# GROUP DELAY BASED MELODY MONOPITCH EXTRACTION FROM MUSIC

*Rajeev Rajan and Hema A. Murthy*

Department of Computer Science and Engineering,
Indian Institute of Technology, Madras, India - 600 036
e-mail:rajeev@cse.iitm.ac.in,hema@cse.iitm.ac.in

## ABSTRACT

In this paper, we propose a modified group delay based method for melodic pitch extraction from heterophonic music. The power spectrum of the music signal is first flattened in order that the system characteristics are annihilated, while the characteristics of the source are emphasized. The modified group delay function of this signal produces peaks at multiples of the pitch period. The first 3 peaks are used to determine the actual pitch period. The performance of the proposed system was evaluated on two datasets ADC-2004, and LabROSA. The performance is comparable to that of other magnitude spectrum based approaches. The algorithms are also applied to heterophonic music, namely Carnatic Music. As ground truth is not available for Carnatic Music, the pitch contours were used to synthesize the music, which was evaluated for correctness by a professional musician.

***Index Terms—*** Group delay, Modified group delay, unwrapped, cepstral smoothing.

## 1. INTRODUCTION

### 1.1. Introduction

Melody extraction is an important component of music transcription. Melodic pitch according to the Music Information Retrieval(MIR) community is *the single (monophonic) pitch sequence that a listener might reproduce when asked to hum or whistle a polyphonic piece of music* [1, 2] . One of the problems in melody extraction is to suppress interference and detect the predominant pitch sequence. In [3] Goto focusses on the estimation of predominant musical voice rather than aiming at the transcription of all sound sources. The main steps in the melody extraction algorithm described in [3] are to select a candidate set of the fundamental frequency based on spectrum peaks, and then use the subharmonic summation method to identify the fundamental frequency from this candidate set. This method estimates the relative dominance of every possible F0 by using MAP (maximum a posteriori probability) estimation. Cao et al. [4] propose a melody extraction method based on the subharmonic summation spectrum and the harmonic structure tracking strategy. They analyze the prominent pitch of the mixture to find stable harmonic structure seeds which are used to estimate pitch. By using the characteristics of vibrato and tremolo Hsu and Jang [5] distingiushes the vocal partial from the music accompaniment partials. In [2], Salamon proposes pitch extraction algorithms that extract the pitch even when the accompaniment is strong. Source separation methods include [6] where an effort is made to separate the melody from polyphonic music. Salience-based algorithms derive an estimation of pitch salience over time and then apply tracking or transition rules to extract the melody line without separating it from the rest [7, 3]. In [1], a data driven approach is used. In this approach, the entire short-time magnitude spectrum is used as training data for a support vector machine classifier.

In this paper, an attempt is made to extract melodic pitch using Fourier transform phase based methods as opposed to the conventional Fourier transform magnitude based methods [1, 7]. When a signal is minimum phase, both Fourier transform phase and magnitude contain source information. In particular, timing information is related to phase rather than magnitude. The phase spectrum is seldom processed to extract source information [8]. The phase spectrum is difficult to process, since any Fourier based processing only yields the wrapped phase. The phase spectrum has to be first unwrapped before any meaningful analysis can be performed. Unwrapping the phase function can be performed using Tribolet's algorithm [9]. Alternatively, the group delay function in which the properties of the phase function are preserved can be processed. Many studies have successfully processed the group delay function to extract formants, features for speaker and speech recognition and spectrum estimation[10, 11]. The group delay function is well-behaved when the signal is minimum phase. Alternatively, the modified group delay function is proposed, which is well-behaved for even nonminimum phase signals.

Several attempts have been made to extract features from the modified group delay function from a segment of mixed phase speech [10]. These features have been gainfully used in applications like speaker verification and speech recognition. The primary motivation for this work arises from the applications of the group delay function in estimating sinusoids from noise [11]. In speech production, the source can be approxi-

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

mately modeled by a periodic function (that is approximately a sinusoid) that modulates the formants, which can be thought of as modulated carriers. If the carrier frequencies (or rather formants) are suppressed in the power spectrum, the periodic source will result in a periodic function in the spectrum[11]. The modified power spectrum can be thought as a sinusoidal signal. This modified power spectrum can then be subjected to sinusoidal analysis. It has already been established in the literature that the modified group delay function emphasizes peaks in spectra well (a property that is exploited in speaker verification) [10]. It has also been shown in [11] that sinusoids in noise can be estimated well using group delay function. Encouraged by the results in [11], the modified power spectrum of speech is processed using the modified group delay function in [12]. In [12], pitch extraction was performed on noisy speech.

