A PITCH DETERMINATION ALGORITHM BASED ON SUBHARMONIC-TO-HARMONIC RATIO

Xuejing Sun

Department of Communication Sciences and Disorders, Northwestern University
2299 N. Campus Dr., Evanston, IL 60208, USA
sunxj@northwestern.edu

ABSTRACT

In the present paper, a pitch determination algorithm (PDA) based on Subharmonic-to-Harmonic Ratio (SHR) is proposed. The algorithm is motivated by the results of a recent study on the perceived pitch of alternate pulse cycles in speech [1]. The algorithm employs a logarithmic frequency scale and a spectrum shifting technique to obtain the amplitude summation of harmonics and subharmonics, respectively. Through comparing the amplitude ratio of subharmonics and harmonics with the pitch perception results, the pitch of normal speech as well as speech with alternate pulse cycles (APC) can be determined. . Evaluation of the algorithm is performed on CSTR's database and on synthesized speech with APC. The results show that this algorithm is one of the most reliable PDAs. Furthermore, superior to most other algorithms, it handles subharmonics reasonably well.

1. INTRODUCTION

Pitch, i.e., fundamental frequency (F0), is an important feature in prosody modeling and many other speech research areas. Numerous pitch determination algorithms (PDA) have been proposed in the past [2]. Unfortunately, to date, pitch determination still remains one of the most difficult problems in speech analysis. The most common errors are pitch doubling and pitch halving. One of the reasons for pitch doubling and pitch halving is the appearance of alternate pulse cycles (APC) in speech signal, which reflects the short-term instability of the vocal fold system [3][4]. Figure 1 shows two schematic examples of alternate pulse cycles. Using a strictly signal processing point of view we would most likely fail to estimate the appropriate pitch when APC occurs in speech. After all, when there is ambiguity, it is the perceived pitch that is the most relevant. In this paper, a perception oriented PDA is proposed, which is particularly aimed at finding a solution to the problem of pitch determination for speech signals with APC, as well as for normal speech in general.

2. THEORETICAL BACKGROUND

In a recent pitch perception study [1], vowels with alternate pulse cycles were modeled through amplitude and frequency modulation, and synthesized using a formant synthesizer. Subjects were asked to determine the pitch of synthesized vowels with various degrees of amplitude or frequency modulation. The index of amplitude modulation and frequency modulation of glottal volume velocity waveform is defined as follows [3][4]:

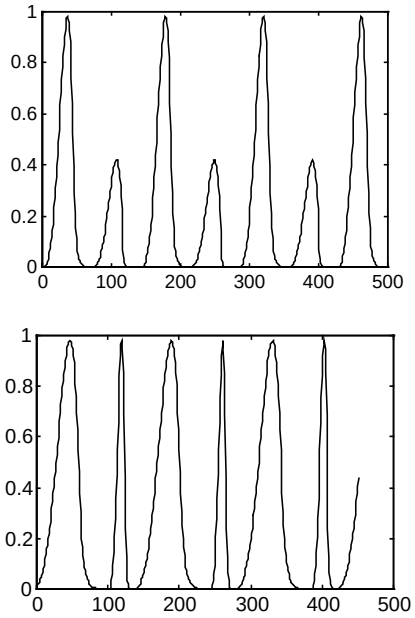

Figure 1: A schematic representation of glottal pulses with alternate pulse cycles (APC). (a). amplitude alternation (b). period alternation.

$$M_{AM} = \frac{A_i - A_{i+1}}{A_i + A_{i+1}} \quad (1)$$

$$M_{FM} = \frac{T_i - T_{i+1}}{T_i + T_{i+1}} \quad (2)$$

Modulation index can be viewed as a time domain descriptor of APC ranging from 0 to 100 percent. On the other hand, in the frequency domain, the modulation is manifested as the presence of subharmonic components (Figure 2). The result of that study shows that pitch perception is closely related to the Subharmonic-to-Harmonic Ratio (SHR), i.e., amplitude ratio between subharmonics and harmonics. When the ratio is small, the perceived pitch remains the same. As the ratio increases above certain threshold, the subharmonics become clearly visible on the spectrum, and the perceived pitch becomes one octave lower than the original pitch. These findings suggest that pitch may be optimally determined by computing SHR and comparing it with the pitch perception data. In the next section, we present the procedures of calculating SHR for speech signal and its application to pitch determination.

