MELODY EXTRACTION USING HARMONIC MATCHING

Vishweshwara Rao
Indian Institute of Technology Bombay
vishu@ee.iitb.ac.in

Preeti Rao
Indian Institute of Technology Bombay
prao@ee.iitb.ac.in

ABSTRACT

This extended abstract describes our submission to the MIREX 2008 evaluation task on Audio Melody Extraction. This algorithm has specifically been designed for vocal F0 extraction in the presence of harmonic interference. The results of the evaluation show high pitch estimation accuracy of our method but lower melodic voice detection accuracy than other submissions.

1. INTRODUCTION

The problem of melody extraction from polyphonic audio, involving the detection and extraction of the pitch contour of the lead melodic instrument, has received considerable attention from researchers in the past; as reflected by the large number of entries for audio melody extraction task in the MIREX 2004, 2005 and 2006 evaluations. Recently however several researchers have shifted focus to the multi-F0 detection problem i.e. the estimation of the melodic contours of multiple instruments that sound simultaneously; as reflected by the lack of a melody extraction task in the MIREX 2007 evaluations. The difference in the above two tasks lies primarily in the nature of the signals operated upon. While the Multi-F0 data can have two or more instruments playing different melodies simultaneously, the melody extraction data has a more clear melody vs. background distinction with the possibility of the background being more complex and richer than that for the multi-F0 problem. However melody extraction is still not a solved problem and in order to assess the performance of current systems this task was resurrected in the 2008 MIREX evaluations.

Our submission to the 2008 audio melody extraction task was primarily developed for north Indian classical vocal music. However it was also found to perform well on publicly available western music datasets (at http://www.ee.columbia.edu/projects/melody/). A brief description of the modules in our system is presented here. This is followed by a description of the evaluation data and a comparative performance evaluation of the different submissions. Conclusions and areas of future work are presented last.

2. ALGORITHM DESCRIPTION

The algorithm comprises two modules, a melody extraction module, which estimates a melodic pitch in every frame of audio, followed by a voice detection module that uses the estimated melodic pitch to determine whether the melodic voice is actually present or not.

2.1. Melody Extraction

The core pitch detection algorithm (PDA) used by the melody extraction module is based on the Two-Way Mismatch (TWM) method [1]. The TWM PDA falls under the category of harmonic matching PDAs that are based on the frequency domain matching of measured spectrum with an ideal harmonic spectrum.

The inputs to the melody extraction module are the magnitudes and frequencies of detected sinusoidal components. These are detected using a slightly relaxed main-lobe magnitude matching criterion [2] on local maxima detected from the magnitude spectrum, which is computed using a high resolution FFT with a fixed data window length of 30 ms.

Unlike typical harmonic matching algorithms that maximize the energy at the expected ideal harmonic locations, the TWM PDA minimizes a spectral mismatch error that is a particular combination of the energy of the partial and its frequency deviation from the ideal harmonic location. The error is computed by comparing the measured peaks in the signal spectrum with a predicted harmonic spectral pattern for each candidate F0. The TWM PDA has been found to be more robust to sparse, but strong, harmonic interferences as compared to other harmonic matching PDAs [3], which makes it suitable for melody extraction of a harmonically rich voice in the presence of tonal accompaniment. In the interest of computational efficiency, the present implementation of the TWM PDA computes the TWM error only at possible trial candidate F0s that are pre-computed from the list of measured sinusoidal components [4] and fall within the F0 search range (between 100 and 1280 Hz).

The TWM PDA is operated within the framework of dynamic programming-based (DP) smoothing. DP uses a combination of suitably defined local measurement and smoothness costs into a global cost, which is optimized

over a continuous voiced segment. Here the measurement cost is the TWM error, normalized to lie in the interval [0, 1]. The smoothness cost is derived from the distribution of inter-frame pitch transition values over a training dataset of clean, sung-voice pitch contours [5].

2.2. Melodic Voice Detection

The TWM-DP PDA produces an estimate of the predominant pitch at every instant irrespective of the underlying signal content. For any useful representation of the melody, it is necessary to find a means to automatically detect frames where the melodic voice is indeed present.

