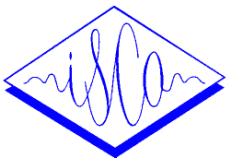
ISCA Archive
http://www.isca-speech.org/archive



3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

A SPECTRAL AMDF METHOD FOR PITCH EXTRACTION OF NOISE-CORRUPTED SPEECH

JaeYeol RHEEM*, MyungJin BAE** and SouGuil ANN*

* *Dept. of Electronics Engr., Seoul National University, Seoul, 151-742, KOREA*
** *Dept. of Telecommunication Engr., SoongSil University, Seoul, 156-743, KOREA*

ABSTRACT

In this paper, we propose a pitch extraction method based on the spectral average magnitude difference function (SAMDF) which is defined as the AMDF of the log-magnitude spectrum of speech. Since the SAMDF is defined on the spectrum of speech signal, the nulls of SAMDF are not affected seriously by the local peaks caused by the additive noise. Furthermore, the proposed method does not need to use any kind of spectral flattening method approach. Experimental result shows that the proposed method is effective for noise-corrupted speech signal and it is adequate for the pitch extraction of female's and child's speech signal.

Keywords: *Pitch Extraction, Spectral AMDF.*

10.21437/Eurospeech.1993-457

1. INTRODUCTION

Over two decades the problem of obtaining an accurate and reliable pitch extraction method has been extensively studied for its various application areas such as speech analysis, speech recognition, speaker verification and identification, and efficient speech coding systems [1]. Generally pitch extraction is known to be a difficult task, even in a noise-free environment. In a noisy environment, the difficulty increases.

Two conventional problems in this field are the effects of noise and formants. To reduce the influence of the formants that results in the nonflat spectral envelope of speech spectrum, spectral flattening methods such as various clipping techniques [1-3], LPC inverse filtering [4][5], and bandpass lifter banks [6] have been investigated. To reduce the effect of noise, frequency-domain techniques that usually utilize the harmonic property of speech spectrum have been developed. For example, they are the harmonic sum and harmonic product technique [7], the peak-picking algorithm [8], the peak-valley detection technique [9], and so on. Although these methods have proven somewhat successful, they are corrupted by the nonflat spectral envelope and they suffer from occasional doubling and halving of the fundamental frequency measurement [5].

In this paper, we propose a pitch extraction method based on the spectral average magnitude difference function (SAMDF). Since the SAMDF is defined as the AMDF of the log-magnitude spectrum of speech data, the local peaks caused by noise in the spectrum are smeared out in the SAMDF and the SAMDF shows deep nulls at the delays corresponding to the multiples of the harmonics of the data spectrum. Furthermore, the suggested method does not need to use any kind of spectral flattening approach to reduce the nonflat spectral envelope of the voiced segment of speech signal. Experimental results are presented for noise free and 0 dB cases. It is shown that the proposed method is effective for noise-contaminated speech signal. It is adequate for the pitch extraction of female's and child's speech and especially, telephone line speech.

2. DISCUSSION ON THE SAMDF

The proposed method is mainly based on the SAMDF(Spectral Average Magnitude Difference Function). In this section, the SAMDF is defined and its properties are discussed.

The SAMDF is defined in the frequency-domain by the following relation, like the AMDF in the time-domain [10],

$$S_d = \frac{1}{N}\sum_{k=0}^{N-1}|X_k - X_{k-d}|, d = 0,1,\ldots,N-1 \qquad (1)$$

where X_k is the N-point log-magnitude spectrum of speech signal $x(n)$, and d is the delay index. The SAMDF, S_d is always zero at delay zero and shows deep nulls at the delays corresponding to the multiples of the harmonics of the spectrum for voiced segment. The deepest null corresponds to the pitch frequency or fundamental frequency of the segment. Since the harmonic structure of log-magnitude spectrum of speech signal, especially noise-corrupted speech signal, is not distinctive in the high frequency region as in the low frequency region, the summation range of the SAMDF in eq. (1) can be reduced. Even in this case, the

positions of nulls of SAMDF do not change in the interesting range of delay and the required computation is reduced.

For pitch extraction, it is not necessary to compute the entire delay range of the SAMDF for each segment of speech data. Since values of pitch frequency generally fall within the range of approximately 66-400 Hz(corresponding to 2.5-15 ms) [5], values of delay falling within such possible range of 2.5-15 ms are computed.

