INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Robust Estimation of Fundamental Frequency using Single Frequency Filtering Approach

Vishala Pannala[1], *G. Aneeja*[2], *Sudarsana Reddy Kadiri*[3], *and B.Yegnanarayana*[4]

Speech and Vision Laboratory,
International Institute of Information Technology, Hyderabad, India
{[1]p.vishala,[2]aneeja.g,[3]sudarsanareddy.kadiri}@research.iiit.ac.in,[4]yegna@iiit.ac.in

Abstract

A new method for robust estimation of fundamental frequency (F_0) from speech signal is proposed in this paper. The method exploits the high SNR regions of speech in time and frequency domains in the outputs of single frequency filtering (SFF) of speech signal. The high resolution in the frequency domain brings out the harmonic characteristics of speech clearly. The harmonic spacing in the high SNR regions of spectrum determine the F_0. The concept of root cepstrum is used to reduce the effects of vocal tract resonances in the F_0 estimation. The proposed method is evaluated for clean speech and noisy speech simulated for 15 different degradations at different noise levels. Performance of the proposed method is compared with four other standard methods of F_0 extraction. From the results it is evident that the proposed method is robust for most types of degradations.

Index Terms: Fundamental frequency, Single frequency filtering, High SNR regions, Harmonics, Root cepstrum.

1. Introduction

Estimation of the fundamental frequency (F_0), i.e., the frequency of vibration of the vocal folds, is essential in many speech processing applications such as synthesis and recognition. Methods for estimation of F_0 involve either time domain or frequency domain or both. In time domain methods, the location of the peak in the correlation sequence computed from a segment of the speech signal or some derived signal (such as linear prediction residual) is estimated [1]. For example, in simplified inverse filter tracking (SIFT) algorithm [2], F_0 is estimated using autocorrelation function of the excitation signal (obtained from inverse filtering of voiced speech). Cepstral-based methods [3, 4] separate the excitation source and vocal tract system in cepstral domain using homomorphic transformation, and F_0 is estimated as the interval to the first dominant peak in the cepstrum (related to excitation signal). Methods such as robust algorithm for pitch tracking (RAPT) [5], yet another algorithm for pitch tracking (YAAPT) [6], Praat [7], estimate F_0 by extracting local maxima of the autocorrelation or crosscorrelation function [8]. Several modifications to the autocorrelation-based methods were carried out to prevent errors in the estimated F_0 as in the YIN algorithm [9]. Frequency domain methods rely on the presence of strong harmonic peaks to estimate F_0 [10]. Examples of this kind are sub-harmonics to harmonics ratio (SHRP) [11], summation of residual harmonics [12], dominant harmonics [13], sawtooth waveform inspired pitch estimator (SWIPE) [14], etc. Also, some approaches [15, 16] combine various techniques of F_0 estimation (e.g., [15] combines harmonic ratios and cepstral analysis) for robustness under degraded conditions.

Typically, in time-frequency domain pitch extraction algorithms, the speech signal is decomposed into multiple frequency bands, and time domain methods are applied on each subband signal. A popular time-frequency domain method is the auditory-model correlogram based algorithm [17]. In this, decomposition is performed using an auditory filter bank, followed by autocorrelation computation on each subband signal. In [18], multi band summary correlogram (MBSC) based pitch detection is proposed, where it uses four wide band FIR filters to capture multiple harmonics in every subband. Different weighting schemes are used to obtain a peak enhanced summary correlogram for robust F_0 estimation.

Some methods [19] use data driven approaches to learn how noise effects the amplitude and location of the peaks in the spectra of speech. In this, the likelihoods of the F_0 candidates are obtained by evaluating the peaks in the spectra using the corresponding models learned from different bands. Also, methods in [20, 21, 22, 23] use statistical approaches to improve F_0 estimation.

There are a few methods which attempt to estimate the glottal closure instants (GCIs), from which the periodicity of the glottal vibration is obtained [1]. In this, the impulse-like nature of excitation (epochs/GCIs) in the sequence of glottal cycles is exploited to derive the instantaneous fundamental frequency from the speech signal directly.

