F0 ESTIMATION FOR NOISY SPEECH BY EXPLORING TEMPORAL HARMONIC STRUCTURES IN LOCAL TIME FREQUENCY SPECTRUM SEGMENT

Dongmei Wang, John H. L. Hansen
Dept. Electrical Engineering, University of Texas at Dallas
800 West Campbell Road, Richardson, Tx. 75080
{dongmei.wang, john.hansen}@utdallas.edu

ABSTRACT

In this paper, we propose a noise robust F0 estimation approach by exploring the temporal harmonic structures in local time-frequency (TF) spectrum segment. Since the speech energy is sparsely distributed on the TF plane, the speech harmonic structures occupied in the higher speech energy TF segment are tending to dominate over noise. Thus, we attempt to derive F0 from such high (signal to noise ratio) SNR TF segments rather than full band signal. Our algorithm comprises of two stages: i) F0 candidate estimation for a series of TF segments; ii) F0 tracking based on the acoustic features of each TF segment as well as the F0 temporal continuity constraints. Experimental results show that our approach outperforms the compared methods in terms of F0 estimation accuracy.

Index Terms— F0 estimation, local TF segment, SNR estimation, temporal continuity constraints

1. INTRODUCTION

Fundamental frequency (F0) is one of the most important characteristics of human speech which represents the vibration rate of the vocal cords during speech production. A promising F0 estimation system will facilitate many speech signal processing areas, such as speech source separation, emotion recognition, speaker/language identification, etc. Recently, F0 estimation has also been applied to assist the mental disease diagnosis [1] [2].

The straightforward way to analyze F0 is either exploring harmonic structures in frequency domain [3] [4] or examine the periodic cues in time domain [5-7]. Correspondingly, autocorrelation function (ACF) and average magnitude difference function (AMDF) are the two basic time domain F0 estimation approaches. Besides, subharmonic summation [3] and comb filter [29] are usually adopted as frequency domain methods. However in adverse conditions, the above traditional F0 estimation methods become ineffective due to both of the temporal periodic cues and harmonic structure are distorted to some degree.

In order to deal with the noisy situation, many efforts have been made by the researchers. For instance, ACF and AMDF are combined together to obtain better periodic peak detection [8] [9]. In addition, various types of adaptive speech representation methods are introduced to enhance the speech component so as to provide a more reliable source for F0 estimation [10-12]. Signal pre-processing is also proposed to attenuate some noise for F0 estimation [13] [14]. Auditory filter bank based F0 estimation is proposed to take advantage of high SNR sub-channels [15-17]. Moreover, the F0 temporal continuity constraints are modeled to ensure more accurate F0 tracking [12] [15] [18]. Recently, statistical and machine learning methods are also widely used for noise robust pitch estimation [19-22].

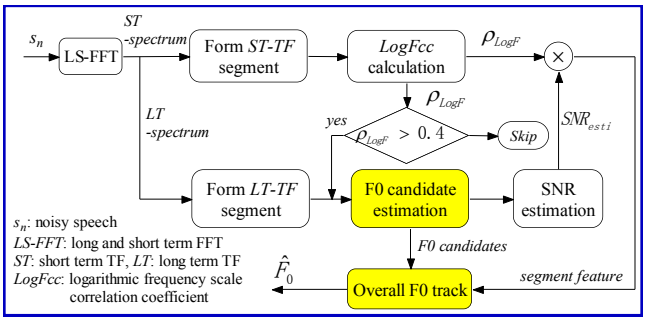

Fig. 1 Algorithm overview

Among the previous studies, temporal harmonic structures have been investigated for noise robust F0 estimation because of the harmonic similarity between adjacent speech frames [12]. Speech sparsity characteristic [23] is also considered that F0 can be estimated from less noise affected channels [15-17] in each frame. However, in seldom cases, temporal harmonic continuity and sparsity are considered simultaneously for F0 estimation. Nevertheless, if the particular spectrum area (TF segment) dominated by continuous frames of harmonic structures are able to be detected for F0 estimation, the performance could be improved.

