F0 estimation in noisy speech based on long-term Harmonic Feature Analysis combined with Neural Network Classification

Dongmei Wang, Philipos C. Loizou, John H. L. Hansen

Dept. of Electrical Engineering, University of Texas at Dallas
{dongmei.wang, john.hansen}@utdallas.edu

Abstract

In this study, we propose a frequency domain F0 estimation approach based on long term Harmonic Feature Analysis combined with artificial neural network (ANN) classification. Long term spectrum analysis is proposed to gain better harmonic resolution, which reduces the spectrum interference between speech and noise. Next pitch candidates are extracted for each frame from the long term spectrum. Five specific features related to harmonic structure are computed for each candidate and combined together as a feature vector to indicate the status of each candidate. An ANN is trained to model the relation between the harmonic features and the true pitch values. In the test phase, target pitch is selected from the candidates according to the maximum output score from the ANN. Finally, post-processing is applied based on average segmental output to eliminate inconsistent or fluctuating decision errors. Experimental results show that the proposed algorithm outperforms several state-of-the-art methods for F0 estimation under adverse conditions, including white noise and multi-speaker babble.

Index Terms: pitch estimation, fundamental frequency, long term harmonic feature, artificial neural network,

1. Introduction

Fundamental frequency (F0) estimation plays an important role in the field of speech/audio signal processing such as speech recognition, speaker identification, music transcription, speech separation, and speech enhancement, etc. Pitch estimation within a complex auditory scene remains a challenging problem and present day solutions/ algorithms are far from satisfactory.

A number of algorithms that address pitch estimation in noisy speech have been reported [1], [2], [3]. It is known that the difficulty of estimating pitch for noisy speech is due to the fact that both periodic cues in the time domain and harmonic cues in the frequency domain are corrupted [4], [5]. Most methods regarding pitch estimation can be classified into three categories, including (i) time domain, (ii) frequency domain and (iii) time-frequency domain methods. Among the time domain methods, YIN is based on the autocorrelation function (ACF) to improve pitch estimation accuracy [6]. The ACF and average magnitude difference function (AMDF) are also jointly used to estimate pitch [7], where the ACF is weighted by the reciprocal of the AMDF. Also, the RAPT algorithm was proposed to generate pitch candidates based on the cross correlation function over voiced speech [8].

Frequency domain pitch estimation methods are primarily based on the harmonic model. Maximum a-posteriori (MAP) probability has been proposed to track the time-varying harmonic parameters to estimate pitch [9]. A statistical approach for F0 estimation named SAFE was proposed by Chu & Alwan [10], which performs well in both clean and noisy environments. In their study, F0 is inferred from the prominent SNR peaks in the speech spectra within a probabilistic framework. Moreover, Huang & Lee proposed a pitch estimation method by using the accumulated peak spectrum and a sparse estimation technique [11]. Their study showed that the accumulated peak spectrum is a robust representation of the pitch harmonics in noise which improves overall pitch estimation accuracy. Furthermore, pitch-scaled harmonic filtering has also been proposed to separate the harmonic and non-harmonic components of the speech signal, which can help in estimation of the prominent pitch for the target speech [12].

Time-frequency associated methods have also been investigated for pitch estimation. One example is based on computational auditory scene analysis (CASAS) [14]. A tandem algorithm [13] was proposed to jointly estimate pitch for the target speech and segregate the target speech simultaneously. In [15], a pitch estimation method was proposed based on the harmonic sinusoidal autocorrelation model and a time-domain matching scheme. Time domain harmonic matching methods [16], [17], [18] have also been investigated for pitch estimation.

Among the aforementioned approaches, most perform well for clean speech, but sensitive and quickly degrade for the noisy speech. One reason stems from the fact that both temporal and spectral features related to pitch are affected by noise in the short term analysis window. In order to obtain noise robust features for pitch estimation, we propose an alternative long term harmonic feature analysis method.

The remainder of this paper is organized as follows. Sec. 2 presents a high level overview of the complete algorithm. Sec. 3 describes the algorithm for pitch candidate extraction. Sec. 4 describes the final pitch selection method. Experiments and results are presented in Sec. 5. Finally, in Sec. 6 conclusions and future directions are presented.
The proposed pitch estimation solution will be performed in the frequency domain, which aims to estimate the target pitch from the observed noisy spectrum. Fig. 1 presents the flow diagram of the proposed algorithm. From Fig. 1, we see that our algorithm consists of two steps: 1) pitch candidate extraction, 2) target pitch selection. One difference for our method from traditional schemes is that long term spectrum analysis is proposed to obtain better frequency resolution for the noisy speech. Spectrum compression and summation is performed on the long term spectrum. Pitch candidates are then extracted from both the original long term spectrum and the corresponding compressed and summed (CS) spectrum. The next step is to calculate the harmonic feature vector for each candidate. Finally, an ANN model is used to initially estimate the target pitch from the pitch candidate list. Post-processing is also performed to eliminate any inconsistent pitch estimation errors.