In this paper, we extend these results for music. The advantage of the proposed approach is that it results in peaks at multiples of melodic pitch. Since the intra-frame pitch period is relatively constant over a framesize of 0.1sec, compared to that of speech, a number of pulses are present in the pulse train. The inter-pulse period corresponds to that of the predominant pitch period in music. The outline of the remainder of paper is as follows. Section 1.2 describes the group delay function and the modified group delay function. In Section 2, melody pitch extraction using the modified group delay function is described. In Section 3, the results and analysis of evaluations is presented. The proposed system is compared with existing pitch extraction algorithms for music analysis. Section 4 discusses the relation to prior work. Finally in Section 5, we conclude with suggestions for further improvement.

### 1.2. Group Delay functions and the Modified Group Delay function(MODGD)

Spectral representation of a signal is complete only when both the phase and magnitude spectrum are available. The magnitude is processed extensively for pitch extraction [13] (via cepstrum), [14] (using autocorrelation) but the phase spectrum is seldom used. Unlike the phase spectrum the group delay function (negative derivative of phase) contains the same information as that of the phase spectrum and has shown to have some useful properties [10]. The group delay function $\tau(\omega)$ of a discrete time signal $x[n]$ and its Fourier transform $X(e^{j\omega})$, is defined as:

$$-\arg(\frac{(dX(e^{j\omega})}{d\omega}) \quad (1)$$

The group delay function can be computed directly from the signal as [15]:

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{\mid X(\omega) \mid^2} \quad (2)$$

where the subscripts $R$ and $I$ denotes the real and imaginary parts. $X(\omega)$ and $Y(\omega)$ are the Fourier transforms of $x[n]$ and $nx[n]$, respectively. Owing to the presence of $\mid X(\omega) \mid^2$ in the denominator, the group delay function is noisy. This is because of zeroes in $\mid X(\omega) \mid^2$ that are caused by the zeroes of the source and convolution with the finite window. finite window length. To overcome the effects of framing and source, the group delay function is modified to give the modified group delay function which is defined [10]:

$$= \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{\mid S(\omega) \mid^{2\gamma}} \quad (3)$$

where $S(\omega)$ is the cepstrally smoothed version of $X(\omega)$. In [11], it is shown that the modified group delay function can be used to process sinusoids in noise. Although for sinusoids the poles are on the unit circle, owing to the finite frame length in digital processing of signals, the finite window length leads to a convolution in the frequency domain. The effect of window zeros is suppressed in the group delay function computed using Equation 3. Figure 1(a) shows a noisy composite signal consisting of a sum of sinusoids. Figure 1(b) shows the magnitude spectrum of the given signal. Figure 1(c) shows the group delay spectrum obtained using the Equation 2. Figure 1(d) shows the modified group delay spectrum obtained using Equation 3. Observe that the two sinusoids are well separated in the modified group delay domain.

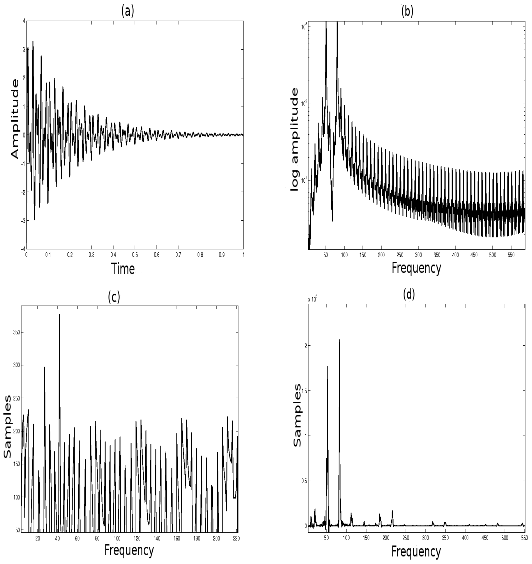

**Fig. 1**. (a)Noisy composite signal,(b) Magnitude Spectrum, (c) Group delay spectrum of (a), (d) Modified group delay spectrum of (a)

The ability of the modified group delay function to resolve sinusoids in noise is explored in the context of extraction of melody for music. This is explained in the next section.