In this work we focus on F0 estimation by exploring temporal harmonic structures in the local TF segment. First, the noisy speech spectrum is decomposed into a series of overlapped TF segments. A F0 candidate contour is estimated for each local TF segment. Subsequently, overall F0 tracking is performed based on Hidden Markov Model (HMM). Two features are proposed to indicate the F0 accuracy in each TF segment, including logarithmic frequency scale correlation coefficients (*LogFcc*) and an estimated SNR. In addition, two dynamic factors are developed to model the F0 temporal continuity constraints, which are inter-frame as well as inter-segment F0 transition probability. A similar F0 estimation algorithm was proposed in our previous paper [30], nevertheless the overall F0 tracking is improved in this work.

This paper is organized as follows. Section 2 describes an overview of the system. The F0 candidate estimation is presented in Section 3. Section 4 illustrates the target F0 tracking. Experiments and results are described in Section 5. Finally, the conclusions are drawn in Section 6.

2. ALGORITHM OVERVIEW

In this section an overall algorithm overview is presented. The general block diagram is shown in Fig. 1. Generally, our algorithm consists of two main stages: i) F0 candidate contour estimation for

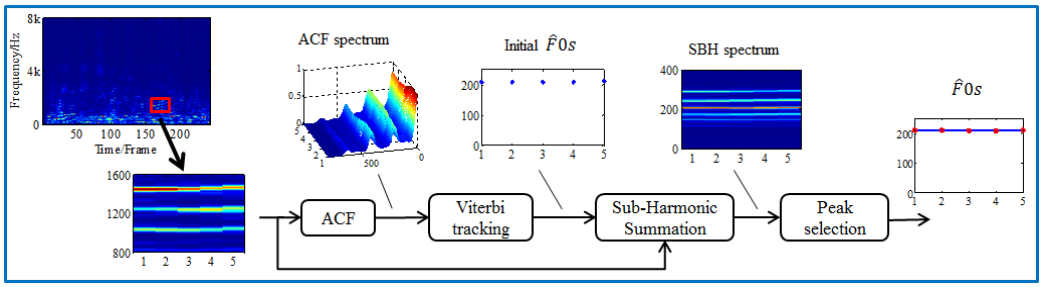

Fig. 2 Overview of F0 candidate estimation

every single TF segment; ii) F0 tracking across the overall TF plane. At first, we analyze the noisy speech signal based on a long-short term associated harmonic model [24]. On one hand, short term spectrum analysis ensures to preserve the short-time stationary property of the speech signal. On the other hand, the long term spectrum analysis is able to obtain a higher frequency resolution, making the speech harmonics more discriminated from noise interference. Each TF segment is formed as 5 frames long in time and 800Hz wide in frequency. The reason we choose 800Hz as the TF segment bandwidth is that at least two harmonic partials are included in such frequency range. A F0 candidate contour with duration of five frames is estimated for each TF segment. After that, the overall F0 tracking is performed based on HMM model. The the observed likelihood of a F0 candidate to be true or false are indicated by two acoustic features: *LogFcc* and an estimated SNR. Moreover all the five F0 candidates located in one TF segment will be assigned the same average likelihood. Meanwhile, the F0 temporal continuity constraints are taken into account by using both the inter-frame and inter-segment based F0 transition probabilities. Finally, Viterbi algorithm is used for F0 decoding.

3. F0 CANDIDATE ESTIMATION

3.1. Initial detection of speech dominated TF segment

The speech harmonic structures usually change more slowly than noise spectrums. The higher the correlation coefficient between two adjacent frames, the more probable the TF segment is dominated by speech. Thus we propose to calculate the *LogFcc* for each short term TF segment to indicate its likelihood of being dominated by speech or not. The computation of *LogFcc* is shown as Eq. (1) - (3)

$$LogFcc = \frac{1}{N-1}\sum_{n}\frac{\left(X^{n}_{\log_F}-\mu_{X_{\log_F}}\right)\left(Y^{n}_{\log_F}-\mu_{Y_{\log_F}}\right)}{\sqrt{\sigma^{2}_{X_{\log_F}}\sigma^{2}_{Y_{\log_F}}}} \quad (1)$$

$$X_{\log_F}(n) = \log_a(X(n+n_{bias})) \quad (2)$$

$$Y_{\log_F}(n) = \log_a(Y(n+n_{bias})) \quad (3)$$

where X and Y are the two neighboring spectrum amplitude vectors in a particular TF segment, N is the sample number of X and Y, n is the index of frequency bin, $n = f \cdot N_{fft} / fs, f \in [1\ \ 800]Hz$, N_{FFT} is the FFT point, fs is the sampling rate, μ and σ^2 are the mean and variance respectively. We set $a = 1.5$, and $n_{bias} = 500 \cdot N_{FFT} / fs$ empirically. For each TF segment, an average *LogFcc* is obtained across five frames.

