



Pitch Determination Using Autocorrelation Function in Spectral Domain

M. Shahidur Rahman, Tetsuya Shimamura

Graduate School of Science and Engineering, Saitama University, Japan

{rahmanms, shima}@sie.ics.saitama-u.ac.jp

Abstract

This paper proposes a pitch determination method utilizing the autocorrelation function in the spectral domain. The autocorrelation function is a popular measurement in estimating pitch in time domain. The performance of the method, however, is effected due to the position of dominant harmonics (usually the first formant) and the presence of spurious peaks introduced in noisy conditions. We applied a series of operations to obtain a noise-compensated and flattened version of the amplitude spectrum which takes a shape of harmonics train. Application of the autocorrelation function on this preconditioned spectrum produces a sequence where the true pitch peak can be readily located. Experiments on long speech signals spoken by a number of male and female speakers have been conducted. Experimental results demonstrate the merit of the proposed method when compared with other popular methods.

Index Terms: Pitch determination, autocorrelation function, spectrum flattening, effect of additive noise

1. Introduction

Pitch (i.e. fundamental frequency) determination is an essential component in a variety of speech processing systems such as the speech analysis-synthesis system, speech coding system, speech and speaker identification system. Though it can be observed by visual inspection, there are some difficulties in automatic determination of pitch. The non-stationarity of speech signal, variety of voices, and environmental conditions are mostly the reasons causing the erroneous estimates [1]. Numerous methods have been proposed to address the issues. The properties of speech signals in either time-domain or frequency-domain, or in both, have been utilized for proposing algorithms for pitch determination [2, 3, 4, 5, 7, 8]. Time-domain estimators operate directly on the speech waveform to estimate the pitch period. Autocorrelation is one of the measurements that is applied frequently in time-domain because of its simplicity and relatively more robustness against white noise. The autocorrelation function (ACF) is also the inverse Fourier transform of the power spectrum of the signal. Thus, if there is a distinct formant structure in the signal, it is maintained in the ACF. Spurious peaks are also sometimes introduced in the spectrum in noisy or even in noiseless conditions. This sometimes makes true peak selection a difficult task. Non-linear operations on the speech signals such as center clipping has been shown to have the effect of flattening the spectrum [1]. Auto-regressive inverse filtering has also been suggested to flatten the signal spectrum [9]. Those preprocessing steps have effects on emphasizing the true period peaks in ACF. However, when few harmonics are present in the spectrum or in noisy environment, the process of obtaining the inverse filter itself is erroneous.

Among many other improvements reported on the ACF method, Talkin proposed a normalized cross correlation based method

[3] that claimed to eliminate many of the above mentioned problems. Shimamura proposed weighting the ACF [5] by the inverse average magnitude difference function [6]. Kawahara used an ACF based difference function [7] in conjunction with some optimization steps. Hasan proposed signal reshaping technique [8] for emphasizing the true peak. In this paper we propose yet another improvement utilizing the ACF in the spectral domain. Unlike the ACF domain, the amplitude spectrum (or the power spectrum) contains a single harmonic (and thus a single candidate) in every pitch duration. If the amplitude (or power) spectrum can be considerably flattened, ACF of the spectrum is expected to be more suitable for pitch determination. This is because determining the first peak is actually required for pitch estimation which does not hold for conventional ACF domain. In this paper, we apply a series of conditioning operations to flatten the spectrum and to reduce the noise effects. The preconditioned spectrum looks like a harmonics train, ACF on which leads to improved estimation accuracy. Experimental results have been presented on speech database and observed to outperform when compared with the competitive methods.

2. Problem Description

The autocorrelation function $R(\tau)$ of the speech signal $x(n)$ is generally defined as Eq.(1)

$$R(\tau) = \frac{1}{N} \sum_{n=0}^{N-1} x(n)x(n + \tau) \quad (1)$$

where N is the length of the underlying speech frame and τ is the lag number. If $x(n)$ is periodic at pitch period T , $R(\tau)$ exhibits peak at $\tau = iT$, where $i = 0, 1, 2, 3, \dots$. As the value of τ increases, $R(\tau)$ tends to decrease which facilitates the use of the second peak (at $\tau = T$) for estimation of the pitch period. According to Wiener-Khinchine theorem [2], the $R(\tau)$ can also be obtained as the inverse Fourier transform of the power spectrum of $x(n)$ as in Eq. (2).