Examples of the SAMDF of possible searching range of delay for several segments of speech are shown in Fig. 1. The dashed lines in Fig. 1 show the SAMDFs of the LPC residual signal of the data segments. The LPC residual signal is obtained from an inverse filtering formed by the 10th-order LPC coefficients that are computed by Burg's method [11]. Considering the SAMDF of entire delay range, it is observed that the SAMDF of the speech signal approximately reflects the spectral envelope of the speech spectrum, while that of the residual signal shows relatively flat envelope that results from the spectral flattening process of LPC inverse filtering.

Comparing the two SAMDFs of the speech and the residual signals, the null positions of them are identical. Within the possible search range of delays, it is observed that the deepest null of SAMDF of the data is always more remarkable than that of the residual signal. Since the envelope of the SAMDF of the speech signal follows the spectral envelope of the speech spectrum approximately in the half way and the interesting search range is in the region of the first formant, the deepest null corresponding to the pitch frequency is the more distinctive as the slope to the first formant frequency is the steeper. Thus it is easier to decide the pitch frequency in the SAMDF of the speech signal than in the SAMDF of the residual signal. This means it is not necessary to use any kind of spectral flattening approach when the pitch extraction is based on the SAMDF.

In Fig. 2, the SAMDF of 0dB noisy speech signal is shown. Since the SAMDF is defined on the log-magnitude of the speech spectrum and the contributions of the noise to the spectrum have no coherent structure [1], the positions of the nulls of the SAMDF are not affected seriously by the additive noise. Thus the decision logic for the pitch extraction can be simple and the phenomena of doubling and halving of the pitch are less occurred, comparing with other pitch extraction methods.

Since the SAMDF utilizes the harmonic structure of speech spectrum and it does not require additional spectral flattening approach, it is effectively useful for the pitch extraction of bandlimited speech signal such as telephone line speech signal.

3. DESCRIPTION OF THE PROPOSED PITCH EXTRACTION METHOD

A. Overview

An overall block diagram of the proposed pitch extraction method is shown in Fig. 3. The input speech is transferred through as antialiasing low pass filter(f_c=4kHz) and sampled at 10 kHz by a 16-bit A/D converter. The algorithm operates on 50 percent overlapped segments of 51.2 ms duration. Successive segments are hamming windowed and processed by a 1024-point FFT routine to obtain the log-magnitude spectrum of the segment.

The position of the maximum of the log-magnitude spectrum is used to hard reject unvoiced segment and calculate the reduced search range of delay of the SAMDF.

For voiced segments, the SAMDF is computed within the pre specified delay range. A decision procedure is applied to the nulls of the SAMDF after the null-picking, where the doubling and halving of the pitch are considered, based on the pitch of the previous segment and the average pitch before the present segment.

B. Hard Rejection of Unvoiced Segment

In the spectrum of voiced segments, it is observed that the first formant is dominant. While in the spectrum of unvoiced segments, the zero-frequency component is dominant, or the dominant formant is located in the higher frequency region than that of voiced segments. Since the position of the maximum of the spectrum can be estimated as the first formant location of voiced segments, the hard rejection of unvoiced segments can be implemented by considering the position of the maximum of the spectrum. Here the hard rejection means that the segments that are sure to be unvoiced are rejected.

The decision logic for hard rejection is as follows: if the position of the maximum of the spectrum corresponds to the zero frequency, or the frequency of more than 2 kHz, the segment is classified into unvoiced segment and rejected from the pitch extraction procedure.

C. Reduction of the Search Range of Delay

To reduce the burden of computation of the SAMDF, the search range of delay can be modified. As the position of the maximum of the log-magnitude spectrum can be used as an estimate of the first formant for the voiced segment, the upper limit of the search range can be replaced by it, if the resulting search range falls within the possible search range. Since if there exists more than one harmonic within the new search range, it is enough to extract the pitch. However, sometimes it is observed that the position of the maximum of the spectrum is so closed to zero-frequency position that there exists less than one harmonic within the resulting search range, especially in the voiced segment of female's and child's speech. In that case, two times of the maximum position is adequate for the upper limit to the new search range of delay.

