Determination of Pitch Range Based on Onset and Offset Analysis in Modulation Frequency Domain

A. Mahmoodzadeh
Speech Proc. Research Lab
ECE Dept.
Yazd University
Yazd, Iran

H. R. Abutalebi
Speech Proc. Research Lab
ECE Dept.
Yazd University
Yazd, Iran &
Idiap Research Institute[1]
Martigny, Switzerland

H. Soltanian-Zadeh
Control and Intelligent Processing Center of Excellence,
University of Tehran
Tehran, Iran &
Image Analysis Lab.
Henry Ford Health System
Detroit, USA

H. Sheikhzadeh
EE Dept.
Amirkabir University of Technology
Tehran, Iran

***Abstract*—Auditory scene in a natural environment contains multiple sources. Auditory scene analysis (ASA) is the process in which the auditory system segregates a scene into streams corresponding to different sources. The determination of range of pitch frequency is necessary for segmentation. We propose a system to determine the range of pitch frequency by analyzing onsets and offsets in modulation frequency domain. In the proposed system, first the modulation spectrum of speech is calculated and then, in each subband onsets and offsets will be detected. Thereafter, the segments are generated by matching corresponding onset and offset front. Finally, by choosing the desired segments, the rage of pitch frequency is determined. Systematic evaluation shows that the range of pitch frequency is estimated with good accuracy.**

Keywords- pitch frequency; onset/offset algorithm; modulation frequency domain

I. INTRODUCTION

The pitch is defined as the fundamental frequency of quasi-periodic or voiced sounds [1]. In the speech signals, the pitch is produced by vibrations of the vocal cords. Pitch determination is a fundamental problem that attracts much attention in speech analysis. A robust pitch detection algorithm (PDA) is needed for many applications including computational auditory scene analysis (CASA), prosody analysis, speech enhancement/separation [2], speech recognition, and speaker identification [3].

Various methods have been proposed for the determination of the pitch frequency. These methods are generally classified into three categories: time-domain, frequency-domain, and time-frequency domain algorithms. Time-domain PDAs directly examine the temporal structure of a signal waveform and estimates the period of the quasi-periodic signal. They use either the autocorrelation function [4],[5], other physical [6][7] and geometric [8] criteria, least-square fitting [9], pattern recognition [10], and neural networks [11]. Frequency domain PDAs utilize the harmonic structure in the short-term spectrum for distinguish the fundamental frequency. These methods include: the harmonic product spectrum; Cepstral analysis, and maximum likelihood. Time-frequency domain algorithms perform time-domain analysis on band-filtered signals obtained via a multichannel front-end.

In these methods, the estimation of pitch frequency is done by framing the speech signal and estimation of pitch frequency in each frame. Then by forming the pitch contour, the range of pitch frequency is determined. In each frame, if the pitch frequency changes, the pitch frequency will not be correctly estimated. Moreover, in these methods we face pitch halving and pitch doubling problems. Also, in a pitch curve, using interpolation for unvoiced regions, some values are applied. In addition, the existence of interference between noise and speech signal deteriorates the performance of such techniques. In recent years, multipitch methods are presented that have some complexities and problems.

In this paper, we propose a system to determine the range of pitch frequency by analyzing onsets and offsets in modulation frequency domain. The proposed method determines the range of pitch frequency of voiced speech without any windowing. At first, modulation spectrum of speech is calculated using the modulation transform. Then, using the onset and offset algorithm, the onset and offset fronts are detected. Thereafter, the segments are generated by matching the corresponding onset and offset fronts. Finally, by choosing the desired segments, the range of pitch frequency is determined. By extending this system, we can determine the range of the multi-pitch frequency of the two speakers. This, in turn, can be used for single channel speech separation.

The fundamental frequency of speech varies from 40 Hz for low-pitched male voices to 350 Hz for children or high-pitched female voices. The pitch frequency of everyone is not constant during the time; however it is bonded in a range. When the range of pitch frequency is known, it may help in single channel speech separation.

This paper is organized as follows. In Section II and III, we propose a working definition for modulation frequency

[1] H. R. Abutalebi has been on sabbatical at Idiap Research Institute during Fall 2010-Summer 2011.

analysis and onset and offset algorithm. In Section IV, we first give a brief description of our system and then present the details of each stage. The results of the system on the determination of range of pitch frequency are reported in Section V. The paper concludes with a discussion in Section VI.

