POLYPHONIC TRANSCRIPTION BASED ON TEMPORAL EVOLUTION OF SPECTRAL SIMILARITY OF GAUSSIAN MIXTURE MODELS

F.J. Cañadas-Quesada, P. Vera-Candeas, N. Ruiz-Reyes, J.J. Carabias-Orti

Telecommunication Engineering, University of Jaén
C/ Alfonso X el Sabio, n 28, 23700, Linares (Jaén), Spain
phone: +34 953648510, fax: +34 953648508, email: fcanadas@ujaen.es
web: www4.ujaen.es/˜fcanadas

ABSTRACT

This paper describes a system to transcribe multitimbral polyphonic music based on a joint multiple-F0 estimation. In a frame level, all possible fundamental frequency (F0) candidates are selected. Using a competitive strategy, a spectral envelope is estimated for each combination composed of F0 candidates under assumption that a polyphonic sound can be modeled by a sum of weighted gaussian mixture models (GMM). Since in polyphonic music the current spectral content depends to a large extent of the immediately previous one, the winner combination is determined taking into account the highest spectral similarity regarding to the past music events which has been selected from a set of combinations that minimize the current spectral distance between input-GMM spectrums. Our system was tested using several pieces of real-world music recordings from RWC Music Database. Evaluation shows encouraging results compared to a recent state-of-the-art method.

1. INTRODUCTION

Polyphonic music transcription is considered as a highly complex task both from a Signal Processing viewpoint and a Music viewpoint since it can only be addressed by the most skilled musician. Finding the polyphony or estimating what pitches are active in a piece of music at a given time is still being an unsolved problem. Multiple-F0 estimation is the most important stage of a polyphonic music transcription system whose aim is to extract a music score from an audio signal. The minimum unit of a music score is a *note-event* which can be described as a temporal sequence, defined by an onset and offset, of the same fundamental frequency. In consequence, multiple-F0 estimation is essential to develop current audio applications as content-based music retrieval, query by humming, enhancing of sound quality, musicological analysis or audio remixing [1][2].

Many polyphonic transcription systems have been proposed in the last years. Goto [3] describes a predominant-F0 estimation method called PreFEst which estimates the relative dominance of every possible F0 by using MAP (maximum a posteriori probability) estimation and considers the F0s temporal continuity by using a multiple-agent architecture. Yeh et al. [4] selects the best combination of candidates based on three physical principles while Pertusa [5] chooses the best one maximizing a criterion based both loudness and spectral smoothness. The system proposed by Li [6] takes into account a hidden Markov model (HMM) which applies an instrument model to evaluate the likelihood of each candidate. Kameoka et al. [7] describes a multipitch estimator based on a two-dimensional Bayesian approach. In [8], Bello et al. considers frequency-time domain information to identify notes in polyphonic mixtures. Klapuri's system [9] uses an iterative cancelation mechanism based on a computational model of the human auditory periphery. Ryynanen [10] reports a combination of an acoustic model for note-events, a silence model, and a musicological model. In [11], Cañadas modifies harmonic decompositions in order to maximize the spectral smoothness for those Gabor-atom amplitudes that belong to the same harmonic structure. Specmurt technique is detailed by Saito et al. [12] which is based on nonlinear analysis using an inverse filtering in the log-frequency domain.

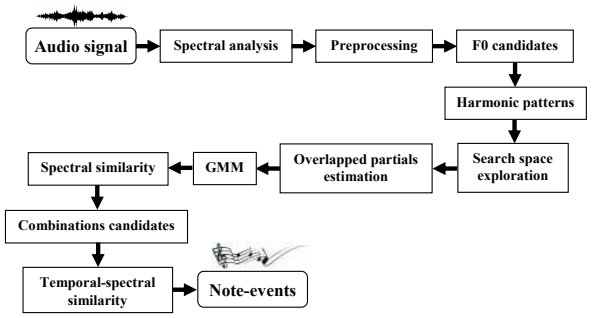


Figure 1: Overview of the proposed polyphonic music transcription system.

In this work, a system to transcribe polyphonic music based on a joint multiple-F0 estimation is described. The system scheme is shown in Fig. 1. The basic idea consists of analyzing the temporal evolution of the spectral envelopes regarding to the estimated GMM spectrums to maximize the spectral similarity between the polyphonic input signal and the estimated models. We rely on the fact that in polyphonic music the current musical events depends to a large extent of the immediately previous ones.

