MELODY EXTRACTION FROM POLYPHONIC MUSIC: MIREX 2011

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 describe our submission for the audio melody extraction task of the Music Information Retrieval Evaluation eXchange (MIREX) 2011 campaign. The system presented here is an updated version of the one submitted to last year's campaign. Following a detailed analysis of each step of our method, system parameters have been optimised for melody extraction and the implementation is now more efficient. Two variants of the system have been submitted, each making use of a different spectral transform, allowing us to asses whether the difference between them is significant for overall performance.

Following the description of the system, we describe the data-sets and metrics used for evaluation. This is followed by a summary of the results and some conclusions.

1. INTRODUCTION

The goal of melody extraction systems is to automatically detect the predominant melodic line of a piece of music, and output a representation of this line. Traditionally, a mid-level representation is used [5], comprised of a sequence of time-stamps and fundamental frequency (F0) values.

In the following sections we describe our melody extraction system, which is an updated version of the one presented in [7]. The system is comprised of four main blocks: sinusoid extraction, salience function computation, pitch contour creation, and melody selection. In [8], the first two blocks of the system were studied in detail. Different signal processing methods were compared for sinusoid extraction, and the parameters of the salience function were optimised for melody estimation. The system presented here incorporates the conclusions reached in the aforementioned study, and in the results section we will asses how the updated system performs compared to last year's submission.

2. METHOD

2.1 Sinusoid Extraction

In the first block of the system we analyse the audio signal and extract spectral peaks (sinusoids) which will be used to construct the salience function in the next block. This process is comprised of three main steps: pre-filtering, transform and frequency/amplitude correction. In the pre-filtering stage we apply the (time-domain) equal loudness filter [1], which was shown in [8] to attenuate spectral components belonging primarily to non-melody sources. Next, we apply a spectral transform and select the peaks of the magnitude spectrum for further processing. For MIREX 2011, two variants were submitted: in the first (SG1), we use the Short-Time Fourier Transform with a 46ms Hann window. In the second variant (SG2), we use the multi-resolution FFT (MRFFT) proposed in [3], combining spectral peaks from windows with of varying lengths (from 5.8ms up to 46ms). In both cases we use a hop size of 2.9ms and a $\times 4$ zero padding-factor. In this way, we are be able to asses whether the difference between a single and multi-resolution transform is indeed significant for melody extraction, at least in the case of our system. In the third step the frequency and amplitude of the selected peaks are re-estimated by calculating the peaks' instantaneous frequency (IF) using the phase vocoder method [4]. The reader is referred to [8] for further details.

2.2 Salience Function

Next the spectral peaks are used to compute a representation of pitch salience over time, a *salience function*. Our salience function is based on harmonic summation with magnitude weighting, and spans a range of almost five octaves from 55Hz to 1760Hz. Further details are provided in [8]. In that study the parameters of the salience function were optimised for melody extraction by evaluating it directly using metrics designed to estimate the predominance of the true melody F0 compared to peaks in the salience function caused by other sources. In the results section we will examine how this optimisation affects the overall performance of the complete system.

2.3 Pitch Contour Creation and Melody Selection

In the next block, the peaks of the salience function are grouped over time using heuristics based on auditory stream-

ing cues [2]. This results in a set of pitch contours, out of which the contours belonging to the melody need to be selected. The contours are automatically analysed and a set of contour characteristics is computed. In the final block of the system, the contour characteristics and their distirbutions are used to filter out non-melody contours. First we remove contours whose features suggest that there is no melody present in this segment of the piece (voicing detection). The remaining contours are used to iteratively calculate an overall melody pitch trajectory, which is used to minimise octave errors and remove pitch outliers. Finally, contour salience features are used to select the melody F0 at each frame from the remaining contours.

3. EVALUATION METHODOLOGY

3.1 Evaluation Collections

Four music data-sets are used for the evaluation, as detailed in Table 1. Note that the excerpts in the MIREX09 data-set were used to create three test collections each using a different signal-to-accompaniment ratio {-5dB, 0dB, +5dB}, resulting in a total of 6 test collections.

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 in three collections: MIREX09 -5dB, MIREX09 0dB and MIREX09 +5dB.

Table 1. Evaluation collections for MIREX 2011.

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 [6]. The algorithms are allowed to return negative pitch values for frames which they determine as non-voiced, which allows us to independently evaluate pitch (and chroma) estimation performance and voicing detection performance.

