International Journal of Conceptions on Electronics and Communication Engineering
Vol. 3, Issue. 1, April' 2015; ISSN: 2357 - 2809

Hybrid High Noise resiliency Pitch Detection Algoritm

Harish Yedla, R Raja Kishore and Dr. M Narsing Yadav
Assistant Professor, Dept. of Electronics & Communication
Malla Reddy Institute of Engineering & Technology
Hyderabad, India
harishyedla@gmail.com, and rajakishore.r@gmail.com

Abstract— **Pitch is one of the essential features in many speech related applications. A pitch detection algorithm (PDA) is an algorithm designed to estimate the pitch or fundamental frequency of a quasiperiodic or virtually periodic signal, usually a digital recording of speech or a musical note or tone. This can be done in the time domain or the frequency domain or both the two domains. Although numerous pitch detection algorithms have been developed, as shown in this paper, the detection ratio in noisy environments still needs improvement. In this paper, we present a hybrid noise resilient pitch detection algorithm named BaNa that combines the approaches of harmonic ratios and Cepstrum analysis. A Viterbi algorithm with a cost function is used to identify the pitch value among several pitch candidates. We use an online speech database along with a noise database to evaluate the accuracy of the BaNa algorithm and several stateof-the-art pitch detection algorithms. Results show that for all types of noises and SNR values investigated, BaNa achieves the best pitch detection accuracy. Moreover, the BaNa algorithm is shown to achieve around 80% pitch detection ratio at 0dB signalto-noise ratio (SNR).**

Keywords- Pitch detection, noise resilience, harmonics, Viterbi algorithm

I. INTRODUCTION

A pitch detection algorithm (PDA) is an algorithm designed to estimate the pitch or fundamental frequency of a quasiperiodic or virtually periodic signal, usually a digital recording of speech or a musical note or tone. This can be done in the time domain or the frequency domain or both the two domains.

PDAs are used in various contexts (e.g. phonetics, music information retrieval, speech coding, musical performance systems) and so there may be different demands placed upon the algorithm. There is as yet no single ideal PDA, so a variety of algorithms exist, most falling broadly into the classes given below.[1]

In the time domain, a PDA typically estimates the period of a quasiperiodic signal, then inverts that value to give the frequency.

One simple approach would be to measure the distance between zero crossing points of the signal (i.e. the Zero-crossing rate). However, this does not work well with complex waveforms which are composed of multiple sine waves with differing periods. Nevertheless, there are cases in which zero-crossing can be a useful measure, e.g. in some speech applications where a single source is assumed. The algorithm's simplicity makes it "cheap" to implement.

More sophisticated approaches compare segments of the signal with other segments offset by a trial period to find a match. AMDF (average magnitude difference function), ASMDF (Average Squared Mean Difference Function), and other similar autocorrelation algorithms work this way. These algorithms can give quite accurate results for highly periodic signals. However, they have false detection problems (often "*octave errors*"), can sometimes cope badly with noisy signals (depending on the implementation), and - in their basic implementations - do not deal well with polyphonic sounds (which involve multiple musical notes of different pitches).

Current time-domain pitch detector algorithms tend to build upon the basic methods mentioned above, with additional refinements to bring the performance more in line with a human assessment of pitch. For example, the YIN algorithm and the MPM algorithm are both based upon autocorrelation.

In the frequency domain, polyphonic detection is possible, usually utilizing the periodogram to convert the signal to an estimate of the frequency spectrum . This requires more processing power as the desired accuracy increases, although the well-known efficiency of the FFT, a key part of the periodogram algorithm, makes it suitably efficient for many purposes.

Popular frequency domain algorithms include: the harmonic product spectrum cepstral analysis and maximum likelihood which attempts to match the frequency domain characteristics to pre-defined frequency maps (useful for detecting pitch of fixed tuning instruments); and the detection of peaks due to harmonic series.

To improve on the pitch estimate derived from the discrete Fourier spectrum, techniques such as spectral reassignment (phase based) or Grandke interpolation (magnitude based) can be used to go beyond the precision provided by the FFT analysis. Another phase-based approach is offered by Brown and Puckette.

Spectral/temporal pitch detection algorithms, e.g. the YAAPT pitch tracking, are based upon a combination of time domain processing using an autocorrelation function such as normalized cross correlation, and frequency domain processing utilizing spectral information to identify the pitch. Then, among the candidates estimated from the two domains, a final pitch track can be computed using dynamic programming. The advantage of these approaches is that the tracking error in one domain can be reduced by the process in the other domain.