Factors that affect the performance of these methods are:
(a) The quasi-periodicity of the waveform
(b) Effect of resonances of the vocal tract
(c) Rapid variation of F_0
(d) Degradations due to environmental factors like noise and reverberation
(e) Artifacts of digital processing such as block processing, windowing and all-pole modeling

In addition, there are many speech sounds where the glottal vibration is inherently aperiodic like in creaky voices and in some expressive voices (Noh voices) [24, 25].

In this paper, we propose a method for estimating F_0, based on deriving the time envelopes of speech signal at each frequency using single frequency filtering (SFF) [26]. The single frequency filtering output of the speech signal gives spectra with high frequency resolution, although there will be some smearing in the time domain. The peaks at the harmonic frequencies are sharp. Also, the SFF output has regions of high SNR in time and frequency domains. These features are exploited for developing a method for F_0 extraction that is robust against degradations.

The paper is organized as follows: Section 2 gives a brief overview of the SFF approach for processing speech signals to obtain spectra at every sampling instant. Section 3 discusses the proposed method for F_0 extraction using the instanta-

http://dx.doi.org/10.21437/Interspeech.2016-1401

neous spectra. Section 4 gives the performance of the proposed method under various types of degraded speech conditions, in comparison with the performance of some standard F_0 extraction methods. Section 5 gives a summary of the paper.

2. Single Frequency Filtering (SFF) of Speech Signal

In the SFF approach, the envelope of the signal is obtained at any desired frequency. The speech signal is shifted in frequency by multiplying with a complex sinusoid, and the frequency shifted signal is filtered by a single pole filter, with the pole located close to the unit circle (i.e., radius $r > 0.99$) at $f_s/2$, where f_s is the sampling frequency. The following are the steps involved in deriving the SFF envelopes [26].

(a) The input speech signal $x(n)$ is multiplied with $e^{j\hat{\omega}_k n}$ to obtain a frequency shifted $x(n)$. That is

$$x_k(n) = x(n)e^{j\hat{\omega}_k n}, \tag{1}$$

where $\hat{\omega}_k = \frac{2\pi}{f_s}\hat{f}_k$

(b) Pass the signal $x_k(n)$ through a single pole filter, with the pole located at $z = -r \approx -1$ in the z-plane. The transfer function of single frequency filter is given by:

$$H(z) = \frac{1}{1 + rz^{-1}} \tag{2}$$

The corresponding filtered signal is given by

$$y_k(n) = -ry_k(n-1) + x_k(n) \tag{3}$$

In this study, r value is chosen as 0.995.

(c) The envelope $v_k(n)$ of $y_k(n)$ is given by

$$v_k(n) = \sqrt{y_{kr}^2(n) + y_{ki}^2(n)}, \tag{4}$$

where $y_{kr}(n)$ and $y_{ki}(n)$ are the real and imaginary parts of $y_k(n)$, respectively.

(d) The desired frequency f_k is related to $\hat{f}_k$ as follows:

$$\omega_k = \frac{2\pi f_k}{f_s} = \pi - \frac{2\pi \hat{f}_k}{f_s} \tag{5}$$

The envelope $v_k(n)$ can be computed at any frequency f_k. In this study, we have chosen frequencies between $0 - f_s/2$ with $\Delta f = 10\ Hz$ intervals. That is

$$f_k = k\ \Delta f, \qquad k = 0, 1, 2, \ldots, N, \tag{6}$$

where $N = \frac{f_s/2}{\Delta f}$. The $v_k(n)$ for different values of k gives the spectrum at the sampling instant n, and hence it is called the instantaneous spectrum. In this paper, sampling frequency is $8\ kHz$ and $\Delta f = 10\ Hz$. Thus we obtain 400 envelopes within $4000\ Hz$.

3. F_0 Extraction from Instantaneous Spectra

The instantaneous spectra plotted at each sampling instant for a $4\ ms$ segment of voiced speech are shown in Fig. 1. Each instantaneous spectrum shows harmonic structure clearly, even in the high frequency regions (as shown in Fig. 3(a) at one sampling instant). The sum of all the values over frequency is considered as energy $E(n)$, and it is plotted as a function of time. That is,

$$E(n) = \sum_{k=0}^{N} v_k(n) \tag{7}$$

Fig. 2(b) shows the energy as a function of time plotted for the segment of voiced speech shown in Fig. 2(a). For the signal over a frame size of $10\ ms$, the instant at which the instantaneous spectrum has maximum of energy is chosen. The F_0 computed from the spectrum at that instant is assigned to that $10\ ms$ frame. It is easier to extract the periodicity information present in terms of harmonics as representation, as the harmonics are highlighted better in SFF spectrum than DFT spectrum. Note that, in DFT spectrum the components at different frquencies are obtained by projecting the signal segments on the basis funcitons, where as in the SFF method, the components are obtained by filtering the signal at each frequency.

