Audio Melody Extraction: Evaluation and Approaches

Enmao Diao

Georgia Institute of Technology, Atlanta, GA, US
diaoenmao@gmail.com

***Abstract*—Although a well-trained musician is able to analyze an audio recording of music, as the quantity and complexity grow, it becomes inefficient and costly to solely rely on human skill. To describe a comprehensive status quo of melody extraction systems, an evaluation of the results provided by MIREX (Music Information Retrieval Evaluation eXchange) through 2009 to 2014 is conducted.**

***Index Terms*—Audio, evaluation, melody extraction, music**

I. INTRODUCTION

Melody attracts listeners not only with joy but also with mystery. Digital signal processing stands out to lift the veal. Researchers have utilized modern technology to analyze polyphonic melody. It may seem trivial for a well-trained musician to transcribe music to pitches at different time frames. However, as the complexity of overlapping and intertwining of harmonics increases, this task, audio melody extraction, becomes challenging.

In 2004, several contests were hosted for melody extraction, genre calassification/artist identification, tempo induction and rhythm calcification. This leads to the establishment of MIREX. Comparing with the result from 2005, the Overall Accrucy has increased about 4% and RPA (Raw Pitch Accuracy) has increased about 10% [1]. From 2009 through 2014, we do not achieve a significant breakthrough as shown in Fig. 1 and 2.

The most basic problem is that for polyphonic audio, four or more notes which come from different sources are overlapping at the same time and their fundamentals may be in integer ratios of each other. The harmonics coincide under the condition that spectral analysis generatges constructive and destructive interference. As a result, it becomes difficult to determine the predominant melody which is a prominent monophonic pitch sequence that can be agreed upon by most human listeners.

Despite the technical issues, resources are also limited to researchers. Although MIREX has been established for 10 years, only one large dataset is available while others have no more than 25 audio clips. Moreover, since MIREX has to hold the contest each year, it only provides the ground truth for ADC04 and MIREX05 database. It is also costly and time-consuming to set up a reliable dataset, as all datasets are manually noted. As a result, researchers cannot frequently adjust their algorithms based on the feedback except the annually held contest. We are still far from practical utilization. This paper aims to provide a summary of the results from 2009 through 2014 among all datasets and an analysis of approaches which are used by those who have achieved decent results.

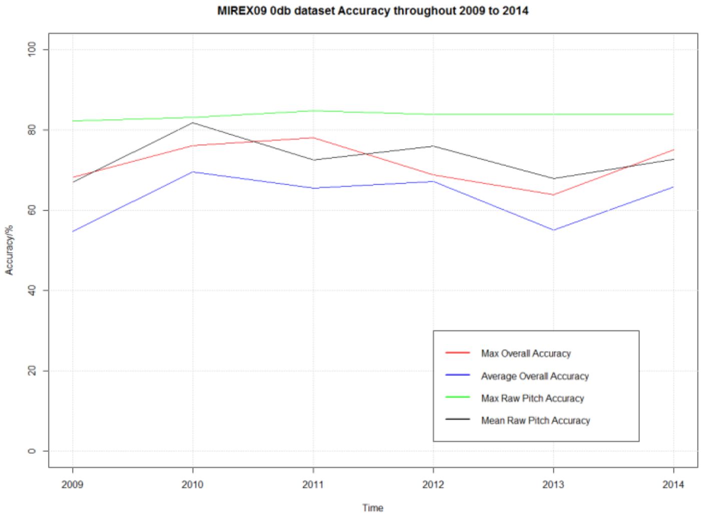

Fig. 1. Accuracy time analysis of dataset MIREX09 0db from 2009 through 2014.

