

PITCH DETERMINATION AND VOICE QUALITY ANALYSIS USING 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

This paper presents an improvement of a previously proposed pitch determination algorithm (PDA). Particularly aiming at handling alternate cycles in speech signal, the algorithm estimates pitch through spectrum shifting on logarithmic frequency scale and calculating the Subharmonic-to-Harmonic Ratio (SHR). The evaluation results on two databases show that this algorithm performs considerably better than other PDAs compared. Application of SHR to voice quality analysis task is also presented. The implementation and evaluation routines are available from <http://mel.speech.nwu.edu/sunxj/pda.htm>.

1. INTRODUCTION

Pitch determination has shown to be one of the most difficult problems in speech analysis. One of the reasons is the occurrence of alternate cycles (alternating in amplitude or period, or both) in speech signal. For normal speech, alternate cycles usually appear in creaky voice or voice with laryngealization, which are often characterized as perceptually rough voices. In pathological voice, alternate cycles can be found even in normal mode of production. The alternate cycles make pitch determination difficult. The solutions of most current algorithms either rely on fine-tuning some threshold parameters based on particular databases or post-processing techniques, such as linear/nonlinear smoothing, dynamic programming, etc. In the present paper, an alternative approach is explored.

Previously, a pitch determination algorithm was proposed particularly aiming at handling the alternate cycles in speech [8]. The algorithm is motivated from two assumptions: (1) human perception can be useful in estimating the pitch of alternate cycles; (2) The alternate cycles in the time domain are manifested by the presence of subharmonics in the frequency domain [11]. More precisely, the magnitude of subharmonics with respect to harmonics reflects the degree of deviation from modal voice. A new parameter named Subharmonic-to-Harmonic Ratio (SHR) was proposed to describe the amplitude ratio between subharmonics and harmonics. A pitch perception study was carried out in an attempt to determine the relationship between perceived pitch and SHR [9]. In the experiment, vowels with alternate cycles were synthesized through amplitude and frequency modulation, which generated subharmonics with lowest frequency of $0.5F_0$. Human listeners were asked to determine the pitch. The results show that pitch perception is closely related to SHR, i.e., amplitude ratio between subharmonics and harmonics. Generally, when the ratio is

smaller than 0.2, the subharmonics do not have effects on pitch perception. As the ratio increases approximately above 0.4, the pitch is mostly perceived as one octave lower that corresponds to the lowest subharmonic frequency. When SHR is between 0.2 and 0.4, the pitch seems to be ambiguous. These findings suggest that pitch could be determined by computing SHR and comparing it with the pitch perception data.

Beside pitch determination, SHR can be also used as a parameter for describing voice quality. As is known, different degrees of “irregularity” can elicit different degrees of “roughness” sensation, and this reflects the status of underlying physiological apparatus [12]. Therefore, it is desirable to have an objective measure to quantify this relationship, which can be used as an index to classify voice production mode for one speaker or compare voice quality for different speakers. Since SHR can describe alternate cycles quantitatively, it in turn can be an indicator for perceptual quality of voice.

The procedure for computing SHR falls in the general category of spectrum compression technique. Since direct estimating the lowest harmonic in the spectrum has shown to be unreliable for pitch determination [4], researchers usually try to take advantage of the harmonic structure. A family of PDAs based on the idea of spectrum compression have been developed, in which the spectrum is compressed along the frequency axis at different ratios and the compressed spectra are added together to make the F_0 peak more prominent (e.g. [3][7]). However, when there are subharmonics, the algorithm may be confused and select the wrong peak. Unfortunately, direct summation of harmonics does not provide a mechanism to control the effect of subharmonics. On the other hand, the current algorithm, instead of looking for one single peak that actually represents the summation of the harmonics and subharmonics, tries to decompose the effects of the harmonics and subharmonics, and examine whether the subharmonics are strong enough to be regarded as pitch candidates.

The basic algorithm has been described in [8]. In this paper, we present the algorithm with modifications and conduct more comprehensive evaluations. The results reveal that it outperforms other methods substantially. We also discuss how SHR can be applied to voice quality analysis.

2. THE ALGORITHM

For each short-term signal, let $A(f)$ represent the amplitude spectrum, and let f_0 , and f_{max} be the fundamental frequency and the maximum frequency of $A(f)$, respectively. Then the *sum of harmonic amplitude* is defined as

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

where N is the maximum number of harmonics contained in the spectrum, and $A(f) = 0$ if $f > f_{max}$. If we confine the pitch search range in $[F0_{min} F0_{max}]$, then $N = \text{floor}(f_{max} / F0_{min})$. In practice, to reduce computational cost, only part of the spectrum is used. Following [3], we set $f_{max} = 1250$ Hz.