In addition, transforming the linear frequency scale into logarithm is to restrain the notable frequency differences between high order harmonic structures in two successive frames.

Accordingly, the TF segment with the average *LogFcc* value smaller than a threshold is considered as noise and is discarded before the further processing. Otherwise, the average *LogFcc* values are saved and used for the overall F0 tracking in next step. Here the threshold is empirically set as 0.4.

3.2. F0 candidate contour estimation

In this subsection, we will perform F0 candidate estimation in the initial detected speech dominated TF segments. Here long term TF segments are used instead of short term one to increase the frequency resolution for F0 estimation. Fig. 2 shows the general flowchart of the F0 candidate estimation. We take a long TF segment as an example. ACF is obtained for each frame and is normalized by dividing the maximum amplitude in each frame. The frequencies of the ACF peaks in each frame are considered as the F0 candidates. Moreover, the amplitudes of the corresponding normalized ACF peaks are considered as observation likelihoods of the candidates belonging to true F0. Meanwhile, the F0 transition probability between two consecutive frames ($p(F0_t / F0_{t-1})$) is learnt from Keele [25] and CSTR [26] databases, both of which provide ground truth F0 values. We assume $p(F0_t / F0_{t-1})$ is equivalent as the probability of the F0 change in logarithmic scale between two neighboring frames, which is shown as Eq. (4)

$$p(F0_t / F0_{t-1}) = p\left(\log_{1.5}\left(\frac{F0_t}{F0_{t-1}}\right)\right) \quad (4)$$

Gaussian mixture model is adopted to model the logarithmic F0 change which will be considered as the F0 transition probability. With the observed likelihood and F0 transition probability, Viterbi algorithm is applied for F0 decoding.

Furthermore, we use the sub-harmonic summation technique [3] [27] [28] to correct some F0 estimation errors caused by ACF based approach. The core technique of sub-harmonic summation is to compress the spectrum vector in each frame along the frequency axis by a series of integer factors and sum the compressed spectrum together. In consequence, multiple harmonics will be coincident enhanced and cause a maximum spectrum peak at fundamental frequency. In our case, the integer factors are equal to the harmonic orders calculated by dividing the TF segment frequency bound by the initial detected F0s from ACF method. The frequency of the maximum peak from above compressed and summed spectrum will be considered as the updated F0 candidate in each frame. The idea behind this is that ACF based method and sub-harmonic summation based method should produce the same F0 results. If conflicts happen, there is a high probability that the estimated F0 might be wrong. IN our case, one typical cause of F0 error by ACF estimation is that some TF segments are occupied by equal distance located noise spectrum peaks. Unfortunately, those frequency distances are easily detected by ACF as F0. However, these spectrum peaks are not harmonically correlated with each other, and their frequencies do not have a common factor. Therefore the sub-harmonic summation technique provides a post processing for error correction.