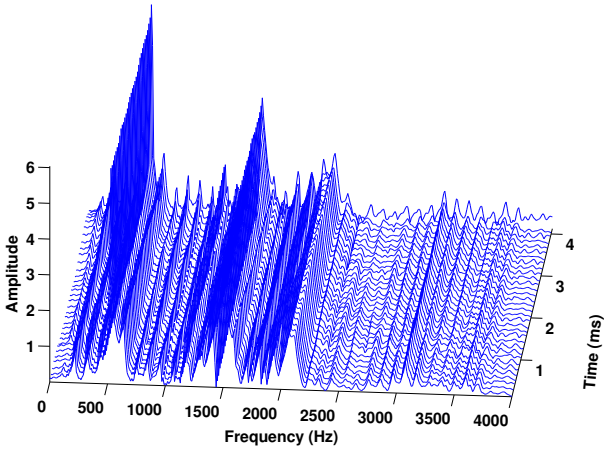

Figure 1: Instantaneous spectra at each sample for a segment of voiced speech.

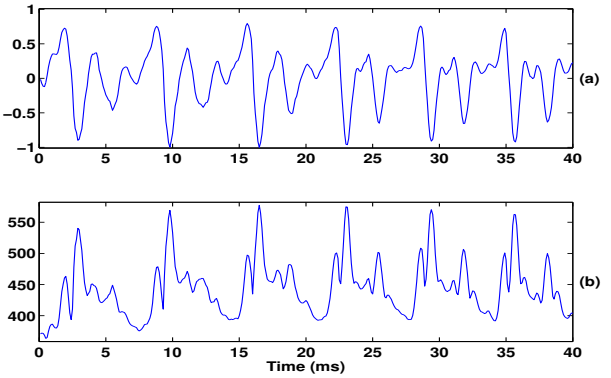

Figure 2: Energy contour for a segment of voiced speech. (a) A segment of voiced speech, and (b) its energy at each sampling instant.

The F_0 is computed using the IDFT of the instantaneous spectrum ($v_k(n)$ as a function of k). The IDFT of $v_k(n)$ may be considered as the root cepstrum ($c_m(n)$) [27], as the spectral values correspond to the values of the envelopes at different frequencies, and not the square of the envelope values. The components corresponding to the spectral envelope (i.e., corresponding to the response of the vocal tract) appear in the low time values of the root cepstrum. The harmonic structure in the spectrum is reflected as the peak in the root cepstrum at the location of the pitch period.

Fig. 3(a) shows the SFF spectrum $v_k(n)$ and the corresponding root cepstrum ($c_m(n)$) is shown in Fig. 3(b). The initial few values in the cepstrum typically represent the vocal tract information. The large peaks present after these initial

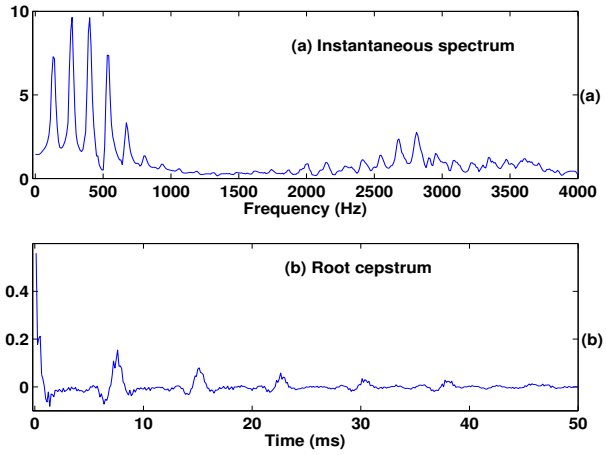

Figure 3: (a) Instantaneous spectrum $v_k(n)$, and (b) Root cepstrum $c_m(n)$ of (a).

values represent the excitation information. The location of the peak in the cepstrum in the range of $2.5-20\ ms$ (corresponding to $20-160$ samples), is considered as the estimated pitch period for that frame. The inverse of T_0 gives F_0. The choice of the interval is made assuming the range of F_0 from $50\ Hz$ to $400\ Hz$. A 5-point median filtering is applied as post-processing for the predicted F_0. The proposed method is further referred to as SFF_CEP. Note that this method is different from normal cepstral based method which uses DFT to obtain the components at different frequencies.