II. Modulation Frequency Analysis

The general modulation frequency analysis framework consists of a filterbank (possibly decimated), followed by subband envelope detection and frequency analysis of the subband envelopes [12]. In its most straightforward form, the filterbank is implemented using the Short-Time Fourier Transform (STFT), envelope detection is defined as the magnitude or magnitude squared of the subband, and subband envelope frequency analysis is performed with the Fourier transform. For a discrete signal $x(n)$, the STFT can be expressed as

$$X_k(m) = \sum_{n=-\infty}^{\infty} h(mM - n)\, x(n) W_K^{kn} \quad \text{for } k = 0,\ldots,K-1 \tag{1}$$

and the envelope detection and modulation frequency analysis as

$$X_l(k,i) = \sum_{m=-\infty}^{\infty} g(lL - m)\left|X_k(m)\right| W_I^{im} \quad \text{for } i = 0,\ldots,I-1 \tag{2}$$

where $W_K = e^{-j(2\pi/K)}$. $h(n)$ and $g(m)$ are the acoustic and modulation frequency analysis windows, respectively. Throughout the paper, we use the shorthand notations

$$T\{x(n)\} = X_l(k,i) \tag{3}$$

and

$$T^{-1}\{X_l(k,i)\} = x(n) \tag{4}$$

to refer to the modulation frequency analysis and synthesis, respectively.

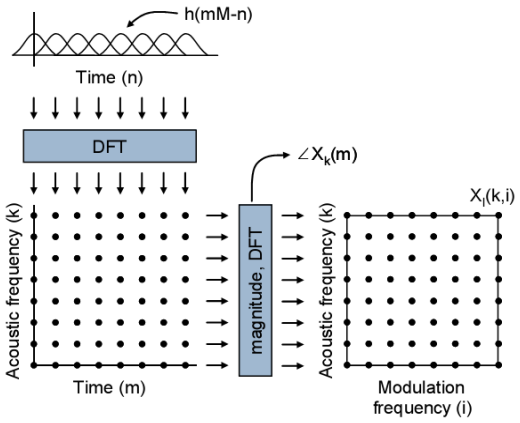

Figure 1. Modulation analysis framework and the modulation spectrogram [12].

The magnitude of the sub-band envelope spectra $\left|X_l(k,i)\right|$ is typically displayed in a modulation spectrogram representation. The vertical axis of this representation is regular acoustic frequency (K), and its horizontal axis is modulation frequency (i), (iii). Gray- or color-scale intensity in the joint acoustic/modulation plane represents modulation spectral energy. The modulation analysis framework is illustrated in Figure 1, and an example of a modulation spectrogram is shown in Figure 2.

III. Onset and Offset

Onsets and offsets correspond to sudden amplitude increases and decreases [13]. A standard way to identify such intensity changes is to take the first-order derivative of the intensity with respect to modulation frequency and then find the peaks and valleys of the derivative. Because of intrinsic intensity fluctuations, many peaks and valleys of the derivative do not correspond to actual onsets and offsets. To reduce such fluctuations, we smooth the intensity over modulation frequency, as is commonly done in edge detection for image analysis. Smoothing can be performed through either a diffusion process or low-pass filtering.

Onsets correspond to the peaks of the derivative above a certain threshold, and offsets are the valleys below a certain threshold. The purpose of thresholdding is to remove peaks and valleys corresponding to insignificant intensity fluctuations. The above procedure is similar to the standard Canny edge detector in image processing [14]. An example of the above procedure is shown in Figure 3.

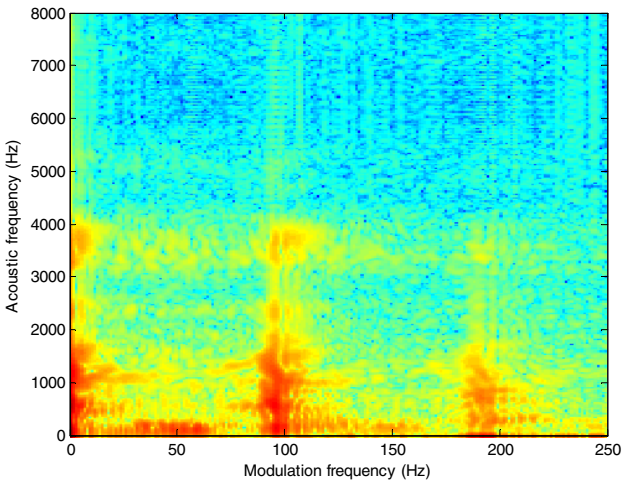

Figure 2. Modulation spectrogram of the male speaker.