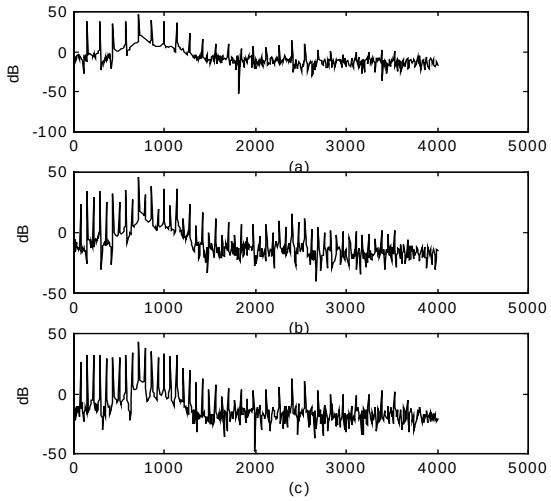

Figure 2: Spectrum of synthetic vowel /a/. (a) without modulation (b) glottal amplitude modulation with index m=50% (c) glottal amplitude modulation with index m=90%.

3. PITCH DETERMINATION

The current algorithm based on SHR is therefore developed in the frequency domain. First, Let $A(f)$ denote the short-term spectrum function, which is obtained by applying the Fourier transform on windowed short-term speech frames. The length of FFT is varied with the sampling rate and frame length. Suppose that the fundamental frequency is f_0, then the *sum of harmonic amplitude* is defined as:

$$SH = \sum_{n=1}^{N} A(nf_0) \tag{3}$$

where N is the maximum number of harmonics to be considered. If we only consider the subharmonic frequency that is at one half of fundamental frequency, the *sum of subharmonic amplitude* is defined as:

$$SS = \sum_{n=1}^{N} A((n-\frac{1}{2})f_0) \tag{4}$$

Consequently, SHR can be obtained by dividing SS with SH:

$$SHR = \frac{SS}{SH} \tag{5}$$

In order to get SS and SH, we could use the direct spectrum compression technique on linear frequency scale as that in Harmonic Product Spectrum (HPS) algorithm [5]. However, because of the numerical problem, a logarithmic transformation on the frequency scale is more preferable, which has been used in Subharmonic Summation algorithm (SHS) [6]. In developing the current algorithm, we adopted this basic approach. Nevertheless, the rationale and detail implementation are quite different, which affects the performance in a significant way. To facilitate our work in log domain, we reformulate the above definitions. Let $LOGA(f)$ denote the short-term log spectrum, and $\log f_0$ denote fundamental frequency on the log scale. Therefore, we have:

$$SH = \sum_{n=1}^{N} LOGA(\log n + \log f_0) \tag{6}$$

$$SS = \sum_{n=1}^{N} LOGA(\log(n-\frac{1}{2}) + \log f_0) \tag{7}$$

The log frequency scale is then linearly interpolated. In order to obtain SH, the spectrum is shifted leftward along the logarithmic frequency abscissa at even orders, i.e., log(2), log(4),… log(2N). These shifted spectra are added together.