Let the lower and the upper limit of the range be d_L and d_U. They correspond to 15 and 2.5 ms, respectively. The range is reduced to $[d_L, d_M]$, where

$$d_M = \min(d_U, \max(P_{\max}, 25)) \quad (2)$$

and P_{max} is the position of the maximum of the spectrum and the constant 25 is empirically chosen.

D. Decision Logic

The decision logic is simple. After null-picking of the SAMDF, the deepest null is found, and the pitch is calculated using the 2nd-order polynomial interpolation[4]. If the deepest null does not exist, the segment is decided to be unvoiced and the pitch is set to be zero. Based on the pitch of the previous segment and the average pitch before the present segment, the test of doubling and halving of the pitch is processed. If a doubling, or a halving occurs, the deepest null is re-determined within the modified range of delay where the correct pitch can exist.

4. EXPERIMENTAL RESULT

For test procedure, a database was constructed with the eye-detected pitch frequency. The speech material used for testing consisted of five utterances spoken by each of three males and two females. The five utterances were selected such that a wide variety of Korean phonemes would be included. Total 4602 overlapped frames(2135 frames from females and 2467 frames from male speakers) were processed by the suggested method. Comparing with the eye-detected pitch frequency, gross errors of 3.8% for noise free speech signal and 8.2% for 0dB speech signal were obtained. The result is shown in Table I. As other frequency domain pitch extraction method shows [6], the result shows some superiority on female speech.

5. CONCLUSION

We have presented a pitch extraction method based on the SAMDF. Since the SAMDF is defined on the spectrum of speech signal, the nulls of SAMDF are not affected seriously by the local-peaks caused by the additive noise. Furthermore, the proposed method does not need to use any kind of spectral flattening approach. The decision logic used in pitch extraction is simple. Experimental result with male and female speech shows its noise immunity. We think that the proposed method is adequate for the pitch extraction of female's and child's speech signal, especially telephone line speech.

REFERENCES

[1] L.R. Rabiner and R.W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[2] M.M. Sondhi, "New methods of pitch extraction," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, No. 2, pp.262-266, June 1968.

[3] L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-25, No 1, pp. 24-33, Feb. 1977.

[4] J.D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, vol. AU-20, pp. 367-377, Dec. 1972.

[5] C.K. Un and S. Yang, "A pitch extraction algorithm based on LPC inverse filtering and AMDF," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-25, No. 6, pp. 565-572, Dec. 1977.

[6] M. Lahat, R.J. Niederjohn, and D.A. Krubsack, "A spectral autocorrelation method for measurement of the fundamental frequency of noise-corrupted speech," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-35, No. 6, pp. 741-750, June 1987.

[7] A.M. Noll, "Pitch determination of human speech by the harmonic product spectrum, the harmonic sum spectrum and a maximum likelihood estimate," in *Proc. Symp. Comput. Processing Commun.*, pp. 779-798, Apr. 1969.

[8] S. Seneff. "Real time harmonic pitch detector," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-26, pp. 358-365, Aug. 1978.

[9] T.V. Screenivas and P.V.S. Rao, "Pitch extraction from corrupted harmonics of the power spectrum," *J. Acoust. Soc. Amer.*, vol. 65, pp.223-228, Jan. 1979.

[10] M.J. Ross, H.L. Shaffer, A Cohen, R. Freudberg, and H.J. Manley, "Average magniitude difference function pitch extractor, *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-22, No. 5, pp. 353-362, Oct. 1974.

[11] J.P. Burg, "Maximum entropy spectral analysis," presented at *37th Annual Int. SEG Meeting*, Oklahoma City, 1967.

[12] L.R. Rabiner, M.J. Cheng, A.E. Rosenberg, and C.A. McGonegal, "A comparative performance study of several pitch detection algorithms, " *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. ASSP-24, pp. 399-417, Oct. 1976.