Fig. 4 illustrates the derived F_0 contours in comparison with ground truth for clean as well as for degraded cases of speech. The clean speech signal and the signal degraded by factory1 noise at $0\ dB$ are shown in Figs. 4(a) and 4(b), respectively. The ground truth of F_0 for this case is shown in Fig. 4(c). The F_0 contour for the clean speech derived by the proposed SFF_CEP method is shown in Fig. 4(d). The F_0 contour in Fig. 4(d) matches well with the one in Fig. 4(c). Figs. 4(e), 4(f) and 4(g) show the F_0 contours for the degraded speech in Fig. 4(b) derived using the SFF_CEP method as well by SWIPE and YIN methods. The F_0 contour in Fig. 4(e) is much closer to the ground truth (Fig. 4(c)), compared to the F_0 contours obtained by SWIPE and YIN methods shown in Figs. 4(f) and 4(g), respectively.

4. Performance Evaluation

In this section, performance of the proposed method of F_0 estimation is evaluated for clean speech signals and for noisy speech signals for various degradations at levels of $0\ dB$ and $10\ dB$. The Keele database [28] along with its reference pitch frequency information is used. The database consists of five male and five female speakers, each speaking for about $35\ s$ duration. All the signals are resampled to $8\ kHz$. The reference pitch frequency for every $10\ ms$ is obtained from the simultaneously recorded electroglottograph (EGG) signal, and is used as ground truth. In the ground truth, the unvoiced frames and uncertain frames are considered as nonvoiced frames. The uncertain frames are the frames with mismatch between the manual marking and the EGG signal.

Degraded data is simulated by adding noises from NOISEX database at levels of $0\ dB$ and $10\ dB$ to the clean speech data from Keele database. All the 15 noises from NOISEX database

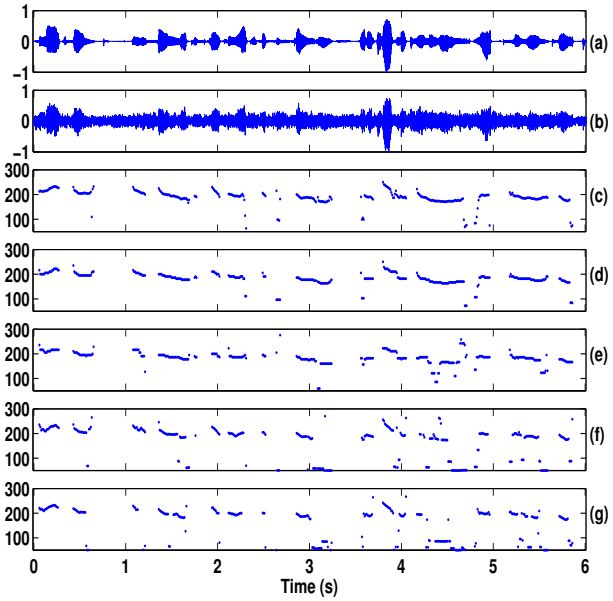

Figure 4: F_0 contour comparison. (a) A segment of clean speech signal. (b) A segment of speech signal simulated with factory1 degradation at $0\ dB$. (c) Ground truth F_0 contour. (d) SFF_CEP F_0 contour of (a). (e) SFF_CEP F_0 contour of (b). (f) SWIPE F_0 contour of (b), and (g) YIN F_0 contour of (b).

are considered here.

The accuracy of the derived F_0 is measured in terms of 3 parameters [1], namely,

- Gross pitch error (GPE): The percentage of voiced frames of estimated F_0, deviating beyond 20% from the reference values.
- Standard pitch deviation (SPD): The standard deviation of the absolute difference between estimated and reference F_0 values.
- Mean pitch deviation (MPD): The mean of the absolute difference between estimated and reference F_0 values.