$$SUMA(\log f)_{even} = \sum_{n=1}^{N} LOGA(\log f + \log(2n)) \tag{8}$$

From Eq. (8), SH is given by:

$$SH = SUMA(\log(0.5 f_0))_{even} \tag{9}$$

Similarly, by shifting the spectrum leftward at log(1), log(3), log(5), …. log(2N-1), we get SS also at $\log(0.5 f_0)$:

$$SUMA(\log f)_{odd} = \sum_{n=1}^{N} LOGA(\log f + \log(2n-1)) \tag{10}$$

$$SS = SUMA(\log(0.5 f_0))_{odd} \tag{11}$$

Next, we obtain the difference function, which is defined as:

$$DA(\log f) = SUMA(\log f)_{even} - SUMA(\log f)_{odd} \tag{12}$$

In so doing, we remove the effect of the contribution of the points around the real peaks, which is equivalent to peak enhancement. Moreover, there are some very interesting properties of the DA(•) function. In ideal cases, if subharmonics do not exist, and ignoring the contribution from the points that are at $\log(nf) \pm \log(0.25 f_0)$, we would have two maximum values at $\log(0.5 f_0)$ and $\log(0.25 f_0)$ from Eq. (12), respectively. The values are:

$$DA(\log(0.5 f_0)) = SH - SS \tag{13}$$

$$DA(\log(0.25 f_0)) = SH + SS \tag{14}$$

where $SS = 0$.

In normal speech without subharmonics, the points at $\log(nf) \pm \log(0.25 f_0)$ cannot be ignored if they are not set to zero or some very small values. Thus, in the current algorithm we need to take these values into account. Let Δ denote the sum of these points, we then have an extra term in Eq. (14):

$$DA(\log(0.25 f_0)) = SH + SS - \Delta \tag{15}$$

where $SS = 0$.

Consequently, we can only obtain one maximum value at $DA(\log(0.5 f_0))$, and the second maximum value at $DA(\log(0.25 f_0))$. At this moment, it becomes obvious that by finding the maximum value of $DA(\cdot)$, we will be able to determine the pitch. However, sometimes the speech signal becomes unstable, and subharmonic appears. When this

happens, the maximum value could occur at either $DA(\log(0.5 f_0))$ or $DA(\log(0.25 f_0))$, depending on the relative magnitude of SS and Δ. Subharmonic-to-Harmonic Ratio (SHR) can be approximated by the following simple formula:

$$SHR \approx 0.5 \frac{DA(\log(0.25 f_0) - DA(\log(0.5 f_0))}{DA(\log(0.25 f_0)) + DA(\log(0.5 f_0))} \qquad (16)$$

$$= \frac{SS - 0.5\Delta}{SH - 0.5\Delta} < \frac{SS}{SH} \qquad \text{if } SS < SH$$

It can be seen that since normally $SS < SH$, we will have an underestimated SHR using Eq. (16).

Based on the above analysis, we perform the following procedures to compute SHR and then determine the pitch: First, we locate the position of the global maximum denoted as $\log f_1$. Then, starting from this point, the position of the next local maximum denoted as $\log f_2$ is selected in the range of $[\log(1.75 f_1), \log(2.25 f_1)]$. Following Eq. (16), SHR can be easily derived:

$$SHR = 0.5 \frac{DA(\log f_1) - DA(\log f_2)}{DA(\log f_1) + DA(\log f_2)} \qquad (17)$$

If SHR is less than a certain threshold value, which is 0.2 in the current implementation, f_2 is chosen as the final pitch. Otherwise, f_1 will be selected.

4. VOICING ESTIMATION AND POSTPROCESSING

The current voicing determination algorithm is rather crude. Before passing the signal to pitch determination module, energy level and zero crossing rate are computed. If the energy level is below certain threshold or the zero crossing rate is too high, the frame will be classified as unvoiced. This certainly will produce many voiced/unvoiced decision errors. Nevertheless, we are more interested in pitch determination in this paper, and more sophisticated voicing estimation algorithm could be readily added in the future. Also, to accommodate to the evaluation strategy adopted in the next section, certain parameter values are used to lower the error rate of classifying voiced frame as unvoiced frame. As a result, more frames will be passed into the pitch determination module, such as fricative. The post-processing technique used in the current implementation is also quite simple, only a seven-point median smoother is employed.

