# MELODY EXTRACTION FROM POLYPHONIC AUDIO OF WESTERN OPERA: A METHOD BASED ON DETECTION OF THE SINGER'S FORMANT

**Zheng Tang**
University of Washington, Department of Electrical Engineering
zhtang@uw.edu

**Dawn A. A. Black**
Queen Mary University of London, Electronic Engineering and Computer Science
dawn.black@qmul.ac.uk

## ABSTRACT

Current melody extraction approaches perform poorly on the genre of opera [1, 2]. The singer's formant is defined as a prominent spectral-envelope peak around 3 kHz found in the singing of professional Western opera singers [3]. In this paper we introduce a novel melody extraction algorithm based on this feature for opera signals. At the front end, it automatically detects the singer's formant according to the Long-Term Average Spectrum (LTAS). This detection function is also applied to the short-term spectrum in each frame to determine the melody. The Fan Chirp Transform (FChT) [4] is used to compute pitch salience as its high time-frequency resolution overcomes the difficulties introduced by vibrato. Subharmonic attenuation is adopted to handle octave errors which are common in opera vocals. We improve the FChT algorithm so that it is capable of correcting outliers in pitch detection. The performance of our method is compared to 5 state-of-the-art melody extraction algorithms on a newly created dataset and parts of the ADC2004 dataset. Our algorithm achieves an accuracy of 87.5% in singer's formant detection. In the evaluation of melody extraction, it has the best performance in voicing detection (91.6%), voicing false alarm (5.3%) and overall accuracy (82.3%).

## 1. INTRODUCTION

Singing voice can be considered to carry the main melody in Western opera. Melody extraction from a polyphonic signal including singing voice requires both of the following: estimation of the correct pitch of singing voice in each time frame and voicing detection to determine when the singing voice is present or not.

The singer's (or singing) formant was first introduced by Johan Sundberg [3] and described as a clustering of the third, fourth, and fifth formants to form a prominent spectral-envelope peak around 3 kHz. It is purportedly generated by widening the pharynx and lowering the larynx. The existence of a singer's formant has been confirmed in the singing voices of classically trained male Western opera singers and some female singers, but it has not yet been found in soprano singers [5] or Chinese opera singers [6]. It has been proposed that singers develop the singer's formant in order to be heard above the orchestra. In Western opera, orchestral instruments typically occupy the same frequency range as the singers. Therefore singers train their vocal equipment in order to raise the amplitude of frequencies at this range.

The LTAS is the average of all short-term spectra in a signal, has been shown to be an excellent tool to observe the singer's formant [7] as can be seen in Figure 1. Characteristics of the singer's formant in the spectral domain include a peak greater than 20 dB below the overall sound pressure level, a peak-location at 2.5-3.2 kHz, and a bandwidth of around 500-800 Hz [5, 7]. However, to date, there has been no method developed to automatically detect the presence of a singer's formant or to quantify its characteristics.

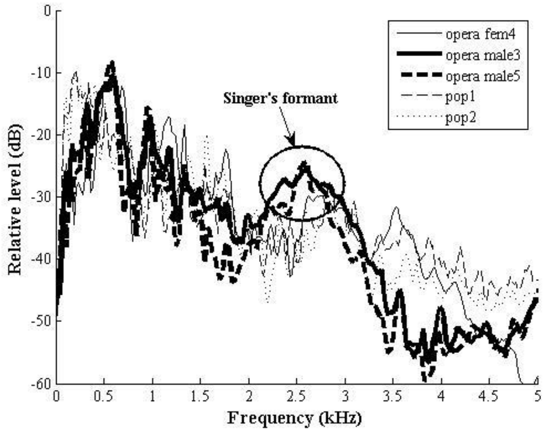

**Figure 1**. Normalized LTAS for 5 audio excerpts from the ADC2004 test collection [1].

### 1.1 Related Work

In 2004, the Music Technology Group of the Universitat Pompeu Fabra organized a melody extraction contest presented at the International Society for Music Information Retrieval Conference. The Music Information Retrieval Evaluation eXchange (MIREX) was set up in 2005 and audio melody extraction has been a highly competitive field ever since. Currently, over 60 algorithms have been submitted and evaluated. So far, none of the approaches consider the presence of the singer's formant.