The fundamental frequency of speech can vary from 40 Hz for low-pitched male voices to 600 Hz for children or high-pitched female voices.

Autocorrelation methods need at least two pitch periods to detect pitch. This means that in order to detect a fundamental frequency of 40 Hz, at least 50 milliseconds (ms) of the speech signal must be analyzed. However, during 50 ms, speech with higher fundamental frequencies may not necessarily have the same fundamental frequency throughout the window. Subjective pitch is defined by the relative highness or lowness of a tone as perceived by the human ear, and is caused by vibrations of the vocal cords. For perfectly periodic speech signals, pitch is the same as fundamental frequency (F0), which is the inverse of the speech signal's largest period. However, due to the aperiodicity of the glottal vibration itself and the movement of the vocal tract that filters the source signal, human speech is not perfectly periodic, making the detection of speech pitch rather difficult. Therefore, pitch detection has always been an important challenge of speech signal analysis. Among the modern state-of-the-art pitch detection algorithms, YIN [1] and Praat [2] are based on the well-known autocorrelation method in the time domain, while the Cepstrum method [3] [4] and Harmonic Product Spectrum (HPS) [5] are based on the spectrum in the frequency domain. YIN uses a difference function to search for the period, and further refines the pitch detection result by two error-reduction steps. Praat, on the other hand, considers the maxima of the autocorrelation of a short segment of the sound as pitch candidates, and chooses the best pitch candidate for each segment by finding the least cost path through all the segments using the Viterbi algorithm. Cepstrum is found by taking the Fourier transform of the log-magnitude Fourier spectrum, which shows a peak corresponding to the period in frequency. HPS multiplies the original signal with downsampled signals, to line up the peak at the pitch value for isolation. A variety of applications can benefit from a more precise and robust pitch detection algorithm. For example, pitch detection is essential in speech recognition, where homophones can be differentiated by recognizing tones [6]. Also, music notation programs use pitch detection to automatically transcribe real performances into scores [7]. Moreover, in emotion detection or other affective measurement, it has been found that prosodic variations in speech are closely related to one's emotional state, and the pitch information is crucial to identification of this state. Some health care providers and researchers even put pitch detectors and other behavior sensing technologies on mobile devices, such as smart phones, for patient monitoring or behavioral studies.

When performing pitch detection in real scenarios, the quality of the input speech signal may be greatly degraded, due to noise introduced by the recording devices or audible background noise. As existing pitch detectors do not perform well for noisy input data, we are motivated to design a noise resilient pitch detection algorithm that is better suited for practical uses. In this paper, we propose a hybrid pitch detection algorithm named BaNa, which combines the idea of using the ratios of harmonic frequencies and the Cepstrum approach to find the pitch from a noisy signal. We test our BaNa algorithm on real human speech samples corrupted by various types of realistic noise. Evaluations show the high noise resiliency of BaNa compared to the state-of-the-art pitch detection algorithms.

II. PROPOSED WORK

A. Preprocessing

Given a digital audio signal, preprocessing is performed before the extraction of the pitch values. In the BaNa algorithm, we filter the speech signal with a bandpass filter. Since human speech is normally higher than 50 Hz, and lower than 600 Hz, the lower bound of the bandpass filter is set to 50 Hz. The upper bound is set to 3000 Hz, which is 5 times the normal range of human speech at 600 Hz, in order to capture enough harmonics that will later be used for pitch detection.

B. Determination of the pitch candidates

Since harmonics are regularly spaced at integer multiple of the fundamental frequency F0 in the frequency domain, we use this characteristic of the speech in the proposed BaNa algorithm to achieve the noise resiliency. If we know the frequency of a harmonic and its ratio to F0, then F0 can be easily obtained. However, even if a harmonic is discovered, its ratio to F0 is unknown. Therefore, we propose a pitch detection algorithm that looks for the ratios of potential harmonics and finds the pitch based on them by applying the following steps.

Step 1: Search for harmonic peaks

Spectral peaks with high amplitudes and low frequencies are preferred to be considered for pitch candidates, since peaks with high amplitudes are less likely to be caused by noise, and peaks with low frequencies are easier to be identified to be harmonics by calculating the ratios. Therefore, we consider the five peaks higher than a certain threshold and with the lowest frequencies to derive pitch candidates.