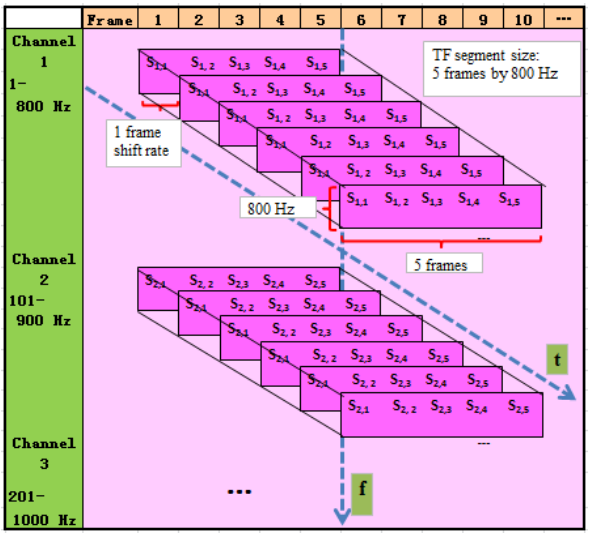

Fig. 3 TF segment status representation

4. OVERALL F0 TRACKING

4.1. Feature extraction for each TF segment

With the estimated F0 candidate contours in each TF segment, we begin to select the optimal pitch via searching those speech dominated TF segment on the overall TF plane. Two parameters are proposed for measuring the likelihood of a specific TF segment is speech dominated or noise dominated. One is logarithmic *LogFcc*, which we described earlier in Section 3, and the other is an SNR value which will be explained here. The SNR is estimated for each TF segment based on harmonics regeneration with estimated pitch candidate contour [22]. First, the harmonic amplitude is obtained by choosing the spectrum peak which is closest to the ideal harmonic frequency (*kF0*) within the predefined deviation range, shown as Eq. (5).

$$A_H^k = \vec{A}_P\left(\lceil kF0 \cdot N_{FFT} / fs \rceil\right) \tag{5}$$

where A_H^k is the selected *kth* order of harmonic amplitude, and $\vec{A}_P$ is the spectrum amplitude peak vector, and $\lceil a \rceil$ represents selecting a existed number that is closest to a. Next the generated harmonic spectrum is obtained by convolving the harmonic peaks with the FFT spectrum of hamming window (with equal size as the short term speech analysis window), see Eq. (6).

$$S_H(n) = A_{ham}(n) * \sum_{k=K_1}^{K_2} A_H^k \cdot \delta(n - F_H^k \cdot N_{FFT}/fs) \cdot e^{jP_H^k} \tag{6}$$

where A_{ham} is the spectrum amplitude of the hamming window, A_H^k, F_H^k and P_H^k are the amplitude, frequency and phase of the *kth* order of harmonics, P_H^k is extracted from the noisy speech directly, and K_1 and K_2 are the lower and upper harmonic order bound of a particular TF segment, here $K_1 = f_l / \hat{F}0$ and $K_2 = f_u / \hat{F}0$, f_l and f_u are the lower and upper frequency bound of that TF segment. In addition, * denotes convolution. Accordingly, the SNR in each frame is calculated as Eq. (7)

$$SNR_{esti} = \max\left(\frac{1}{L}\sum_{j=1}^{L} \frac{\sum_n |S_H^j(n)|^2}{\sum_n \left(|S_N^j(n)| - |S_H^j(n)|\right)^2}, 10\right) \tag{7}$$

where j is the index of frame, L is the total frame number in a TF segment, and $|S_N(n)|$ is the noisy speech spectrum. Finally, the observation likelihood of a specific TF segment containing the true F0 candidate is obtained as:

$$p_l = SNR_{esti} \cdot \bar{\rho}_{LogF} \tag{8}$$

4.2. F0 tracking

The F0 tracking step is to select the best F0 candidate from the candidates list for each frame. Here we model all of the F0 candidates as states in a hidden Markov model (HMM). The TF segment features obtained in section 4.1 is used as the observation likelihood of the F0 candidate states. In addition, we proposed a F0 transition probability for the model that contains two different dynamic factors. One is F0 changing over consecutive time frames, and the other is F0 changing over the adjacent TF segments. The former one is obtained as the same procedure in section 3.2, while the latter one is defined in Eq. (9)

$$p(S_{i',j'} / S_{i,j}) = \begin{cases} 0.7, i' = i, j' = j+1 \\ 0.2, j' = 1, j = 5 \\ 0.1, others \end{cases} \tag{9}$$

where $S_{i,j}$ and $S_{i',j'}$ are the TF segment status of the previous and current F0 candidate respectively, i denotes the frequency channel index which starts from one to the total number of channels, j represents the frame index in each TF segment and it starts from one to the overall frame number for a TF segment. Fig. 3 shows an example of the TF segment status. Each purple horizontal bar represents a TF segment. In fact the TF segments are overlapped in both time and frequency. However, we display the overlapped TF segment separately in different frequency channel in Fig. 3. The TF segment states are shown on the bar frame by frame. Since the F0 candidates are estimated from the TF segments which are overlapped both in time and frequency, the optimal F0 tracking might switch between different TF segments. Nevertheless, it is essential to guarantee that the F0 tracking go through the particular entire TF segment in most of the cases, avoiding the frequent halfway hopping between adjacent TF segments. Therefore, we assign a higher probability for the F0 transition of inner TF segment, and lower probability for other cases.