Assuming the lowest subharmonic frequency is one half of f_0 , the *sum of subharmonic amplitude* is defined as

$$SS = \sum_{n=1}^N A((n-1/2)f_0) \quad (2)$$

Note that the current algorithm is extendable to other subharmonic frequencies.

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

$$SHR = \frac{SS}{SH} \quad (3)$$

In practice, however, estimating *SHR* directly using Eq. (3) is not trivial. Thus an alternative way is proposed. Firstly, we transform the linear frequency scale to logarithmic scale as described in [3]. Let *LOGA*(\bullet) denote the spectrum with log frequency scale, then we can represent *SH* and *SS* as

$$SH = \sum_{n=1}^N \text{LOGA}(\log(nf_0)) = \sum_{n=1}^N \text{LOGA}(\log(n) + \log(f_0)) \quad (4)$$

$$SS = \sum_{n=1}^N \text{LOGA}(\log(n-1/2) + \log(f_0)) \quad (5)$$

To obtain *SH*, the spectrum is shifted leftward along the logarithmic frequency abscissa at even orders, i.e., $\log(2)$, $\log(4)$, ... $\log(4N)$. These shifted spectra are added together and denoted by

$$\text{SUMA}(\log f)_{\text{even}} = \sum_{n=1}^{2N} \text{LOGA}(\log f + \log(2n)) \quad (6)$$

Since $\text{LOGA}(\log f) = 0$ when $f > f_{max}$, from Eqs. (4)–(6) we have:

$$\text{SUMA}(\log(1/2 f_0))_{\text{even}} = SH \quad (7)$$

$$\text{SUMA}(\log(1/4 f_0))_{\text{even}} = SH + SS \quad (8)$$

Similarly, by shifting the spectrum leftward at $\log(1)$, $\log(3)$, $\log(5)$, ... $\log(4N-1)$, we have

$$\text{SUMA}(\log f)_{\text{odd}} = \sum_{n=1}^{2N} \text{LOGA}(\log f + \log(2n-1)) \quad (9)$$

$$\text{SUMA}(\log(1/2 f_0))_{\text{odd}} = SS \quad (10)$$

$$\text{SUMA}(\log(1/4 f_0))_{\text{odd}} = \Delta \quad (11)$$

where Δ represents the sum of the values at $\log(nf_0) \pm \log(1/4 f_0)$.

Next, we define a difference function as

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

From Eqs. (7)(8)(10)(11), we can easily get

$$DA(\log(1/2 f_0)) = SH - SS \quad (13)$$

$$DA(\log(1/4 f_0)) = SH + SS - \Delta \quad (14)$$

Figure 1 gives an example for each function defined above, i.e., *LOGA*, *SUMA_{even}*, *SUMA_{odd}*, *DA*.

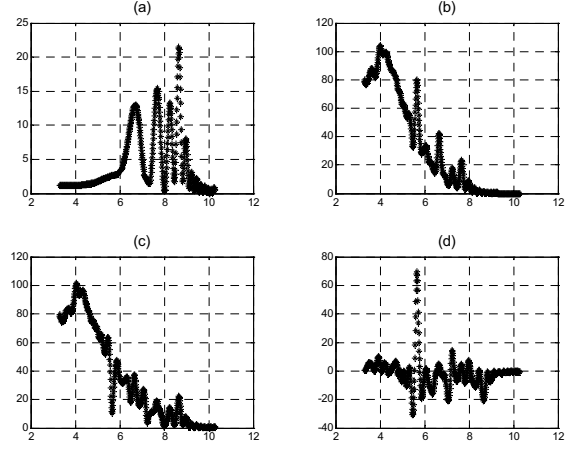


Figure 1. Schematic representations of four functions for calculating *SHR*. (a) *LOGA* (b) *SUMA_{even}* (c) *SUMA_{odd}* (d) *DA*

In “regular” speech where $SS \approx 0$, the maximum value of *DA*(\bullet) would be at $\log(1/2 f_0)$. On the other hand, if the magnitude of subharmonics becomes substantial, the maximum value of *DA*(\bullet) would be most likely at $\log(1/4 f_0)$ since now $\Delta \approx 0$, and the second maximum value is at $\log(1/2 f_0)$. We can then calculate *SHR* approximately using Eqs. (13)(14):

$$\frac{DA(\log(1/4 f_0)) - DA(\log(1/2 f_0))}{DA(\log(1/4 f_0)) + DA(\log(1/2 f_0))} = \frac{SS - 1/2 \Delta}{SH - 1/2 \Delta} \approx SHR \quad (15)$$