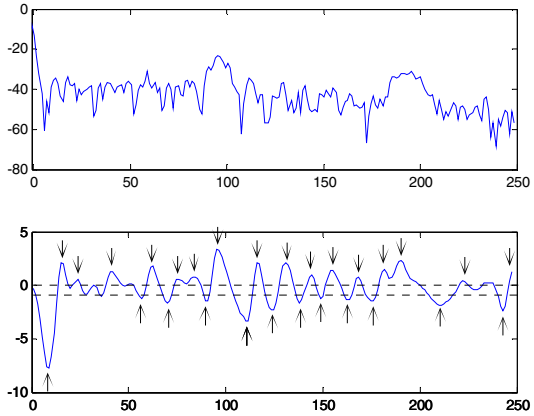

Figure 3. The upper panel shows the response intensity, and the lower panel shows the results of onset and offset detection using low-pass filter.

IV. System Description

This research estimates the range of the pitch frequency. The proposed system estimates this range via an analysis of signal in modulation frequency domain using onset and offset detection algorithm for one speaker. Although the proposed system is capable of determining the range of pitch frequency of one speaker, by its expansion we can present a new system for determination of multipitch frequency of speakers in one channel.

The place of pitch energy in modulation frequency spectrogram is an important feature for determination of the range of the pitch frequency of speech. Figure 4 shows a block diagram of our proposed system. In the first stage, the modulation spectrum of the speech signal is calculated. Then segments are generated in the modulation domain using the onset and offset algorithm and finally in the decision-making stage, the range of pitch frequency is determined. A detailed description of the stages is as follows:

A. Cochlear filtering and modulation transform

At first the spectrum of speech signal is calculated using cochlear filtering and STFT transform. Then, using the modulation transform, the modulation spectrum of speech signal is calculated. The vertical axis of modulation spectrum is acoustic frequency, and its horizontal axis is modulation frequency. Colour intensity in the joint acoustic/modulation plane represents modulation spectral energy.

B. Smoothing

Smoothing corresponds to low-pass filtering. Our system smoothes the intensity over modulation frequency with a low-pass filter. Let $v(c,f,0,0)$ denote the initial intensity at modulation frequency f in filter channel c. We have

$$v(c,f,0,s_f) = v(c,f,0,0) * h(s_f) \tag{5}$$

where $h(s_f)$ is a low-pass filter (in the modulation frequency domain with passband $[0,s_f]$ in Hertz). Here, "$*$" denotes convolution. The parameter (s_f) indicates the degree of smoothing. The smaller (s_f) is, the smoother $v(c,f,0,s_f)$ is.

C. Maximum detection:

By detecting the onsets and offsets and forming the onset and offset front, the modulation spectrum of speech signal is segmented. A few of these segments consist of information about the range of pitch frequency. The speaker's pitch ranges have to be [60,350] Hz (for men, women and children). Therefore, in every subband (for each acoustic frequency), the maximums in [60,350] Hz and above a certain threshold are founded.

Onset/offset detection and matching: onset and offset candidates are detected by marking peaks and valleys of the modulation frequency derivative of the smoothed intensity

$$\frac{d}{df}v(c,f,0,s_f) = \frac{d}{df}\left[v(c,f,0,0) * h(s_f)\right] \tag{6}$$

In each sub-band we select the onsets and offsets that occurred around any specified maxima. After determining onset and offset of each subband, onset and offset fronts should be found and finally the bounds of segments will be identified.