With the observation likelihoods and F0 transition probabilities, a Viterbi algorithm is performed to decode the overall F0 contour by maximizing the likelihood, shown as Eq. (10).

$$Q_T = \arg \max_{1 \le i \le N_c, 1 \le j \le N_F} \left[p(F0_t) \cdot p(F0_t / F0_{t-1}) \cdot p(S_{i',j'} / S_{i,j})\right] \tag{10}$$

where $p(F0_t)$ is the observed probability of current F0 candidates, which is equals to p_l that is obtained in Eq. (8), and $p(F0_t / F0_{t-1})$ is the frame based F0 state transition probability.

5. EXPERIMENTS AND RESULTS

We use the Keele [25] and CSTR [26] database for the performance evaluation which provides ground truth pitch labels and can be used as a reference for performance assessment. Keele database contains 10 long sentences spoken by five female and five male native British English speakers with total duration of 9 mins. The CSTR database contains 50 English utterances, spoken by both one female and one male English native speaker, with the duration of 7 mins. Six types of daily life noise are used to simulate the noisy environments, including airport, babble, exhibition, restaurant, street, and train noise. Seven SNR levels are set from -10dB to 20dB. Three other state-of-the-art non

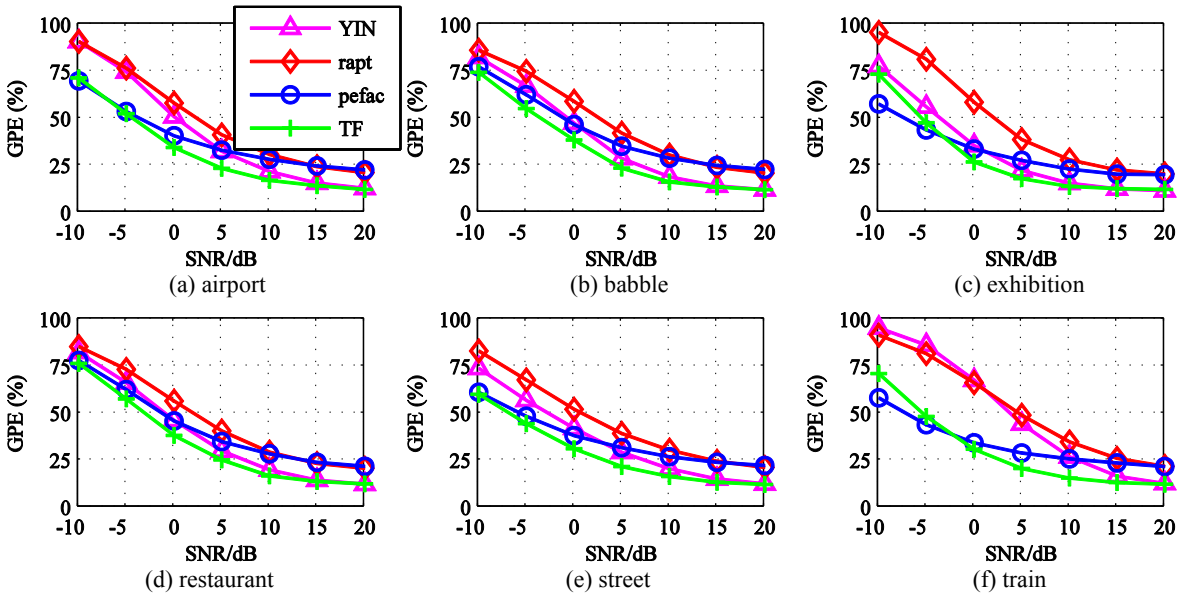

Fig. 4 GPE results for Keele database

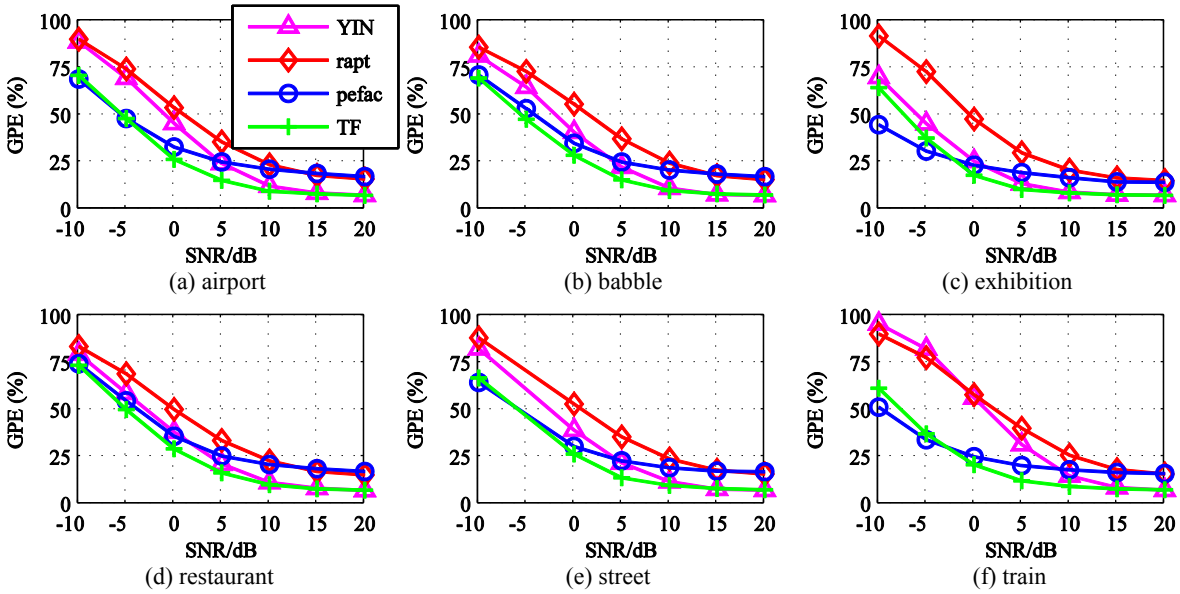

Fig. 5 GPE results for CSTR database

-parametric F0 estimation algorithms are used for performance comparison: RAPT [5], YIN [8] and PEFAC [13]. Our algorithm is denoted as "TF". Both the proposed and the reference algorithms do not require any prior voiced/unvoiced decision. Global pitch error (GPE) is used as the evaluation metric, which defines that the estimated pitch out of 5% of the ground truth value is considered as incorrect [13]. Fig. 4 and Fig. 5 show the GPE results for the Keele and CSTR database respectively. From Fig. 4 and Fig. 