This paper is organized as follows. In section 2, the proposed joint multiple-F0 estimation method is introduced. In section 3, Gaussian mixture model is depicted in detail. In section 4, our selection criterion based on temporal-spectral similarity between polyphonic spectrums is described. In section 5, experimental results are shown. Finally, the conclusions and future work are presented in section 6.

2. PROPOSED MULTIPLE-F0 ESTIMATION METHOD

The spectrum $X(k)$ computed by the Short Time Fourier Transform (STFT) of the signal $x(n)$ is detailed in eq. (1),

$$X(k) = \sum_{d=\frac{-N}{2}}^{\frac{N}{2}-1} x(nh+d)w(d)e^{-j\frac{2\pi}{N}dk} \quad (1)$$

, where $w(d)$ is a N samples Hamming window, a $\frac{N}{4}$ samples time shift h and a sampling frequency fs. The size of the windowed frame is increased, by a factor of 8, using a zero-padding method to achieve better estimation of the new lower spectral bins [5].

2.1 Preprocessing

A preprocessing stage must be applied to the magnitude $X(k)$ because often it contains a high amount of spurious peaks which obstruct each fundamental frequency extraction. The resultant spectrum, $X_{th}(k)$, is composed of significant spectral harmonic peaks which describes most of specific spectral characteristics of harmonic instruments which belong to. Our peak-picking algorithm is based on adaptive-per-frame threshold T_u which selects the most prominent logarithmically weighting peaks $\bar{P}_m$ from $X(k)$. This thresholding, based on empirical tests using the University of Iowa Musical Instrument Samples [13], presents a good performance discriminating harmonic and noise peaks. The value β (see. eq. 2) is related to noise and weak-harmonics tolerance level.

$$T_u = \beta \log_2 \bar{P}_m \quad (2)$$

$$X_{th}(k) = \begin{cases} |X(k)| & |X(k)| \geq T_u \\ 0 & |X(k)| < T_u \end{cases} \quad (3)$$

2.2 Selection of F0 candidates

Each F0 candidate represents a possible active pitch in the analyzed frame. A F0 candidate is whatever frequency bin k from $X_{th}(k)$ whose frequency is located from $C2$ (65.4 Hz or MIDI number 36) to $B6$ (1976.0 Hz or MIDI number 95) in a well-tempered music scale.

This system cannot detect a note-event with missing fundamental because does not exist its F0 candidate. We do not use information from musical instrument modeling to estimate octave note-events [14]. In our system, an octave 2F0 candidate can exist only if the amplitude of the octave fundamental is higher than 2 times the amplitude of the non-octave F0 candidate.

2.3 Construction of spectral harmonic patterns

For each F0 candidate, a spectral harmonic pattern is estimated in the log-frequency domain. This log-domain exhibits the following advantage respect to linear-domain which minimizes the loss of harmonics due that spectral location of these ones regarding to its fundamental frequency is constant [12]. As consequence, a more accurate harmonic pattern construction is achieved to handle a major number of non-overlapped partials to resolve the overlapped partials.

$H^O_{F_0}$ is defined as the harmonic pattern of linear fundamental frequency F_0 and order O. The partial n^{th}, represented by the frequency bin $k^n_{F_0}$, is found searching the nearest frequency bin from non-inharmonicity harmonic within a spectral range $U^n_{F_0} = [log_{10}F_0+log_{10}n\text{-}log_{10}2^{\frac{1}{24}}, log_{10}F_0+log_{10}n+log_{10}2^{\frac{1}{24}}]$, that is, around $\pm\frac{1}{2}$ semitone from the n^{th} non-inharmonicity harmonic belonging to the fundamental frequency F_0. The partial n^{th} is considered as non-existing partial if no frequency bin is found in $U^n_{F_0}$ limits.

Our system establishes an upper frequency F_H to group partials belonging to a harmonic pattern. All spectral content located above F_H is discarded because the magnitude of these partials is considered as negligible information.

