# Pitch Estimation From Speech Using Grating Compression Transform on Modified Group-Delay-gram

Jilt Sebastian, P A Manoj Kumar and Hema A Murthy
Indian Institute of Technology Madras
Chennai, India

***Abstract*—This work presents an approach for pitch extraction based on Grating Compression Transform (GCT) on harmonically-enhanced Modified Group-Delay-gram (Modgdgram). The work explores the use of peakedness and high resolution properties of the group delay functions and the ability of GCT to smear harmonically related components in the spectrum and to track pitch across frames. The power spectrum of the signal is divided by a cepstrally smoothened version of the power spectrum to obtain flattened spectrum. Owing to the picket-fence harmonics due to pitch in the flattened spectrum, the spectrum resembles a sinusoid that is corrupted by noise. This signal is treated as a sinusoidal signal and modified group delay based analysis is performed. Localized time-frequency regions of Modgdgram are used for GCT computation. Peak picking is performed on the resulting rate-scale domain and pitch dynamics are used to finalize the pitch values. The proposed algorithm without any post processing is compared with the traditional GCT computed on the magnitude spectrum and the modified group delay alone. Both natural and synthetic speech are considered for evaluation and an overall improvement of 27% is obtained in the error measures. Finally, two commonly used advanced algorithms which include post processing steps are also considered and the results obtained are comparable.**

## I. Introduction

Pitch estimation is an essential task in speech applications including recognition, synthesis and analysis. *Pitch* is defined as the fundamental frequency caused by the vibrations of vocal folds in the voiced regions of speech. Although a well researched subject, pitch extraction continues to be a challenging task, owing to deviations from idealism in the production of voiced sounds in the glottis. Various approaches have been proposed to estimate pitch from speech segments. They can be broadly classified as spectral, temporal or a combination of both. A major challenge in pitch extraction is to suppress noise interference and system information and detect the predominant pitch sequence. Yin [1], RAPT [2] and YAAPT [3] approaches use correlation based methods with a set of post processing steps to find the pitch values, where as [4] uses cepstral representation of magnitude spectrum for pitch estimation. Salamon et al [5] use pitch contour characteristics for pitch estimation from music signals. In [6], an approach for melody extraction in the presence of pitched accompaniment in polyphonic music is proposed and an interface for melodic pitch extraction from polyphonic music is discussed in [7].

Though phase spectrum has rarely been employed in pitch extraction from speech [8], modified group delay is found to possess useful properties which make it convenient for this task [9]. The phase spectrum is difficult to process, because no meaningful analysis can be performed directly on it since it is in wrapped form. Unwrapping can be done using Tribolet's algorithm [10]. The group delay function, which is defined as the negative derivative of phase spectrum can also be used alternatively. It has been used to extract formants as well as features for tasks such as speech and speaker recognition and spectrum estimation [11], [12]. However, due to windowing, zeros are introduced close to the unit circle which appear as large peaks in the group delay spectrum. The modified group delay function was proposed to reduce this effect. Features extracted from the modified group delay function from a segment of mixed phase speech have been used in speaker verification and speech recognition applications [11]. Using an empirical analysis, it was shown that the modified group delay function is robust in [12], while [13] establishes theoretically the robustness of group delay processing. The modified power spectrum is processed for pitch estimation from noisy speech using the modified group delay function in [9] based on the results in [12]. In [4], the concept is extended for pitch extraction from music signals.

In [14], application of 2-D processing of speech to pitch estimation is proposed. 2-D Fourier transform of localized regions of spectrogram are processed to obtain Grating Compression Transform (GCT) of patches. It was shown to have better performance as compared to sine wave based pitch estimator [15], though it did not develop into a stand-alone algorithm. However, multi pitch estimation

978-1-4799-6619-6/15/$31.00 © 2015 IEEE

techniques [16] have been proposed on GCT owing to the discriminative ability of GCT in the rate-scale domain. A detailed analysis of some of the pitch estimation techniques can be found in [17].