5 we can see that our proposed algorithm outperforms the reference algorithms in most of the noise conditions. However, there are still several noise scenes (e.g., exhibition, train) at low SNR levels (-10dB) where PEFAC performs better than the proposed algorithm. The reason is probably that in low SNR levels, fewer speech dominated TF segments stand out over noise, which brings down the pitch candidate contour estimation accuracy. In this case, a full band spectrum with enough redundancy is preferable for pitch estimation. When the SNRs are above 0dB, our algorithm is comparable with all of the reference methods.

6. CONCLUSIONS

We presented a study on noise-robust F0 estimation by exploring the temporal harmonic structures in local TF segments. First, a series of F0 candidate contours are estimated from different TF segments. Second, F0 tracking is performed across the overall TF plane to select the best F0. The speech dominated TF segments have a better SNR level than full band signal. And hence the harmonic structures in these high SNR TF segments provide a more reliable source for F0 estimation in noise. Experiments and results have shown that our algorithm substantially outperforms the compared state-of-the-art methods in terms of pitch estimation accuracy.

8. REFERENCES

[1] M. Asgari, A. Bayestehtash, I. Shafran, "Robust and accurate features for detecting and diagnosing autism spectrum disorders". In: Proc. INTERSPEECH. Lyon, France, pp. 191–194, 2013.

[2] Y. Yang, C. Fairbirn, J. F. Cohn, "Detecting depression severity from vocal prosody". IEEE Trans. Audio Speech Lang. Process. , vol. 4, no. 2, 142– 150, 2013.

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

[4] H. Duifhuis, L. F. Willems, R. J. Sluyter, "Measurement of pitch in speech: An implementation of goldsteins theory of pitch perception,". J. Acoust. Soc. Am., vol. 71, no. 6, pp. 1568 -1580, 1982.