2.4 Search space exploration

The search space ψ, composed of all possible F0 candidates combinations C_ψ, increases exponentially when a new F0 candidate is added. The number of combinations can be seen as a Combinatorics without repetition problem where its size $S_{C_\psi} = \Sigma_{n=1}^{P_{max}} C^n_m = \Sigma_{n=1}^{P_{max}} \binom{m}{n} = \Sigma_{n=1}^{P_{max}} \frac{m!}{n!(m-n)!}$, being m the total number of candidates, n the number of simultaneous candidates at a time and P_{max} the maximum polyphony considered in the analyzed signal. In order to reduce C_ψ, only the most E prominent harmonic patterns are considered (P_{max}=E).

3. GAUSSIAN MIXTURE MODEL ESTIMATION

We assume that a polyphonic magnitude spectrum is additive, in other words, can be seen as a sum of GMM spectrums. $GMM^O_{n_t}(k)$ is a GMM model, related to n^{th} combination of F0 candidates within the search space ψ at the frame t using O normal gaussian functions (see eq. 4), weighted by amplitudes $A^i_{F_0}$, centered in frequencies determined by the spectral pattern $H^O_{F_0}$ and a full width at half maximum $FWHM$ equal to $\frac{1.5fs}{N} < \frac{4fs}{N}$ in order to capture most of the energy belonging to a harmonic peak and avoid interference out of the window spectral main-lobe. The weights $A^i_{F_0}$ (see eq. 5) belonging to a GMM model are composed of non-overlapped $A^j_{F0_{NOV}}$ and/or overlapped $A^m_{F0_{OV}}$ partial amplitudes.

$$GMM^O_{n_t}(k) = \sum_{i=1}^{O} A^i_{F_0} e^{(\frac{-2(k-k^i_{F_0})Ln(2)}{FWHM})^2} \quad (4)$$

$$A^i_{F_0} = A^j_{F0_{NOV}} \bigcup A^m_{F0_{OV}}, \quad i = j \cup m \quad (5)$$

Since non-overlapped partials are not interfered by other F0 candidates, their amplitudes $A^j_{F0_{NOV}}$ are considered as credible information. From this information, we estimate overlapped partial amplitudes $A^m_{F0_{OV}}$ by means of linear interpolation using the nearest neighboring non-overlapped partials, as in [5]. Fig. 2 shows the multitimbral magnitude spectrum of a frame composed of five instrument sounds from [13] ($F0_1$ Tenor Trombone, $F0_2$ Bassoon, $F0_3$ Flute, $F0_4$ Bb Clarinet and $F0_5$ Eb Clarinet), and F0 candidates combinations using GMM spectrums estimated by our system. It can be observed that a correct multiple-F0 estimation increases the spectral similarity between input-GMM modeling.

4. TEMPORAL-SPECTRAL SIMILARITY

Our assumption is that a current polyphonic music note-event depends to a large extent of the previous one. Tak-

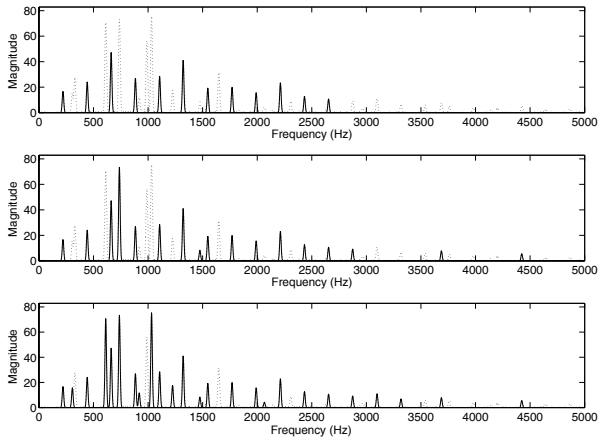


Figure 2: Magnitude spectrum $X(k)$ (dashed line) of an analyzed frame and GMM combinations (solid line) estimated by our system. The input spectrum $X(k)$ is composed of five different instrument sounds ($F0_1^{MIDI57}$=220.0 Hz, $F0_2^{MIDI63}$=311.1 Hz, $F0_3^{MIDI64}$=329.6 Hz, $F0_4^{MIDI78}$=740.0 Hz and $F0_5^{MIDI84}$=1047.0 Hz). In top plot, GMM composed of one harmonic sound $F0_1$. In middle plot, GMM composed of two harmonic sounds $F0_1$ + $F0_4$. In bottom plot, GMM composed of four harmonic sounds $F0_1$ + $F0_2$ + $F0_4$ + $F0_5$.