5. EVALUATION

5.1. Evaluation on Normal Speech

CSTR's database[1] is used for the performance evaluation [7]. This database contains 50 sentences each from one male and one female speaker, which in total is about five minutes long. The speech signal is sampled at 20KHz using a 16-bit A/D converter. The reference pitch values are provided by simultaneously recorded laryngeal frequency contour. To compare with the results obtained by Bagshaw [7], we use 38.4 ms for frame length, and compute F0 at 6.4 ms interval; The F0 range is limited to 50Hz-250Hz for male speaker, and 120Hz-400Hz for female speaker. In the present evaluation, only gross errors are taken into account, since smaller errors are considered to be trivial in comparison [6]. The gross error, i.e., pitch doubling and pitch halving, is defined as when the determined F0 value is 20% higher or lower than the reference F0 value [7]. Table 1 shows the gross error rate (GER) for nine PDAs. The results of the first seven PDAs are obtained from Bagshaw's study [7]. The eighth is the result of a modified AMDF-based PDA with probabilistic error correction [8]. The last one is the results of current SHR algorithm.

PDAs	Male		Female	
	GER (%)		GER (%)	
	High	Low	High	Low
CPD	4.09	0.64	0.61	3.97
FBPT	1.27	0.64	0.60	3.55
HPS	5.34	28.2	0.46	1.61
IPTA	1.40	0.83	0.53	3.12
PP	0.22	1.74	0.26	3.20
SRPD	0.62	2.01	0.39	5.56
eSRPD	0.90	0.56	0.43	0.23
mAMDFp	1.94	2.33	0.63	2.93
SHR	**1.29**	**0.78**	**0.75**	**1.69**

Table 1: Comparison of 9 different PDAs.

- Cepstrum pitch determination (CPD)[9]
- Feature-based pitch tracker (FBPT)[10]
- Harmonic product spectrum (HPS)[5]
- Integrated pitch tracking algorithm (IPTA)[11]
- Parallel processing method (PP)[12]
- Super resolution pitch determinator (SRPD)[13]
- Enhanced version of SRPD (eSRPD)[7]
- Modified AMDF-based PDA with probabilistic error correction (mAMDFp) [8]
- Pitch determination algorithm based on Subharmonic-to-Harmonic Ratio (SHR)

It can be seen that the performance of current algorithm is inferior only to that of eSRPD algorithm, but better than those of all other algorithms. This is a promising result in that it shows SHR is indeed a very reliable parameter for pitch determination. It should be noted that because of the interpolation and harmonic summation process, the fine error rate is often higher. Nevertheless, given that the current algorithm is still under development, and more elaborate pre- or post-processing techniques and other fine modification could be applied later, it is our belief that the current algorithm based SHR estimation will be very accurate and robust.

[1] The author would like to thank Dr. Bagshaw for providing the database and the evaluation results freely available.

5.2. Evaluation on Speech with APC

Note that eSRPD is a time-domain approach, while the current SHR algorithm is a frequency domain approach. One of the main strength of the current approach lies in its way of handling speech with alternate pulse cycles (APC). On the other hand, a time-domain approach, such as eSRPD, is more sensitive to speech with APC and thus more likely to make errors. Thus, it is more interesting to compare the current SHR based algorithm to those similar frequency domain approaches, such as HPS and SHS. From Table 1, we can see that HPS is one of the most unreliable PDAs. As for SHS, it was unfortunately not evaluated in Bagshaw's study. Thus, we performed an informal comparison on SHS and SHR using the Praat program (Developed by Paul Boersma and David Weenink), which includes an implementation of SHS algorithm. We only used synthesized speech with APC, and leave a vigorous evaluation on pathological voice to future work. The synthesized speech samples contain both amplitude modulation and frequency modulation signal, and have been used in the pitch perception study [1]. Preliminary results show that the current algorithm performs much better than SHS, which is more sensitive to APC. Figure 3 presents the pitch contours generated by the two algorithms on the vowel /a/. The vowel is synthesized with fundamental frequency at 140 Hz and glottal amplitude modulation at 20 percent. The perception results show that it still should be perceived as 140Hz [1]. Clearly, by computing SHR we can have a more reasonable pitch contour, whereas SHS cannot predict pitch reliably in this case.

