Interspeech 2018
2-6 September 2018, Hyderabad

# Estimation of Fundamental Frequency From Singing Voice using Harmonics of Impulse-like Excitation Source

*Sudarsana Reddy Kadiri and B. Yegnanarayana*

Speech Processing Laboratory,
International Institute of Information Technology, Hyderabad, India
sudarsanareddy.kadiri@research.iiit.ac.in, yegna@iiit.ac.in

## Abstract

This paper focuses on the problem of estimating fundamental frequency from singing voice. Estimation of fundamental frequency is a well studied topic in the speech research community. From the recent studies on fundamental frequency estimation from singing voice with state-of-art methods proposed for speech, there exists a significant gap in accuracy for singing voice. This is mainly because of the wider and rapid variations in pitch in singing voice compared to that in speech. To overcome this, in this paper we propose a method to derive the fundamental frequency from singing voice by exploiting the harmonics of impulse-like excitation in sequence of glottal cycles. The proposed method is compared with the eight state-of-art methods such as YIN, SWIPE, YAAPT, RAPT, SRH, SFF_CEP, PEFAC and SHRP on the LYRICS singing database. From the experimental results, it is observed that the accuracy of fundamental frequency by the proposed method is better than many state-of-art methods in various singing categories and laryngeal mechanisms.

**Index Terms**: Fundamental frequency, Singing voice, Excitation source, Glottal closure instants.

## 1. Introduction

Fundamental frequency ($F_0$) is defined as the inverse of the pitch period caused by the periodic vibration of the vocal folds in voiced speech/singing. $F_0$ is of particular interest in several speech/singing voice processing applications such as analysis, modification/conversion, recognition and synthesis. Although $F_0$ estimation is a well studied topic in the area of speech processing, this is not up to the extent in the case of singing voice. Even though both speech and singing are produced by the same vocal apparatus, transporting the speech processing approaches to singing may not be straight forward. This is mainly because of the significant coupling between source and system in singing compared to speech, which is neglected mostly. Also, rapid and wider variations in pitch, greater dynamic range, prolonged voice sounds make the singing voice processing techniques difficult compared to speech. Apart from these, the diversity of singing techniques, categories makes difficult to consider the singing voice as a whole and systematic approach. Because of these factors, speech and singing research fields have evolved side by side by sharing several approaches. In studies [1–3], attempts were made to see the effectiveness of various speech processing techniques such as glottal closure instants (GCIs), $F_0$ and vocoding techniques for singing voice. From these studies, it was found that the usage of speech processing techniques may not be guaranteed for singing voice. Hence, there is a need for developing sophisticated methods for processing singing voice.

Methods of estimating $F_0$ can be classified according to the features they rely on. Based on this, existing methods can be grouped into following three categories.

1. Methods using time-domain properties.
2. Methods using frequency-domain properties.
3. Methods using time-frequency domain properties.

The periodicity information can also be processed in a deterministic way or using a statistical approach.

The time domain methods exploit the periodicity information present in the speech signal or approximated excitation signal (such as LP residual) after decomposing the signal into excitation and vocal tract system [4]. In these methods, the location of the peaks in the autocorrelation or cross correlation sequence is used for estimation of pitch period and there by $F_0$ [5]. For example, PRAAT [6] uses the local maxima in autocorrelation or cross correlation computed on the speech signal. Similarly methods such as RAPT [7] and YAAPT [8] estimate pitch by extracting local maxima of the normalized cross correlation function of the speech signal. Several modifications on the autocorrelation based methods were carried out in YIN [9] with post processing to prevent errors.

Instead of speech signal, some methods use excitation signal to estimate fundamental frequency. In SIFT [10], fundamental frequency is estimated using autocorrelation function of excitation signal where, excitation signal is obtained by applying inverse filtering. Cepstral methods [11, 12] separates the excitation and vocal tract contributions in the cepstral domain using homomorphic transformation. In this, pitch period is computed by measuring the interval to the first dominant peak in the cepstrum. In TEMPO method [13], fundamental frequency is estimated by evaluating the fundamentalness of speech which achieves a maximum value when the AM & FM modulation magnitudes are minimized. Also a few attempts were made for the estimation of periodicity [4] using impulse-like nature of excitation i.e., epochs/GCIs. The interval between consecutive GCIs refers to the pitch period and inverse of it refers to fundamental frequency.