We use a measure of voicing based on the signal energy associated with the predominant pitch estimate, called normalized harmonic energy (NHE) [3]. It is defined as the sum of the energies of individual harmonics corresponding to the predominant pitch. All frames are labeled as having/missing the melodic voice by applying a static threshold to this feature.

The frame level labels are further smoothened over homogenous segments as determined by an automatic boundary detection method [6] based on detecting abrupt but stable changes in the harmonic energy feature. Grouping is done by a process of majority voting, i.e. the segment assumes the label of that class (voiced/unvoiced) into which the majority of the frames in that segment have been classified.

3. EVALUATION

In the MIREX 2008 evaluation of audio melody extraction, the tasks of melodic voice detection and melodic pitch estimation were evaluated independently. This was ensured by allowing algorithms to assign negative pitch values to frames that were labeled as unvoiced. The metrics used to quantify the performance of melodic voice detection are Voicing Recall, Voicing False Alarms and Voicing d-prime, while those used to measure the quality of melodic pitch estimation are Raw Pitch Accuracy and Raw Chroma Accuracy [7]. The Overall Accuracy represents the combined performance of the pitch estimation and voicing detection tasks.

3.1. Evaluation datasets

Two of the datasets used in the current evaluation were previously used for the MIREX 2005 and 2006 Audio melody extraction tasks (ADC 2004 and MIREX 2005 datasets). These consist of short excerpts of audio from a variety of genres such as Rock, Pop, R&B, Jazz, Classical and Opera. The melodic signal in a majority of cases was the human singing voice.

In addition to the above two datasets, a third new data set (MIREX 2008 dataset) was also included for this year's evaluations. Part of this dataset consists of north Indian classical vocal music excerpts that we contributed with permission of the data owners (National Centre for the Performing Arts, Mumbai).

For all data, ground truth pitch estimates are available every 10 ms, with unvoiced pitch estimates being set to 0.

3.2. Evaluation results

Evaluation results for the MIREX 2008 and 2005 datasets are shown in Tables 1 and 2 respectively. Bold numbers indicate the best performance in each column. For the former, the overall accuracy of our system is ranked fifth. However the melodic pitch estimation is found to give the best performance as indicated by the highest values of raw pitch and raw chroma accuracy values. The low values of overall accuracy are not disturbing because the particular threshold for the NHE feature chosen in the system results in significantly lesser number of false alarms than all other systems, but also lower voicing recall. The recall can easily be increased, at the expense of increasing the false alarms, by lowering the threshold for the NHE feature. This will result in similar voicing behavior and overall accuracy values[1] as those of the other systems.

For the MIREX 2005 data our system was ranked third in terms of overall accuracy because the performance of the melodic voice detector using the same threshold as before is significantly better than for the 2008 data (as indicated by the higher values of Voicing d-prime). Here the raw pitch accuracy of our system was ranked fourth but the raw chroma accuracy was still the highest. This indicates that our system is susceptible to making octave errors for this data, which can possibly be attributed to the biasing of the TWM error function towards pitches that lie in the lower and middle regions of the F0 search range. As expected our algorithm is found to perform better on sung rather than instrumental melodic signals. This is evident from the lower values of pitch and chroma accuracy for the non-vocal data (56.9 and 69.9 respectively) as compared to the vocal data (76.1 and 79.2 respectively), from individual evaluations over the sub-categories. This trend is also exhibited by the other submissions.

In terms of runtime, our algorithm operating time was found to be between 0.3 and 0.4 times real-time. This is not far behind the fastest algorithm, by Cao, Li, Liu and Yan, operating at 0.1 times real-time while showing significantly higher melodic pitch estimation accuracy.

[1] Although the overall accuracy depends upon both, the correctly identified voiced frames with correctly estimated pitch and correctly identified unvoiced frames, a reduction in the NHE threshold will lead to a more significant increase in the former than decrease in the latter due to the larger vocal bias in the data.