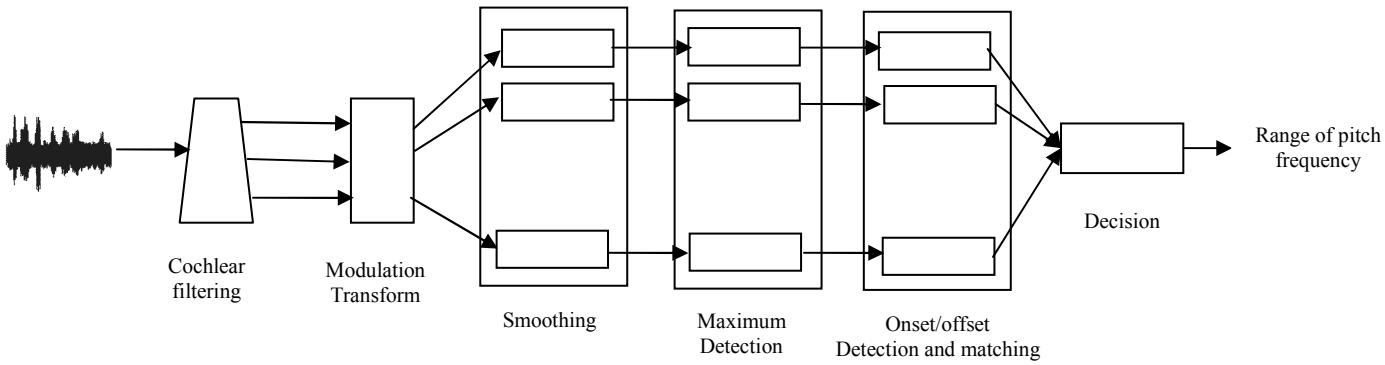

Figure 4. The block diagram of the proposed system.

D. *Decision-making:*

Obviously, in the frequency domain of speech signal, there are some peaks in the pitch frequency and its harmonic. Based on this, in the decision stage, we select the segments whose ranges of modulation frequency are the harmonics of each other. In accord to the onset and offset fronts of the desired segment, and by calculating the mean of onset and offset of these fronts, we can find the beginning and ending of the range of pitch frequency.

V. RESULTS

We now present experimental results demonstrating the robustness and the accuracy of our method compared to the least square harmonic model [15] in different SNR's . In [15] an algorithm for harmonic decomposition of time-variant signals is derived from a least squares harmonic (LSH) technique. The estimates of harmonic amplitudes and phases are formulated as the solution of a set of linear equations that minimize the mean square error.

The experimental results in [15] demonstrate the robustness and the accuracy of LMS method relative to standard algorithms such as RAPT [16] and the maximum a posterior estimator (MAP) of [17].

The signal frequency is modeled by a linear or quadratic polynomial and obtained via a local search over polynomial coefficients. An initial estimate of signal frequency is necessary to reduce computation time.

To evaluate the accuracy of the proposed algorithm in pitch range estimation, we choose signals from a corpus of 10 speech signal (male and female speeches), that commonly used for CASA research [18] and TIMIT database. At first we selected the clean speech from the TIMIT database, $x(n)$ (speaker: "one two three four five … "). The speech was sampled at 16 kHz. The algorithm parameters were set to $M = 16$, $K = 512$, $L = 38$, $I = 512$, and $h(n)$ and $g(m)$ were a 48-point and 78-point Hanning window. The additive noise is a white noise with zero mean and unity variance.

Figures 2 and 5(a) show the signal and modulation spectrum of the initial 0.8s of the speech ("one"). The applied filter for smoothing of each subband is FIR low-pass filter. By selecting a proper threshold, only the maxima above the certain value in each subband are chosen. In this way, we can avoid the production of undesired segments.

The obtained pitch contours of the speech using the LSH model is shown in Figure 5(b) for different SNR's. The exact value and the obtained range of pitch frequency using the proposed system and LSH model for cleaned speech is reported in Table 1. The performance of the proposed system and the LSH model in terms of the range of the pitch frequency for different SNR's is reported in Table 2. Table 3 shows the obtained mean error percentage of pitch range estimation for 10 speech signals after using the proposed system and LSH model for different SNR's.

According to Figure 5(b), we can observe that the LSH method can not accurately estimate the pitch frequency in the transition region between voiced and unvoiced. According to these results and by comparing with the exact range of pitch frequency, one may deduce that for low SNR's, the LSH model faces the error and its accuracy reduces, while using the proposed system, in low SNR's the range of pitch frequency is estimated accurately.

VI. CONCLUSION AND DISCUSSION

We demonstrated that modulation frequency localization of pitch energy is an important feature for determination of range of pitch frequency of speech in modulation spectrogram that can be exploited for single channel speaker separation. We presented a new approach for determination of range of pitch frequency based on modulation frequency analysis and onset/offset algorithm. The proposed method is purposely simple and accurate. Also, the results show that the accuracy of the proposed algorithm is acceptable in noisy conditions and for different SNR's.

By expansion of proposed system, we can present a new system for determination of multipitch frequency in modulation frequency domain with using pitch energy.