ing into account C_ψ combinations of spectrums $GMM_{n_t}^O(k)$, $n \in [1, S_{C_\psi}]$, instead of using spectral features of harmonic sounds as occurs in [4][5], our system attempts to replicate the input polyphonic signal. Therefore, we consider that the most likely combination c_{winner} will exhibit the highest spectral similarity regarding to immediately past music event. This combination c_{winner} is selected from a subset $C_{candidates}$, where $C_{candidates} \subset C_\psi$, which minimizes the current spectral distance related to the current input spectrum $X(k)$. Next, our selection criterion is detailed.

4.1 First stage. Similarity in spectral domain

Considering the temporal frame t, our system calculates the spectral Euclidean distance DC_{n_t} (see eq. 6) for each combination n. This spectral similarity attempts to explain most of the harmonic peaks present in the analyzed signal.

$$DC_{n_t} = \sum_k (|X(k)| - GMM_{n_t}^O(k))^2, \quad n_t \in C_\psi \qquad (6)$$

4.2 Second stage. Similarity in temporal domain

Spectral information is not sufficient to perform an accurate multiple-F0 estimation since it is common that part of a note-event often is missed because of several reasons such as high polyphony, harmonic relations between overlapped partials or low energy notes-events. To overcome this problem, we assume that in polyphonic music a note-event depends to a large extent of the immediately previous one. In this way, we select a subset of combinations ($C_{candidates}$) which minimize the spectral similarity regarding to the current analyzed frame. A temporal window of Υ previous frames is considered in order to add temporal information. Temporal information allows to compare similarities between the last winner combinations and the $C_{candidates}$ combinations estimated in the current frame (see eq. 7).

$$DP_{n_t}^\Upsilon = \prod_\Upsilon \sum_k (GMM_{n_t}^O(k) - GMM_{c_{winner_{t-\Upsilon}}}^O(k))^2 \qquad (7)$$

where $n_t \in C_{candidates}$

4.3 Third stage. Combination of temporal-spectral similarity

The combination c_{winner} (eq. 9) is determined maximizing the temporal-spectral similarities, in other words, minimizing the distance $DT_{n_t}^\Upsilon$.

$$DT_{n_t}^\Upsilon = DC_{n_t} DP_{n_t}^\Upsilon \qquad (8)$$

$$c_{winner} = arg\ min_{n_t \in C_{candidates}} DT_{n_t}^\Upsilon \qquad (9)$$

5. EXPERIMENTAL RESULTS

Our system was tested using 5 excerpts of real-world monaural polyphonic music signals from RWC Music Database [15]. These excerpts represents 36% of evaluation test used in [12] which were chosen randomly. For each excerpt, approximately the first 20 seconds were selected for the analysis. The parameters used by our system are shown in Table 1. In order to minimize spurious events, we only consider events which present a significant musical time duration $t > T_{min}$.

