A COMPARATIVE ANALYSIS OF TIME-FREQUENCY DECOMPOSITIONS IN POLYPHONIC PITCH ESTIMATION

F. J. Cañadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J. Carabias, P. Cabañas and F. Rodriguez
Telecommunication Engineering Department, University of Jaén, Polytechnic School, Linares, Jaén, Spain

Keywords: Polyphonic signal, Time-frequency decomposition, Fundamental frequency, Constant Q Transform, STFT, Note-event, Candidate, Overlapped partial, Spectral modeling.

Abstract: In a monaural polyphonic music context, time-frequency information used by most of the multiple fundamental frequency estimation systems, extracted from temporal-domain of the polyphonic signal, is mainly computed using fixed-resolution or variable resolution time-frequency decompositions. This time-frequency information is crucial in the polyphonic estimation process because it must clearly represent all useful information in order to find the set of active pitches. In this paper, we present a preliminary study analyzing two different decompositions, Constant Q Transform and Short Time Fourier Transform, which are integrated in the same multiple fundamental frequency estimation system, with the aim of determining what decomposition is more suitable for polyphonic musical signal analysis and how each of them influences in the accuracy results of the polyphonic estimation considering low-middle-high frequency evaluation.

1 INTRODUCTION

Multiple fundamental frequency (Multiple-F0) estimation has long been an interesting subject in signal processing and it is still being a challenging task in the field of monaural polyphonic musical signals. The goal of a multiple-F0 estimation system is to find both the number of active pitches (polyphony) and the frequencies associated to these active pitches in a piece of music at a given time. Multiple-F0 estimation systems cover a wide range of recent audio applications: musical transcription (Marolt, 2004) (Poliner and Ellis, 2007) (Emiya et al., 2008), bass-melody detection (Goto, 2004), sound manipulation (Neubacker, 2009), content-based music retrieval (IDMT, 2009) or sound source separation (Burred and Sikora, 2007) (Li et al., 2009).

Most of the multiple-F0 estimation systems perform a preprocessing stage in which time-frequency decomposition is computed from the input signal using the time-domain. This time-frequency decomposition provides useful information in order to estimate the spectral content of a polyphonic signal and it is usually computed using a transformation with fixed resolution (Klapuri, 2003) (Yeh et al., 2005) (Bello et al., 2006) (Carabias et al., 2008) (Cañadas Quesada et al., 2008) or variable resolution (Kameoka et al., 2007) (Saito et al., 2008) (Smaragdis, 2009).

In this paper, we present a comparative study analyzing two different time-frequency decompositions, specifically, Constant Q Transform (variable resolution) and Short Time Fourier Transform (fixed resolution), which are integrated in a joint multiple fundamental frequency estimation system with the aims of determining what decomposition is more suitable to estimate the set of active pitches in a polyphonic musical signal and analyzing the performance of time-frequency decomposition in low-middle-high frequency regions.

The remainder of this paper is structured as follows. In Section 2 we briefly review the Constant Q Transform and Short Time Fourier Transform describing the advantages/disadvantages of each of them. In Section 3 the multiple-F0 estimation system is described. Experimental results are shown in Section 4. Finally, conclusions are presented in Section 5.

J. Cañadas-Quesada F., Vera-Candeas P., Ruiz-Reyes N., Carabias J., Cabañas P. and Rodriguez F. (2010).
A COMPARATIVE ANALYSIS OF TIME-FREQUENCY DECOMPOSITIONS IN POLYPHONIC PITCH ESTIMATION.
In *Proceedings of the International Conference on Signal Processing and Multimedia Applications*, pages 145-150
DOI: 10.5220/0002955601450150

2 DECOMPOSITIONS OF MUSICAL SIGNALS

2.1 Short Time Fourier Transform of Musical Signals

The Short Time Fourier Transform (STFT) is the standard tool to perform signal analysis in the frequency domain. Considering a frame-by-frame evaluation, the STFT of the signal $x[n]$ related to the t^{th} frame can be formulated as (see eq. 1),

$$X^{STFT}(t,k) = \sum_{d=0}^{M-1} x\Big[(t-1)J+d\Big]w[d]e^{-j\frac{2\pi}{M}dk}, \quad (1)$$

k = 0,...,M-1, where $w[d]$ is a N samples Hamming window and parameter J is a $\frac{N}{8}$ samples time shift. Nevertheless, the length N of each frame can be extended to M using a zero padding technique. Specifically, we have used three values of M which are referred as STFT-1N (M=N), STFT-2N (M=2N) and STFT-8N (M=8N). Each STFT has been computed using a sampling rate fs=44100 Hz, N = 4096 (92.9 ms) and J = 512 (11.6 ms).