Frequency domain methods use the presence of harmonic peaks in the spectrum. In this category, methods are proposed based on the idea of using summation of harmonics in the spectrum [14–17]. In [14], $F_0$ is estimated using subharmonic summation. Instead of using spectrum of speech, in [15], authors used spectrum of LP residual signal so as to reduce the effect of vocal tract system in the resultant spectrum. Methods that use the harmonics of LP residual signal, rely on source filter decomposition. It is known that source filter decomposition through LP analysis fails drastically, for the voices such as singing voice, due to significant coupling between source and system.


10.21437/Interspeech.2018-2495

Some methods combine various techniques such as combination of harmonic ratios and cepstral analyses [18, 19]. In methods that use time-frequency domain, the speech signal is decomposed into multiple frequency bands and then time domain methods such as autocorrelation are applied on each of the sub bands [20]. Also, different weighting schemes on sub-bands are proposed for robust $F_0$ estimation [21]. Some methods use statistical or data-driven approaches to learn the degradation effects on the strength of location of the harmonics in speech using models such as HMM, DNN and CNN [22–27]. These data-driven approaches yield robust pitch estimation when the test data has characteristics that are similar to the data used in training.

Most of the existing techniques assume periodicity in the successive glottal cycles which is limited to generic human pitch range $60 - 400Hz$. However due to rapid and wider variations in pitch of singing voice, there is a need for methods that can handle these variations.

In summary, factors that affect the performance of the $F_0$ estimation methods are:

- Significant source filter coupling
- Effect of vocal tract resonances
- Rapid and wider variations in pitch
- Usage of thresholds such as setting range of $F_0$

In this paper, we propose a method for fundamental frequency estimation for singing voice, based on the harmonics of impulse-like excitation source, derived from modified zero frequency filtering (ZFF) method. The modified ZFF gives the impulse-like sequence of excitation with their corresponding strengths. The method can handle rapid and wider variations in pitch. The organization of the paper is as follows: Section 2 gives a method of extraction of impulse-like excitation sequence from singing voice. Section 3 presents the proposed method of estimation of pitch using the harmonics of impulse-like excitation. The experimental protocol is described in Section 4. The proposed method is compared with several standard methods of pitch estimation in Section 5. Finally Section 6 gives a summary.

## 2. Motivation for the study

The following two ideas motivated for the present study. One is the method of accurate estimation of glottal closure instants (GCIs) from singing voice which can handle rapid variations in pitch [28, 29]. Other is the summation of harmonics in the spectrum [15]. In this study, we are exploiting these two ideas for the estimation of pitch in singing voice. The summation of harmonics method uses the property that the periodicity in time domain is reflected as harmonics in frequency domain. Obtaining periodicity in time domain is a harder task as the vocal tract resonance influences the signal, especially in the presence of significant source filter coupling. To overcome this problem, we are proposing to use impulse-like sequence representation of excitation source derived directly from the speech signal.

## 3. Extraction of Impulse-like Excitation Source

For extracting the impulse-like excitation characteristics from singing voice, we used modified zero frequency filtering method. In this method, the differenced speech signal is passed through a resonator (given in Eqn. (1)), and the trend in the resonator output ($y_0[n]$) is removed by using a moving average filter (given in Eqn. (2)) [28, 30].

$$y_0[n] = -\sum_{k=1}^{2} a_k y_0[n-k] + x[n], \tag{1}$$

where $a_1 = -2$ and $a_2 = 1$.