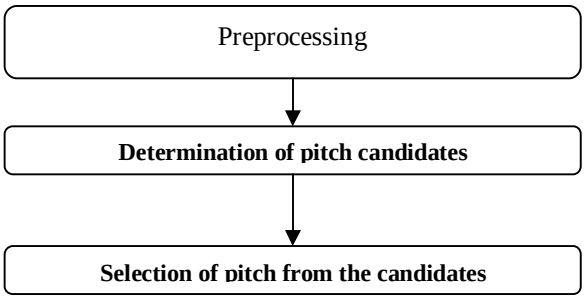
Step 2: Calculate pitch candidates

Note that due to the imperfect periodicity of human speech, the harmonics may not be exactly on integer multiples of F0, and we observed that higher order harmonics have even larger drift than lower order harmonics in practice. Therefore, we set a smaller ratio tolerance range for lower order harmonics, and we set a larger ratio tolerance range for higher order harmonics. In total, C2^5 = 10 ratio values are calculated between every pair of ^ F. Since both ratios of F1/F0 and F3/F1 are equal to 2, it is not trivial to differentiate to which harmonics this ratio belongs. In our algorithm, we assume it belongs to F1/F0 and

calculate the pitch candidate based on that. In addition, we include the peak with the smallest frequency value as one of the pitch candidates, since we have noticed that in some cases only the F0 peak is high enough to be detected. In the BaNa algorithm, we also include the pitch value found by the Cepstrum method as an additional candidate to the ones derived by the harmonic ratio analysis. The reason is that the five spectral peaks we choose mainly belong to low frequency values. For some rare cases, the higher order harmonics (e.g., 5th to 10th harmonics) are found to yield higher spectral peak values compared to the low order harmonics. In that case, the spectral peaks at low frequencies are more vulnerable to noise. However, since cepstrum depicts the global periodicity of the spectrum, and considers all spectral peaks, it can help to detect the pitch in those rare cases. In Section III, we show the benefit of including the detected pitch from cepstrum as a candidate.

C. Selection of the pitch from the candidates

In II-B, the distinctive candidates of each frame are obtained independently. However, the pitch values of neighboring frames may correlate, since the pitch values of human speech exhibit a slow time variation, and hence, large pitch jumps among subsequent frames are rare. Therefore, we use the Viterbi algorithm [11] for post-processing to go through all the candidates in order to correct pitch detection errors. We aim to find a path that minimizes the total cost. The cost consists of two parts: the frequency jumps between the candidates of two consecutive frames, and the inverse of the confidence score of each distinctive candidate.



Let ~ Fn i denote the ith pitch candidate of frame n, and let ~ Fn+1 j denote the jth pitch candidate of the next frame. Let Nframe denote the number of frames in the speech segment. For every frame n, pn is the index of the chosen candidate. Thus, fpnj1 _ n _ Nframeg defines a path through the candidates. For each path, the path cost is defined to be

$$PathCost\left(\{p_n\}\right) = \sum_{n=1}^{N_{frame}-1} Cost\left(\tilde{F}_i^n, \tilde{F}_j^{n+1}\right),$$

where Cost is used to calculate the cost of adjacent frames. We define the function Cost by using the pitch differences between the adjacent frames and the confidence score of the candidates. Since the perceived pitch difference has a logarithm relation with frequency difference, as defined by the Mel scale for pitch, we also model that in the cost function. The larger the pitch difference, the higher the Cost should be. Also, we should assign a lower cost to candidates with higher confidence score, thus we use the inverse of the confidence score in the expression of the cost. A weight w is introduced to balance the two parts. We set its value to 0.2 as determined by tests. Then, Cost is defined mathematically as

$$Cost\left(\tilde{F}_i^n, \tilde{F}_j^{n+1}\right) = \left|\log_2 \frac{\tilde{F}_i^n}{\tilde{F}_j^{n+1}}\right| + w \times \frac{1}{V_i^n}.$$

We use the Viterbi algorithm to find the minimum cost path, i.e., the path that reduces the pitch jumps the most while giving priority to the pitch candidates with higher confidence scores.

The optimal path is found for each voiced part in the speech. Whenever the Viterbi algorithm meets an unvoiced part or irregularly voiced portion of the speech (diplophony, creak), the path cost is reset to 0 and the Viterbi algorithm starts all over again from the next voiced part.