Several researchers claim that the STFT is not suitable to analyze musical signals because each frequency bin is computed on a linear frequency scale (fixed resolution) which provides little resolution at low frequencies and tends to concentrate too much needed information at high frequencies. Consider as the music unit a semitone which presents a constant ratio of $2^{\frac{1}{12}}$ between the fundamentals. Using a fixed frequency resolution $\frac{f_s}{N}$=10.8 Hz, if we analyze the event-note C2 (65.4 Hz) and C6 (1047.0 Hz), we can observed that there is sufficient resolution to discriminate $F0_{C6} - F0_{C\#6}$=62 Hz but it is not sufficient to discriminate $F0_{C2} - F0_{C\#2}$=3.9 Hz. In this last case, all information belonging to three event-notes ($C2$, $C\#2$ and $D2$) is represented in the same frequency bin.

2.2 Constant Q Transform of Musical Signals

The Constant Q Transform (CQT) (Brown, 1991) uses a varying time window to calculate the logarithmic frequency spectrum. Each octave is composed of the same number of frequency bins b. In this manner, each frequency f_k can formulated as (see eq. 2),

$$f_k = f_{min}2^{\frac{k}{b}}\ Hz \quad (2)$$

k = 0,...,N-1. The number of frequency bin k associated to f_k can be calculated using equation 3,

$$k = \frac{b}{log_{10}2}log_{10}\left(\frac{f_k}{f_{min}}\right) \quad (3)$$

Moreover, in order to achieve a constant Q factor, both the resolution Δf_k and the number of samples N_k of the window vary inversely with the frequency f_k,

$$\Delta f_k = f_{k+1} - f_k = f_k(2^{\frac{1}{b}-1})\ Hz \quad (4)$$

$$Q = \frac{f_k}{\Delta f_k} = \frac{1}{(2^{\frac{1}{b}-1})} \quad (5)$$

$$N_k = \frac{Qf_s}{f_k} \quad (6)$$

Considering the t^{th} frame, the constant Q Transform of the signal $x[n]$ is defined in equation 7,

$$X^{CQT}(t,k) = \frac{1}{N_k}\sum_{d=0}^{N_k-1} x\Big[(t-1)J+d\Big]w[d,k]e^{-j\frac{2\pi}{N_k}dk} \quad (7)$$

In this paper, we have evaluated the direct calculation of CQT transformation (Brown, 1991) with the following parameters: b = 48 (an octave-tone resolution) and f_{min}= 32.7 Hz.

Although CQT is more computationally expensive than STFT, it is an interesting and desirable tool to apply to musical signals because exhibits two main advantages. First, it exhibits more resolution in the low frequency range. Second, the spectral localization of a note-event depends on the spectral localization of the fundamental frequency (F0) but the relative spectral localizations of the harmonics with respect to each other are invariant drawing the same spectral pattern for all note-events (see Figure 1).

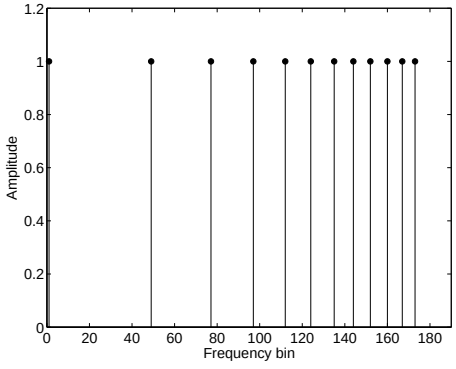

Figure 1: Spectral pattern computed by CQT of an harmonic signal composed of 12 harmonics with equal amplitude.

