MELODY EXTRACTION FROM POLYPHONIC MUSIC AUDIO

Justin Salamon
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
justin.salamon@upf.edu

Emilia Gómez
Music Technology Group
Universitat Pompeu Fabra, Barcelona, Spain
emilia.gomez@upf.edu

ABSTRACT

In this paper we provide a description of our melody extraction algorithm submitted to the MIREX 2010 competition. We start with a brief introduction to the task of melody extraction, followed by an overview of our proposed method. The data collections and metrics used for the evaluation are described, followed by the presentation of the results and some comments on the performance of our approach.

1. INTRODUCTION

Over recent years we have seen substantial growth in the distribution and consumption of digital audio. With musical collections reaching vast numbers of songs, we now require novel ways of describing, indexing, searching and interacting with music. One approach to address this issue is through the automatic extraction of the melodic line of a piece, as the melody is often recognised as the *'essence'* of a musical piece [8]. Melody estimation has many potential applications, an example being the creation of large databases for music search engines based on Query by Humming (QBH) or by Example (QBE) [3]. In addition to retrieval, melody extraction can form a core component in other music computation tasks involving transcription and classification. What is more, determining the melody of a song could be used as an intermediate step towards the determination of semantic labels from musical audio, thus helping to bridge the *semantic gap* [11].

Much effort has been devoted to melody extraction from polyphonic music, a task made difficult both by the interference of other audio sources in the polyphonic music audio signal and the challenge of determining which of the detected sources is the melody line [2]. One of the first to demonstrate successful melody extraction from real world audio signals was Goto with his well-known PreFEest system [6]. In this work he proposes to represent the melody as time dependent sequences of fundamental frequency values, which has become the standard representation in melody estimation systems [8]. A testimony to the increasing cumulative effort at automatic melody extraction over the past years is the melody extraction task at the Music Information Retrieval Evaluation eXchange (MIREX) competition, this year being the sixth repetition of the event.

2. METHOD

The overall structure of our algorithm is based on the one proposed by Dressler in [5], incorporating ideas from other successful approaches such as the ones presented by Cancela [2] and Klapuri [7], as detailed below.

2.1 Spectral Analysis

We start by applying an equal loudness filter to the signal in an attempt to better approximate human perception of loudness as dependent on frequency [9]. The signal is then split into frames using the Short-Time Fourier Transform (STFT) using a Hann windows with a window size of 46ms. The peaks of the spectrum are obtained, and we use the phase spectrum to compute their instantaneous frequency (IF) as proposed in [4], which gives us a refined estimation of the peak frequency. The phase spectrum is also used to detect quasi-stationary sinusoidal peaks and filter out the rest, obtaining a "cleaner" representation.

2.2 Salience Function Generation

The remaining spectral peaks are used to generate a "Salience Function", a representation of pitch salience over time. This is done using harmonic summation similar to the one performed in [10], extended to a five octave range (55Hz–1760Hz).

2.3 Segment Creation and Filtering

The peaks of the salience function are then used to create *segments*, continuous sequences of peaks formed by grouping the peaks over time and frequency using proximity and continuity rules based on the concepts of auditory stream analysis introduced by Bregman [1]. Once created, the salience of each segment is refined by recomputing the harmonic summation of its constituent peaks using an approach similar to [7]. Voicing (determining whether the melody is present or not) is handled by filtering out weak segments. We then estimate the pitch mean for the melody from the remaining segments using an iterative process which involves detecting and removing octave errors, removing pitch outliers (segments far from the most recently

estimated pitch mean), and weighting the mean computation by exploiting segment characteristics such as energy, pitch height, contour, the presence of vibrato and proximity to the previously calculated pitch mean.

2.4 Melody Selection

After several iterations a pitch mean is produced which is used for selecting the final melody. Octave errors and pitch outliers with respect to the mean are removed, and for every frame the peak belonging to the most salient segment present is selected. This means that the pitch mean estimation process is a crucial part of the algorithm, and the success of the algorithm depends on how well the estimated pitch mean approximates the true melodic line.