Gross pitch errors are not considered in determining SPD and MPD.

Performance of the proposed method is compared with four standard methods with their default parameters. The four standard methods are SWIPE [14], YIN [9], RAPT [5] and SHRP [16]. Table 1 shows the performance comparison of various methods in terms of GPE, obtained by averaging over all types of degradations (excluding clean speech) at SNR levels of $0\ dB$ and $10\ dB$. From the table, it can be seen that the SFF_CEP method performs better than other methods.

Table 1: Comparison of methods in terms of gross pitch error (GPE %) using average performances across all 15 noises, for clean speech, at $0\ dB$ and at $10\ dB$ SNR cases.

Method	clean	$10\ dB$	$0\ dB$
SFF_CEP	**2.533**	**5.194**	**19.736**
SWIPE	2.647	7.930	28.365
YIN	5.004	12.073	37.327
RAPT	5.779	17.589	50.711
SHRP	6.993	13.577	28.980

Table 2 shows the results for clean speech data and simulated degraded data at $0\ dB$ in terms of GPE, SPD and MPD. From Table 2, the following observations can be made. In general, the SFF_CEP and SWIPE methods perform better than

Table 2: Performance comparison (in terms of GPE %) of various F_0 estimation methods for clean speech and for 15 types of degraded speech at 0 dB.

Degradation	Method	GPE	SPD	MPD
clean	SFF_CEP	**2.533**	4.981	3.808
	SWIPE	2.647	2.322	1.046
	YIN	5.004	3.018	1.506
	RAPT	5.779	2.903	1.179
	SHRP	6.993	4.126	2.293
white	SFF_CEP	**11.130**	5.699	3.984
	SWIPE	11.710	3.106	1.390
	YIN	14.990	3.672	1.673
	RAPT	61.227	1.858	0.547
	SHRP	21.341	4.273	2.126
babble	SFF_CEP	**23.507**	5.991	3.677
	SWIPE	26.591	4.562	1.826
	YIN	37.212	4.289	1.687
	RAPT	43.468	4.475	1.531
	SHRP	33.340	5.038	2.210
machinegun	SFF_CEP	**14.267**	5.226	3.446
	SWIPE	18.312	2.609	1.022
	YIN	27.588	3.152	1.276
	RAPT	25.064	3.326	1.166
	SHRP	21.273	4.069	1.983
hfchannel	SFF_CEP	10.539	5.714	4.011
	SWIPE	**5.326**	3.429	1.611
	YIN	13.911	3.809	1.744
	RAPT	58.364	2.193	0.668
	SHRP	19.061	4.495	2.266
pink	SFF_CEP	**19.071**	5.415	3.618
	SWIPE	29.807	2.754	1.084
	YIN	36.759	3.128	1.197
	RAPT	59.903	2.170	0.581
	SHRP	29.919	4.247	1.931
volvo	SFF_CEP	16.197	4.802	3.363
	SWIPE	**15.889**	2.112	0.871
	YIN	31.796	2.905	1.156
	RAPT	35.601	2.578	0.834
	SHRP	18.001	3.933	1.991
buccaneer1	SFF_CEP	**23.482**	5.642	3.403
	SWIPE	38.188	2.743	1.046
	YIN	40.116	3.215	1.185
	RAPT	63.001	2.189	0.578
	SHRP	34.377	3.981	1.761
buccaneer2	SFF_CEP	**20.078**	5.677	3.618
	SWIPE	27.099	3.318	1.325
	YIN	36.408	3.354	1.277
	RAPT	61.199	2.367	0.617
	SHRP	30.554	4.206	1.911
destroyerengine	SFF_CEP	22.749	5.641	3.528
	SWIPE	**18.023**	3.263	1.449
	YIN	26.319	3.633	1.557
	RAPT	35.981	4.192	1.623
	SHRP	26.491	4.274	2.026
destroyerops	SFF_CEP	**19.512**	5.620	3.668
	SWIPE	28.885	4.050	1.654
	YIN	42.763	3.949	1.476
	RAPT	42.413	4.406	1.514
	SHRP	32.995	4.611	2.009
factory1	SFF_CEP	**19.119**	5.498	3.621
	SWIPE	32.044	3.233	1.244
	YIN	39.561	3.185	1.208
	RAPT	55.110	2.702	0.771
	SHRP	31.539	4.289	1.936
factory2	SFF_CEP	**21.004**	5.161	3.493
	SWIPE	31.388	2.702	1.024
	YIN	42.949	3.041	1.111
	RAPT	52.258	2.609	0.765
	SHRP	28.186	4.435	2.041
f16	SFF_CEP	**19.120**	5.586	3.738
	SWIPE	37.211	2.394	0.901
	YIN	42.145	3.095	1.141
	RAPT	57.366	2.575	0.738
	SHRP	30.230	4.380	1.990
m109	SFF_CEP	**29.522**	4.873	3.032
	SWIPE	64.638	1.752	0.475
	YIN	71.422	2.118	0.556
	RAPT	58.326	3.080	0.826
	SHRP	40.469	3.479	1.462
leopard	SFF_CEP	**26.747**	5.39	3.139
	SWIPE	40.363	3.157	1.140
	YIN	55.964	3.167	0.996
	RAPT	51.385	3.908	1.202
	SHRP	36.928	4.258	1.772