The majority of algorithms presented at MIREX are salience based [2]. These assume that the fundamental frequency of the melody is equivalent to the most salient pitch value in each frame. The Short-Time Fourier Transform (STFT) is often chosen to compute pitch salience [7, 8]. In 2008, Pablo Cancela proposed the Fan Chirp Transform (FChT) method, combined with Constant Q Transform (CQT) in music processing. The FChT is a time-warped version of the Fourier Transform that provides better time-frequency resolution [4, 9]. Although the STFT provides adequate resolution in the majority of cases, it fails to generate a satisfying outcome when dealing with Western opera signals. This is because opera typically exhibits complex spectral characteristics due to vocal ornamentations such as vibrato [1]. Vibrato is a regular fluctuation of singing pitch produced by singers. This increases the difficulty in tracking the melody. With better resolution, the fast change of pitch salience can be better observed and tracked by using FChT.

It has been proposed that the singer's formant may cause octave errors [2]. The presence of a spectral peak (the singer's formant) at a higher frequency may cause the fundamental frequency to be confused with the frequency at the centre of the singer's formant. To address this, Cancela developed a method called 'subharmonic attenuation' that can minimize the negative effects of ghost pitch values at the multiple and submultiple peaks of a certain fundamental frequency [2, 9].

Voicing detection typically receives much less attention than pitch detection, to the extent that some previous melody extraction algorithms did not contain this procedure [10]. The most common approach is to set an energy threshold, which might be fixed or dynamic [9]. However, this technique is too simplistic since the loudness of musical accompaniment in Western opera may fluctuate considerably. It is therefore impossible to define an appropriate threshold. An alternative technique is to use a trained classifier based on a Hidden Markov Model (HMM) [11] but it is time-consuming to create a large dataset for training and there are always exceptions beyond the scope of the training set. In 2009, Regnier and Peeters proposed a voicing detection algorithm based on extraction of vocal vibrato [12], but has not been applied to melody extraction. In general, the high rate of false positives when detecting voiced frames limits the overall accuracy of melody extraction algorithms and a reduction of this is beneficial [2, 13].

This paper is organized as follows. In Section 2, we describe the design and implementation of our proposed algorithm for melody extraction. Starting with a general workflow of the system, the function and novelty of each component is explained in detail. Section 3 explains the evaluation process and presents a comparison of existing algorithms. The creation of the new dataset is also presented in this section. Finally, we draw conclusions from the results and give suggestions for future work.

## 2. DESCRIPTION OF THE ALGORITHM

### 2.1 General Workflow

Figure 2 shows an overview of our system. In order to extract the pitch of singing voice from polyphonic audio, we must first determine whether the audio contains singing voice. The presence of a singer's formant would indicate the presence of a classically trained singer. The LTAS is used to determine whether a singer's formant exists in the audio, and hence determines whether our method can be applied.