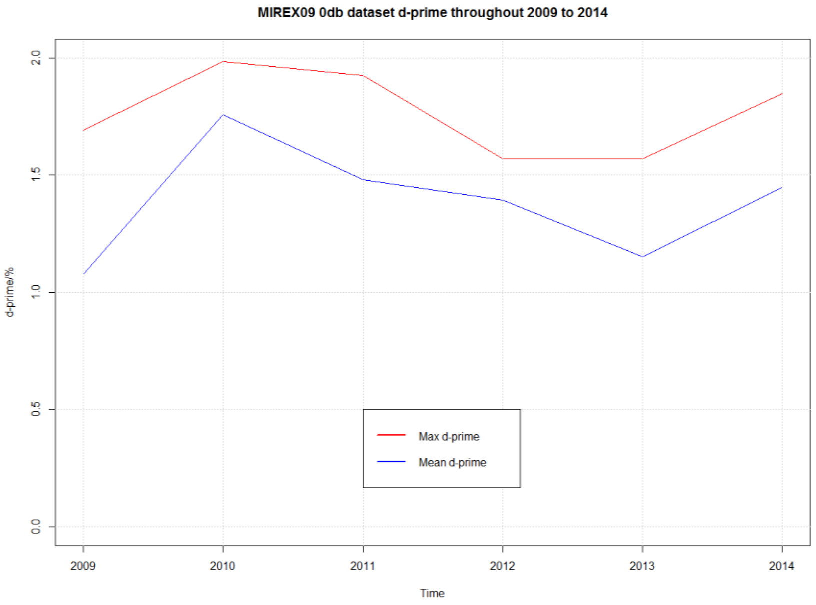

Fig. 2. Discrimination d' time analysis of dataset MIREX09 0db from 2009 through 2014.

II. Evaluation

Pitch identification is part of the most important aspects of audio digital signal processing. Specially, pitch identification of polyphonic audio recording is rather tougher. All approaches need to identify the dominant melody through different harmonics at each time frame. Moreover, it becomes clear that voice detection plays a significant role in improving overall accuracy on the basis of MIREX evaluation metrics. Since the dominant melody can be tuned on and off, most approaches are not capable of detecting voice acutely. Submission results from MIREX through 2009 to 2014 are chosen to be analyzed in this paper, on the grounds that the evaluation datasets are unchanged through these years.

A. Evaluation Metrics

All the algorithms submitted are required to detect the predominant frequency of the dominant melody at each regular time frame. The evaluation system can be divided into two parts. One is pitch identification which involves in RPA (Raw Pitch Accuracy) and RCA (Raw Chroma Accuracy). The other part involves in voicing dectection which includes Voice Detection Rate and Voicing False Alarm Rate which will be introduced later. Discriminability d' and Runtime are no longer offered after 2006. However, d' is still provided in this paper in order to determine the effectiveness of Voice Detection. The voice detection system of melody extraction is based on signal detection theory shown in Table I.

TABLE I. Evaluation Procedure

		Detected as		Sum
		Unvoiced	Voiced	
Ground Truth	Unvoiced	TN	FP	GU
	Voiced	FN	TP	GV
Sum		DU	DV	TO

- **Overall Accuracy**: A measurement that combines both the voicing detection and the pitch detection. It gives the proportion of time frames that were correctly labeled with both pitch and voicing, i.e. (TPC + TN)/TO[1].

[1] TN: Truth Negative, FN: False Negative, FP: False Positive, TP: Truth Positive, DU: Detected as Unvoiced, DV: Detected as Voiced, GU: Ground Truth Unvoiced, GV: Ground Truth Voiced, TO: Total, C: correctly labeled pitch, I: incorrectly labeled pitch, ch: chroma

- **RPA**: The probability of a correct pitch value (to within ± ¼ tone) in each frame is estimated within all the time frames. This includes the pitch guesses for frames that were judged unvoiced i.e. (TPC + FNC)/GV.

- **RCA**: The probability that the chroma (i.e. the note name) is correct over the voiced frames. This ignores errors where the pitch is wrong by an exact multiple of an octave (octave errors). Thus, this value is often greater than RPA. It is (TPCch + FNCch)/GV

- **Voicing Detection Rate**: The probability that a frame which is truly voiced is labeled as voiced i.e. TP/GV (also known as "hit rate").

- **Voicing False Alarm Rate**: The probability that a frame which is not actually voiced is none the less labeled as voiced i.e. FP/GU.

