A GENETIC ALGORITHM FOR POLYPHONIC TRANSCRIPTION OF MUSIC

Gustavo Reis
Polytechnic Institute of Leiria
Portugal
`gustavo.reis1@ipleiria.pt`

Francisco Fernandéz
University of Extremadura
Spain
`fcofdez@unex.es`

Aníbal Ferreira
University of Porto
Portugal
`ajf@fe.up.pt`

ABSTRACT

This paper describes our method, submitted to MIREX 2011 task “Multiple Fundamental Frequency Estimation & Tracking”. This task restricted the problem of Multiple F0 Estimation and Tracking to three cases: i) Estimate active fundamental frequencies on a frame-by-frame basis; ii) Track note contours on a continuous time basis (as in audio-to-midi); iii) Track timbre on a continuous time basis. The presented method is a genetic algorithm approach to polyphonic music transcription. Two versions of our method have been submitted: RFF1 and RFF2. RFF1 is intended for transcription of piano music. RFF2 is RFF1 with support to other kinds of pitched instruments.

1. SYSTEM OVERVIEW

Our system consists of three phases: (1) audio segmentation, (2) transcription process, and (3) note duration adjustment (see Algorithm 1).

Algorithm 1 System overview

```
Detect onsets
Split audio stream in segments between subsequent onsets
for each segment do
    Apply 50 generation Genetic Algorithm
end for
Apply a Hill-Climber to adjust the duration of notes transversing several segments
return MIDI file
```

During the Audio segmentation an onset detector is applied on the input signal to extract onset information. Afterwards, the audio signal is divided into several audio segments, according to the detected onsets. Each interval between two consecutive onsets is considered a segment. Then, for each segment, a *thread* is launched running a 50 generation genetic algorithm to perform the corresponding transcription. The search for the most-likely combination of F0s to model the audio input signal is aided by an internal database of previously recorded piano samples. The genetic algorithm also adapts the spectral envelope of the used piano samples in order to best match the power spectrum of the corresponding audio segment [3]. During this process, the spectral envelope of the residual noise is also dynamically estimated to favor the search process towards the desired solution. The results obtained on each audio segment are then merged in one whole transcription. Finally, a hill-climber algorithm is applied on the global transcription to adjust the duration of several notes or merge notes that transverse several segments. The output of the system is the final result of the hill-climber.

2. PROPOSED GENETIC ALGORITHM

It is important to emphasize that the main idea behind a genetic algorithm [2] is to have a set of candidate solutions (individuals) to a problem evolving towards the desired solution. In each iteration (generation) those candidate solutions are evaluated according to their quality (fitness). The worst solutions are then discarded and the best will generate new candidate solutions, resulting from the combination of their parent’s characteristics (genes) and minor variations (mutation). This way, candidate solutions with better quality tend to live longer and to generate better solutions, improving the robustness of the algorithm.

2.1 Genotype

Since the problem being solved is the automatic transcription of an audio segment, a candidate solution must be a candidate transcription. We consider a transcription as a set of musical notes, where each note has four attributes: start time, duration, MIDI note and also MIDI velocity. Therefore an individual is encoded as a chromosome with a set of genes, where each gene corresponds to a musical note.

The genotype includes additional chromosomes for both the spectral envelope of the original piano and the residual noise.

2.2 Fitness Evaluation

To evaluate candidate transcriptions, first we need to render them to an audio signal and then compare the corresponding audio signals to the input audio segment. Transcrip-

	BD2	BD3	LYC1	RFF1	RFF2	YR1	YR3
Ave. F-Measure Onset-Offset	0.2036	0.2077	0.2076	0.1767	0.1414	**0.3493**	0.3392
Ave. F-Measure Onset Only	0.4465	0.4506	0.3862	0.4078	0.3564	**0.5601**	0.5465
Ave. F-Measure Chroma	0.2307	0.2438	0.2573	0.2029	0.1655	**0.3579**	0.3470
Ave. F-Measure Onset Only Chroma	0.5026	0.5232	0.4649	0.4566	0.3986	**0.5647**	0.5519