The method proposed in this work uses a new variant of modified group delay function (after spectral harmonic reinforcement) to build a time frequency representation. We refer to it as the Modified group delay gram (Modgdgram), on which GCT is computed. It differs from earlier applications of GCT which were applied on the magnitude spectrum and also from pure modified group delay for pitch estimation. Moreover, pitch dynamics property [14] is exploited which corrects octave errors.

The rest of the paper is organized as follows - Section II provides a brief overview of modified group delay functions and Grating Compression Transform. Section III discusses the proposed algorithm for pitch estimation. Sections IV and V discuss the experiments and conclusion respectively.

## II. Theory

### A. Modified Group Delay functions

Magnitude and phase spectrum are used for spectral representation of a signal. Magnitude spectrum has been widely used for pitch extraction from speech [18], [1]. However, the phase spectrum is seldom used since it does not reveal any relevant properties in unwrapped form. Group delay function, which is the negative derivative of phase also has the same information but alongwith some distinct properties [11]. The modified group delay function of a discrete time signal $x[n]$ with its Fourier transform $X(\omega)$ can be computed as [19]:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^2} \tag{1}$$

where $Y(\omega)$ is the Fourier transform of $nx[n]$ and $|S(\omega)|$ is a cepstrally smoothed version of $|X(\omega)|$. Figure **1** shows the resolving power of group delay functions for a signal that is a sum of two sinusoids and additive noise.

When the magnitude spectrum of a frame of speech is flattened, the system information is removed and the resulting spectrum corresponds to the spectrum of the source, which is noisy owing to window effects and additive noise.

Assuming that the source is a train of impulses with period $T_0$, the $Z$-transform of the source can be written as:

$$E(z) = 1 + z^{-T_0} + z^{-2T_0} + ... + \tag{2}$$

The power spectrum of the source (assuming $e[n]$ is real) is given by

$$E(z)E^*(z) = (1 + z^{-T_0} + z^{-2T_0} + ... +)(1 + z^{T_0} + z^{2T_0} + ... +) \tag{3}$$

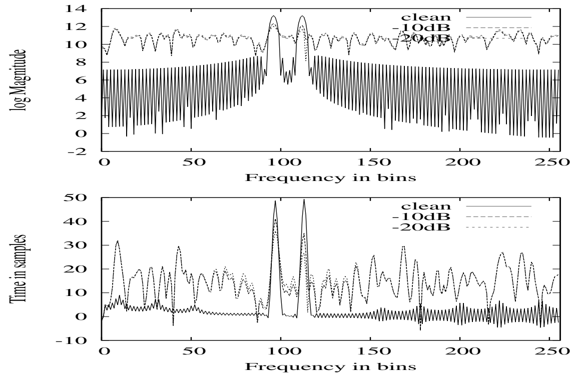

Fig. 1: (a) log Magnitude spectrum of sinusoids and additive noise (b) modified group delay function of sinusoids at different noise levels

Restricting to two impulses per frame, and evaluating the power spectrum on the unit circle we have:

$$\begin{aligned} |E(e^{j\omega}|^2 &= (1 + e^{-j\omega T_0})(1 + e^{+j\omega T_0}) \\ &= (1 + \cos\omega T_0)^2 + (\sin\omega T_0)^2 \\ &= (2 + 2\cos\omega T_0) = 2(1 + \cos\omega T_0) \end{aligned} \tag{4}$$

Equation 4 only introduces sinusoids with frequencies at multiples of $\omega T_0$. Instead of using $|E(e^{j\omega})|^2$, $|E(e^{j\omega})|^{2\gamma}$ is used, where $0 < \gamma < 1$

$$|E(e^{j\omega})|^{2\gamma} = 2^\gamma(1 + \cos\omega T_0)^\gamma \tag{5}$$

The second factor in this equation can be expanded using binomial series as $\cos\omega T_0$ can be at most 1.

$$\begin{aligned} |E(e^{j\omega})|^{2\gamma} &= 2^\gamma(1 + \gamma\cos\omega T_0 + \\ &\quad \frac{\gamma(\gamma-1)}{2}\cos^2\omega T_0 + \\ &\quad \frac{\gamma(\gamma-1)(\gamma-2)}{3!}\cos^3\omega T_0 + ... +) \end{aligned} \tag{6}$$