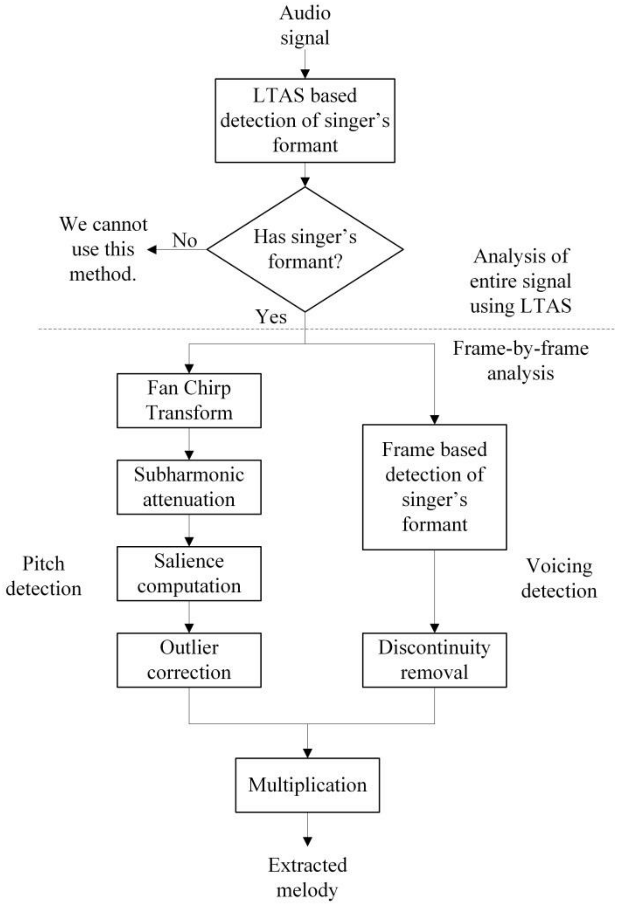

**Figure 2**. System overview.

Once the presence of a singer is confirmed the spectrum is analysed on a frame-by-frame basis. Two decisions are made for each frame: firstly, does the frame contain singing and hence a salient pitch? Secondly, what is the salient pitch of that frame?

We examine the spectral content of each frame to establish the presence of a singer's formant in that frame. If present, that frame is designated 'voiced' and assumed to contain melody carried by the singer's pitch.

Each frame is also transformed to the frequency domain using the FChT and further processed by subharmonic attenuation to obtain the pitch.

### 2.2 Singer's Formant Detection and Voicing Detection

Based on the characteristics of the singer's formant (see Section 1) we introduce a novel algorithm to automatically detect the presence of a singer's formant (and hence the presence of a classically trained singer). Using Monson's method to compute the LTAS of the input audio signal [14] the presence of a singer's formant would be confirmed if the LTAS exhibited the following properties:

1. There exists a spectral peak which has an amplitude greater than 20 dB less than the overall sound pressure level.
2. The peak is located between 2.5 and 3.2 kHz.
3. The peak has a bandwidth of around 500-800 Hz.

However, these properties were observed through analysis of singing voice in the absence of musical accompaniment [7]. When analysing singing with accompaniment, these criteria had to be modified in the following ways: the amplitude threshold of the spectral peak was found to be lower than the theoretical value and thus the first criteria becomes:

1. The spectral peak has an amplitude greater than 30 dB less than the overall sound pressure level.

The LTAS exhibited irregular fluctuations that made accurate identification of the singer's formant peak problematic. We therefore smoothed the LTAS (20 point average) and used polynomial fitting of degree 30. This smoothing and polynomial fitting will shift the location of the spectral peak and hence the range of the peak must be expanded. The second criteria is therefore modified to:

2. The peak is located between 2.2 and 3.4 kHz.

Similarly, we observe that the polynomial bandwidth may be slightly different from the LTAS curve. Therefore the bandwidth of the singer's formant is set to be larger than the original value:

3. The peak has a bandwidth larger than 600 Hz.

We must then add another criteria to ensure the significance of the peak. In order to measure the significance, we employed the first-order and second-order derivatives of the LTAS to measure the LTAS curvature and, from empirical evidence, designate significance to be a peak with a curvature greater than 0.01:

4. The curvature exceeds 0.01 at the location of the spectral peak.

In order to illustrate the criteria, we present the following figures. Figure 3 shows the fitting polynomials of smoothed LTAS for 5 samples from the MIREX ADC2004 test collection [1]. The singer's formant can be clearly observed for the male opera samples. Presented in Figure 4 is the second-order derivative of LTAS. This is negative when the curve is convex and hence can be used to determine the formant bandwidth. Our constraint that the bandwidth be at least 600 Hz is illustrated. In Figure 5, we show that the constraint on curvature can ensure the degree of convexity of the curve. It is clear from all plots that the opera signals sung by male singers contain the singer's formant but others do not.