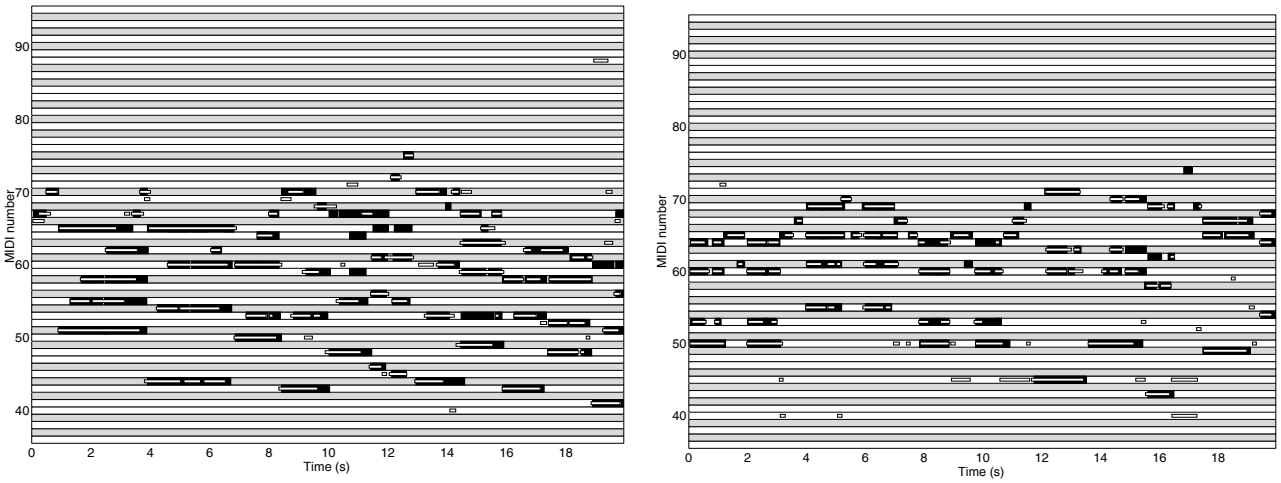
fs (Hz)	44100
N (samples)	4096 $\approx$ (92.9 ms)
h (samples)	1024 $\approx$ (23.2 ms)
O (partials)	12
F_H (Hz)	5000
E (candidates)	5
$FWHM$ (Hz)	16
$C_{candidates}$	5
Υ	1
$T_{min}(ms)$	100

Table 1: Parameters of the proposed system

The MIDI files, from RWC Music Database, used for the evaluation test have been manually corrected because present temporal inaccuracies regarding to onsets and offsets of the reference note-events which drastically decrease the estimated accuracy.

Accuracy measure was calculated in a frame level matching reference and transcribed events using the metrics proposed in [12]. In Table 2, we only present one accuracy measure because this one is the unique measure provided in [12]. In order to provide more helpful information about our system performance, additional error measures (total error E_{tot}, substitution error E_{sub}, miss error E_{miss} and false alarm error E_{fa}) using the metrics proposed in [2] are depicted in Table 3. These last measures are more suitable for polyphonic music transcription because provide information about possible weaknesses of the evaluated system.

The results, in percentages (%), of comparing our system and a recent state-of-the-art system [12] are shown in Table



(a) RWC-MDB-J-2001 No.7

(b) RWC-MDB-J-2001 No.9

Figure 3: Polyphonic transcription of the first 20 seconds of two excerpts from RWC Music Database. *x-axis* indicates time in seconds. *y-axis* indicates MIDI events from MIDI number 36 to MIDI number 95. Each white and gray row represents a white and black key of a standard piano. Reference note-events (black rectangles) and transcribed note-events (white rectangles) are displayed.

RWC identifier	Instruments	Proposed	Specmurt [12]
RWC-MDB-J-2001 No.7	G	**69.6%**	68.1%
RWC-MDB-J-2001 No.9	G	68.8%	**77.5%**
RWC-MDB-C-2001 No.35	P	61.1%	**63.6%**
RWC-MDB-J-2001 No.12	F + P	38.3%	**44.9%**
RWC-MDB-C-2001 No.12	F + VI + VO + CE	41.9%	**48.9%**
	Average result	55.9%	**60.6%**

Table 2: Accuracy measure based on the metrics proposed in [12]. Specmurt analysis uses a β=0.2. Instruments: Guitar (G), Piano (P), Flute (F), Violin (VI), Viola (VO), Cello (CE)

	Proposed				
RWC identifier	*Acc*	E_{tot}	E_{sub}	E_{miss}	E_{fa}
RWC-MDB-J-2001 No.7	69.6%	30.5%	8.2%	17.3%	5.0%
RWC-MDB-J-2001 No.9	68.8%	31.2%	6.3%	14.1%	10.8%
RWC-MDB-C-2001 No.35	61.1%	38.8%	8.4%	23.0%	7.4%
RWC-MDB-J-2001 No.12	38.3%	61.7%	16.2%	44.4%	1.1%
RWC-MDB-C-2001 No.12	41.9%	58.0%	15.2%	3.0%	39.8%

Table 3: Accuracy and error measures based on the metrics proposed in [2] regarding to the results shown in Table 2.