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

[6] Y. Gong, J. Haton, "Time domain harmonic matching pitch estimation using time-dependent speech modeling, " IEEE Trans. Acoust., Speech, Signal, Process., vol. ASSP-35, no. 10, pp. 1386-1400, Oct. 1987.

[7] W. Hess, Pitch Determination of Speech Signals. Spring - Verlag, Berlin, Germany, 1983.

[8] A. Cheveigne, H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., vol. 111, no. 4, pp. 1917-1930, 2002.

[9] T. Shimamura, H. Kobayashi, "Weighted autocorrelation for pitch extraction of noisy speech, "IEEE Trans. Speech, Audio Processing, vol. 9, no. 7, pp. 727-730, Oct. 2001.

[10] F. Huang, T. Lee, "Pitch estimation in noisy speech using accumulated peak spectrum and sparse estimation technique," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 21, no. 1, pp. 99-109, 2013.

[11] D. Liu, C. Lin, "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," IEEE Trans. Acoust., Speech, Signal, Process., vol. 9, no. 6, pp. 609-621, Sep. 2001.

[12] J. L. Roux, H. Kameoka, N. Ono, A. Cheveigne, S. Sagayamma, "Single and multiple F0 contour estimation through parametrix spectrogram modeling of speech in noisy environments," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 15, no. 4, pp. 1135–1145, 2007.

[13] S. Gonzalez, M. Brookes, "PEFAC - A pitch estimation algorithm robust to high levels of noise," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 22, no. 2, pp. 518–530, 2014.

[14] H. Boril, P. Pollak, "Direct time domain fundamental frequency estimation of speech in noisy conditions," in Proc. Eurospeech, 2004, vol. 2, pp. 1003–1006.

[15] M. Wu, D. Wang, "A multipitch tracking algorithm for noisy speech," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 11, no. 3, pp. 229-241, 2003.

[16] B. S. Lee, D. P. W. Ellis, "Noise robust pitch tracking by subband autocorrelation classification," in Proc. Interspeech 2012, Sep. 2012, Portland.

[17] L. N. Tan, A. Alwan, "Multi-band summary correlogram-based pitch detection for noisy speech," Speech Communication vol. 55, pp. 841–856, 2013.

[18] M. Mauch, S. Dixon, "PYIN: A fundamental frequency estimator using probabilistic threshold distributions," ICASSP 2014, May, 2014, Florence.

[19] W. Chu, A. Alwan, "SAFE: A statistical approach to F0 estimation under clean and noisy conditions," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 20, no. 3, pp. 933–944, 2012.

[20] K. Han, D. Wang, "Neural network based pitch tracking in very noisy speech," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 22, no. 12, pp. 2158–2168, 2014.

[21] E. Terhardt, "Calculating virtual pitch," Hearing Research, vol. 1, pp. 155-182, 1979.

[22] D. Wang, P. C. Loizou, J. H. L. Hansen, "F0 estimation in noisy speech based on long-term harmonic feature analysis combined with neural network classification," in Proc Interspeech 2014, Sep. 2014, Singapore.

[23] M. Cooke, "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am., vol. 119, no. 3, pp. 1562-1573, 2005.

[24] Q. Huang, D. Wang, "Single channel speech separation based on long-short frame associated harmonic model," Digital Signal Processing, vol. 21, pp. 497-507, Mar., 2011.

[25] F. Plante, G. Meyer, and W. A. Ainsworth, "A pitch extraction reference database," in Proc. Eurospeech, 1995, pp. 837–840.

[26] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching," in Proc. Eurospeech, 1993, vol. 2, pp. 1003–1006.

[27] E. Terhardt, "Pitch, consonance, and harmony," J. Acoust. Soc. Am., vol. 55, pp. 1061-1069, 1974.

[28] E. Terhardt, G. Stoll, M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals," J. Acoust. Soc. Am., vol. 71, pp. 679-688, 1982.

[29] M. Gainza, B. Lawlor, E. Coyle, "Multi pitch estimation by using modified IIR comb filters," in Proc. International Symposium focused on Multimedia Systems and Applications (ELMAR), Zadar, 2005.

[30] D. Wang, J. H. L. Hansen, E. Tobey, "F0 estimation for noisy speech based on exploring local time frequency segment," in Proc. WASPAA-2015, Oct. 2015, New Paltz.