$$R(\tau) = \frac{1}{M} \sum_{k=0}^{M-1} |X(k)|^2 e^{2\pi\tau k/M} \quad (2)$$

where $|X(k)|$ is the amplitude spectrum and M is the number of DFT (Discrete Fourier Transform) points. Unfortunately, if there is distinct formant structure in the signal, it persists in the ACF. Thus, the ACF is not robust to higher harmonic or subharmonic error. This can be seen in Figs. 1 and 2. Only the first 256 points of the spectrum are shown here for visual clarity. Clipping operations in the time domain has shown to be useful in flattening the spectrum. This has effect in increasing the prominence of actual period peaks in ACF. Auto-regressive inverse filtering has also effectiveness as a preprocessing step to flatten

10.21437/Interspeech.2010-246

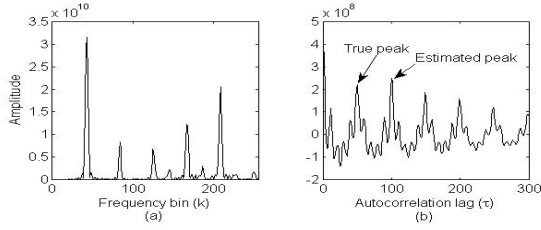


Figure 1: a) Spurious harmonics in the power spectrum, b) Period doubling in autocorrelation function.

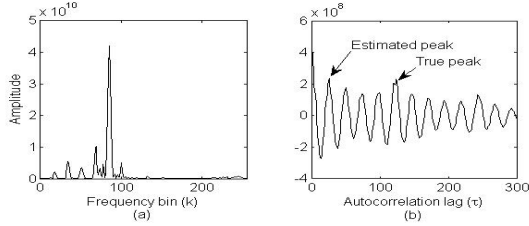


Figure 2: a) Dominant formant in the power spectrum, b) Error in period estimation using the autocorrelation function.

the signal spectrum. When speech signal is rich in harmonics and the SNR (Signal to Noise Ratio) is high, this method is beneficial in removing formant effect. However, when only a few harmonics are present and SNR goes low, inverse filtering is no longer effective. In this paper, we apply the autocorrelation measurement on a preconditioned amplitude spectrum. Unlike conventional ACF, the amplitude spectrum has only one pitch candidate (i.e. harmonics) in every pitch duration. Even if there is any spurious harmonics as in Fig. 1, it will be deemphasized in the ACF of spectrum due to its lack of correlation with the typical harmonics. Figs. 3 and 4 show the preconditioned amplitude spectra and their autocorrelations corresponding to the signals used in Figs. 1 and 2, respectively (the process will be detailed in Section 3). In both the cases, the spectral domain ACF determines the pitch accurately.

3. The Proposed Method

The proposed method can be described in two steps: spectral conditioning and autocorrelation on the preconditioned spectrum. The whole method is outlined in the block diagram of Fig. 5 and illustrated graphically in Fig. 6.

3.1. Spectral Conditioning

This step aims to enhance the spectrum in terms of flatness and robustness against additive noise. Since the ACF is the inverse DFT of the signal power spectrum, a more flattened version of the same can be obtained by taking DFT of the half-wave rectified ACF. Again, since the amplitude spectrum $|X(k)|$ is more compressed in magnitude than the power spectrum $|X(k)|^2$ and an autocorrelation-like sequence $R_X(k)$ (as shown in Fig. 6(c)) can also be derived by taking the inverse DFT of the amplitude spectrum, we used the amplitude spectrum instead of the power spectrum. Half-wave rectification of $R_X(k)$ (as shown in Fig. 6(d)) results in a more flattened spectrum when DFT is applied on the rectified version of the sequence (as shown in Fig. 6(e)). Another clipping operation is then applied on the modified amplitude spectrum to eliminate

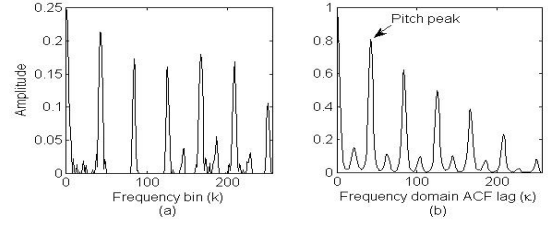


Figure 3: a) Preconditioned amplitude spectrum (corresponding to that in Fig. 1(a)), b) Autocorrelation of the amplitude spectrum in (a).