$$y[n] = y_0[n] - \frac{1}{2N+1} \sum_{m=-N}^{N} y_0[n+m], \tag{2}$$

where $2N + 1$ corresponds to the number of samples used for computing the trend. This process is repeated twice i.e., passing the resultant signal ($y[n]$) through a resonator and removing the trend. It is to be noted that, this repetition operation is different from passing the signal through three resonators and removing the trend. The resulting signal oscillates according to variation of local pitch period [28], and is referred to as modified ZFF signal. The instants of negative-to-positive zero crossings (NPZCs) correspond to the major significant excitations called as epochs/GCIs. Let the epochs be denoted by $\mathcal{E} = \{e_1, e_2, ..., e_M\}$, where $M$ is the number of epochs. A measure of the strength of impulse-like excitation around the GCI and is given by

$$e[l] = |y[e_l + 1] - y[e_l - 1]|. \qquad l = 1, 2, ..., M. \tag{3}$$

For illustration, Fig. 1 (a) shows the segment of a baritone singing voice, (b) is the modified zero-frequency filtered signal, (c) is the strength of impulse-like excitation sequence, and (d) is the differenced electroglottographic (EGG) signal . It can be seen that, there is a close agreement between the locations of the strong negative peaks of the differenced EGG signal and the instants of NPZCs derived from the modified ZFF signal.

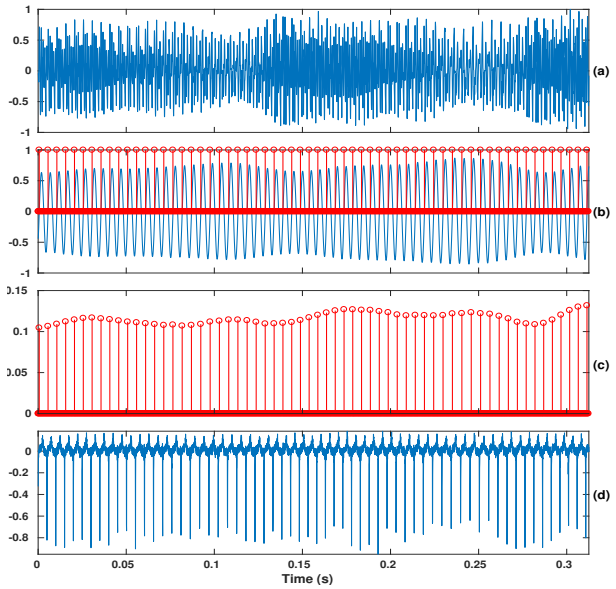


Figure 1: *(a) A segment of baritone singing voice. (b) Modified zero-frequency filtered signal along with the epoch locations. (c) Strength of impulse-like excitation sequence, and (d) Differenced EGG signal for reference.*

The interval between successive GCIs gives the pitch period. However, due to is a shift in the estimated GCIs or due to spurious GCIs, the estimation of pitch is effected. To overcome this, we use the harmonics of strength of impulse-like excitation sequence.

## 4. Fundamental Frequency Estimation based on Harmonics of Impulse-like Excitation Source

The method relies on the analysis of the strength of impulse-like excitation sequence. As the strength of impulse-like excitation at GCI has the higher value, we are computing its spectrogram to highlight the periodicity information present in it. Similar to the methods proposed in [15], the present method also focuses on the harmonicity information. But it is to be noted that, this method does not use information of source filter decomposition so as to obtain the periodicity information in the excitation signal. Instead, it uses one important characteristic of the excitation signal namely, strength of the impulse-like excitation. In this study, the harmonic criterion is similar to that proposed in [15]. From the spectrogram of the strength of impulse-like excitation sequence (Fig. 2), it can be seen that, for the voiced segments of speech, peaks are present at harmonics of $F_0$. The strength of impulse-like excitation sequence is denoted as $e(t)$, which is $e[l]$ at epochs and at other time instants as zero. The amplitude spectrum of $e(t)$ is computed for each hamming windowed frame, covering several glottal cycles and is denoted as $E(f)$. From the spectrum $E(f)$, in the frequency range $[F_0min,\ F_0max]$, the summation of impulse characteristics (SIH) is computed as

$$SIH(f) = E(f) + \sum_{k=2}^{N_h}[E(k.f) - E((k-1/2).f)]. \quad (4)$$

In equation (4), the term $E(k.f)$ in the summation, takes the contributions of the first $N_h$ harmonics into the account. It is expected that this expression reaches a maximum value for $f = F_0$. However, this is also true for the harmonics present in the range $[F_0min,\ F_0max]$. To overcome this effect, the subtraction of $E((k-1/2).f)$ allows the significant reduction of relative importance of maximum of SIH at even harmonics. The estimated $F_0$ value for a given frame is thus the sequence that maximizes $SIH(f)$. Figure 2 illustrates the spectrogram of impulse-like excitation sequence for a segment of baritone singing voice shown in Fig. 3(a).