Figure 2 and Figure 3 exhibits the visual differences between STFT and CQT when an excerpt of monophonic signal composed of nine event-notes (piano) is analyzed. Figure 2 indicates that the spectral

spacing between harmonics changes in function of each fundamental frequency. It can be observed that this spacing grows with the frequency. However, Figure 3 shows how the logarithmic frequency spectrum maintains the relative spacing between harmonics independently of the fundamental frequency, creating the same frequency distribution for all note-events.

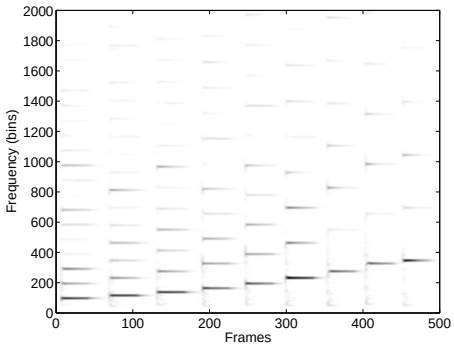

Figure 2: Short Time Fourier Transform (STFT). The energy of each frequency is represented by the gray level.

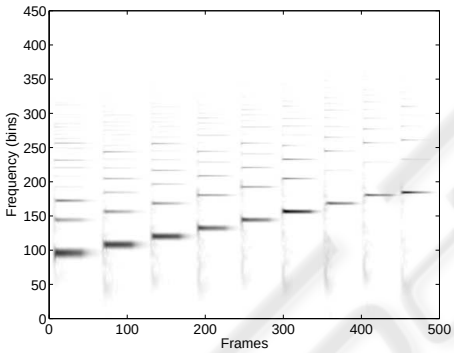

Figure 3: Constant Q Transform (CQT). The energy of each frequency is represented by the gray level.

3 MULTIPLE-F0 ESTIMATION SYSTEM

In order to evaluate the two time-frequency transformations (STFT and CQT), we have simplified the multiple-F0 estimation system proposed in (Cañadas Quesada et al., 2010) removing both temporal information extracted from λ previous frames and the HMM stage. The reason is because we only are interested to analyze the performance of the transformation in the multiple-F0 estimation, at each frame, without any kind of other complementary information. Next, the most relevant stages of this system are described. For more details of the multiple-F0 estimation system, see (Cañadas Quesada et al., 2010).

The block diagram of the integrated multiple-F0 estimation system is shown in Figure 4.

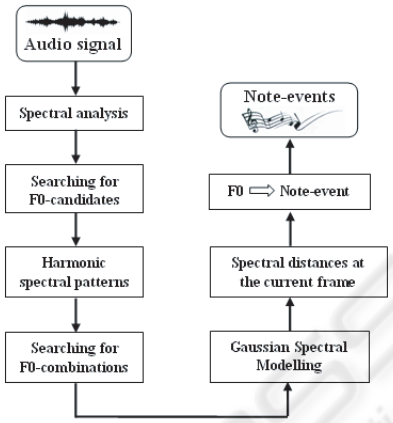

Figure 4: Multiple-F0 estimation system.

In the spectral analysis stage, both STFT and CQT are calculated to search meaningful spectral peaks from the input musical signal. To facilitate the multiple-F0 estimation, a high amount of spurious peaks is removed, using a frequency-dependent threshold (Every and Szymanski, 2006), achieving a new spectrum from the original input spectrum which is composed of only significant spectral peaks.

The next stage is the searching F0-candidates. From the previous stage, a significant spectral peak is regarded as F0-candidate if its frequency is included in the interval defined by note-events ranging from *C*2 (MIDI note 36) to *B*6 (MIDI note 95) in a well-tempered music scale. This interval has been selected because it is a typical analysis interval considered in multiple-F0 estimation systems (Klapuri, 2003) (Emiya et al., 2008) (Cañadas Quesada et al., 2008). For each F0-candidate, a harmonic spectral pattern is estimated in the logarithmic frequency domain.

After all possible harmonic patterns belonging to F0-candidates have been defined at frame level, an exhaustive search for all possible combinations of these patterns is performed.