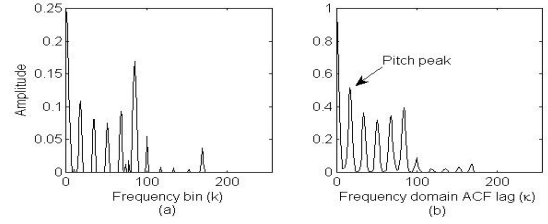


Figure 4: a) Preconditioned amplitude spectrum (corresponding to that in Fig. 2(a)), b) Autocorrelation of the amplitude spectrum in (a).

the contribution of noise and spurious peaks which makes it robust especially to additive noise (as shown in Fig. 6(f)). A 10-15% clipping is seen to be safe and beneficial. Finally, a simple yet useful flattening operation defined by Eq. (3) is employed on the clipped spectrum $|X_C(k)|$ that produces the ultimate preconditioned spectrum $|X_P(k)|$.

$$X_P(k) = 0.5 |X_C(k)| (1 - 0.5 |X_C(k)|) \quad (3)$$

where $|X_C(k)|$ is normalized to unit amplitude before applying Eq. (3). The factor $(1 - 0.5 |X_C(k)|)$ when multiplied with $(0.5 |X_C(k)|)$ has the effect of emphasizing the lower-amplitude harmonics and deemphasizing the higher-amplitude harmonics resulting in more flattened spectrum. Since the spurious components are clipped out before the flattening operation, it essentially has the effect only on the harmonics. As seen in Fig. 6(g), the preconditioned spectrum now takes the shape of a harmonics train.

3.2. Autocorrelation in the Spectral Domain

The autocorrelation in the spectral domain is determined by Eq.(4) as

$$\Psi(\kappa) = \frac{1}{L} \sum_{n=0}^{L-1} X_P(n)X_P(n + \kappa) \quad (4)$$

where L is the length of integration window. We used the first 256 frequency lags. The pitch is then calculated from the frequency lag of the first peak. For higher pitched speech (e.g. female speech), the first peak is, in fact, the only peak in the search range which is from 50~400 Hz (as in Fig. 6(h)). For lower pitched speech (e.g. male speech), on the other hand, if the spectrum is considerably flat there will be multiple peaks in the search range with decreasing amplitudes. However, if the dominant harmonic is preceded by harmonic(s) with much lower amplitude, the first peak (in $\Psi(\kappa)$) might not be the

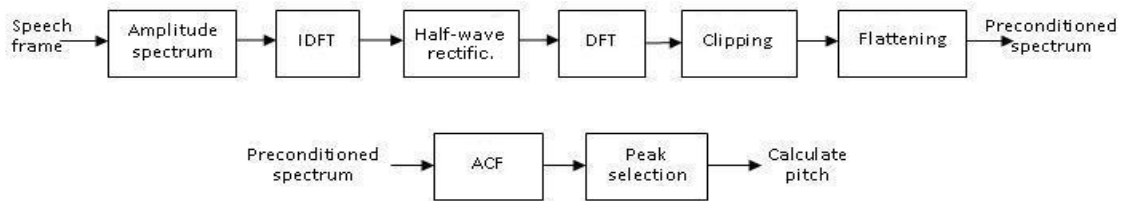


Figure 5: Block diagram of the proposed method.