- **Voicing d-prime** *d'*: is a measure of the sensitivity of detecting to factor out the overall bias towards labeling any frame as voiced (which can move both hit rate and false alarm rate up and down in tandem). It converts the hit rate and false alarm into standard deviations away from the mean of an equivalent Gaussian distribution, and reports the difference between them. A larger value indicates a detection scheme with better discrimination between Voicing Detection Rate and Voicing False Alarm Rate.

B. Datasets Introduction

Datasets remain unchanged from 2009 through 2014. This offers a basic foundation for data analysis. The evaluation of the results carried out in this paper is isolated among different datasets. Since four available datasets are very distinct from each other, and they yield diverse results among different datasets.

- **ADC04 database**: Dataset from the 2004 Audio Description Contest. 20 excerpts of about 20s each.

- **MIREX05 database**: 25 phrase excerpts of 10-40 sec from the following genres: Rock, R&B, Pop, Jazz, Solo classical piano.

- **MIREX08 database**: 4 excerpts of 1 min. from "north Indian classical vocal performances", instruments: singing voice (male, female), tanpura (Indian instrument, perpetual background drone), harmonium (secondary melodic instrument) and tablas (pitched percussions). There are two different mixtures of each of the 4 excerpts with differing amounts of accompaniment for a total of 8 audio clips.

- **MIREX09 database**: 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 1122 audio clips. Instruments: singing voice (male, female), synthetic accompaniment. The Ground Truth pitch of each clip is human labeled, with a frame size of 40ms, a hop size of 20 ms. Note that the center of the first frame is located at 20ms starting from the very beginning of a clip. The human labeled pitch is then interpolated to have a hop size of 10ms. Thus the time sequence of the pitch vector are 20ms, 30ms, 40ms, 50ms, and so on.

C. Evaluation

Evaluation is conducted from 2009 through 2014 and datasets based. MIREX09 database is less biased because of its relatively larger quantity. Among three subsets of MIREX09 database, +5dB yields highest overall accuracy and -5db yields the lowest. Because submissions which do not perform Voice Dectection are allowed, those, which yield both Voicing Dection Rate and Vocing False Alarm Rate greater than 90% are excluded in the calculation involving Voice Detection and Overall Accuracy. Recall that although d' is no longer offered officially, it is calculated using the built-in R programming function [2]. Fig. 3 and 4 displays statistical summary of the overall accuracy and raw pitch accuracy among all the datasets. In order to address different approaches used by those with leading result in Section III. The histogram of overall accuracy is illustrated in Fig. 5 and raw pitch accuracy in Fig. 6.

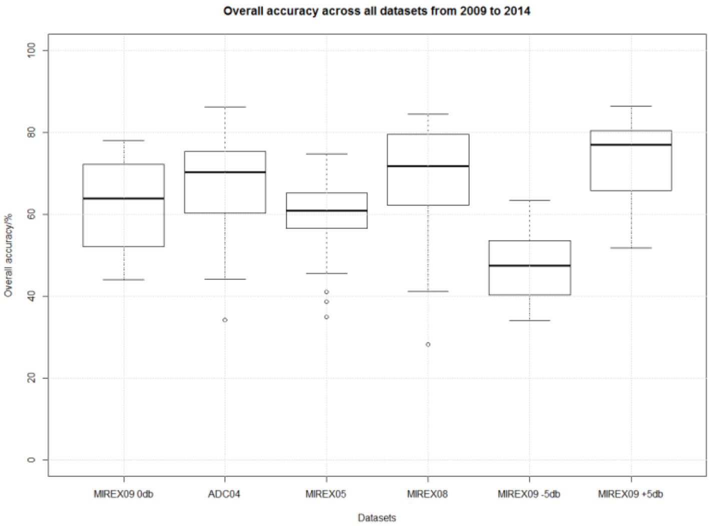

Fig. 3. Statistical summary of the overall accuracy throughout all datasets using boxplot.