In this study, we use $N_h = 5$, window length of 50 $ms$ with shift of 10 $ms$ and $F_0$ range is set to $60 - 1500Hz$ to account for wider variations in pitch. As there exists fluctuations in $F_0$, especially at the transition regions, a 5 point median filtering is carried out.

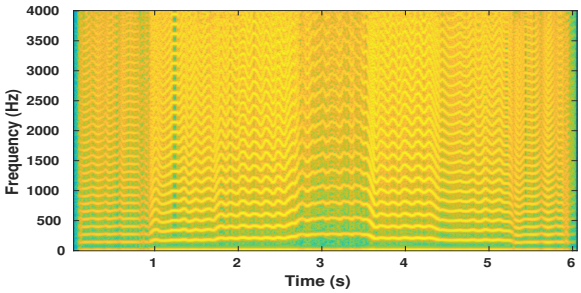


Figure 2: *Spectrogram for the strength of impulse-like excitation sequence for a segment of baritone singing voice shown in Fig. 3(a).*

Fig. 3 illustrates the derived $F_0$ contours in comparison with ground truth for baritone singing voice shown in Fig. 3(a). The ground truth of $F_0$ for this case is shown in Fig. 3(b). The $F_0$ contour derived by the proposed method (SIH) is shown in Fig. 3(c). The $F_0$ contour in Fig. 3(c) matches well with the one in Fig. 3(b). Figs. 3(d), 3(e) and 3(f) show the $F_0$ contours derived using the SRH, SWIPE and YIN methods, respectively.

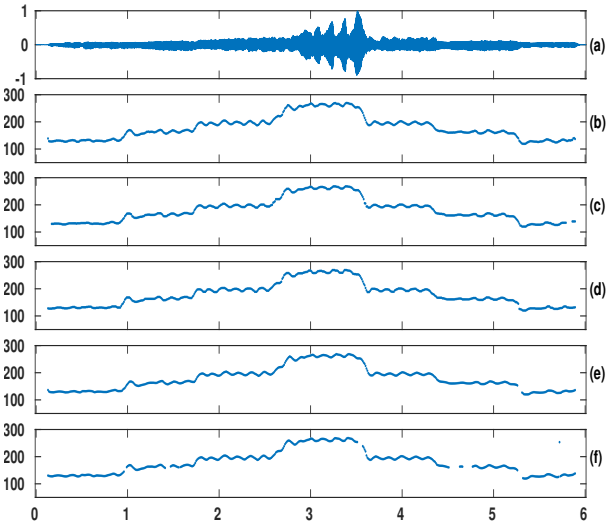


Figure 3: *$F_0$ contour comparison. (a) A segment of baritone singing voice. (b) Ground truth $F_0$ contour. (c) SIH $F_0$ contour. (d) SRH $F_0$ contour. (e) SWIPE $F_0$ contour, and (f) YIN $F_0$ contour.*

Table 1: *Averaged performance comparison of various $F_0$ estimation methods for total database which consists of different types of singing categories and laryngeal mechanisms.*

| Method | GPE | SPD | MPD |
|---|---|---|---|
| SFF_CEP | 6.04 | 6.95 | 12.44 |
| SRH | 21.31 | 2.17 | 2.17 |
| **SIH** | **1.72** | 2.77 | 2.72 |
| SWIPE | 1.78 | 2.17 | 2.25 |
| YAAPT | 25.97 | 4.19 | 5.67 |
| YIN | 5.95 | 1.2 | 0.94 |
| RAPT | 7.56 | 2.16 | 2.03 |
| PEFAC | 27.64 | 4.4 | 3.61 |
| SHRP | 23.26 | 1.97 | 1.62 |

## 5. Experimental Protocol

This section describes the singing database used, methods used for comparison and the evaluation metrics for pitch estimation.