From 6 it is observed that only terms with frequency at multiples of $T_0$ are introduced by this operation. This signal is therefore equivalent to that of a sum of sinusoids in noise. Thus this signal can be processed using modified group delay. For practical signals, the sinusoids are not everlasting, the window function introduces a fixed bandwidth corresponding to the main lobe width of Fourier transform of the window. When this signal is taken as the time domain signal and processed using the modified group delay function, peaks appear at multiples of $T_0$ of constant height. This is illustrated in Figure 2. A segment of natural speech is chosen for analysis (Figure 2(a)). Figure 2(b) shows the spectrum that is obtained by smoothening the speech spectrum using root cepstral processing. Figure 2 shows the modified group delay spectrum for a value of $\gamma = 0.9$. As the starting point for the processing is the spectral domain, the $X$-axis in the group

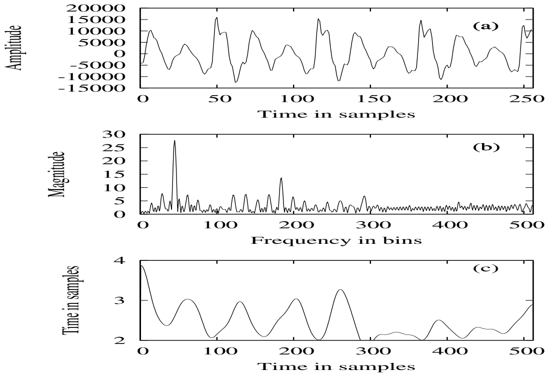

Fig. 2: (a) A frame of speech (b) flattened spectrum (c) modified group delay spectrum of (b)

delay function corresponds to time. Peaks at multiples of $T_0$ are observed in the group delay domain.

### B. Grating compression Transform(GCT)

2-D analysis for pitch estimation is proposed by Quatieri, [14] which considers the spectrogram representation as a 2-D sinusoidal function $s[n,m]$ sitting on a flat pedestal.

$$s[n,m] = K + cos(\omega_s \Phi[n,m]) \tag{7}$$

where $\omega_s$ is the sinusoid frequency, $n$ and $m$ are temporal and frequency variables, $\Phi[n,m]$ denotes the spatial orientation and $K$ is an additive constant. The 2-D Fourier transform of (7) is shown to be

$$\begin{aligned} S(\omega,\Omega) = 2\pi K\delta(\omega,\Omega) + 2\pi\delta(\omega + \omega_s sin(\theta), \Omega - \omega_s cos(\theta)) \\ + 2\pi\delta(\omega - \omega_s sin(\theta), \Omega + \omega_s cos(\theta)) \end{aligned} \tag{8}$$

with $\theta$ denoting the orientation of the sinusoids with respect to time axis. Hence, 2-D Fourier transform of localized regions of spectrogram consists of an impulse at the origin corresponding to flat pedestal and impulses at $\pm\omega_s$ corresponding to sine wave as shown in Figure 3.

Since in the GCT domain (hereafter referred to as rate-scale domain) harmonically related components of the spectrogram are smeared together, it is exploited in pitch estimation. Pitch is estimated by peak picking in the rate-scale domain using the equation:

$$f_0 = \frac{1}{N_{FFT}} \frac{2\pi F_s}{\omega_s cos(\theta)} \tag{9}$$

$f_0$ is the pitch estimate, $N_{FFT}$ is the order of FFT used to compute the spectrogram, $F_s$ is the sampling rate.

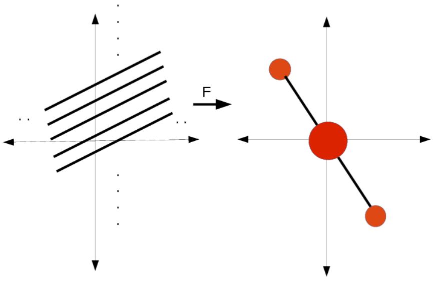

Fig. 3: Schematic of harmonic line structure and it's 2-D Fourier transform for increasing pitch where rotated lines are uniformly spaced over localized time-frequency plane. F denotes 2-D Fourier transform.