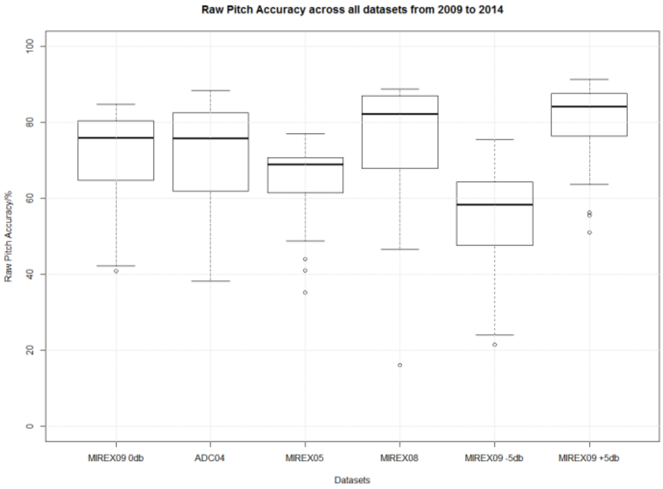

Fig. 4. Statistical summary of the raw pitch accuracy throughout all datasets using boxplot.

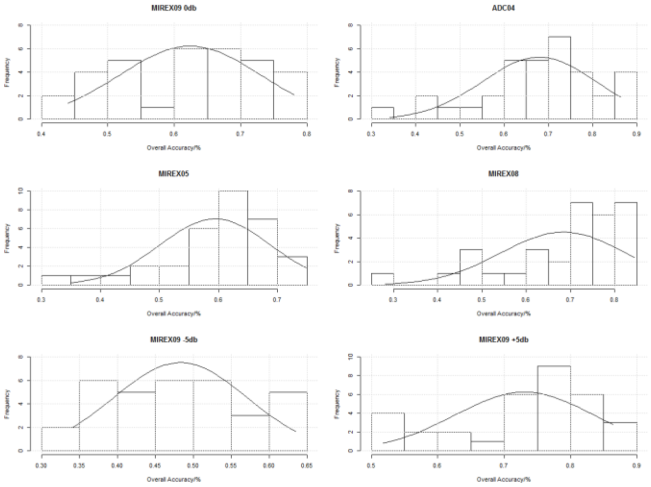

Fig. 5. Statistical summary of the overall accuracy throughout all datasets using histogram.

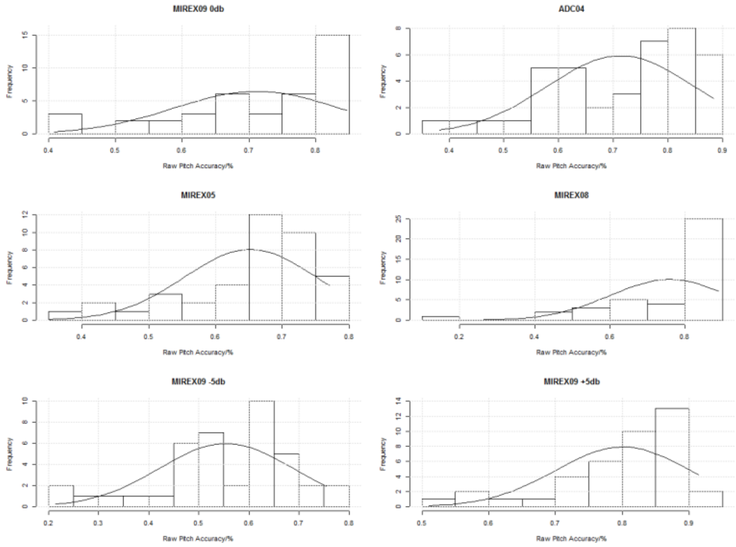

Fig. 6. Statistical summary of the raw pitch accuracy throughout all datasets using histogram.

As displayed in Fig. 3, dataset MIREX05 which has the most musical instrumental clips yields the worst overall accuracy except the MIREX09 -5db dataset. It has 9 clips of musical instrument including saxophone, guitar and synthesized piano and 15 clips of the human voice. Dataset ADC04 has one saxophone clip and 4 MIDI instruments other than the human voice. It yields relatively decent result, but it is also the shortest dataset which generates the least Voicing False Alarm Rate comparing to the others because the clips may not involve frequent dominant melody halts. MIREX08 yields the best performance. However, it contains the fewest clips including 4 different instruments. Thus, it is not practically effective [1]. If we solely concentrate on the human voice as the predominant melody, MIREX09 is the least biased considering the large number of audio clips. All other three datasets are not well-established representative samples for data analysis.