Rank	Participant	Voicing recall	Voicing False alarm	Voicing d-prime	Raw pitch accuracy	Raw chroma accuracy	Overall Accuracy	Runtime (sec)
1	Durrieu, Richard and David 1	**98.7 %**	68.9 %	**1.74**	85.8 %	88.3 %	**76.0 %**	4 days
2	Rynnanen and Klapuri	90.4%	48.5%	1.34	83.5%	83.8%	75.3%	100
3	Durrieu, Richard and David 2	96.6%	57.1%	1.65	81.8%	82.6%	75.0%	2318
4	Cancela	95.7%	65.4%	1.32	83.9%	84.0%	73.3%	47031
5	Rao and Rao	68.3%	**21.1%**	1.28	**88.2%**	**88.6%**	66.7%	184
6	Cao, Li, Liu and Yan 1	77.6%	43.4%	0.92	54.7%	55.3%	51.4%	49
7	Cao, Li, Liu and Yan 2	86.1%	67.8%	0.62	54.7%	55.3%	49.7%	50

Table 1. MIREX 2008 Audio Melody Extraction results (**2008 Dataset**)

Rank	Participant	Voicing recall	Voicing False alarm	Voicing d-prime	Raw pitch accuracy	Raw chroma accuracy	Overall Accuracy	Runtime (sec)
1	Cancela	83.9%	**19.9%**	**1.83**	71.0%	72.8%	**69.8%**	61470
2	Durrieu, Richard and David 2	**92.0%**	49.2%	1.42	**72.4**%	76.2%	66.0%	3294
3	Rao and Rao	85.1%	23.8%	1.75	69.7%	**76.5%**	64.9%	200
4	Rynnanen and Klapuri	85.6%	44.4%	1.21	71.2**%**	74.4%	63.5%	115
5	Cao, Li, Liu and Yan 2	77.8%	46.7%	0.85	68.9%	72.0%	61.4%	57
6	Durrieu, Richard and David 1	91.9%	60.4%	1.14	57.4%	65.3%	52.2%	3 days
7	Cao, Li, Liu and Yan 1	58.2%	42.2%	0.40	68.9%	72.0%	48.9%	56

Table 2. MIREX 2008 Audio Melody Extraction results (**2005 Dataset**)

4. CONCLUSIONS AND FUTURE WORK

This article described our submission to the MIREX 2008 audio melody extraction evaluation task. The performance of our method in terms of melodic pitch estimation was found to be on par, if not the best, with the best performing methods. Here too the pitch estimation performance was found to perform better for sung melodies rather than instrumental ones.

The melodic voice detection performance was found to vary across different signal sets. In general the system was found to give fewer false alarms than other submissions at the expense of a lower voicing recall. A more generic melodic voice detection method that is independent of the source melodic signal content is a definite area of improvement to be explored in the future.

ACKNOWLEDGEMENTS

The authors would like to express their heartfelt thanks to the group at IMIRSEL for their considerable effort to run the independent evaluations. We would also like to thank the NCPA, Mumbai for giving us permission to contribute their Indian classical music data to the evaluation.

REFERENCES

[1] R. Maher and J. Beauchamp, "Fundamental Frequency Estimation of Musical Signals using a Two-Way Mismatch Procedure," *J. Acoustical Soc. America*, vol. 95, no. 4, pp. 2254-2263, 1994.

[2] D. Griffin and J. Lim, "Multiband Excitation Vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223 – 1235, 1994.

[3] V. Rao and P. Rao, "Vocal melody detection in the presence of pitched accompaniment using harmonic matching methods," in *Proc. of the 11th International Conference on Digital Audio Effects (DAFx-08)*, Espoo, Finland, 2008.

[4] P. Cano, "Fundamental frequency estimation in the SMS analysis," in *Proc. of COST G6 Conf. on Digital Audio Effects 1998*, Barcelona, Spain, 1998.

[5] A. Bapat, V. Rao and P. Rao, "Melodic contour extraction of Indian classical vocal music," in *Proc. Intl. Workshop on Artificial Intelligence and Music (Music-AI '07)*, Hyderabad, India, January 2007.

[6] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *Proc. IEEE Intl. Conf. Multimedia and Expo (ICME)*, vol. 1, pp. 452-455, 2000.

[7] G. Poliner, et. al., "Melody Transcription from Music Audio: Approaches and Evaluation," *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no.4, pp. 1247-1256, May 2007.