## III. The Proposed Algorithm

The algorithm for pitch estimation using modified group delay was first proposed in [20]. In subsection II-B, GCT was shown as a powerful transform to discriminate pitch information from system information. The proposed algorithm performs GCT on the modified group delay gram. A search is made within the expected pitch values on the modified group delay spectrum in every frame. In case of occurrence of a single peak, its harmonic property is reinforced by substituting the value of the single peak. This additional step helps in better visibility of the peaks in rate-scale domain. The pitch estimate can be obtained from modified group delay frames as $1/T_0$. The time-frequency window for computing GCT is centered at this peak location. Since GCT is performed on Modgdgram, it is important to note that the $Y$-axis in the Modgdgram is also time. Hence, in the rate-scale domain, the vertical distance is *directly* proportional to the value of the pitch in samples. In Figure 3, it is seen that the slope of the impulse from the origin is a measure of rate of change of pitch for the selected region. This slope measure is obtained as,

$$\theta = \tan^{-1}\left(\frac{d_{horizontal}}{d_{vertical}}\right) \tag{10}$$

and is used to compute the mean pitch period for each segmented region of Modgdgram.

### A. Examples

Figure 4 illustrates the pitch output obtained for a synthetic signal. The synthetic signal is created using one varying frequency cosinusoid. Two regions are selected for showing the GCT representations, which show peak locations as well as slope from the vertical axis, illustrating that the GCT captures the pitch dynamics.

In a second example, spectrogram and Modgdgram of a short duration of synthetic speech signal is shown (Figure 5)

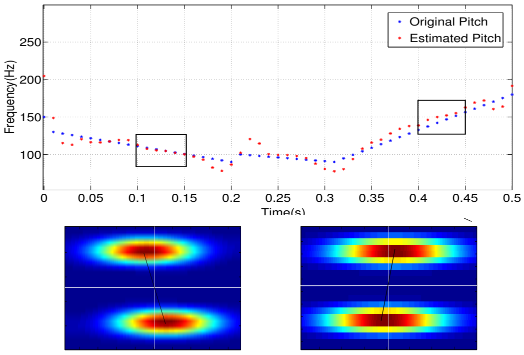

Fig. 4: Ground truth and pitch obtained by proposed method for synthetic signal.Two specific regions of the output and corresponding rate-scale representations are also shown

with the pitch trajectory superimposed. Observe that in the Modgdgram, pitch tracks are emphasized at $kT_0$, where $k$ is an integral. This illustrates the power of Modgdgram and GCT in tandem for pitch extraction.

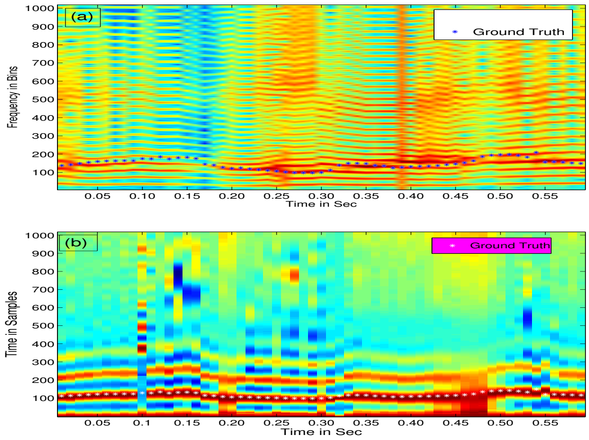

Fig. 5: Ground Truth plotted on (a) Spectrogram and (b) Modgdgram for synthetic speech

## IV. Experimentation

Experiments are performed to corroborate the theoretical findings on synthetic and natural speech dataset. The synthetic dataset has been generated as follows:

Assuming a source system model for speech production, a time dependent vocal tract filter was obtained using a formant vocoder model given by:

$$V(z)=\prod_{1}^{3}\frac{1-2e^{-\alpha_k T}\cos(2\pi F_k T)+e^{-2\alpha_k T}}{1-2e^{-2\alpha_k T}\cos(2\pi F_k T)z^{-1}+e^{-2\alpha_k T}z^{-2}} \tag{11}$$

Figure 6(a) shows the formant contour used. The bandwidths corresponding to $\alpha_k$ were set to 10, 30, 20% of the formant frequency, respectively. This contour was excited by the pitch contour (impulse and Rosenberg's glottal pulse [21]) given in Figure 6(b). The frame length was fixed at 256 samples and the sampling rate was set to 16 kHz.

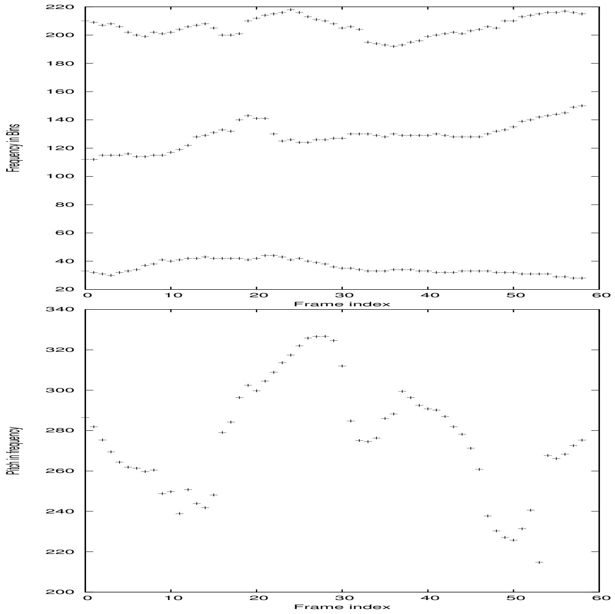

Fig. 6: Formant and pitch used in synthetic speech

The second dataset is the Keele pitch dataset [22] which consists of 5 samples each for men and women uttering the same sentence. The length of each file is over 30 seconds. The files are sampled at 20 kHz with a frame length of 25.6ms and frame shift of 10ms. The pitch reference provided also contains information about voiced and unvoiced frames and has been treated as ground truth. Pitch ranges for the synthetic dataset was fixed at 200-350Hz and at 60-400Hz for Keele dataset. GCT is computed over segments of 50ms duration (patches of 5 frames) with 10ms shift and a frequency range of three times pitch period index. Pitch period index is a rough estimate of the initial peak and is computed over the modified group delay function. Number of FFT points for computing GCT is taken as 4096 along frequency axis and 512 along time axis.

### A. Performance Measures

The proposed method is compared with other algorithms based on the following error measures [23]:
The *Gross Pitch Error (GPE)* is the fraction of frames, where the decisions of both the pitch tracker algorithm and the ground truth are voiced, and for which the relative error of f0

| | Synthetic Database | | | | Keele Database | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **GPE** | **V-UV Error** | **UV-V Error** | **FPE** | **GPE** | **V-UV Error** | **UV-V Error** | **FPE** |
| *ModGD* | 0.038 | 0.000 | 0.000 | 3.857 | 0.031 | 0.046 | 0.213 | 4.520 |
| $Mag+GCT$ | 0.051 | 0.780 | 0.000 | 5.240 | 0.124 | 0.322 | 0.113 | 5.089 |
| **ModGD+GCT** | 0.000 | 0.076 | 0.000 | 3.620 | 0.015 | 0.033 | 0.184 | 4.165 |

TABLE I: Error values for the proposed algorithm vs existing algorithms using modified group delay and GCT

| | Synthetic Database | | | | Keele database | | | |
|---|---|---|---|---|---|---|---|---|
| **Method** | **GPE** | **V-UV Error** | **UV-V Error** | **FPE** | **GPE** | **V-UV Error** | **UV-V Error** | **FPE** |
| *Get* $f0$ | 0.000 | 0.362 | 0.000 | 1.175 | 0.008 | 0.029 | 0.033 | 2.874 |
| *Praat* | 0.017 | 0.009 | 0.000 | 1.901 | 0.009 | 0.061 | 0.050 | 3.251 |
| **ModGD+GCT** | 0.000 | 0.076 | 0.000 | 3.620 | 0.015 | 0.033 | 0.184 | 4.165 |

TABLE II: Comparison of the proposed approach with two standard pitch extraction algorithms for speech

is higher than a threshold of 20%.