### 5.1. Database used and ground truth

In this study, we use LYRICS singing database which consists of 13 trained singers [3, 31]. The database comprises of 7 bass-baritones (B1 to B7), 3 sopranos (S1 to S3), and 3 countertenors (CT1 to CT3). The recordings were carried out in a soundproof room. Acoustic and electroglottographic (EGG) signals were recorded simultaneously and the acoustic signal was recorded at a distance of 50 cm from the singer's mouth. The singing tasks comprises of ascending and descending glissandos, crescendos-decrescendos and arpeggios and sustained vowels. Depending on possibility, singers sing in laryngeal mechanisms M1, M2 and Mx (laryngeal mechanism smoothly switches from one to the other).

In order to objectively assess the performance of pitch trackers, a ground truth is required. In this study, we use pitch contours created by the authors in [3, 31], which are shown to be reliable. The authors derived pitch from EGG signal and then applied a manual verification process by visually compar-

Table 2: *Performance comparison of various $F_0$ estimation methods for different types of singing categories (baritone, countertenor and soprano).*

| Singing Category | **Method** | **GPE** | **SPD** | **MPD** |
|---|---|---|---|---|
| Baritone | SFF_CEP | 9.55 | 3.74 | 4.94 |
| | SRH | 29.2 | 1.54 | 1.51 |
| | **SIH** | **2.72** | 2.19 | 2.11 |
| | SWIPE | 3.21 | 1.64 | 1.62 |
| | YAAPT | 7.17 | 3.36 | 4.91 |
| | YIN | 9.02 | 0.88 | 0.6 |
| | RAPT | 5.6 | 1.65 | 1.54 |
| | PEFAC | 7.91 | 3.89 | 3.76 |
| | SHRP | 16.04 | 1.37 | 1.2 |
| Countertenor | SFF_CEP | 3.22 | 8.07 | 13.34 |
| | SRH | 15.85 | 2.73 | 2.77 |
| | **SIH** | 0.92 | 3.3 | 3.3 |
| | SWIPE | **0.08** | 2.42 | 2.68 |
| | YAAPT | 27.08 | 5.78 | 7.9 |
| | YIN | 3.51 | 1.38 | 1.14 |
| | RAPT | 5.57 | 2.38 | 2.44 |
| | PEFAC | 37.32 | 5.39 | 4.23 |
| | SHRP | 18.47 | 2.5 | 2.28 |
| Soprano | SFF_CEP | 0.26 | 13.85 | 30.48 |
| | SRH | 7.38 | 3.17 | 3.19 |
| | **SIH** | 0.09 | 3.63 | 3.63 |
| | SWIPE | **0.05** | 3.23 | 3.35 |
| | YAAPT | 72.49 | 4.55 | 5.12 |
| | YIN | 0.86 | 1.82 | 1.57 |
| | RAPT | 14.76 | 3.2 | 2.79 |
| | PEFAC | 66.96 | 4.59 | 2.55 |
| | SHRP | 46.97 | 2.91 | 1.95 |

ing each contour to the spectrogram of the EGG signal.

### 5.2. Evaluation metrics

It is to be noted that the proposed method is unsupervised in nature and hence no training is involved. For assessing the performance, the accuracy of the derived $F_0$ is measured in terms of 3 parameters [4], namely, gross pitch error (GPE), standard pitch deviation (SPD) and mean pitch deviation (MPD). GPE is the percentage of voiced frames of estimated $F_0$ deviating beyond 20% from the ground truth $F_0$ values. SPD and MPD are the standard deviation and the mean of the absolute difference between estimated and ground truth $F_0$ values. For a better performance method, all the values should be low.

### 5.3. Methods for comparison

Performance of the proposed method is compared with eight standard methods. The eight standard methods are SWIPE [32], YIN [9], RAPT [7] and SHRP [16], YAAPT [8], SRH [15], PEFAC [19] and SFF-CEP [33]. For all the methods, $F_0$ search range was set between $60 - 1500\ Hz$ according the study on singing voice in [3]. For evaluation purposes, the frame shift is fixed to $10\ ms$ for all the methods.

## 6. Results and Discussion

Table 1 shows the performance comparison of various methods obtained by averaging all types of singing voices for entire database. It can be seen that, the proposed method (SIH) outperforms all the standard methods in terms of GPE. Among standard methods, RAPT, SFF_CEP and YIN provide better results than SRH, YAAPT, PEFAC and SHRP. The SWIPE method is significantly better than all of these. It is to be noted that, although SFF_CEP method was shown to be better than SWIPE in case of speech signals, it is not as efficient as SWIPE for singing voices.