other methods for clean as well as degraded speech cases. The SFF_CEP method performs better than SWIPE method for clean speech data and across 12 types of degradations. The SFF_CEP method exhibits significantly better performance (beyond 10%) in noises such as m109, f16, leopard, factory1, factory2, pink. At the same time, the difference between SWIPE and SFF_CEP methods is least seen ($< 1\%$) for clean data and for data degraded with white noise. The difference between SWIPE and SFF_CEP methods is small in cases where the performance of SWIPE is better than SFF_CEP.

5. Summary and Conclusion

A new method for robust estimation of F_0 is presented in this paper. The method is based on deriving the envelopes of the signal at different frequencies using SFF approach. The SFF approach gives high SNR regions in time and frequency regions due to correlation of signal samples and lack of correlation of noise samples both in time as well as in the frequency domains. This enables us to choose only the high SNR regions for F_0 estimation. Moreover, the high frequency resolution in the spectrum due to SFF gives sharp harmonics, which help in estimating F_0. The robustness of the method is demonstrated by considering speech affected by various degradations. The performance of the proposed method is compared with the performance of some standard methods for F_0 extraction.

The advantages of SFF approach can be exploited by choosing appropriate single-pole filter at $f_s/2$ to obtain good temporal resolution of the envelopes of the signal at different frequencies, but at the expense of resolution in the frequency domain. The F_0 can then be estimated from the autocorrelation sequence computed from segments of time envelopes. The results of F_0 estimation from both time and frequency domains can be combined to obtain highly robust F_0 estimation under several practical cases of degradations.

6. References

[1] 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.

[2] 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.

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

[4] 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.

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

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

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

[8] T. Shimamura and H. Kobayashi, "Weighted Autocorrelation for Pitch Extraction of Noisy Speech," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 7, pp. 727–730, Oct 2001.

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

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

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

[13] 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.

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

[15] 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.

[16] 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.

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

[18] 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.

[19] W. Chu and A. Alwan, "SAFE: A Statistical Approach to F0 Estimation under Clean and Noisy Conditions," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 3, pp. 933–944, 2012.

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

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

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

[23] 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.

[24] O. Fujimura, K. Honda, H. Kawahara, Y. Konparu, M. Morise, and J. C. Williams, "Noh voice quality," *Logopedics Phoniatrics Vocology*, vol. 34, no. 4, pp. 157–170, 2009.

[25] V. K. Mittal and B. Yegnanarayana, "Study of characteristics of aperiodicity in noh voices," *J. Acoust. Soc. Am.*, vol. 137, no. 6, pp. 3411–3421, 2015.

[26] G. Aneeja and B. Yegnanarayana, "Single Frequency Filtering Approach for Discriminating Speech and Nonspeech," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 705–717, Apr 2015.

[27] T. Nagarajan, V. K. Prasad, and H. Murthy, "Minimum phase signal derived from root cepstrum," *Electronics Letters*, vol. 39, no. 12, pp. 941–942, 2003.

[28] F. Plante, G. F. Meyer, and W. A. Aubsworth, "A pitch extraction reference database," in *Proc. European Conf. on speech comm. (Eurospeech)*, Sep 1995, pp. 827–840.