In searching for the maximum value, we first 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 $[\log(1.9375f_1), \log(2.0625f_1)]$. However, some special cases need to be treated differently:

- (1) If $DA(\log(f)) \leq 0$ for all f , the frame is regarded as unvoiced.
- (2) If $1.9375f_1 > F0_{max}$, only f_1 is returned.
- (3) If $DA(\log(f_1)) > 0$ and $DA(\log(f_2)) \leq 0$, only f_1 is returned.

After the two peaks are located, *SHR* can be easily derived following Eq. (15):

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

If *SHR* is less than a certain threshold value, it indicates that subharmonics are weak and we should favor the harmonics. Thus, f_2 is selected and the final pitch value is $2f_2$. Otherwise, f_1 is selected and the pitch is $2f_1$. Based on the pitch perception results [9] mentioned in Introduction, 0.2 is selected as the threshold value. But any other values in $[0.2, 0.4]$ should not make a significant difference. For specific tasks, threshold value outside this range could also be used. For example, using higher threshold value will encourage the algorithm to favor the harmonics even though the pitch of this frame should be represented by the subharmonics when listened in isolation. This is usually desirable in intonation modeling for speech synthesis where a globally smooth contour is wanted.

To detect voicing, the noise floor is estimated first which is usually three times of the energy of the first frame. If the energy of a frame is higher than the noise floor, it is passed into the pitch determination module. Using the estimated F0 value, we derive the fundamental period and select two periods of signal from this frame, one before the middle point and one after it. If the correlation between the two segments is higher than 0.2 and the zero-crossing rate of either segment is higher than 3500 Hz, the frame is classified as voiced, otherwise unvoiced.

3. EVALUATION

We first evaluate the accuracy of pitch determination. All the databases, PDAs (except for “GET_F0” from ESPS package) and the evaluation routines used in this task are freely available from the Internet, so that other researchers can replicate the current results and conduct new comparisons easily in the future. Second, we present a simple voice quality analysis using estimated SHR.

3.1 Databases

Two databases are used for evaluation. The first is the CSTR database [1], which contains five minutes of speech from one male and one female speaker. The speech signal is sampled at 20KHz with 16-bit resolution. The reference pitch values are provided which are estimated from simultaneously recorded EGG signals. Another one is the Keele pitch extraction reference database [6]. This database contains speech from 10 speakers, with five males and five females. The sampling rate is 20K Hz and the resolution is 16 bits. The reference pitch values provided by the database are determined by an autocorrelation method at 10 ms frame rate.

3.2 Evaluation of pitch determination

To evaluate pitch determination alone, we try to disable the voicing estimation and post-processing modules for all the algorithms. This is because most algorithms make voicing estimation errors, such as misclassifying many “difficult” voiced segments as unvoiced, which could result in a seemingly better pitch determination. Specifically, we compare the present algorithm (SHRP) with the followings:

eSRPD: Enhanced super resolution pitch determinator has shown to be superior to other seven algorithms in [1].

PDA: “PDA” program is included in Edinburgh Speech Tool Library <<http://www.cstr.ed.ac.uk>>, which is described as an implementation of super resolution pitch determinator (SRPD) [1][5]. An example command is “pda m1nw0000.wav -otype ascii -o cstr_f0/m1nw0000.f0 -fmin 50 -fmax 550 -shift 0.010 -length 0.040 -L”

GET_F0: “GET_F0” is included in ESPS package, which is an implementation of RAPT algorithm [10]. We use the default setting of the algorithm specified in the ESPS document except that we set “voice_bias=1” to encourage the algorithm to make more voicing hypotheses.

PRAAT: Praat software <<http://www.praat.org>> includes several PDAs, and we use the one described in [2] which usually gives very good results according to our experience. An example command is “To Pitch (ac)... 0.01 50 15 1 0 0 0.01 0.35 0 550”.

Note that even though we try to make the algorithms yield a F0 value for each frame, we were unable to disable voicing

detection completely. For “PDA”, we basically used the default setting, which includes a full voicing determination module. For “GET_F0” and “PRAAT” we got a smaller number of frames classified as unvoiced. Thus, for these algorithms we perform analysis only for the frames classified as voiced in both reference and the estimated data.