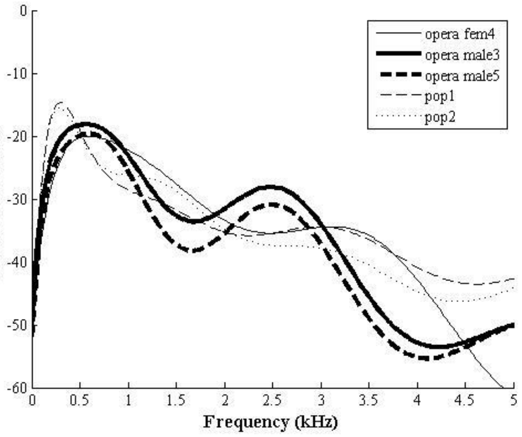

**Figure 3**. The fitting polynomials of smoothed LTAS for 5 audio excerpts from ADC2004 [1].

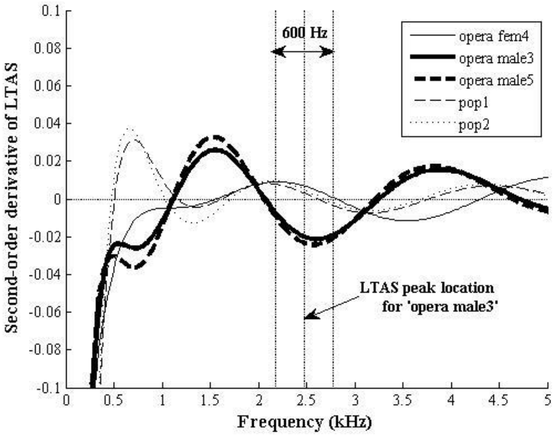

**Figure 4**. The second-order derivatives of LTAS for 5 audio excerpts from ADC2004 [1].

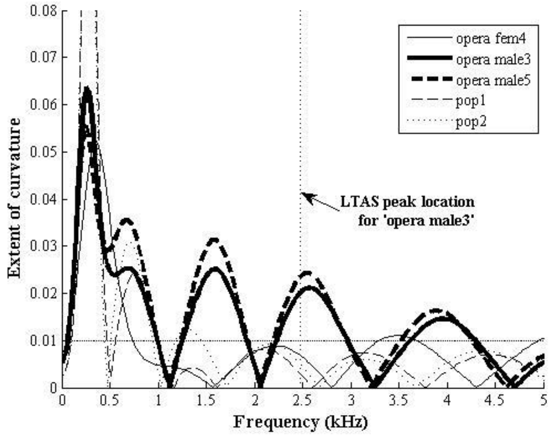

**Figure 5**. The curvatures of LTAS for 5 audio excerpts from ADC2004 [1].

If the LTAS satisfies all four criteria the audio is presumed to contain a trained singer. Use of the same criteria to analyse the spectrum of a single audio frame can indicate whether the frame is voiced (contains singing) or not. For a single frame, only the second and third criteria are applied, as the other two criteria are more influenced by observed amplitude variations in individual short-term spectra. The output of this stage is a two-value sequence whose length is the number of frames, with '1' indicating a voiced frame and '-1' unvoiced. Subsequently, when considering points of discontinuity causing false detections, single values within a sequence are removed.

### 2.3 Pitch Detection

If a frame is classified as voiced it can be expected to contain a clearly defined pitch. Vibrato in singing can cause pitch ambiguity. We therefore adopt Cancela's method to perform FChT since it exhibits optimal time-frequency resolution. This chirp-based transform is based on an FFT performed in a warped time domain. It is combined with CQT in order to guarantee high resolution even when the fan chirp rate is not ideal. More details can be found in [4] and [9].

In Western opera the singer's formant will cause peaks at frequencies higher than the fundamental [2]. The algorithm from Cancela provides subharmonic attenuation - an effective solution to this problem. It will suppress multiple and submultiple pitch peaks of the fundamental frequency. Then, we can perform salience computation to detect the pitch in each frame.

In the outlier correction stage, to improve Cancela's method, we compute two additional peaks per frame as candidate substitutes for the wrong pitch. Firstly, the most salient pitch peaks are compared with those from adjacent frames. If a difference of more than 2 semitones occurs on both sides, the estimated pitch in this frame is considered as a wrong detection. In this case we substitute the pitch for this frame with the pitch among the three candidates which is closest to the average of the two adjacent estimations. Due to subharmonic attenuation, the influence of the subharmonics of the top peak is reduced when calculating the other pitch candidates.