The *Fine Pitch Error (FPE)* is defined as the standard deviation (in percent) of the relative error of f0 for which this error is below a threshold of 20%. These are the regions of the speech waveform during which pitch actually gets tracked.

The *V-UV Error* is the number of Voiced positions which are obtained as Unvoiced from the pitch tracker.

The *UV-V Error* is the number of unvoiced positions which are estimated as voiced by the pitch estimator.

### *B. Methods compared in this work*

The proposed method ($ModGD+GCT$) is compared with two algorithms with similar approaches and also with two advanced algorithms for pitch extraction. The existing algorithms chosen are methods involving:

$Mag+GCT$: This is the implementation of algorithm proposed by [14] for single pitch extraction using grating compression transform.

*Mod GD*: This is the algorithm used for pitch extraction from speech [11] as well as music [4] using modified group delay function.

The proposed method is then compared with following two methods of pitch estimation and tracking:

*Get* $f0$: This is included in the ESPS package, this method is an implementation of the RAPT algorithm [2].

*Pitch Listing* ($PRAAT$): This is an implementation of PRAAT algorithm [24]. In PRAAT, pitch extraction can be performed using auto-correlation, cross-correlation, SPINET (Spatial PItch NETwork) or subharmonic summation. We have used cross-correlation method which was having better error performance.

### *C. Results*

Table I shows the performance of the proposed algorithm in comparison with similar approaches. The proposed algorithm performs much better than the raw magnitude based approach. The proposed algorithm hence exploits the high resolution property of modified group delay functions and property of GCT to smear harmonically related impulses. Moreover, the Modgdgram is harmonically reinforced and pitch dynamics are explored for selecting pitch track in GCT domain. Figure 7(a) illustrates the performance of the algorithms on a synthetic dataset. It closely follows the ground truth and has an improvement of 32% for synthetic dataset and 21% for Keele dataset.

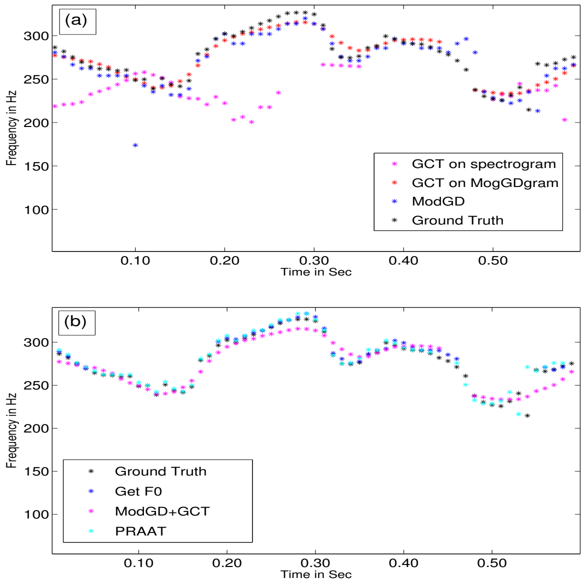

Fig. 7: Pitch estimated on synthetic signal. Proposed Method is compared with similar approaches in (a) and two advanced algorithms in (b)