For the CSTR database, we use the same evaluation procedures described in [1]. That is, 38.4 ms for frame length, 6.4 ms for frame interval, 50Hz-250Hz for male speaker, and 120Hz-400Hz for female speaker for F0 range. The gross error rate (GER), i.e., pitch doubling and pitch halving, is defined as when the estimated F0 value is 20% higher or lower than the reference F0 value. The evaluation programs are also from the original database. Since the original results in [1] contain both voicing and pitch estimation, we also report the evaluation results with voicing detection for our algorithm (SHRPv).

Tables 1 and 2 show the evaluation results for male and female speech, respectively, which include unvoiced error rate (UER), voiced error rate (VER), GER, and absolute deviation. Due to some mislabelings in the database, the voiced error rate (VER) is not zero even though we set all the frames as voiced. After correction we do see an improvement. But for a fair comparison, we only list the results using the original database.

PDAs	Male					
	Voice (%)		GER (%)		Absolute Deviation (Hz)	
	UER	VER	High	Low	Mean	SD
eSRPD	4.63	12.07	0.90	0.56	1.40	1.74
GET_F0	155.55	0.22	1.14	2.50	2.76	3.54
PRAAT	206.47	0.15	1.23	1.53	2.22	3.42
SHRP	218.78	0.02	0.58	1.21	1.90	3.07
SHRPv	18.24	6.46	0.26	0.70	1.64	2.40

Table 1: PDA comparisons on the CSTR male speech.

PDAs	Female					
	Voice (%)		GER (%)		Absolute Deviation (Hz)	
	UER	VER	High	Low	Mean	SD
eSRPD	2.73	9.13	0.43	0.23	4.17	5.13
GET_F0	94.88	1.63	1.09	0.86	5.43	7.10
PRAAT	249.11	0.04	1.21	0.88	4.82	6.76
SHRP	250.58	0.02	0.89	0.85	4.95	6.72
SHRPv	5.64	10.17	0.39	0.40	4.16	5.24

Table 2: PDA comparisons on the CSTR female speech.

From Tables 1 and 2, it can be seen that the performance of current algorithm is better than that of eSRPD algorithm when the voicing estimation error is within the same scale. The overall mean GER for eSRPD is 0.53%, whereas ours is 0.44%.

For the Keele database, we use 40 ms for frame length, 10 ms for update interval and [50 550] for F0 range. Tables 3 and 4 list GER for each speaker and mean GER for each gender.

PDAs	GER (%)					
	M1	M2	M3	M4	M5	Mean
PDA	5.17	10.22	3.40	3.16	5.15	5.42
GET_F0	1.49	11.36	2.74	2.59	1.59	3.95
PRAAT	3.36	8.32	1.30	2.96	1.59	3.30
SHRP	4.29	4.49	0.41	0.55	0.68	2.08

Table 3: PDA comparisons on the Keele male speakers.

PDA's	GER (%)					
	F1	F2	F3	F4	F5	Mean
PDA	7.28	4.97	4.22	14.06	4.48	7.00
GET F0	11.23	6.15	6.62	7.15	2.74	6.78
PRAAT	4.31	2.21	2.98	4.66	1.08	2.99
SHRP	2.22	1.63	1.66	2.61	0.59	1.74

Table 4: PDA comparisons on the Keele female speakers.

The above evaluation results on two databases show that the present approach indeed yields better pitch estimation by considering subharmonic effects. Note that our algorithm does not employ any post-processing techniques in this evaluation, whereas others usually have an integrated post-processing module. Although it could be argued that we have developed and evaluated the algorithm based on the same databases so the results may be biased, we believe the bias effect is not significant. This is because the parameter values in our pitch determination module are not derived by fine-tuning using the databases but rather from perception results or commonly used by other algorithms.

3.3 Voice quality analysis

To relate SHR to pitch determination and voice quality, we perform some simple analyses. For each speaker, we select all the voiced frames and calculate the SHR frequency distribution (see Tables 5 and 6).

