonic tone complexes, randomized partial distributions, and partial distributions that mimic musical instruments. Base frequencies started between 40 Hz and 1 kHz, and modulated by one octave either continuously or at different musical step sizes with different rates of change. As a whole, the synthetic signals were designed to contain signals similar to both musical applications and human speech.

The single instrument recordings were excerpts from the MIREX[7] dataset, and their base frequencies were annotated by hand using the provided MIDI tracks. They were edited to contain no pauses. All instrument recordings and synthetic signals were mixed with white noise, pink noise, and two kinds of bandpass noise, at signal-to-noise ratios between 100 dB and -20 dB.

To evaluate the accuracy of each algorithm, the Normalized Musical Correctness (NMC) was used

$$\mathrm{NMC} = \frac{1}{M}\sum_{m=0}^{M}\begin{cases}1 \text{ if } \left|\frac{f_m}{\hat{f}_m} - 1\right| < 2^{\frac{1}{24}} \\ 0 \text{ otherwise}\end{cases} \tag{6}$$

where f_m is the true base frequency track, and $\hat{f}_m$ is the estimated modulation frequency track, normalized to the true base frequency by the median of the quotient between the estimated track and the true track. This normalization does not change the shape of the frequency track, but multiplies its magnitude by a fixed value. This makes the frequency-less modulation track of Stretch-Correlation comparable to the base frequency estimates from YIN and PEFAC.

Figure 1 shows the algorithms' performance in comparison to YIN and PEFAC. On the left side, Stretch-Correlation outperforms YIN by about 10 dB SNR, and is significantly more accurate than PEFAC. Only at very low SNRs and instrument recordings can PEFAC show better accuracy than Stretch-Correlation.

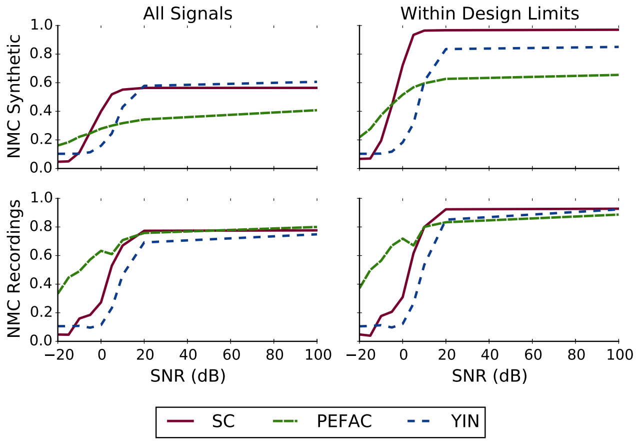

Figure 1: Evaluation of Stretch-Correlation's performance in comparison to YIN and PEFAC for synthetic signals and recordings, within and beyond the algorithms' design limits.

The right side of Figure 1 shows the algorithms' performances if they only operate on signals they were designed for. This excludes non-harmonic tone complexes for YIN and PEFAC and base frequencies beyond human speech for PEFAC. Stretch-Correlation only excludes very low base frequencies. If a higher FFT length is chosen, even that limitation can be avoided. With these limits in place, Stretch-Correlation achieves almost 100% accuracy for positive SNRs. YIN is still inferior by about 20 dB SNR, and PEFAC still has an edge for music recordings and very low SNRs.

Conclusion

Stretch-Correlation is a new algorithm that can estimate the frequency modulation of tone complexes with arbitrary partial structures. It was shown to outperform two well-known base frequency estimators for synthetic and real signals in a variety of noises. More importantly, its accuracy approaches 100% for positive SNRs over a wide variety of signals. Stretch-Correlation achieves this by modelling of the frequency modulation of tone complexes as spectral stretching, which is a powerful concept that should also be applicable to a wide range of problems beyond base frequency estimation.

References

[1] Judith C. Brown. Musical fundamental frequency tracking using a pattern recognition method. *The Journal of the Acoustical Society of America*, 92(3):1394–1402, 1992.

[2] Sira Gonzalez and Mike Brookes. PEFAC - a pitch estimation algorithm robust to high levels of noise. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 22(2):518–530, February 2014.

[3] Anssi Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *in ISMIR*, pages 216–221, 2006.

[4] A. Noll. Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate. In *Proceedings of the Symposium on Computer Processing in Communcations*, volume 19, pages 779–797, New York, 1970. Polytechnic Press.

[5] M. R. Schroeder. Period histogram and product spectrum: New methods for fundamental-frequency measurement. *The Journal of the Acoustical Society of America*, 43(4):829–834, 1968.

[6] Alain de Cheveigné and Hideka Kawahara. YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 2002.

[7] International Music Information Retrieval Systems Evaluation Laboratory at the Graduate School of Library and University of Illinois at Urbana-Champaign Information Science. MIREX: Musical Information Retrieval Evaluation eXchange. Web Download, 2014. `http://music-ir.org/mirex/wiki/MIREX_HOME`.