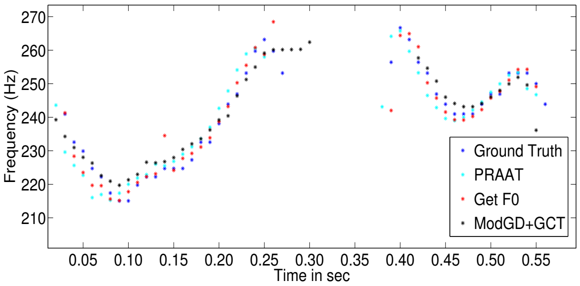

Fig. 8: Pitch estimated on a segment of female speaker from Keele database. Proposed Method is compared with two advanced algorithms

In comparison with two of the advanced algorithms, the error values for the proposed algorithm are nearly comparable even without post processing as shown in Table II. Figure 7(b) shows the pitch track for synthetic dataset and Figure 8 compares the track for a segment of speech from a female speaker from Keele dataset. Even without any voiced activity detector, the proposed algorithm has the same voiced error values. Hence, it has the potential to develop into a robust pitch estimation algorithm.

## V. Conclusion

In this paper, we proposed an algorithm for pitch estimation that employs the high spectral resolution of group delay functions and the pitch tracking capabilities of the grating compression transform. Both properties are individually highlighted by comparing the performance with a magnitude spectrum based approach and a purely modified group delay based approach. Further, we show that performance is comparable to existing advanced techniques and show that the proposed method can be developed into a stand-alone pitch estimation algorithm.

## References

[1] A. D. Cheveigne and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, p. 111(4):19171930, 2002.

[2] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Klein and K. K. Palival, Eds. Elsevier, 1995.

[3] K. Kasi and S. A. Zahorian, "Yet another algorithm for pitch tracking," in *Acoustics, Speech, and Signal Processing (ICASSP), IEEE International Conference on*, vol. 1, May 2002, pp. I–361–I–364.

[4] R. Rajan and H. A. Murthy, "Group delay based melody monopitch extraction from music," in *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, May 2013, pp. 186–190.

[5] 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, pp. 1759–1770, 08/2012 2012.

[6] V. V. Rao and P. Rao, "Vocal melody extraction in the presence of pitched accompaniment in polyphonic music," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2145–2154, 2010.

[7] V. R. S. Pant and P. Rao, "A melody detection user interface for polyphonic music," in *National Conference on Communications 2010 (NCC-2010)*, Chennai, India, January 2010.

[8] R. Smits and B. Yegnanarayana, "Determination of instants of significant excitation in speech using group delay function," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 5, pp. 325–333, Sep 1995.

[9] H. A. M. B. Yegnanarayana and V. R. Ramachandran, "Processing of noisy speech using modified group delay functions," pp. pp.945–948, May 1991.

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

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

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

[13] S. Parthasarathi, R. Padmanabhan, and H. A. Murthy, "Robustness of group delay representations for noisy speech signals," *International Journal of Speech Technology*, vol. 14, no. 4, pp. 361–368, 2011.

[14] T. F. Quatieri, "2-d processing of speech with application to pitch estimation," in *7th International Conference on Spoken Language Processing, ICSLP - INTERSPEECH, Denver, Colorado, USA*, September 2002.

[15] R. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Acoustics, Speech, and Signal Processing, ICASSP-90., IEEE International Conference on*, Apr 1990, pp. 249–252 vol.1.

[16] T. T. Wang and T. F. Quatieri, "2-d processing of speech for multi-pitch analysis," in *INTERSPEECH, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom*, September 2009, pp. 2827–2830.

[17] M. M. Sondhi, "New methods of pitch extraction," *Audio and Electroacoustics, IEEE Transactions on*, vol. 16, no. 2, pp. 262–266, 1968.

[18] A. M. Noll, "Cepstrum pitch determination," *Journal of the Acoustical Society of America*, pp. 179–195, 1967.

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

[20] H. A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, no. 3, pp. 259 – 267, 1991.

[21] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[22] F. Plante, G. F. Meyer, and W. A. Ainsworth, "A pitch extraction reference database." in *Proceedings of EUROSPEECH*, September 1995, pp. 837–840.

[23] L. R. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 399–418, Oct 1976.

[24] P. Boersma, "Praat, a system for doing phonetics by computer." *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.