4. RESULTS AND COMMENTS

In Table 2 we present the overall accuracy results for all participating algorithms (our submissions are SG1 and SG2). The best score achieved for each test-set is highlighted in bold. In addition to the per test-set results the table provides the mean overall accuracy averaged over the six test-sets. We compute both the unweighted mean and a weighted mean where the overall accuracy obtained for each test-set is weighted by its total playtime.

We see that both variants of our algorithm achieve the highest overall accuracy in four of the six test-sets. Consequently, our method also achieves the highest mean overall accuracy for both unweighted and weighted cases.

It is also interesting to observe that the unweighted and weighted means are practically the same for our method, indicating relatively consistent performance across all data-sets. The only exception to this is the MIREX 2005 data-set, where the performance of our method is relatively low [1] . The cause for this is most likely the higher proportion of instrumental excerpts (i.e. the melody is played by an instrument rather than sung) for which our method does not perform as well.

Next, we see that there is no significant difference in performance between our two submissions (SG1 and SG2). This reinforces the conclusions reached in [8], in which it was shown that for data-sets containing a varied selection of excerpts from different genres, the difference between a single resolution transform (STFT) and a multi-resolution one (in our case the MRFFT [3]) is not significant for melody extraction.

Finally we compare our results with those obtained by our algorithm in last year's MIREX evaluation, provided in Table 3. We see that for all data-sets there is a significant improvement in the overall accuracy. This confirms that the parameter optimisation carried out in [8] indeed results in a significant improvement in the overall performance of our method [2] .

5. ACKNOWLEDGEMENTS

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, COFLA (P09-TIC-4840-JA) and DRIMS (TIN2009-14247-C02-01-MICINN).

6. REFERENCES

[1] Equal loudness filter, July 2011.

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

[3] 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.

[1] The performance for the MIREX09 (-5dB) test-set is expected to be lower than the rest due to the signal to accompaniment mixing ratio.

[2] The parameter optimisation in [8] was done using a different data-set from the ones used for evaluation in MIREX, meaning the results are not affected by parameter overfitting.

Algorithm	ADC 2004	MRX 2005	MRX 2008	MRX 09 (0dB)	MRX 09 (-5dB)	MRX 09 (+5dB)	Mean (Unweighted)	Mean (Weighted)
TY3	0.47	0.51	0.70	0.52	0.41	0.56	0.53	0.50
TY4	0.47	0.51	0.70	0.52	0.41	0.56	0.53	0.50
TOS1	0.59	0.57	0.72	0.74	**0.62**	0.82	0.68	0.72
LYRS1	0.73	0.59	0.72	0.47	0.36	0.54	0.57	0.47
HCCPH1	0.44	0.45	0.64	0.50	0.39	0.59	0.50	0.50
CWJ1	0.73	0.57	0.69	0.53	0.40	0.62	0.59	0.52
YSLP1	**0.85**	0.65	0.73	0.52	0.39	0.66	0.63	0.53
PJY1	0.81	0.65	0.71	0.74	0.54	0.83	0.71	0.70
SG1	0.74	0.66	0.83	**0.78**	0.61	**0.85**	0.74	**0.75**
SG2	0.74	**0.68**	**0.84**	**0.78**	0.61	**0.85**	**0.75**	**0.75**

Table 2. Overall accruacy results: MIREX 2011.

Algorithm	ADC 2004	MRX 2005	MRX 2008	MRX 09 (0dB)	MRX 09 (-5dB)	MRX 09 (+5dB)	Mean (Unweighted)	Mean (Weighted)
SG (2010)	0.70	0.62	0.78	0.74	0.58	0.81	0.70	0.71
SG1 (2011)	0.74	0.66	0.83	0.78	0.61	0.85	0.74	0.75

Table 3. Comparison of MIREX 2010 and MIREX 2011 results.

[4] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell Systems Technical Journal*, 45:1493–1509, 1966.

[5] 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.

[6] 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.

[7] J. Salamon and E. Gómez. Melody extraction from polyphonic music audio. In *6th Music Information Retrieval Evaluation eXchange (MIREX)*, extended abstract, Utrecht, The Netherlands, August 2010.

[8] J. Salamon, E. Gómez, and J. Bonada. Sinusoid extraction and salience function design for predominant melody estimation. In *Proc. 14th Int. Conf. on Digital Audio Effects (DAFX-11)*, Paris, France, September 2011.