Our method is novel to Cancela's in the following ways: (1) The algorithm designed by Cancela extracts multiple salient peaks simultaneously and these are viewed as separate melodies. We introduce the correction block so that the less salient peaks are taken as substitutes of wrong pitch detection in a single estimation of melody. (2) We improve the voicing detection by considering the singer's formant. (3) Cancela's method is not specifically designed for opera items and its potential for dealing with vibrato and other spectrum characteristics has not been explored.

Finally, the estimated pitch sequence is multiplied by the two-value voicing detection sequence. The output of our algorithm follows the standard format of MIREX and records the time-stamp and estimated frequency of each frame.

## 3. EVALUATION

### 3.1 The Dataset

The dataset we used for evaluation is a combination of the ADC2004 test collection and our own dataset[1]. Details of the dataset can be found in Table 1.

Among the existing test collections in MIREX, only ADC2004 contains 2 excerpts in the genre of Western opera. In order to evaluate the performance of melody extraction algorithms upon sufficient amount of opera samples for meaningful comparison, we created a new dataset. Nine students from the Central Academy of Drama in Beijing were recorded. All had received more than 5 years of classical voice training except for an amateur Western opera male singer. Their singing voices were recorded in a practice room, about 10×5×5 m with moderate reverberation. The equipment included a Sony PCM-D50 recorder and an AKG C5 microphone. The accompaniments played by orchestra were recorded separately. All the signals were digitized at a sample rate of 44.1 kHz with bit depth 16. We normalized the maximum amplitude of the singing voices to be -1.25 dB. The signal-to-accompaniment ratio is set to 0 dB. The ground truth for melody extraction was generated by a monophonic pitch tracker in SMSTools with manual adjustment [2] using the vocal track only. The frame size was 2048 samples with a step size of 256 samples.

We conducted two evaluations based on this combined dataset. The test set for melody extraction consists of 18 excerpts of 15s-25s duration sung by classically trained Western opera tenors. For the evaluation of singer's formant detection, we will compare them with 14 excerpts sung by trained Western opera sopranos, trained Peking opera singers, pop singers, and a single unprofessional Western opera male singer.

| Test set | Singing type | No. of songs | Expectation/ detection of singer's formant |
|---|---|---|---|
| ADC2004 | Tenor, Western | 2 | Yes/ Yes |
| | Soprano, Western | 2 | No/ No |
| | Popular music | 4 | No/ No |
| The dataset recorded at the Central Academy of Drama | Tenor, Western | 16 | Yes/ Yes |
| | Soprano, Western | 2 | No/ Yes |
| | Amateur, Western | 2 | No/ Yes |
| | Laosheng, Peking | 2 | No/ No |
| | Qingyi, Peking | 2 | No/ No |

**Table 1.** Test dataset for the evaluations of melody extraction and singer's formant detection.

[1] This database is available for download under a creative commons license at http://c4dm.eecs.qmul.ac.uk/rdr/ all usage should cite this paper.

### 3.2 Melody Extraction Comparison

Of the many melody extraction algorithms submitted to MIREX, few are freely available. We present five algorithms for comparison. We were limited in our choice by availability, but the methods are representative of the majority of algorithms submitted to MIREX in that they cover common approaches and best performance. Each method is briefly introduced next.

Cancela's algorithm was submitted in 2008 [9]. He used FChT combined with CQT to estimate the pitch in each time frame. Voicing detection is conducted through the calculation of an adaptive threshold, but this procedure is not included in the open-source code provided online. For the purposes of comparison, we added a common voicing detection function utilizing an adaptive energy threshold as described in [9].

Salamon's algorithm was introduced in 2011 [8]. It has been developed into a melody extraction vamp plug-in: MELODIA. This algorithm achieved the best score in MIREX 2011. It applies contour tracking to the salience function calculated by STFT to remove all the contours except for the melody. The voicing detection step is carried out by removing the contours that are not salient.

The algorithm developed by Sutton in 2006 [11] innovatively combines two pitch detectors based on the features of singing voice including pitch instability and high-frequency dominance. A modified HMM processes the estimated melodies and determines the voicing.