## 2. THEORY OF MELODIC PITCH EXTRACTION USING MODIFIED GROUP DELAY FUNCTION

Assuming a source-system model for production of sounds in music, melody in music corresponds to the periodicity and

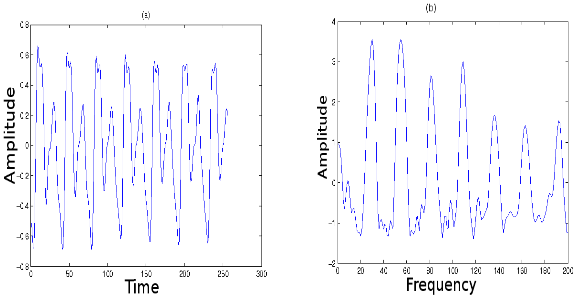

**Fig. 2**. (a) Time domain Signal, (b) Flattened spectrum

amplitude of the source while the timbre information corresponds to that of the instrument or vocal tract. Periodicity of a signal is related to timing. The periodicity of the source manifests as picket fence harmonics in the power spectrum of the signal. If the timbral information can be suppressed, the picket fence harmonics are essentially pure sinusoids (see Figure 2). The sinusoids can be thought of as sinusoids in noise. Referring to the source system in Figure 2, assume that the excitation is periodic with some period $T_o$. Consider the $Z$-transform of two impulses separated by $T_o$. Then,

$$E(z) = 1 + z^{-T_o} \tag{4}$$

The root Fourier transform magnitude spectrum $|E(z)|^\gamma$ is given by.

$$\mid E(\omega) \mid^\gamma = |2 + 2cos(\omega T_O)|^\gamma \tag{5}$$

where $0 < \gamma \leq 2$. The parameter $\gamma$ controls the flatness of the spectrum. In the frequency domain $\mid E(\omega) \mid^2$ has a periodic component with period $\frac{1}{T_o}$. If the spectral components corresponding to the period component is emphasised, the problem of pitch extraction reduces to that of estimation of a frequency of a sinusoid with period $\frac{1}{T_0}$ in the frequency domain. The cepstral approach uses this idea to estimate pitch[13]. The cepstral approach is successful when the length of the frame is large. For values of $\gamma$ other than 2, peaks are introduced at multiples of $\omega_0$ and are useful for reinforcing the value of pitch. We now replace $\omega$ by n and $T_o$ by $\omega_o$ in Equation 5 and remove the dc component to obtain a signal which is ideally a sinusoid:

$$s[n] = cos(n\omega_o), n = 0, 1, 2, 3.......N - 1 \tag{6}$$

The $Z$-transform of this signal is given by

$$S(z) = \frac{1 - 2cos\omega_o(N-1)z^{-N} + z^{-2N}}{1 - 2cos(\omega_o)z^{-1} + z^{-2}} \tag{7}$$

In the previous Section it was shown that the modified group delay function is quite effective in estimating sinusoids in noise. We first flatten the power spectrum using a root cepstral based smoothing technique[12]. The high signal noise ratio regions of this spectrum are processed using the modified group delay function.

We apply the algorithm to derive the modified group delay function $\tau(\omega)$ described earlier, for this signal s[n]. The numerator polynomial of Equation-7 corresponds to the zeros due to the finite window applied in the time domain. Figure 3(a) shows the modified group delay function of the flattened spectrum for a frame of music. Observe that prominent peaks at multiples of the pitch period. The multiplicity of the pitch period is used to reinforce the estimate of the pitch by folding over. The height of the peak is taken as a measure of salience. The first salient peak is obtained. To reinforce the value of pitch obtained, a search is made for peaks at multiples of this pitch period. A voting rule is used to determine the value of the pitch. Dynamic programming is used to ensure the consistency across frames in the pitch tracking. Melody pitch extracted using the proposed method is compared with ground truth for *daisy1.wav* in Figure 3(b). In Figure 4 melodic pitch extracted for a Carnatic music excerpt of *Kamboji* raga(Melody) [16] using the proposed algorithm is plotted along with YIN pitch. Since ground truth is not available, pitch contours were used to synthesize the music, which was evaluated for correctness by a professional musician. In the next Section, we present the results of the given approach.