### 3. Pitch Candidate Extraction

In this stage, pitch candidates are extracted from the noisy speech spectrum in each frame. Long term (80 ms) spectrum analysis is proposed to obtain higher harmonic resolution in the presence of noise. This can be deduced from the Discrete Fourier transform (DFT) with a resolution $\Delta f = 1/T$, (where $T$ is the frame length), as well as the spectral smearing introduced by the main-lobe width and side-lobes of the analyzing window. When the analysis window is longer, more precise harmonic resolution will be obtained. In addition, the long term noisy spectrum is compressed along the frequency axis with a series of integer factors and then summed to generate the CS spectrum [3]. Pitch candidates are then extracted from both the original and CS spectra.

### 4. Target Pitch Selection

In this section, we propose to select the target pitch from the pitch candidate list based on specific harmonic features and an artificial neural network. Five harmonic features are calculated for each $F_0$ candidate. They collectively determine the potential likelihood of each candidate. As each of these features may be affected differently by the noise, these effect can partially compensate for each other in feature fusion. The features may be affected differently by the noise, these effect can partially compensate for each other in feature fusion. The contribution weight of each feature for candidate selection should be estimated in a cooperative manner. Here, an ANN provides a practical method for obtaining the learning function from the examples [22]. In our case, an ANN is trained to model the relation between the input harmonic features and the ground truth pitch value. Once training is completed, the ANN model is ready to perform target pitch selection for unseen test data. The backpropagation algorithm is used to learn the

![Fig. 2. An example of harmonic resolution compared between (a) short-frame spectrum, versus (b) long-frame spectrum](image1)

![Fig. 3. CDF of frequency difference within 80ms](image2)
weights for the ANN model. The ANN algorithm block diagram is presented in Fig. 4.

4.1. Long term harmonic feature set extraction

In order to help decide which pitch candidate should be chosen from the candidate list, we characterize the property of each candidate using five harmonic features. This feature set works together to decide the accuracy of each pitch candidate. In addition, these features are extracted from the long term spectrum to obtain more precise frequency resolution [19]. The five harmonic features are listed as follows:

i) \( H_d \): Harmonic deviation. This feature describes the average frequency deviation of the detected harmonic from the ideal harmonic frequency \( kF_0 \) in a harmonic group. Usually, the true pitch candidate leads to a smaller \( H_d \) when the target speech signal has prominent energy.

ii) \( E_r \): The ratio of the detected harmonic energy over the entire spectrum energy. The harmonic spectrum for each \( F_0 \) candidate is generated by multiplying the harmonic peak amplitude vector with the spectrum of the short time Hamming window (30 ms). Here, \( E_r \) is calculated as,

\[
E_r = \frac{\sum_{k=1}^{N} S_d(k)}{\sum_{k=1}^{N} S_y(k)}
\]

(4)

where \( S_d(k) \) is the generated harmonic spectrum, and \( S_y(k) \) is the original noisy spectrum. The value of \( E_r \) is dependent on the SNR level and the specific pitch candidate in each frame.

iii) \( O_2e \): The energy ratio between odd harmonics and even harmonics. This feature measures the ratio between the detected odd harmonic and even harmonic energies:

\[
O_2e = \frac{\sum_{k=1}^{N} S_d(k) \cdot \text{odd}}{\sum_{k=1}^{N} S_y(k) \cdot \text{even}}
\]

(5)

where \( S_d(k) \) is the generated harmonic spectrum.

The \( O_2e \) is able to effectively control half-pitch errors. For the speech signal, the \( O_2e \) value should be in a specific range. That is to say, the odd order harmonic energy should be similar to that of the even order harmonics. If the \( O_2e \) is very low (which means the odd harmonic energy is very low compared to the even harmonic energy), it usually indicates that the corresponding candidate is a half-pitch error. In this case, the noise spectrum peaks at the half-pitch-harmonic position were mistaken for harmonics.

iv) \( R_h \): The ratio between the number of detected harmonics and total number of harmonic order in a specific frequency range (e.g. 0–4000 Hz). When the speech signal has prominent energy over the noise, \( R_h \) should be close to 1.

v) \( R_c \): The ratio between the CS spectrum amplitude at the \( F_0 \) candidate position and the maximum CS amplitude value. As noted in Sec. 2, for clean speech the frequency of the maximum compressed spectrum peak is usually due to the true pitch frequency. However, when speech is contaminated by noise, the maximum peak may not occur at the exact fundamental frequency position. In fact, the maximum compressed spectrum peak may appear at a frequency corresponding to half or one third of the true pitch value. Thus, \( R_c \) is not always equal to 1, but a true pitch value usually has a relatively higher \( R_c \) value compared with other candidates.