As for the three subsets of MIREX09 database, as shown in Fig. 4 and Fig. 6, -5db version yields a significant shift to lower accuracy. It results from the fact that most of the algorithms are severely dependent on the energy or salience of the harmonics.

Table II displays the best algorithm with respect to the benchmarks for all datasets. Recall that less representative database ADC04 and MIREX08 also yield higher accuracy. Voice detection and pitch detection are not isolated problems, because most algorithms which have max Raw Pitch Accuracy also give out Best Voicing Detection Rate.

TABLE II.

RESULTS OF BEST PERFORMANCE AND ITS AUTHOR OVER SIX DATASETS

Dataset	Max Overall Accuracy	Max Overall Accuracy Author	Max Raw Pitch Accuracy	Max Raw Pitch Accuracy Author	Best Voicing Detection Rate(d-prime)	Best Voicing Detection Rate Author
MIREX09	78.07%	Salamon	84.71%	Salamon	1.986	Hsu
ADC04	86.30%	Dressler	88.36%	Dressler	2.385	Dressler
MIREX05	74.84%	Dressler	77.03%	Dressler	1.584	Dressler
MIREX08	84.44%	Salamon	88.85%	Salamon	2.251	Salamon
MIREX09 -5db	63.49%	Tachibana	75.49%	Tachibana	1.483	Hsu
MIREX09 +5db	86.38%	Doreso[2]	91.39%	Salamon,	2.605	Doreso[2]

[2] A company

III. APPROACHES|

A total of 44 submissions have been made since 2009. One author may submit twice or more in the same year and some of them submit very similar algorithms in different years. They are treated as one submission and selected the best performing algorithm to be analyzed. Among the 40 submissions which are considered not duplicated, some of them are determined to ignore voice detection if their Voicing Dection Rate and Vocing False Alarm Rate are both greater than 90%. It appears that 33 submission take voice detection into account. On the basis that MIREX09 database is the least biased database, MIREX09 0db dataset is chosen to be the standard for approaches evaluation. The statistical result is given in Table III. Submissions which perform above the threshold are chosen to be evaluated.

TABLE III. STATISTICAL EVALUATION FOR MIREX09 0DB DATASET

	Average	Threshold
RPA	71.69%	75%
d’	1.401	1.5
Overall Accuracy	62.70%	-

Most researchers agree on the multi-pitch extraction stage. Short-time Fourier transform (STFT) is proved to be the most widely utilized approach. Among those best performing algorithms, some apply additional processes like Multi-stage Harmonic/Percussive Sound Separation (MHPSS) [3] and multi-resolution Fast Fourier Transform (MR-FFT) [4][5][6]. Harmonic/Percussive Sound Separation (HPSS) does not utilize harmonic structure of sound and the previous knowledge of percussions. Instead, it treats time domain as harmonic structure and frequency domain as percussions. MHPSS utilizes one short window (15-50ms) and one long window (100-500ms). Hsu uses HPSS considering that the long window length may damage harmonic structure with strong temporal variability [7]. However, both MHPSS and HPSS achieve high up to 84% RPA. Most well performing algorithms either choose to use a fixed to about 50ms Hanning window or use variable Hanning window ranging from 5.8ms to 46ms, but there is no clue to show that variable window length is better. Moreover, according to Salamon 2011 submissions, MR-FFT is proved not to play a significant role in their system since both STFT and MR-FFT yield the same overall accuracy [4]. MHPSS or HPSS can help with preprocessing spectrum. STFT more like serves a fundamental rule for pitch identification which severely determines the RPA.