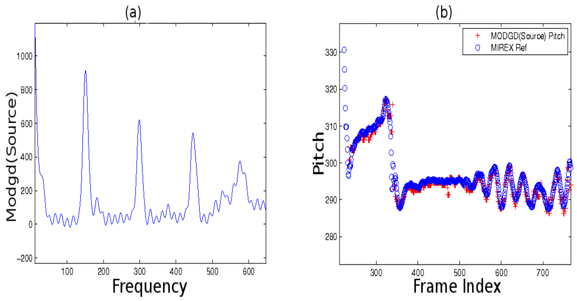

**Fig. 3**. (a)MODGD(Source) plot of a frame, (b) Melody Pitch extraction for *'daisy1.wav'* using MODGD(Source)

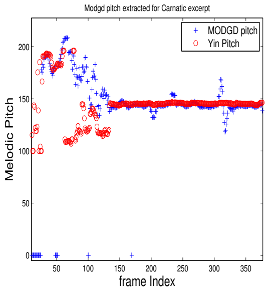

**Fig. 4**. Melody Pitch extraction using MODGD(Source) for Carnatic music excerpt

## 3. EVALUATION

### 3.1. Data set

The proposed algorithm was evaluated using ADC 2004 Dataset(20 audio clips)[17] and LabROSA training set [18]. LabROSA training set is a referential dataset of MIREX, provided by LabROSA of Columbia University,consists of 13 audio files and ground truth $F(0)$ data for each audio. We have also shown the result on one Carnatic music excerpt using the proposed method.

### 3.2. Evaluation method

Evaluation is performed under the assumption that preprocessing using VAD for Voiced/Unvoiced classification has been carried out. We evaluate the performance of the melody pitch extraction for voiced frames only. The reference frequencies of an unvoiced frame is considered to be 0 Hz. The estimated pitch of a voiced frame will be considered correct when it satisfies the following condition:

$$| F_r(l) - F_e(l) | \leq \frac{1}{4} tone(50cents) \tag{8}$$

where $F_r(l)$ and $F_e(l)$ denote reference frequency and estimated pitch frequency on the $l^{th}$ frame respectively. The performance of the proposed algorithm is evaluated using two metrics:[19, 20].

*The Raw Pitch Accuracy(RPA)*: It is defined as the ratio between the number of the correct pitch frames in voiced segments and the number of all voiced frames. A larger raw pitch accuracy means better performance.

*The Raw Chroma Accuracy(RCA)* : It as same as raw pitch accuracy, except that both the estimated and ground truth F0 sequences are mapped into a single octave, in this way ignoring octave errors in the estimation of the correct pitch frames in voiced segments and the number of all voiced frames.

*The Standard deviation of the pitch detection* $\sigma_e$: it is defined as:

$$\sigma_e = \sqrt{(\frac{1}{N} \sum (p_s - p_s')^2 - e^2)} \tag{9}$$

where $p_s$ is the standard pitch, $p_s'$ is the detected pitch, N is the number of correct pitch frames and e is the mean of the fine pitch error. e is defined as:

$$e = \frac{1}{N} \sum (p_s - p_s') \tag{10}$$

### 3.3. Results and Analysis

Table-1 compares the raw pitch accuracy (RPA) for many methods submitted for MIREX evaluation in 2012 with that of proposed system using ADC-2004 Dataset [21].Table-2 compares the standard deviation of pitch detection of proposed method with that of YIN and Wavesurfer. From the table it can be observed that the results obtained using modified group delay function are comparable to that of magnitude spectrum based methods and even better than some of the proposed methods in Table-1.

## 4. RELATION TO PRIOR WORK

This paper uses phase based features for melody extraction. Similar to many algorithms that exist in the literature, the starting point for the proposed algorithm is the frequency spectrum [13]. The power spectrum is first flattened to yield a signal that is rich in harmonics. The harmonically rich spectrum is then subjected to modified group delay processing for the estimation of sinusoids in noise as in [11]. Given the rich harmonic nature of music the pitch accuracy is quite accurate. This method of pitch extraction was proposed for speech earlier with little success owing to the fact that intra-frame pitch changes were significant.