Assuming that the amplitude spectrum of a polyphonic music signal is additive, each combination of harmonic patterns is modeled as a sum of weighted Gaussian spectral models in which non-overlapped partials are used to infer overlapped partials interpolating linearly the nearest non-overlapped partials.

The next stage consists of computing a spectral distance measure between the current audio frame and the Gaussian spectral models for each combination. Although there are many other distances, we decided to use the Euclidean spectral distance because other distances, specifically the Kullback-Liebler (KL) di-

vergence, did not show significant differences (referred to accuracy results) between the KL and Euclidean distances. Moreover, the Euclidean distance provides a higher computational efficiency because it requires a lower number of operations.

The optimum combination, composed of the most likely set of active pitches, minimizes the Euclidean spectral distance, that is, achieves the highest spectral similarity which explains most of the harmonic peaks present in the signal.

4 EXPERIMENTAL RESULTS

The performance of four time-frequency decompositions (STFT-1N, STFT-2N, STFT-8N and CQT) has been evaluated using 3 excerpts: F1 (Sonata No. 8 C minor Pathetique), F2 (Piano Sonata in G major Hoboken XVI:40) and F3 (Sonata No. 13 Bb major KV 333) of monaural polyphonic music signals played by a Yamaha Disklavier playback grand piano (Poliner and Ellis, 2007). For each excerpt, the first 20 seconds were analyzed taking into account three frequency regions: low (MIDI 36 - MIDI 60), middle-high (MIDI 60 - MIDI 95) and low-middle-high (MIDI 36 - MIDI 95) frequency regions. Accuracy and error measures were computed using the metrics proposed in (Poliner and Ellis, 2007). A user interface has been implemented to show the polyphonic estimation (a screen-shot of the GUI editor can be seen in Figure 5).

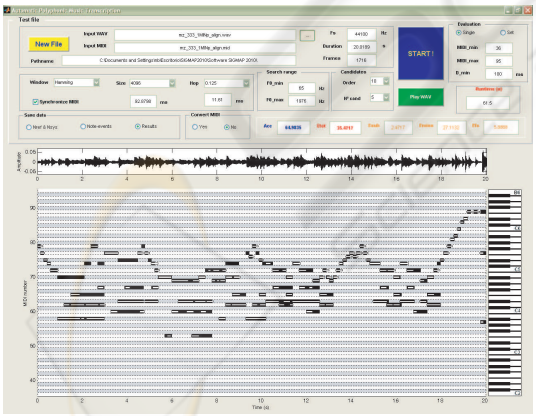

Figure 5: GUI for visualization of input parameters and pitch estimations of the system. Each black rectangle represents a reference note-event. Each white rectangle represents an estimated note-event provided by our system.

For each frequency range, accuracy results are shown in Figure 6. As can be seen in Figure 6, the STFT-8N presents the best accuracy performance with respect to the other STFT evaluated, followed by the STFT-2N and the STFT-1N. Although the STFT-2N and the STFT-8N exhibit similar results, we only compare the STFT-8N and the CQT. Figure 6(a) indicates that the CQT, compared to STFT-8N, achieves better accuracy rates in low frequency range but this improvement does not exist in the middle-high frequency range (see Figure 6(b)). A possible reason is because the CQT uses variable resolution which provides higher spectral discrimination in the low frequency range where the fundamental frequencies of event-notes are closer. This variable resolution avoids allocating adjacent event-notes in the same frequency bin. Nevertheless, analyzing the overall frequency range (see Figure 6(c)) we can observe that the performance between the CQT and the STFT-8N is approximately the same.

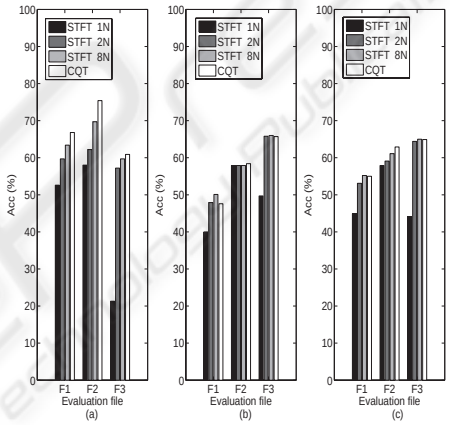

Figure 6: Frame-level accuracy results. (a) Low frequency range. (b) Middle-high frequency range. (c) Low-middle-high frequency range.