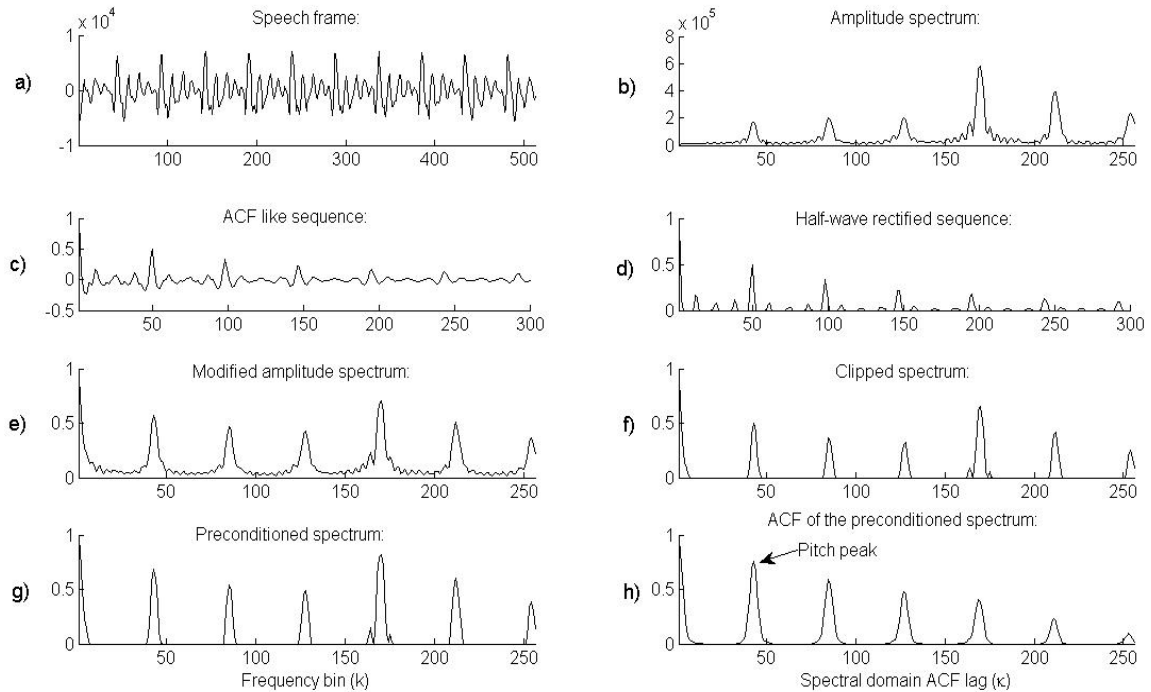


Figure 6: Illustration of the steps described in the block diagram in Fig. 5

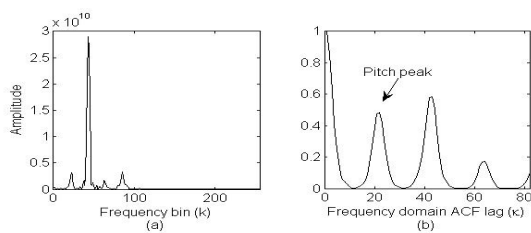


Figure 7: a) Power spectrum with strong formant. b) Autocorrelation of the spectrum after preconditioning.

largest one, as seen in Fig. 7. As we have already mentioned, determining the first peak is only important in this method. It is regardless whether it is maximum or not. The first peak is decided by the minimum value of the autocorrelation lag κ corresponding to the peaks with greater than 40% of the amplitude of the largest peak. The Voicebox [10] implementation of a robust peak-picking algorithm is used in this work.

4. Experiments and Results

The performance of the proposed method is examined on natural speech spoken by four Japanese female and four male speakers. Speech materials are 11 sec-long sentences spoken by every speaker sampled at 10 kHz rate taken from NTT [11] database. The reference file is constructed by computing the fundamental frequencies every 10 ms using a semi-automatic technique based on visual inspection. White Gaussian noise is used to corrupt the speech signal. Pitch estimation error is calculated as the difference between the reference and estimated fundamental frequency. Gross Pitch Error (GPE) is mainly used as a measure of errors in estimating pitch frequency. If the estimated pitch for a frame deviates from the reference by more than 10%, we recognize the error as GPE. The possible sources of GPE is pitch doubling, halving, inadequate suppression of formants as to affect the estimation. The number of GPEs found in determining the pitch using NCCF (without the candidate refinement part in [3]), YIN, and the proposed method are summarized in Table 1. The NCCF is very similar to the ACF, but is better able to follow the rapid changes in pitch and amplitude. The Matsig

(a Matlab library for signal processing) implementation of YIN algorithm is used here [12] for the comparative study without any changes. The routine is verified to work properly. Pitch is determined from every 51.2 ms frame after weighting by rectangular window at 10 sec interval. The pitch range is set to 50~400 Hz. The integration window size for YIN is 31.2 ms (=51.2 ms - 20 ms). The number of DFT point is 2048. The proposed method searches the pitch within 11~82 lag value in the spectral domain ACF. A 15% clipping in the amplitude spectrum is used for all speech signal at all SNR conditions. The numbers in parenthesis in the first column represent the number of voiced frame spoken by the respective speaker. Average

Table 1: Number of gross pitch errors for four female and four male speakers at different SNR conditions.