Table 3: *Performance comparison of various $F_0$ estimation methods for laryngeal mechanisms (LM1, LM2 and LMx).*

| Laryngeal Mechanism | **Method** | **GPE** | **SPD** | **MPD** |
|---|---|---|---|---|
| Laryngeal Mechanism 1 | SFF_CEP | 9.43 | 3.74 | 4.88 |
| | SRH | 30.03 | 1.49 | 1.46 |
| | **SIH** | **2.47** | 2.15 | 2.09 |
| | SWIPE | 2.73 | 1.66 | 1.69 |
| | YAAPT | 6.77 | 3.31 | 4.89 |
| | YIN | 9.08 | 0.88 | 0.62 |
| | RAPT | 5.04 | 1.67 | 1.59 |
| | PEFAC | 7.31 | 3.93 | 3.8 |
| | SHRP | 15.7 | 1.37 | 1.21 |
| Laryngeal Mechanism 2 | SFF_CEP | 0.31 | 12.44 | 25.4 |
| | SRH | 6.67 | 3.35 | 3.4 |
| | **SIH** | 0.39 | 3.83 | 3.82 |
| | SWIPE | **0.03** | 3.04 | 3.21 |
| | YAAPT | 58.96 | 5.67 | 6.95 |
| | YIN | 0.53 | 1.76 | 1.5 |
| | RAPT | 11.88 | 3.01 | 2.76 |
| | PEFAC | 61.75 | 5.24 | 3.33 |
| | SHRP | 36.24 | 3.02 | 2.32 |
| Laryngeal Mechanism x | SFF_CEP | 3.77 | 7.88 | 11.03 |
| | SRH | 12.36 | 1.71 | 2.05 |
| | **SIH** | **5.36** | 1.99 | 2.09 |
| | SWIPE | 9.8 | 2.17 | 2.37 |
| | YAAPT | 19.76 | 6.63 | 9.07 |
| | YIN | 11.64 | 1.02 | 0.92 |
| | RAPT | 9.61 | 2.11 | 2.39 |
| | PEFAC | 48.19 | 3.41 | 2.48 |
| | SHRP | 27.78 | 1.73 | 1.51 |

Table 2 gives the performances based on singing categories: baritone, countertenor and soprano. Out of the three singing categories considered, the SWIPE method performs better than any other standard methods. Although the proposed method is significantly better than SWIPE, its performance is marginally low in the case of countertenor. Even in the case of laryngeal mechanisms, the proposed method performs better than standard methods, except for a marginal difference with SWIPE in LM2. This can be seen in Table 3.

## 7. Summary and Conclusion

In this paper, we proposed a simple method of fundamental frequency estimation by exploiting the impulse-like nature of excitation source. A criterion based on the summation of impulse harmonics (SIH) is proposed for fundamental frequency estimation methods. A comparison with eight state of the art methods is performed for various singing categories and laryngeal mechanisms. The proposed method performed better than the existing methods in many cases.

## 8. Acknowledgments

The first author would like to thank Tata Consultancy Services (TCS), India for supporting his PhD program.

# 9. References


[1] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, and T. Dutoit, "A quantitative comparison of glottal closure instant estimation algorithms on a large variety of singing sounds," in *INTERSPEECH*, 2013, pp. 1702–1706.

[2] O. Babacan, T. Drugman, T. Raitio, D. Erro, and T. Dutoit, "Parametric representation for singing voice synthesis: A comparative evaluation," in *ICASSP*, 2014.

[3] O. Babacan, T. Drugman, N. D'Alessandro, N. Henrich, and T. Dutoit, "A comparative study of pitch extraction algorithms on a large variety of singing sounds," in *ICASSP*, 2013, pp. 7815–7819.

[4] B. Yegnanarayana and K. S. R. Murty, "Event-Based Instantaneous Fundamental Frequency Estimation From Speech Signals," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 614–624, 2009.

[5] W. Hess, "Pitch determination of speech signals springer-verlag," 1983.