**Table 1**. Melody extraction results(ADC-2004)(in %)

| | **RPA** | **RCA** |
|---|---|---|
| V.Arora et al | 64.23 | 71.21 |
| Sam Meyer | 81.41 | 85.92 |
| Bin Liao et al(1) | 55.87 | 66.71 |
| Bin Liao et al(2) | 48.32 | 59.90 |
| Bin Liao et al(3) | 48.32 | 59.90 |
| **MODGD** | **61.96** | **69.51** |

**Table 2**. Comparison of standard deviation of pitch detection.

| Method | $\sigma_e$ | |
|---|---|---|
| | ADC | LabROSA |
| ESPS method(Wavesurfer) | 3.25 | 2.76 |
| YIN | 3.12 | 3.01 |
| MODGD | 3.15 | 2.84 |

## 5. CONCLUSION

In this paper melody pitch is extracted using the modified group delay function as opposed to the conventional magnitude spectrum based approach. The results indicate that the proposed approach is comparable to that of magnitude spectrum based approaches. The range of pitch values is restricted to 50-900Hz. Ideally the range must be higher for music given that the range of pitch in music can span three octaves. The technique does result in pitch doubling and halving occasionally. Dynamic programming is used to ensure the consistency across frames.

## 6. ACKNOWLEDGEMENT

We would like to thank Ashwin Bellur for going through the manuscript carefully and critically analysing the work that is proposed.

## 7. REFERENCES

[1] G. Poliner, D Ellis, A Ehmann, E Gomez, S Streich, and B Ong, "Melody transcription from music audio:approaches and evaluation," *In Proc. of the IEEE Int. Conf. on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1247–1256, May 2007.

[2] Justin Salamon and Emilia Gomez, "Melody extraction from polyphonic music signals using pitch contours characteristics," *In IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 6, pp. 1759–1770, August 2012.

[3] M Goto and S Hayamizu, "A real-time music scene description system: Detecting melody and bass lines in audio signals," *Working Notes of the IJCAI-99 Workshop on Computational Auditory Scene Analysis*, pp. 31–40, May 1999.

[4] C Cao, M Li, J Liu, and Y Yan, "Singing melody extraction in polyphonic music by harmonic tracking," *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 373–374, 2007.

[5] C L Hsu and J S Jang, "Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion," *Proc. International Society for Music Information Retrieval (ISMIR)*, pp. 525–530, May 2010.

[6] G Richard, J L Durrieu, and B David, "Singer melody extraction in polyphonic signals using source separation methods," *In Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 169–172, 2008.

[7] R Paiva, T Mendes, and A Cardoso, "Melody detection in polyphonic music signals," *Comput. Music J.*, vol. 30, no. 4, pp. 80–98, 2006.

[8] Roel Smits and B Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *IEEE Trans. on Speech Audio Processing*, vol. 3, no. 5, pp. 325–333, September 1985.

[9] J.M. Tribolet, "A new phase unwrapping algorithm," *IEEE Trans. Acoustics Speech and Signal Processing*, vol. ASSP-, no. 2, pp. 170–179, 1979.

[10] Hema A. Murthy and B. Yegnanarayana, "Group delay functions and its application to speech processing," *Sadhana*, vol. 36, no. 5, pp. 745–782, November 2011.

[11] B Yegnanarayana and Hema A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Trans. Signal Processing*, vol. 40, no. 9, pp. 2281–2289, September 1992.

[12] B Yegnanarayana, H A Murthy, and V.R.Ramachandran, "Processing of noisy speech using modified groupdelay functions," *In Proc. of the IEEE Int. Conf. on Audio, Speech and Signal Processing*, pp. 945–948, May 1991.

[13] A. M. Noll, "Cepstrum pitch determination," in *J. Acoust. Soc. Amer.*, 1967, pp. 179–195.

[14] A De Cheveigne and H Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, pp. 111(4):1917–1930, 2002.

[15] A V Oppenheim and R W Schafer, *Discrete Time Signal Processing*, Prentice Hall, Inc, New Jersey, 1990.

[16] "http://en.wikipedia.org/wiki/kambhoji," .

[17] S Joo, S Jo, and C D Yoo, "Melody extraction from polyphonic audio signal mirex-2010," *MIREX-2010*, 2010.

[18] H Tachibana, T Ono, N Ono, and S Sagayama, "Melody line estimation in homophonic music audio signals based on temporal-variability of melodic source," *In Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 425–428, April 2010.

[19] S Jo, S Joo, and C D Yoo, "Melody pitch estimation based on range estimation and candidate extraction using harmonic structure model," *In Proc. INTERSPEECH*, pp. 2902–2905, 2010.

[20] T Cheng, W Xu, Y Tian, and X Hei, "Extracting singing melody in music with accompaniment based on harmonic peak and subharmonic summation," *In Proc. 4th IET international conference on Wireless,Mobile,Multimedia Networks(ICWMMN-2011)*, pp. 200–205, 2011.

[21] "www.music-ir.org/mirex/wiki/2012:mirex2012-results," .