Table 1. Note Tracking Mixed Set Overall Summary Results

	BD2	BD3	LYC1	RFF1	RFF2	YR1	YR3
Ave. F-Measure Onset-Offset	0.1003	0.1136	0.1926	0.1941	0.1550	**0.2127**	0.1913
Ave. F-Measure Onset Only	0.5263	0.5890	0.5260	0.5205	0.4435	**0.6055**	0.5881
Ave. F-Measure Chroma	0.1098	0.1205	0.2068	**0.2261**	0.1944	0.1966	0.1800
Ave. F-Measure Onset Only Chroma	0.5400	**0.5996**	0.5412	0.5645	0.4930	0.5547	0.5391

Table 2. Note Tracking Piano-Only Overall Summary Results

tions whose audio is similar to the audio input are close to the desired solution and, thus, have less errors. The comparison between the candidate transcriptions and the input audio segment is done in the frequency domain.

The current fitness function is based on the Log Spectral Distance or Log Spectral Distortion, which corresponds to the spectral distance (expressed in dB) between two spectra:

$$D_{LS} = \sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\frac{P(\omega)}{\widehat{P}(\omega)}\right]^2 d\omega} \qquad (1)$$

where $P(\omega)$ is the original magnitude spectrum and $\widehat{P}(\omega)$ is the model spectrum, which is guaranteed to be non-zero. The fitness function is defined by Equation 2. $|X(n,k)|$ is the magnitude of the kth bin from the nth frame of the original spectrum, $|\widehat{X}(n,k)|$ represents the magnitude of the kth bin of the nth frame of the model spectrum (candidate solution being evaluated). N is the size of the Hamming window, which is 93 ms (i.e. $N = 4096$ with 44100 Hz sampling rate). k starts in 2 because it is the bin corresponding to the frequency of the first piano note (A_0 = 27,5 Hz).

$$\sum_{n=1}^{nMax}\sqrt{\sum_{k=2}^{\frac{N}{2}}\left(\left[10\log_{10}\frac{|X(n,k)|}{\left|\widehat{X}(n,k)\right|}\right]^2\times\left[\log_2\left(1+\frac{1}{k}\right)\right]\right)} \qquad (2)$$

The multiplication by $\log_2\left(1+\frac{1}{k}\right)$ happens to normalize the weight of the bins of each octave so that, when summed, all the octaves have the same weighted sum equal to 1.

3. RESULTS

If we look into Table 2, RFF1 achieves the second best algorithm on Onset-Offset on transcribing Piano Music. If we take into account that he YR1 is currently the best Automatic Music Transcription Algorithm, our results seems encouraging. Onset-Offset metric takes into account MOR which means that our algorithm have good subjective quality because it has good phrasing similarity with the original pieces. On the other hand, on the Onset Only metric our algorithm performs as the worst algorithm.In Chroma evaluations, F0's are mapped to a single octave before evaluating. Here our algorithm performs as the best on Onset-Offset Chroma and second best on Onset Only Chroma. This means that our approach was efficient in transcribing non-harmonic related notes. Compared to the system RFF1 [1] submitted for the MIREX08 task, this system reports an F-Measure increase of 25.25%.

RFF2 has worst results because it was blindly adapted to multi-timbral. Since RFF2 deals with a bigger database of internal samples, the size of the search space is much greater and this requires different parameters.

4. REFERENCES

[1] G. Reis, F. Fernandez, and A. Ferreira, "Genetic Algorithm Approach to Polyphonic Music Transcription for MIREX 2008," in *Music Information Retrieval Evaluation eXchange*, Philadelphia, USA, September, 2008.

[2] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley Professional, January 1989.

[3] G. Reis, F. Fernandez, and A. Ferreira, "Transcripcin de música multi-timbre mediante algoritmos genéticos," in *MAEB 2009 VI Congreso Espaol sobre Metaheursticas, Algoritmos Evolutivos y Bioinspirados*, February 2009.