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

[7] D. Talkin, "Robust algorithm for pitch tracking," *Speech Coding and Synthesis*, pp. 497–518, 1995.

[8] K. Kasi and S. Zahorian, "Yet another algorithm for pitch tracking," *ICASSP*, vol. 1, pp. 361–364, 2002.

[9] Alain de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 111, no. 4, pp. 1917–1930, Apr 2002.

[10] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Transactions on Audio and Electroacoustics*, vol. 20, no. 5, pp. 367–377, Dec 1972.

[11] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, vol. 41, no. 2, pp. 293–309, 1967.

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

[13] H. Kawahara, H. Katayose, A. de Chevigne, and R. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity," *Eurospeech*, vol. 6, pp. 2781–2784, 1999.

[14] X. Sun, "Pitch Determination and Voice Quality Analysis using Subharmonic-to-Harmonic Ratio," in *ICASSP*, vol. 1, 2002, pp. 333–336.

[15] T. Drugman and A. Alwan, "Joint Robust Voicing Detection and Pitch Estimation Based on Residual Harmonics," *Interspeech*, pp. 1973–1976, 2011.

[16] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, no. 1, pp. 257–264, Jan 1988.

[17] T. Nakatani and T. Irino, "Robust and Accurate Fundamental Frequency Estimation based on Dominant Harmonic Components," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3690–3700, Dec 2004.

[18] N. Yang, H. Ba, W. Cai, I. Demirkol, and W. Heinzelman, "BaNa: A Noise Resilient Fundamental Frequency Detection Algorithm for Speech and Music," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1833–1848, Dec 2014.

[19] S. Gonzalez and M. Brookes, "PEFAC - A Pitch Estimation Algorithm Robust to High Levels of Noise," *IEEE/ACM Transactions Audio, Speech and Language Processing*, vol. 22, no. 2, pp. 518–530, Feb 2014.

[20] A. de Cheveigne, "Speech F0 extraction based on Lickliders pitch perception model," *ICPhS*, pp. 218–221, 1991.

[21] L. N. Tan and A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," *Speech Communication*, vol. 55, no. 7-8, pp. 841–856, 2013.

[22] L. H. J. G. S. Ying and C. D. Michell., "A probabilistic approach to AMDF pitch detection," *ICSLP*, vol. 2, pp. 1201–1204, 1996.

[23] I. J. W. Y. R. Wang and T. C. Tsao, "A statistical pitch detection algorithm," *ICASSP*, vol. 1, pp. 357–360, 2002.

[24] F. Sha, J. Burgoyne, and L. Saul, "Multiband statistical learning for f0 estimation in speech," *ICASSP*, vol. 5, pp. 661–664, 2004.

[25] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum aposteriori probability pitch tracking in noisy environments using harmonic model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 1, pp. 76–87, Jan 2004.

[26] K. Han and D. Wang, "Neural network based pitch tracking in very noisy speech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 2158–2168, Dec 2014.

[27] H. Su, H. Zhang, X. Zhang, and G. Gao, "Convolutional neural network for robust pitch determination," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 579–583.

[28] S. R. Kadiri and B. Yegnanarayana, "Analysis of singing voice for epoch extraction using zero frequency filtering method," in *ICASSP*, Apr. 2015, pp. 4260–4264.

[29] S. R. Kadiri and B. Yegnanarayana, "Epoch extraction from emotional speech using single frequency filtering approach," *Speech Communication*, vol. 86, pp. 52 – 63, 2017.

[30] K. S. R. Murty and B. Yegnanarayana, "Epoch extraction from speech signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 8, pp. 1602–1613, Nov. 2008.

[31] N. Henrich, C. dAlessandro, M. Castellengo, and B. Doval, "Glottal open quotient in singing: Measurements and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency," *J. Acoust. Soc. Am.*, vol. 117, no. 3, pp. 1417–1430, 2005.

[32] H. J. Camacho, A., "A sawtooth waveform inspired pitch estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 124, pp. 1638–1652, 2008.

[33] V. Pannala, G. Aneeja, S. R. Kadiri, and B. Yegnanarayana, "Robust estimation of fundamental frequency using single frequency filtering approach," in *INTERSPEECH*, 2016, pp. 2155–2159.