Pitch identification is the most controversial part. Some researchers attempt to utilize probability model like Gaussian function [3][8][9] and Hidden Markov Model (HMM) [10][11][12]. Some researchers utilize Normalized Sub-harmonic Summation (NSHS) which is originally developed by Hermes [13]. Some researchers also take advantage of timbre model or the properties of melody in order to set up experimental rules for pitch identification [3][14][15]. Probability models prove to generate an average accuracy. If algorithms only apply STFT in multi-pitch extraction stage, Gaussian model turns out to be a little bit better than HMM. However, a combination of NSHS and HMM achieve up to 83% RPA. Solely relying on STFT and NSHS or Sub-harmonic Summation (SHS) only brings about 50% RPA. Algorithms which rely on the rules on the basis of timbre model or the properties of melody display an extreme variance of RPA. Salamon utilizes salience function which is used to estimate the mean pitch during several iterations. Different salience functions Salamon submitted in 2010 and 2011 achieve 80% and 85% RPA respectively, although both submissions have similar overall structure. Other algorithms which set up rules can only achieve around 60% RPA. Pitch identification requires much more work than multi-pitch extraction. A possible stable solution may be a combination of NSHS and HMM, while timbre model or the properties of melody can help with limiting uncertainty factors and as a result boost up the accuracy.

Voice detection, a rather tougher task, has not been taken into account specifically by most researchers. Most researchers pay very limited amount of effort in voice detection, since this topic not only welcome algorithms which ignore voice detection but also avoid emphasizing overall accuracy as the ultimate benchmark. Most researchers only demonstrate their results of RPA and RCA in their papers. However, most of the best performing RPA algorithms also yield decent result of d’, vice versa. The most widely applied approach for voice detection is Mel-frequency cepstral coefficients (MFCC). Some researchers take advantage of probability model like Gaussian mixture model (GMM), HMM [16][17] and Mahalanobis distance [3]. Salamon also relies on salience function and contour estimation to determine the voice regarding the claim that vibrato in a contour proves it is in a melody contour. The most well performing algorithm for voice detection by Hsu utilize several approaches including Multilayer Perception (MLP), MFCC, STFT, and Cepstral mean subtraction (CMS) [7]. However the algorithm can still be improved a lot because its Voice Recall rate lies on about 82%. Voice detection is still an elusive topic comparing to previous two stages.

As for post processing, most researchers realize that smoothing is necessary. Salamon removes the outliers and estimate the blank space according to contour estimation [18]. Hsu utilize trend estimation considering that the locality of contour smoothness may not be sufficient during a short period of time [19]. Besides smoothing process, Salamon and Dressler utilize multiple streaming to identify the prominent melody line by calculating salience function and rating. The statistical evaluation for approaches utilized above threshold is shown in Table IV. RPA threshold is related to multi-pitch extraction, pitch identification and window length while d’ threshold indicates voice detection.

Some researchers also attempt to discard spectral analysis [15][20]. Chien's 2014 submission improves RPA from 60% to 78% after adding Q-transform to original 2011 submission. Solely relying on self-developed models on the basis of timbre information and the properties of melody do not guarantee a decent result regarding that musical standard is complex and subjective. However, Musical Instrument Digital Interface (MIDI) data mining applied by Myer yield results above the average.

TABLE IV.

STATISTICAL EVALUATION FOR APPROACHES ABOVE THRESHOLD

Multi-pitch Extraction		Pitch Identification		Window Length		Voice Detection	
STFT	10	Salience function	2	Variable	5	MFCC	5
(M)HPSS	2	(N)SHS	1	Fixed	3	Salience function	2
MR-FFT	4	HMM	2			HMM	1
Q-transform	1	Gaussian function	2			Mahalanobis distance	1
		Timbre model	1			NHE[3]	1
						Power estimation	1
Total	17		9		8		11

IV. CONCLUSION

The contest conducted throughout these years provided a platform for individual researchers to make a quantitative comparison and evaluation of their approaches. Although we do not achieve a significant breakthrough from 2009, considering the large quantity of MIREX 09 database, for human voicing audio, we can reach up to 85% RPA and 78% Overall Accuracy.

At practical level, Query-by-humming (QbH) has already been commercially utilized and received positive feedback from the public. However, audio melody extraction is far from being practice. We may suggest that a large and standardized database can contribute to diagnostic analysis for researchers. Moreover, we can also simulate practical condition by emphasizing voice detection. We can also take advantage of other topics like audio onset detection and audio key detection. We can also take running time into account which is eliminated after 2006. Regardless of the fact that audio melody extraction is still far from practical use, the path is clear.