The final two algorithms were both proposed by Vincent in 2005 [10]. One makes use of a Bayesian harmonic model to estimate the melody, and the other is achieved via loudness-weighted YIN method. Vincent assumed that the melody was continuous throughout the audio, and voicing detection was not included in his algorithm.

### 3.3 Results

The evaluation results of singer's formant detection can be found in Table 1. Among the 32 audio files in the dataset, the assumption is that only the 18 excerpts sung by Western opera tenors possess the singer's formant, while the others do not. The results show that 28 of the files (87.5%) meet our expectation. The singer's formant is also detected in the excerpts of the Western opera amateur and sopranos in our dataset. The amateur singer is from the Acting Department at the Central Academy of Drama (Beijing) and declares that he has not received any formal training in opera. However he used to take courses in vocal music due to a requirement of the school. Thus, there is a possibility that the presence of singer's formant only requires a short period of training. Although sources state that there is no singer's formant present in soprano singing [5, 7], the mean pitch of the two excerpts in our dataset is at the low end of the range for sopranos (550.43 Hz). The presence of a singer's formant is pitch related. The higher the pitch, the less likely a singer's formant is present. A precise study of this relationship is a topic for future work.

Table 2 shows the melody extraction results of the 6 algorithms. Voicing detection measures the probability of correct detection of voiced frames, while voicing false alarm is the probability of incorrect detection of unvoiced frames. Raw pitch accuracy and raw chroma accuracy both measure the accuracy of pitch detection, with the latter ignoring octave errors. The overall accuracy is the proportion of frames labeled with correct pitch and voicing. Since Vincent's algorithms did not perform voicing detection, their voicing metrics and overall accuracy are inapplicable.

| First author/ completion year | Voicing detection | Voicing false alarm | Raw pitch accuracy | Raw chroma accuracy | Overall accuracy |
|---|---|---|---|---|---|
| Vincent (Bayes)/ 2005 | N/A | N/A | 64.8% | 68.6% | N/A |
| Vincent (YIN)/ 2005 | N/A | N/A | 69.5% | 72.2% | N/A |
| Sutton/ 2006 | 89.3% | 51.9% | 87.0% | 87.6% | 76.9% |
| Cancela/ 2008[1] | 72.6% | 39.3% | 83.9% | 84.8% | 62.4% |
| Salamon/ 2011 | 62.3% | 21.8% | 25.4% | 30.1% | 31.3% |
| Our method | 91.6% | 5.3% | 84.3% | 85.1% | 82.3% |

**Table 2.** Results of the audio melody extraction evaluation.

Our algorithm ranks highest in overall accuracy. We also achieve the highest voicing detection rate as 91.6% and the lowest voicing false alarm rate as 5.3%, which proves that voicing detection based on the singer's formant is extremely effective for male Western opera. The improvement in raw pitch accuracy by outlier correction when compared to Cancela's method is not large. This allows us to hypothesise that the melody in Western opera may be so prominent that the influence of any accompaniment can be disregarded.

Sutton's method also has excellent performance on our dataset. That success might be attributed to his similar focus on the characteristics of singing voice. He also makes use of the vibrato feature to estimate the pitch of melody. Due to the application of a high-frequency correlogram, Sutton's algorithm may indirectly benefit from the presence of a singer's formant. However, the method we propose for voicing detection is much more convenient than the use of an HMM. Moreover, Sutton's algorithm exhibits a much higher voicing false alarm rate.

The poor performance of Salamon's algorithm on our dataset can be explained by the fact that it fails to estimate the pitch in detected unvoiced frames accurately.

We also evaluated the 4 audio files that contradicted our expectation in singer's formant detection (two West-

[1] The voicing detection part of this algorithm is implemented by us and cannot represent the original design of the author.

ern soprano singers and one amateur male Western opera singer). The performance of our algorithm declines significantly with a voicing detection rate of 53.1% and an overall accuracy of 53.7%. This may be due to the fact that the singer's formant, although present, is not as pronounced or stable as the Western opera tenor's.