III. RESULTS

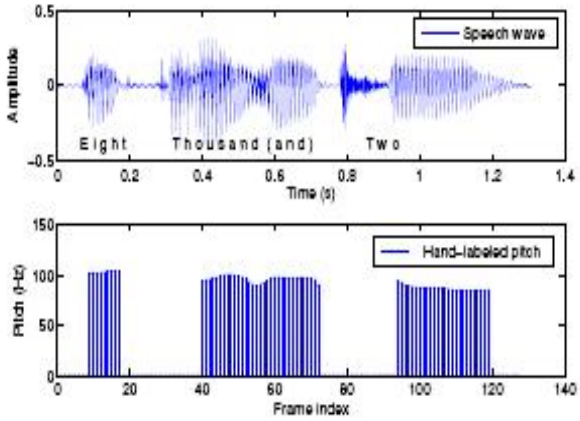


Fig. 1: Speech waveform and hand-labeled pitch values.

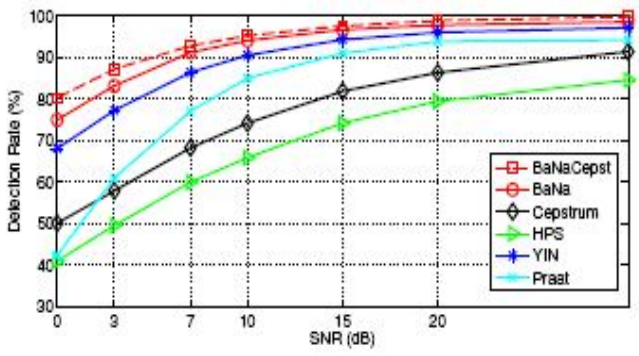


Fig. 2: Accuracy of different algorithms, averaged over all 8 types of noise.BaNa-NoCepst refers to the BaNa algorithm without Cepstrum as a pitchcandidate. Inf represents the clean speech with no added noise.

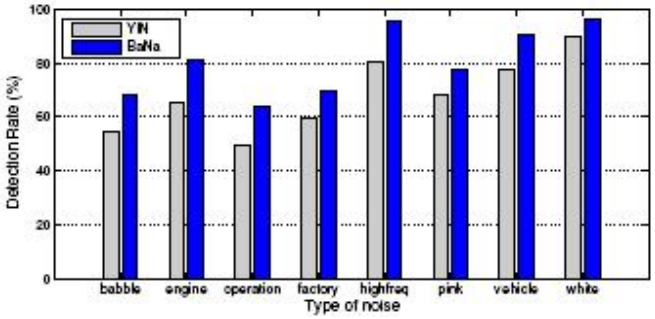


Fig. 3: Accuracy of BaNa and YIN for 8 types of noise at 0dB

IV. CONCLUSION

In this paper, we presented BaNa, a noise resilient hybrid pitch detection algorithm. BaNa was designed to detect pitch in a noisy environment, for example on a mobile phone.This would enable the wide deployment of voice-based applications, such as the ones that use emotion detection. We were able to show that BaNa achieves the best detection rate, among all the algorithms investigated from the literature, for different types of background noise, and under different SNR levels from 0dB to 20dB. Even for the very noisy scenario of 0dB SNR, BaNa can still correctly detect about 80% of the pitch values, outperforming the most competitive state-of-theart reference algorithm YIN by 12%.

REFERENCES

[1] A. de Cheveigné and H. Kawahara. YIN, a fundamental frequency estimator for speech and music. The Journal of the Acoustical Society of America, 111:1917, 2002. doi:10.1121/1.1458024

[2] P. McLeod and G. Wyvill. A smarter way to find pitch. In Proceedings of the International Computer Music Conference (ICMC'05), 2005.

[3] Hayes, Monson (1996). *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, Inc. p. 393. ISBN 0-471-59431-8.

[4] A. Michael Noll, "Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate," Proceedings of the Symposium on Computer Processing in Communications, Vol. XIX, Polytechnic Press: Brooklyn, New York, (1970), pp. 779-797.

[5] Michael Noll, "Cepstrum Pitch Determination," Journal of the Acoustical Society of America, Vol. 41, No. 2, (February 1967), pp. 293-309.

[6] Mitre, Adriano; Queiroz, Marcelo; Faria, Régis. Accurate and Efficient Fundamental Frequency Determination from Precise Partial Estimates. Proceedings of the 4th AES Brazil Conference. 113-118, 2006.

[7] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound," in Proceedings of the Institute of Phonetic Sciences 17, 1993, pp. 97–110.