TABLE I. GROSS ERROR(%) BY THE PROPOSED PITCH EXTRACTION METHOD

	Female	Male	Overall
Noise free	3.5	4.1	3.8
O dB	7.4	8.9	8.2

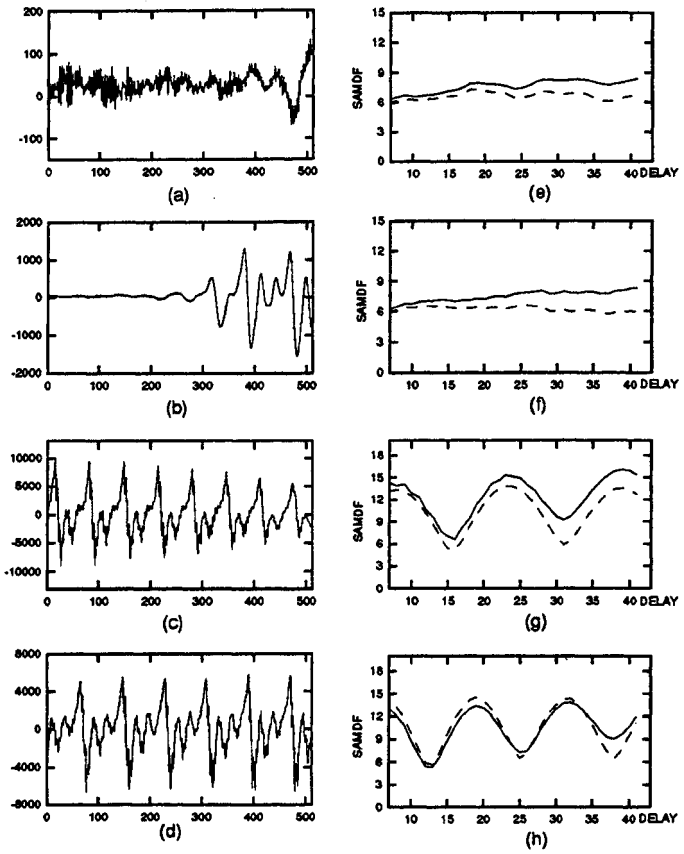


Fig. 1. SAMDF examples of various speech segments; (a)-(d) are speech data, (e)-(h) are resulting SAMDFs where the dashed lines are those of the LPC residual signals.

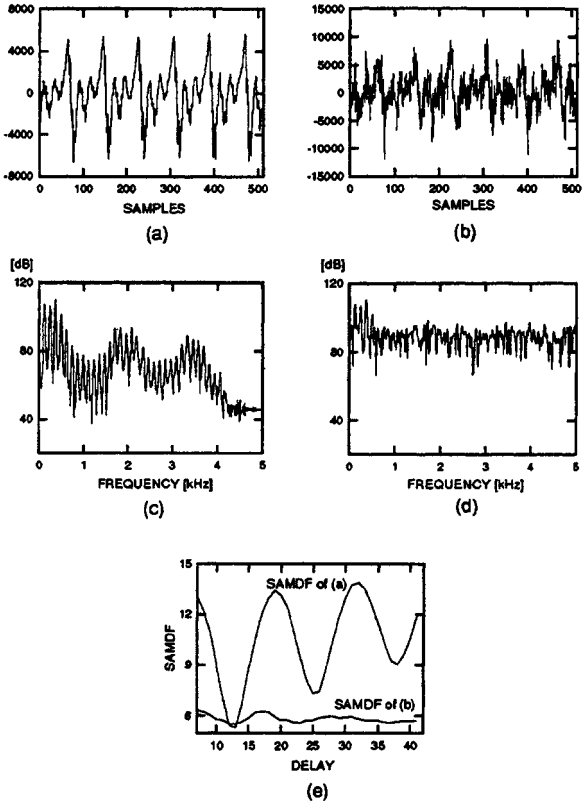


Fig. 2. SAMDF examples of noisy data; (a) clean data (b) 0 dB data, (c) spectrum of (a), (d) spectrum of (b), (e) SAMDFs of (a) and (b).

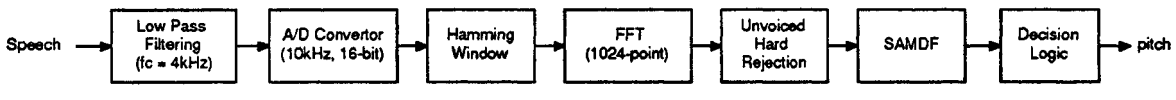


Fig. 3. Block diagram of overviewing the proposed method for pitch extraction.