## 4. CONCLUSION AND FURTHER WORK

In this paper, we have presented a novel melody extraction algorithm based on the detection of singer's formant. This detection relies on 4 criteria modified from previously proposed characteristics of the singer's formant. The pitch detection step of our algorithm is achieved using FChT and subharmonic attenuation to overcome the known difficulties when detecting the melody in opera. We also improved the algorithm so it is capable of removing outliers in pitch detection.

From the evaluation results, it can be seen that our algorithm can detect the singer's formant accurately. Melody extraction evaluation on our dataset confirms that our algorithm provides a clear improvement in voicing detection. Furthermore, its overall accuracy is comparable to state-of-the-art methods when dealing with Western opera signals.

In the future, we plan to study the performance of this algorithm on signals in other genres and expand its scope of application. Additionally, the possible effects of performing environments and accompaniment music to the usage of singer's formant will also be explored.

## 5. ACKNOWLEDGMENTS

This paper is based upon a research collaboration with the Department of Peking Opera at the Central Academy of Drama. Thanks to Prof. Ma Li and his students for their recording samples and professional advice on traditional opera. Besides, we would like to express our thanks to Pablo Cancela, Justin Salamon, Emilia Gómez, Christopher Sutton and Emmanuel Vincent for contributing their algorithm codes.

## 6. REFERENCES

[1] E. Gómez, S. Streich, B. Ong, R. P. Paiva, S. Tappert, J. M. Batke, G. Poliner, D. Ellis, and J. P. Bello: "A quantitative comparison of different approaches for melody extraction from polyphonic audio recordings," Univ. Pompeu Fabra, Barcelona, Spain, 2006, Tech. Rep. MTG-TR-2006-01.

[2] J. Salamon, E. Gómez, D. Ellis, and G. Richard: "Melody extraction from polyphonic music signals: approaches, applications and challenges," *IEEE Signal Processing Magazine*, Vol. 31, No. 2, pp. 118-134, 2013.

[3] J. Sundberg: "Articulatory interpretation of the 'singing formant'," *The Journal of the Acoustical Society of America*, Vol. 55, No. 4, pp. 838-844, 1974.

[4] L. Weruaga, and M. Képesi: "The fan-chirp transform for non-stationary harmonic signals," *Signal Processing*, Vol. 87, No. 6, pp. 1504-1522, 2007.

[5] R. Weiss, Jr, W. S. Brown, and J. Moris: "Singer's formant in sopranos: fact or fiction?" *Journal of Voice*, Vol. 15, No. 4, pp. 457-468, 2001.

[6] J. Sundberg, L. Gu, Q. Huang, and P. Huang: "Acoustical study of classical Peking Opera singing," *Journal of Voice*, Vol. 26, No. 2, pp. 137-143, 2012.

[7] J. Sundberg: "Level and center frequency of the singer's formant," *Journal of voice*, Vol. 15, No. 2, pp. 176-186, 2001.

[8] J. Salamon, and E. Gómez: "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, No. 6, pp. 1759-1770, 2012.

[9] P. Cancela, E. López, and M. Rocamora: "Fan chirp transform for music representation," *Proceedings of the 13th Int Conference on Digital Audio Effects DAFx10 Graz Austria*, 2010.

[10] E. Vincent, and M. D. Plumbley: "Predominant-F0 estimation using Bayesian harmonic waveform models," *2005 Music Information Retrieval Evaluation eXchange (MIREX)*, 2005.

[11] C. Sutton: "Transcription of vocal melodies in popular music," Report for the degree of MSc in Digital Music Processing, Queen Mary University of London, 2006.

[12] L. Regnier, and G. Peeters: "Singing voice detection in music tracks using direct voice vibrato detection," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1685-1688, 2009.

[13] G. E. Poliner, D. P. Ellis, A. F. Ehmann, E. Gómez, S. Streich, and B. Ong: "Melody transcription from music audio: Approaches and evaluation," *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 4, pp. 1247-1256, 2007.

[14] B. B. Monson: "High-frequency energy in singing and speech," Doctoral dissertation, University of Arizona, 2011.