TABLE I. THE EXACT AND THE OBTAINED RANGE OF PITCH FREQUENCY USING THE PROPOSED SYSTEM AND LSH MODEL FOR CLEANED SPEECH

Clean speech		
True value	***Proposed system***	***LSH model***
[91,125]	[92,126]	[93,126]

TABLE II. OBTAINED RANGE OF PITCH FREQUENCY USING PROPOSED SYSTEM AND LSH MODEL FOR DIFFERENT SNR'S

SNR	**Proposed system**	**LSH model**
0	[85,117]	[80,135]
5	[92,123]	[79,126]
10	[92,124]	[93,126]
15	[92,125]	[93,127]

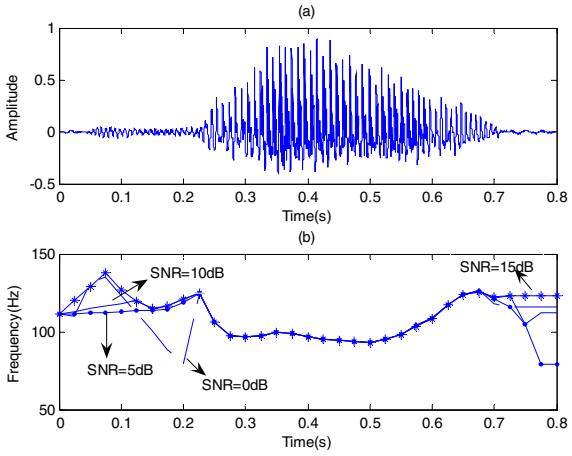

Figure 5. (a) Speaker's signal ("one"). (b) Pitch contours of speech signal obtained with using least square harmonic model for different SNR's.

TABLE III. THE OBTAINED MEAN ERROR PERCENTAGE OF PITCH RANGE ESTIMATION FOR 10 SPEECH SIGNAL

SNR	Proposed system	LSH model
0	39.12	51.31
5	8.41	35.62
10	5.02	9.11
15	2.85	5.23

REFERENCES

[1] P. Vary and R. Martin ,Digital Speech Transmission Enhancement, Coding And Error Concealment, Wiley, 2006.

[2] C. Breithaupt, T. Gerkmann, and R.Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," Proceedings of ICASSP, pp. 4897–4900, April 2008.

[3] C. MadsGraesboll and A. Jakobsson, Multi-Pitch Estimation-Synthesis Lectures on Speech and Audio Processing, M & C Publishers, 2009.

[4] Lawrence R. Rabiner, "On the use of autocorrelation analysis for pitch detection," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 25, no. 1, pp. 24–33, February 1977.

[5] A. de Cheveign´e 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.

[6] J. C. Brown and M. S. Puckette, "A high resolution fundamental frequency determination based on phase changes of the furier transform," Journal of the Acoustical Society of America, vol. 94, no. 2, pp. 662–667, 1993.

[7] John E. Lane, "Pitch detection using a tunable IIR filter," Computer Music Journal, vol.14, no.3, pp.46–59, Fall 1990.

[8] D. Cooperand Kia C. Ng, "A monophonic pitch-tracking algorithm based on waveform periodicity determinations using landmark points," Computer Music Journal, vol. 20, no. 3, pp. 70–78, Fall 1996.

[9] A. Choi, "Real-time fundamental frequency estimation by least-square fitting," IEEE Transactions on Speech and Audio Processing, vol. 5, no. 2, pp. 201–205, March 1997.

[10] J. C. Brown, "Musical fundamental frequency tracking using a pattern recognition method," Journal of the Acoustical Society of America, vol. 92, no.3, pp. 1394–1402, September 1992.

[11] H. Sano and B. Keith Jenkins, "A neural network model for pitch perception," Computer Music Journal, vol. 13, no. 3, pp. 41–48, Fall 1989.

[12] S. Schimmel, L. Atlas, and K. Nie, "Feasibility of signal channel speaker separation based on modulation frequency analysis," Proceedings of ICASSP, pp. 605–608, 2007.

[13] G. Hu, and D.L. Wang, "Auditory segmentation based on onset and offset analysis," IEEE Trans. Audio, Speech, Lang. Process. Vol. 15, no. 2, pp. 396–405, February 2007.

[14] J. Canny, " A computational approach to edge detection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 8, pp. 679–698, 1986.

[15] Q. Li, L. Atlas, "Time-variant least-squares harmonic modelling" , Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, Hong-Kong, April 2003.

[16] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in Speech Coding and Synthesis, W. B. Klein and K. K. Paliwal, Eds.NewYork, NY: Elsevier, pp.495–518, 1995.

[17] J. Tabrikian, S. Dubnov, and Y. Dickalov, "Maximum a posterior probability pitch tracking in noisy environments using harmonic model," IEEE Trans. Speech Audio Process., vol. 12, pp. 76–87, 2004.

[18] G. J. Brown, and M. P. Cooke, "Computational auditory scene analysis," Comput. Speech Lang., vol. 8, pp. 297–336, 1994.