3. EVALUATION

3.1 Evaluation Collections

Four music collections are used for the evaluation, as detailed in Table 1.

Collection	Description
ADC2004	20 excerpts of roughly 20s in the genres of pop, jazz and opera.
MIREX05	25 phrase excerpts of a 10-40s duration in the genres of Rock, R&B, Pop, Jazz and Solo classical piano.
MIREX08	Four 1 minute long excerpts from north Indian classical vocal performances.
MIREX09	374 Karaoke recordings of Chinese songs. Each recording is mixed at three different levels of Signal-to-Accompaniment Ratio {-5dB, 0dB, +5 dB} for a total of 1,122 audio clips.

Table 1. Evaluation collections for MIREX 2010.

3.2 Evaluation Metrics

The algorithms are evaluated in terms of voicing recall, voicing false alarm, raw pitch, raw chroma, and overall accuracy which combines both pitch and voicing performance. Further details on the evaluation metrics can be found in [8]. It should be noted that algorithms are allowed to return negative pitch values for frames which are determined by the algorithm as non-voiced, thus allowing to evaluate the pitch and chroma performance independently of the voicing performance (however results will also depend on how coupled the voicing and pitch detection parts of each algorithm are).

3.3 Results

The results for all participating algorithms are presented in Table

$$TABLE$$

Genereal comments about the results go here.

4. CONCLUSION AND FUTURE WORK

Comments about our results and some conclusions. Things we want to improve.

5. ACKNOWLEDGMENTS

We would like to thank the IMIRSEL team at the University of Illinois at Urbana-Champaign for running MIREX. The research is funded by the Programa de Formación del Profesorado Universitario (FPU) of the Ministerio de Educación de España.

6. REFERENCES

[1] A. Bregman. *Auditory scene analysis*. MIT Press, Cambridge, Massachussetts, 1990.

[2] P. Cancela. Tracking Melody in Polyphonic Audio. In *Proc. MIREX*, 2008.

[3] R. B. Dannenberg, W. P. Birmingham, B. Pardo, N. Hu, C. Meek, and G. Tzanetakis. A Comparative Evaluation of Search Techniques for Query-by-Humming Using the MUSART Testbed. *Journal of the American Society for Information Science and Technology*, February 2007.

[4] K. Dressler. Sinusoidal Extraction using an Efficient Implementation of a Multi-resolution FFT. In *Proc. of the Int. Conf. on Digital Audio Effects (DAFx-06)*, pages 247–252, Montreal, Quebec, Canada, Sept. 2006.

[5] Karin Dressler. Audio melody extraction for mirex 2009. In *5th Music Information Retrieval Evaluation eXchange (MIREX)*, 2009.

[6] M. Goto. A real-time music-scene-description system: predominant-f0 estimation for detecting melody and bass lines in real-world audio signals. *Speech Communication*, 43:311–329, 2004.

[7] A. Klapuri. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proc. 7th International Conference on Music Information Retrieval*, Victoria, Canada, October 2006.

[8] G. E. Poliner, D. P. W. Ellis, F. Ehmann, E. Gómez, S. Steich, and O. Beesuan. Melody transcription from music audio: Approaches and evaluation. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4):1247–1256, 2007.

[9] D. W. Robinson and R. S. Dadson. A re-determination of the equal-loudness relations for pure tones. *British Journal of Applied Physics*, 7:166–181, 1956.

[10] J. Salamon and E. Gómez. A Chroma-based Salience Function for Melody and Bass Line Estimation from Music Audio Signals. In *Sound and Music Computing Conference*, pages 331–336, Porto, Portugal, July 2009.

[11] X. Serra, R. Bresin, and A. Camurri. Sound and Music Computing: Challenges and Strategies. *Journal of New Music Research*, 36(3):185–190, 2007.