Spkr	Method	Clean	20 dB	10 dB	5 dB	0 dB	-5 dB
F1 (581)	NCCF	3	3	5	8	28	101
	YIN	4	5	22	53	128	235
	PRO	1	1	1	1	6	39
F2 (554)	NCCF	0	0	0	2	33	99
	YIN	5	6	30	117	188	296
	PRO	0	0	0	0	5	15
F3 (594)	NCCF	6	7	7	8	19	98
	YIN	11	13	20	101	176	281
	PRO	0	0	0	1	4	18
F4 (570)	NCCF	2	2	3	7	17	87
	YIN	6	7	17	39	123	251
	PRO	0	0	1	0	3	20
M1 (541)	NCCF	8	8	8	8	30	89
	YIN	11	12	13	23	82	166
	PRO	3	3	4	5	16	53
M2 (572)	NCCF	27	24	25	30	49	138
	YIN	41	39	45	61	100	183
	PRO	9	9	9	17	35	100
M3 (545)	NCCF	2	2	2	2	4	32
	YIN	2	2	5	16	42	123
	PRO	1	1	1	1	2	15
M4 (674)	NCCF	7	7	5	5	15	68
	YIN	7	7	10	23	71	185
	PRO	2	2	2	3	8	29

number of GPEs for female and male speakers are shown in Figs. 8 and 9, respectively. From the figures and Table 1, it is obvious that the proposed method outperforms the NCCF and YIN methods. The YIN method is seen to perform poorly in noisy conditions because the threshold 0.1 used in the method is more suitable for clean and strongly voiced cases. For highly corrupted speech, pitch halving is observed to take place frequently. The proposed method performs much better for female voice than male voice because of wider harmonics in the frequency domain.

5. Conclusions

Accurate pitch estimation is a difficult problem in speech analysis especially in noisy environments. A common problem is that the estimated pitch is one octave lower or upper than the actual pitch. The proposed technique has reduced the occurrence of octave errors significantly even in highly noisy conditions. The improvement comes out due to the nature of the ACF function. In spectral domain ACF, we always determine the first peak which is not the case in other two methods.

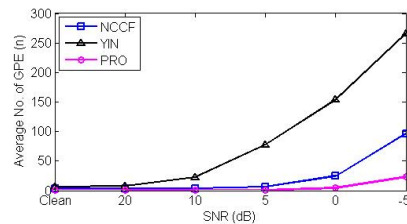


Figure 8: Average gross pitch errors for four female speakers at different SNR conditions.

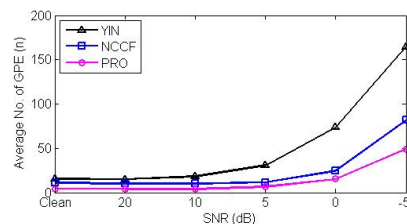


Figure 9: Average gross pitch errors for four male speakers at different SNR conditions.

6. Acknowledgments

This work has been supported by the Japan Society for the Promotion of Science.

7. References

- [1] Hess W. J., Pitch Determination of Speech Signals. Berlin, Germany: Springer-Verlag, 1983.
- [2] Rabiner L.R., Shafer R.W., Theory and Applications of Digital Speech Processing, Prentice Hall, 2010.
- [3] Talkin D.(Ed.), A robust algorithm for pitch tracking RAPT, in Speech Coding and Synthesis, Elsevier, 495-518, 1995.
- [4] Liu D. and Lin C., "Fundamental frequency estimation based on the joint time-frequency analysis of harmonic spectral structure," IEEE Trans. Speech and Audio Processing, 9(6):609-621, 2001.
- [5] Shimamura T. and Kobayashi H., "Weighted autocorrelation for pitch extraction of noisy speech," IEEE Trans. Speech and Audio Processing, 9(7):727-730, 2001.
- [6] Ross M. J., Shaffer H. L., Cohen A., Freudberg R., and Manley H. J., Average magnitude difference function pitch extractor, IEEE Trans. Acoust., Speech, Signal Processing, 22:353-362, 1974.
- [7] Cheveigne A. and Kawahara H., "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., 111(4):1917-1930, 2002.
- [8] Hasan M. K., Hussain S., Hossain M.T., Nazrul M. N., "Signal reshaping using dominant harmonic for pitch estimation of noisy speech," Signal Processing, 86:1010-1018, 2006.
- [9] Markel J., The SIFT algorithm for fundamental frequency estimation, IEEE Trans. Audio Electroacoust., 20:367-377, 1972.
- [10] D. M. Brookes, VOICEBOX: A speech processing toolbox for MATLAB, 1997. Online: <http://www.ee.imperial.ac.uk/hp/staff/dmb/voicebox/voicebox.html>, Accessed on April 30, 2010.
- [11] "Multilingual Speech Database for Telephony, NTT Advance Technology Corp., Japan, 1994.
- [12] Online: <http://www.sourceforge.net/projects/matsig>, Accessed on April 30, 2010.