2. Our proposed system presents a promising performance since achieves an average accuracy of 55.9% versus 60.6% by Saito's system [12]. Moreover, our system is able to transcribe multitimbral polyphonic music because exhibits a robust behavior independently of the spectral characteristics of the harmonic instruments which compose the mixture signal. Table 3 suggests that most of the errors are due to miss note-events. Fig. 3(a) and Fig. 3(b) indicate that most of reference note-events are correctly estimated while octave note-events are missed.

6. CONCLUSIONS AND FUTURE WORK

This paper presents a system to transcribe polyphonic music based on a joint multiple-F0 estimation. The main idea consists of combining temporal and spectral similarities of GMM spectrums in order to replicate the polyphonic input signal under assumption that a current musical event depends to a large extent of the immediately previous one.

Our system shows encouraging results achieving an average accuracy of 55.9% versus 60.6% of a recent state-of-the-art system [12]. Moreover, the proposed system is able to transcribe multitimbral polyphonic music because exhibits a robust behavior independently of the harmonic instruments which compose the mixture signal.

Our future work will be focused on a more accurate overlapped partials estimation to minimize misses due to octave events.

REFERENCES

[1] Alonso, M., Richard, G. & David, B., "Extracting note onsets from musical recordings," In Proceedings of IEEE International Conference on Multimedia and Expo, Amsterdam, The Netherlands, 2005.

[2] Poliner, G., Ellis, D., "A discriminative model for polyphonic piano transcription," EURASIP Journal on Advances in Signal Processing, vol. 8, pp. 19, 2007.

[3] Goto, M., "A real-time music-scene-description system: predominant-F0 estimation for detecting melody and bass lines in real-world audio signals," Speech Communication, vol. 43, no.4, pp.311-329, September 2004.

[4] Yeh, C., Robel, A., & Rodet, X., "Multiple fundamental frequency estimation of polyphonic music signals," in IEEE, Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, USA, 2005.

[5] Pertusa A., Inesta J.M., "Multiple Fundamental Frequency estimation using Gaussian smoothness," Proc.

of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP 2008, pp.105-108, Las Vegas, USA, 2008.

[6] Li, Y., Wang, D.L., "Pitch detection in polyphonic music using instrument tone models," Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, ICASSP, pp. 481-484, Hawaii, USA, 2007.

[7] Kameoka, H., Nishimoto, T., & Sagayama, S., A Multipitch Analyzer Based on Harmonic Temporal Structured Clustering," IEEE Trans. Audio, Speech and Language Processing, vol. 15, no. 3, pp. 982994, 2007.

[8] Bello, J., and Daudet, L. & Sandler, M., Automatic piano transcription using frequency and time-domain information, IEEE Trans. Acoustic, Speech and Signal Processing, vol. 14, no. 6, pp. 22422251, November, 2006.

[9] Klapuri, A., "Multipitch analysis of polyphonic music and speech signals using an auditory model," IEEE Trans. Audio, Speech and Language Processing, vol. 16. no. 2, pp. 255-266, February, 2008.

[10] Ryynnen, M., Klapuri, A., "Polyphonic music transcription using note event modeling, " in Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, New York, October, 2005.

[11] Cañadas, F.J., Vera, P., Ruiz, N., Mata, R. & Carabias, J., "Note-event detection in polyphonic musical signals based on harmonic matching pursuit and spectral smoothness," Journal of New Music Research, vol. 37, no. 3, pp- 167-183, December, 2008.

[12] Saito, S., Kameoka, H., Takahashi, K., Nishimoto, T., & Sagayama, S., "Specmurt Analysis of Polyphonic Music Signals," IEEE Trans. on Audio, Speech and Language Processing, vol.16, no. 3, pp. 639-650, 2008.

[13] The University of Iowa Musical Instrument Samples, http://theremin.music.uiowa.edu/MIS.html [Online]

[14] Monti, G., Sandler, M., "Automatic Polyphonic Piano Note Extraction Using Fuzzy Logic in a Blackboard System," Proc. of the 5^{th} Int. Conference on Digital Audio Effects (DAFX), Hamburg, Germany, September, 2002.

[15] Goto, M., Hashiguchi, H., Nishimura, T., & Oka, R., "RWC music database: Popular, classical, and jazz music database," in Proc. Int. Symp. Music Inf. Retrieval, pp. 287-288, Oct. 2002.