The above five features are combined together to form the input feature vector \((H_d \ E_r \ O_2e \ R_h \ R_c)\) for the ANN model designated in the following step.

4.2. Neural network setting

Considering an ANN structure, three layers of sigmoid units are sufficient to express a rich variety of target functions [20]. In our algorithm, we use three layers with sigmoid activation functions and one input layer with a linear activation function. We set 10 linear units in the input layer, 6 sigmoid units in the first hidden layer, 5 sigmoid units in the second hidden layer, and one output neuron.

After establishing the network structure, we train the neural network with prepared training data to obtain the optimal weight values. Once the training step is completed, we proceed with testing. During the test step, in each frame, the candidate with the maximum output value is selected as the best candidate and the initial estimated pitch.

The output of the ANN indicates the status of the input pitch candidate as true or false. The status is defined based on the difference between ground truth and the corresponding candidate, where for the difference within 20% of the ground truth value, the status is set as true, otherwise it is false. In order to accommodate for properties of the sigmoid activation function and backpropagation algorithm, we map the true status to 0.99 and false to 0.01 (rather than 1 and 0).

The backpropagation algorithm [22] learns the weights for a multilayer network, given a network with a fixed set of units and interconnections.

4.3. Pitch tracking

An overview of pitch tracking is presented in Fig. 5. Pitch tracking is performed to increase accuracy. The objective of our algorithm is to first detect the major pitch frequency range (MPFR) from the initial estimated pitch, and then re-estimate the pitch within that detected frequency range. The MPFR of a particular utterance is obtained by calculating which frequency band has the maximum summation of the ANN output value.
The re-estimated pitch is obtained by selecting the entry with the maximum ANN output value within the MPFR. Pitch tracking is performed on re-estimated pitch based on continuity in time. The average ANN output value is computed for each pitch track, and the one with maximum average output value is chosen as the final estimated pitch.

5. Experiments and Results

5.1. Evaluation database

In order to evaluate performance of the proposed algorithm, we use the Keele [20] and CSTR database [23]. Both databases provide ground truth pitch labels, which can be used as reference for performance assessment. Two types of noise are used to simulate the noisy environment, including white noise and babble noise from the NOISEX 92 database [24]. The noisy environment is simulated by adding noise into the speech signal using FaNT tool [25].

5.2. Parameter settings

All noisy speech signals were resampled into 16 kHz. Each sentence is decomposed into overlapping frames with a long frame length of 80ms, and a sequential overlap of 10ms. A 16000 point FFT is applied to calculate the frequency spectrum which generates a frequency resolution of 1Hz/1(FFT point). In the ANN training phase, we use Keele sentences for training, and CSTR sentences for testing.

5.3. Experimental results

Two different evaluation metrics are used to determine the final pitch estimation performance, including gross pitch error (GPE), and standard deviation of the fine pitch error (SDFPE) [1]. The GPE results are presented in Fig. 6. We denote the proposed method as “LSFANN” in this section. From Fig. 6 we see that LSFANN outperforms all of the comparing algorithms in terms of GPE result in the condition of babble noise. However, in the white noise condition, one of the latest developed algorithms SAFE has an overall best performance among all comparing methods. At the SNR level of -5dB, LSFANN has about 6% higher GPE than SAFE, and at 0dB, LSFANN has about 2% higher GPE than SAFE. From 5 dB to 20 dB of white noise condition, our algorithm performs as well as SAFE.

The SDFPE results are presented in Tables 1. SDFPE measures the absolute deviation of the correctly estimated pitch from the ground truth value. From Table 1 shows that our algorithm is competitive to all other compared methods at all SNR level in the condition of white noise. In addition, our results are most competitive versus almost all other results, except for the isolated condition of -5dB and 0 dB.

6. Conclusions

In this study, an approach for noise robust pitch estimation was proposed based on a set of long term harmonic features combined with artificial neural network classification. Our method consisted of two steps, where pitch candidate are first extracted, followed by final pitch selection. The long term spectrum analysis used here increases the harmonic resolution for the target speech. Pitch candidates are extracted from noisy spectrum peaks, as well as the CS spectrum peak. A five feature set related to harmonic characteristics were calculated for each pitch candidate. These harmonic features described the frequency factor and energy factor of each potential harmonic group generated by each pitch candidate. ANN is used to decide the weight of each harmonic feature for target pitch selection, which correlates all five features together to achieve a composite final decision. Segmental based post processing helps track more accurate pitch contours. Experimental results demonstrated that our proposed LSFANN algorithm yielded substantially better performance (lower GPE) than other state-of-the-art algorithms (e.g., RAPT, YIN) in low SNR and also a latest sophisticatedly developed method (SAFE).

7. Acknowledgements

Research supported by Grant No. R01 DC010494 from NIDCD/NIH1.
8. References