Figure 7, Figure 8 and Figure 9 show the error distribution into different categories for each frequency range and time-frequency transformation. The CQT provides the best substitution error E_{sub} and miss error E_{miss} rates in low frequency range (see Figure 7(a) and Figure 8(a)) but it does not ocurr in middle-high frequency range. However, CQT provides the best E_{sub} and E_{miss} rates when low-middle-high frequency regions are analyzed (see Figure 7(c) and Figure 8(c)). Considering false alarm error E_{fa} rates, CQT exhibits worse performance than STFT-8N in low frequency range (see Figure 9(a)). A possible reason for these results can be derived by the higher resolution of CQT in low frequency which generate a higher amount of spurious F0-candidates.

In order to compare the complexity of each transformations, a set of files of random duration have been transformed in time-frequency domain by means of the analyzed transformations. Results report that the CQT requires the most computational cost with re-

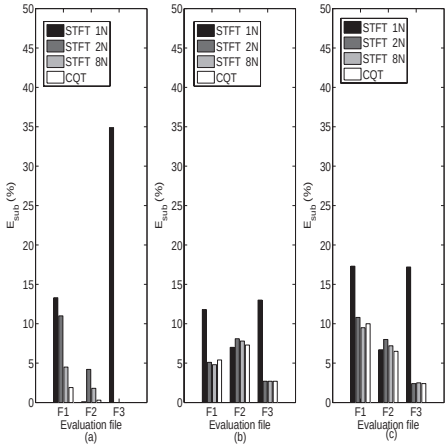

Figure 7: Frame-level substitution error results. (a) Low frequency range. (b) Middle-high frequency range. (c) Low-middle-high frequency range.

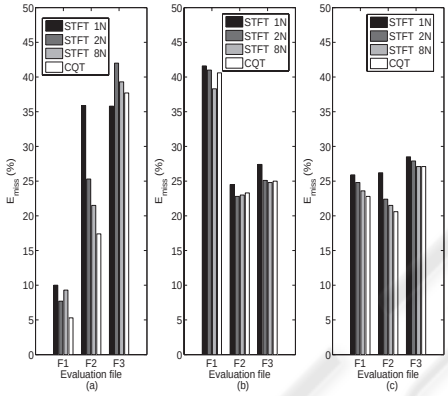

Figure 8: Frame-level miss error results. (a) Low frequency range. (b) Middle-high frequency range. (c) Low-middle-high frequency range.

spect to the others evaluated transformations. Specifically, the computational cost of the CQT is proportional to the duration of the input signal multiplied by a factor of approximately 50. However, this factor decreases to perform the computation with the other transformations: 0.26 (STFT-1N), 0.38 (STFT-2N) and 0.78 (STFT-8N).

5 CONCLUSIONS

This paper presents a preliminary analysis of two different time-frequency decompositions, Constant Q Transform and Short Time Fourier Transform, in a polyphonic pitch estimation context applied to monaural music signals. As shown in the results, the CQT provides better accuracy rates in low frequency range which it is attractive to apply in applications fo-

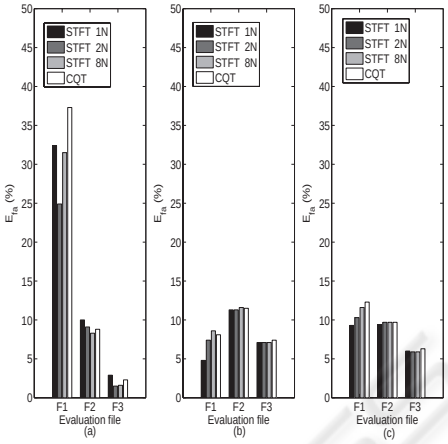

Figure 9: Frame-level false alarm error results. (a) Low frequency range. (b) Middle-high frequency range. (c) Low-middle-high frequency range.

cused on low frequency regions as accompaniment or bass music lines detections. However, this improved performance is approximately the same of the STFT-8N when the overall frequency range (low-middle-high) is analyzed. In order to analyze polyphonic signals in a wider spectral range, the STFT-8N is more suitable because exhibits the best trade-off between accuracy and computational cost, allowing the implementation of possible real-time applications.