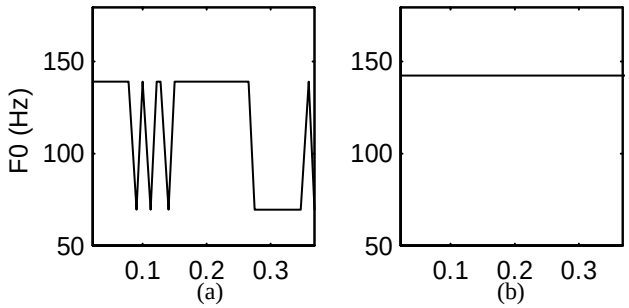

Figure 3: Comparison of SHS and SHR algorithm on a synthesized vowel /a/ with glottal amplitude modulation at 20%. (a). pitch contour generated by SHS (b). pitch contour generated by the current SHR based algorithm.

6. SUMMARY

In this study, a pitch determination algorithm based on Subharmonic-to-Harmonic Ratio is proposed and tested. By combining with pitch perception results, this algorithm can effectively reduce the gross error rate resulting from subharmonics, i.e., speech with APC. It is our belief that the current algorithm based on calculating SHR employs more information that can be expected from a short-term speech segment than the previous similar approaches, such as HPS and SHS, and is therefore more reliable. This algorithm is also superior to time-domain algorithms in dealing with speech with APC. In future research we will try to improve the current algorithm and employ more sophisticated voicing estimation and post-processing techniques. In addition, an important work is to apply the present algorithm for pitch extraction in pathological voice and intonation modeling research.

7. ACKNOWLEGEMENT

This work is partially supported by a Graduate Research Grant from the University Research Grants Committee at Northwestern University and by NIH grant DC03902. The author wishes to thank Yi Xu for helpful comments on the manuscript.

8. REFERENCES

1. Sun, X., and Xu, Y. "The Perceived Pitch of Synthesized Vowels with Alternate Pulse Cycles" (Forthcoming)
2. Hess, W. J. "Pitch and Voicing Determination," In *Advances in Speech Signal Processing*, edited by S. F. a. M. M. Sondhi, Marcel Dekker, Inc., New York, NY: 3-48, 1991.
3. Titze, I. R. *Workshop on Acoustic Voice Analysis-Summary Statement*, National Center for Voice and Speech, Denver, 1995.
4. Titze, I.R. *Principles of Voice Production*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1994.
5. Schroeder, M.R. "Period histogram and product spectrum: New methods for fundamental frequency measurement," *J. Acoust. Soc. Am.*, 43(4):829-834, 1968.
6. Hermes, D. J. "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, 83, 257-264, 1988.
7. Bagshaw, P.C. "Automatic prosody analysis," *PhD thesis*, University of Edinburgh, Scotland, UK, 1994.
8. Ying, G. S., Jamieson, L. H., and Mitchell, C. D. "A Probabilistic Approach to AMDF Pitch Detection," *Proceedings of the 1996 International Conference on Spoken Language Processing*, Philadelphia, PA, Oct. 1201-1204, 1996.
9. Noll, A.M. "Cepstrum pitch determination," *J. Acoust. Soc. Am.*, 41(2):293-309, 1967.
10. Phillips, M.S. "A feature-based time domain pitch tracker," *J. Acoust. Soc. Am.*, 77:S9-S10 (A), 1985.
11. Secrest, B.G., and Doddington, G.R. "An integrated pitch tracking algorithm for speech systems," In *Proc. IEEE ICASSP*, 1352-1355, Boston, 1983.
12. Gold, B., and Rabiner, L.R. "Parallel processing technique for estimating pitch period of speech in the time domain," *J. Acoust. Soc. Am.*, 46(2, part 2):442-448, 1969.
13. Medan, Y., Yair, E., and Chazan, D. "Super resolution pitch determination of speech signals," *IEEE Trans. ASSP*, 39:40-48, Jan. 1991.