REFERENCE

[1] G. E. Poliner, D. P. W. Ellis, A. F. Ehmann, E. Gomez, S. Streich, and B. Ong, "Melody Transcription From Music Audio: Approaches and Evaluation," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[2] C. Pallier, "Computing discriminability and bias with the R software," *URL http//www. pallier.org/ressources/aprime/aprime.* pp. 1–6, 2002.

[3] H. Tachibana, T. Ono, N. Ono, and S. Sagayama, "Melody Extraction In Music Audio Signals By Melodic Component Enhancement And Pitch Tracking," 6th Music Information Retrieval Evaluation eXchange (MIREX), 2010.

[4] J. Salamon and E. Gomez, "Melody extraction from polyphonic music: MIREX 2011," 7th Music Information Retrieval Evaluation eXchange (MIREX), 2011.

[5] K. Dressler, "Audio Melody Extraction For Mirex 2014," 10th Music Information Retrieval Evaluation eXchange (MIREX), 2014.

[6] 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), Montreal, Quebec, Canada, Sept. 18–20, 2006, pp. 247–252.

[3] Normalized Harmonic Energy

[7] C. Hsu and J. R. Jang, “Singing Pitch Extraction At Mirex 2010,” 6th Music Information Retrieval Evaluation eXchange (MIREX), 2010.

[8] S. Jo and C. D. Yoo, “Melody Extraction From Polyphonic Audio Signal,” 6th Music Information Retrieval Evaluation eXchange (MIREX), 2010

[9] S. Jo, C. D. Yoo, and S. Park, “Melody Extraction From Polyphonic Audio Signal,” 7th Music Information Retrieval Evaluation eXchange (MIREX), 2011.

[10] S. Jo and C. D. Yoo, “Melody Extraction From Polyphonic Audio Signal,” 5th Music Information Retrieval Evaluation eXchange (MIREX), 2009.

[11] T. Yeh, M. Wu, and J. Jang, “A Hybrid Approach To Singing Pitch Extraction Based On Trend Estimation And Hidden Markov Models,” *Acoust. Speech Signal Process.*, pp. 457–460, 2012.

[12] L. Song and M. Li, “Bayesian Framework-Based Vocal Melody Extraction For Mirex 2014,” 10th Music Information Retrieval Evaluation eXchange (MIREX), 2014.

[13] Dik Hermes. “Measurement of pitch by subharmonic summation,” Journal of Acoustic of Society of America, vol.83, pp.257-264,1988.

[14] C. Cao, M. Li, J. Liu, and Y. Yan, “Singing Melody Extraction In Polyphonic Music By Harmonic Tracking,” 5th Music Information Retrieval Evaluation eXchange (MIREX), 2009.

[15] J. Yoon, C. Song, S. Lee, and H. Park, “Extracting Predominant Melody of Polyphonic Music based on Harmonic Structure,” 7th Music Information Retrieval Evaluation eXchange (MIREX), 2011.

[16] C. Hsu, J. R. Jang, and L. Chen, “Singing pitch extraction at mirex 2009,” 5th Music Information Retrieval Evaluation eXchange (MIREX), 2009.

[17] Y. Ikemiya, “Mirex2014 : Audio Melody Extraction, ” 10th Music Information Retrieval Evaluation eXchange (MIREX), 2014.

[18] J. Salamon and E. Gómez, “Melody Extraction From Polyphonic Music Signals Using Pitch Contour Characteristics,” *Audio, Speech, Lang. Process.*, vol. 20, no. 6, pp. 1759–1770, 2012.

[19] C. Hsu, D. Wang, and J. Jang, “A Trend Estimation Algorithm For Singing Pitch Detection In Musical Recordings,” *Acoust. Speech Signal Process.*, pp. 247–252, 2011.

[20] Y. Chien, H. Wang, and S. Jeng, “Vocal Melody Extraction Based On An Acoustic-Phonetic Model Of Pitch Likelihood,” 7th Music Information Retrieval Evaluation eXchange (MIREX), 2011.