Speakers	SHR distribution (%)					
	0	(0,0.2)	[0.2,0.4]	(0.4,0.6]	(0.6,0.8]	[0.8,1]
M1	38.39	0.88	2.53	12.65	24.37	21.18
M2	72.50	4.27	1.95	2.82	8.40	10.06
M3	85.69	0.48	0.34	1.10	3.70	8.69
M4	89.53	0.12	0.12	0.37	2.89	6.96
M5	41.67	0.58	1.26	10.28	23.90	22.31

Table 5: SHR distribution of the Keele male speakers

Speakers	SHR distribution (%)					
	0	(0,0.2)	[0.2,0.4]	(0.4,0.6]	(0.6,0.8]	[0.8,1]
F1	62.38	14.24	1.11	1.37	7.84	13.06
F2	71.14	12.72	1.10	2.47	3.68	8.89
F3	72.85	7.68	0.93	1.79	5.76	10.99
F4	52.52	10.37	5.16	10.15	11.31	10.48
F5	56.30	16.85	0.48	6.78	9.74	9.85

Table 6: SHR distribution of the Keele female speakers

As discussed in Introduction, when SHR is in the medium range, especially [0.2,0.4], perceived pitch becomes ambiguous. For our data, in Table 3 M1 and M2 have higher GERs whereas M3 has the lowest. Correspondingly, in Table 5, M1 and M2 have higher SHR percentage in the range of [0.2,0.4] among five speakers, whereas M4 has the lowest. For the female speakers, F4 and F5 represent the worst and best cases. Visual inspection and listening to the speech waveform confirm that M1 and M2 indeed contain more "irregular" speech cycles and appears to have low and rough voice, whereas M3's speech seems to be much more "regular". Similarly F4 has more creaky voice than F5 despite the average pitch of F4 is much higher. Tables 5 and 6 also show that the female speakers have greater number of SHRs in the range of (0,0.2) compared with male speakers. This indicates that female speech might have greater amount of small amplitude or period fluctuations, which, however, are not significant enough to affect pitch perception.

4. SUMMARY

In this paper, a pitch determination algorithm based on Subharmonic-to-Harmonic Ratio estimation is proposed and tested. Performance analysis indicates that it is superior to other algorithms being evaluated. Simple voice quality analysis based SHR is also presented. There is still much room for improvement. Future work should include searching for optimal parameter values for different tasks such as intonation modeling and voice quality analysis.

5. ACKNOWLEDGMENTS

This work is supported by NIH grant DC03902. The author wishes to thank Yi Xu for helpful comments on the manuscript. Thanks also to George Meyer for Keele database setup and to Herbert Griebel for bug fixing in the program. The evaluation databases can be downloaded from:

CSTR: <http://www.cstr.ed.ac.uk/~pcb/>

Keele: <ftp://ftp.cs.keele.ac.uk/pub/pitch/>.

6. REFERENCES

- [1] P. C. Bagshaw, S. M. Hiller, and M. A. Jack, "Enhanced pitch tracking and the processing of f0 contours for computer aided intonation teaching," *Eurospeech'93*, Berlin, 1993, pp. 1003–1006.
- [2] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", *Proceedings of the Institute of Phonetic Sciences* 17, pp. 97-110, 1993.
- [3] D. J. Hermes, "Measurement of pitch by subharmonic summation," *J. Acoust. Soc. Am.*, vol. 83, pp. 257-264, 1988.
- [4] W. J. Hess, "Pitch and Voicing Determination," in *Advances in Speech Signal Processing*, S. F. a. M. M. Sondhi, Ed. New York, NY: Marcel Dekker, Inc., 1991, pp. 3-48.
- [5] Y. Medan, Yair, E., and Chazan, D. "Super resolution pitch determination of speech signals," *IEEE Trans. ASSP*, 39:40-48, Jan. 1991.
- [6] F. Plante, Meyer, G. and Ainsworth, W.A., "A pitch extraction reference database," *Eurospeech'95*, Madrid, Spain, 1995, pp.837–840.
- [7] M. R. Schroeder, "Period histogram and product spectrum: New methods for fundamental frequency measurement," *J. Acoust. Soc. Am.*, vol. 43, pp. 829-834, 1968.
- [8] X. Sun, "A pitch determination algorithm based on subharmonic-to-harmonic ratio," the *6th International Conference of Spoken Language Processing*, Beijing, China, 2000, 4, pp. 676-679.
- [9] X. Sun and Y. Xu, "Perceived Pitch of Synthesized Voice with Alternate Cycles", *submitted*.
- [10] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds.: Elsevier Science, 1995, pp. 495–518.
- [11] I. R. Titze, *Workshop on Acoustic Voice Analysis-Summary Statement*. Denver: National Center for Voice and Speech, 1995.
- [12] I. R. Titze, *Principles of Voice Production*. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1994.