ACKNOWLEDGEMENTS

This work was supported in part by the Spanish Ministry of Education and Science under Project TEC2009-14414-C03-02 and the Andalusian Council under project P07-TIC-02713.

REFERENCES

Bello, J., Daudet, L., and Sandler, M. (2006). Automatic piano transcription using frequency and time-domain information. *IEEE Transactions on Speech and Audio Processing*, 14(6):2242–2251.

Brown, J. (1991). Calculation of a constant q spectral transform. *Journal of the Acoustical Society of America*, 89(1):425–434.

Burred, J. and Sikora, T. (2007). Monaural source separation from musical mixtures based on time-frequency timbre models. *Proc. International Conference on Music Information Retrieval (ISMIR)*. Vienna, Austria.

Cañadas Quesada, F., Ruiz-Reyes, N., Vera-Candeas, P., Carabias-Orti, J., and Maldonado, S. (2010). A multiple-f0 estimation approach based on gaussian

spectral modeling for polyphonic music transcription. *accepted to appear in Journal of New Music Research*.

Cañadas Quesada, F., Vera-Candeas, P., Ruiz-Reyes, N., Mata-Campos, R., and Carabias-Orti, J. (2008). Note-event detection in polyphonic musical signals based on harmonic matching pursuits and spectral smoothness. *Journal of New Music Research*, 89(8):1653–1660.

Carabias, J., Vera, P., Ruiz, N., Mata, R., and Canadas, F. (2008). Polyphonic piano transcription based on spectral separation. *124thAudio Engineering Society (AES)*. Amsterdam, The Netherlands, 2008.

Emiya, V., Badeau, R., and David, B. (2008). Automatic transcription of piano music based on hmm tracking of jointly-estimated pitches. *Proc. European Conference on Signal Processing (EUSIPCO)*.

Every, M. and Szymanski, J. (2006). Separation of synchronous pitched notes by spectral filtering of harmonics. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1845–1856.

Goto, M. (2004). A real-time music-scene-description system: Predominant-f0 estimation for detecting melody and bass lines in real-word audio signals. *Speech Communications*, 43(4):311–329.

IDMT, F. (2009). Musicline. http://www.musicline.de/de/melodiesuche/input.

Kameoka, H., Nishimoto, T., and Sagayama, S. (2007). A multipitch analyzer based on harmonic temporal structured clustering. *IEEE Trans. on Audio, Speech and Language Processing*, 15(3):982–994.

Klapuri, A. (2003). Multiple fundamental frequency estimation by harmonicity and spectral smoothness. *IEEE Trans. Speech and Audio Processing*, 11(6):804–816.

Li, Y., Woodruff, J., and Wang, D. (2009). Monaural musical sound separation based on pitch and common amplitude modulation. *IEEE Trans. on Audio, Speech and Language Processing*, 17(-):1361–1371.

Marolt, M. (2004). A connectionist approach to automatic transcription of polyphonic piano music. *IEEE Transactions on Multimedia*, 6(3):439–449.

Neubacker, P. (2009). Celemony. http://www.celemony.com.

Poliner, G. and Ellis, D. (2007). A discriminative model for polyphonic piano transcription. *EURASIP Journal on Advances in Signal Processing*, 2007(1):154–162.

Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T., and Sagayama, S. (2008). Specmurt analysis of polyphonic music signals. *IEEE Trans. on Audio, Speech and Language Processing*, 16(3):639–650.

Smaragdis, P. (2009). Relative pitch tracking of multiple arbitrary sounds. *Journal of the Acoustical Society of America*, 125(5):3406–3413.

Yeh, C., Roebel, A., and Rodet, X. (2005). Multiple fundamental frequency estimation of polyphonic music signals. *in Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. Philadelphia, USA.