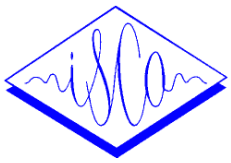

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

PROPOSAL AND EVALUATION OF A NEW SCHEME FOR RELIABLE PITCH EXTRACTION OF SPEECH

Hiroya Fujisaki, Keikichi Hirose and Shigenobu Seto

Faculty of Engineering, University of Tokyo
Bunkyo-ku, Tokyo, Japan

ABSTRACT

Analysis using short frame length is necessary in order to realize correct tracking of time-varying features of quasi-periodic signals such as speech. However, when the frame length is reduced for the analysis of rapidly changing signal characteristics, the analysis results are strongly affected by the position of the frame and sometimes may lead to gross errors. In the pitch extraction schemes using the conventional definition of the short-time autocorrelation function, the value of its peak indicating the fundamental period varies with the frame position. In order to reduce these variations, we present a new definition for the normalized short-time autocorrelation function that does not require the selection of frame length. A new scheme for pitch extraction of speech is proposed that assures high accuracy of results without adjusting the frame length for each speaker. The validity of the proposed scheme is confirmed by the experiments using speech materials recorded by both male and female radio announcers.

1. INTRODUCTION

As is well known, the voice fundamental frequency plays an important role in the transmission of such prosodic information as word accent, sentence structure, discourse structure, and speaker's intention. The automatic extraction of this acoustic parameter (henceforth pitch extraction), therefore, is a task that is quite important in analysis, synthesis, coding, as well as in automatic recognition/understanding of speech, and a number of methods have been already devised. Because of inherent complexities of the speech signal itself, and of the imperfect representation of the signal due to inappropriate choice of analysis methods, however, none of the existing methods are known to work perfectly for a wide variety of voices and environments [1, 2].

Since the speech signal is quasi-periodical and its fundamental frequency gradually changes as a function of time, analysis of shorter frame length is necessary to realize better temporal resolution. In most of the existing methods of pitch extraction based on the conventional definition for the short-time autocorrelation function, however, serious errors may occur especially when using frame length as low as two fundamental periods. This is because the value of the major peak of the autocorrelation function, whose position on the time axis representing fundamental period, varies with the frame position and sometimes becomes smaller than those of other peaks. In order to reduce these variations, we present a new definition for the normalized autocorrelation function whose peak values are exactly equal to 1 for perfectly periodic signals, and propose a new scheme for pitch extraction of speech using this function. The scheme's validity is tested by comparing with other schemes using conventional definitions for the normalized short-time autocorrelation function and autocovariance function.

2. ERRORS DUE TO ANALYSIS FRAME LENGTH AND ITS POSITION

The short-time autocorrelation function, when combined with an appropriate pre-processing, has been shown to be a powerful means to facilitate detection of the fundamental period of a signal, and thus has been widely adopted in many schemes for pitch extraction of speech. When the analysis frame length is fixed, however, the value of its peak indicating the fundamental period varies with the frame position, and this variation may cause errors in pitch extraction. Figure 1 shows

1) short-time autocorrelation function using the conventional definition (for frame lengths L's of a)18ms, b)20ms, c)24ms), and
2) short-time autocovariance
(with integration periods I's of b)10ms, c)14ms, d)44ms),

of the prediction residual obtained by a running PARCOR analysis for a synthesized vowel /a/ with 10ms fundamental period. The peak corresponding to the fundamental period takes its maximum value of 1 irrespective of frame position only when the integration period is equal to an integral multiple of the fundamental period as in the case of panel b). The large variation of the peak values due to the frame position requires elaborate procedures to correctly select the peak for fundamental period from several candidate peaks. The variation also makes it difficult to decide the voiced/unvoiced intervals correctly.

In conventional schemes of pitch extraction, two different lengths of analysis frame are usually selected, i.e., one for male voice and the other for female voice. This technique can reduce gross errors due to inappropriate frame length, but cannot fully cope with the individual variations of a large number of speakers and the temporal variations of fundamental frequency in continuous speech. Adaptive control of the frame length is desirable but is rather difficult to realize.

3. PROPOSAL OF A NEW PITCH EXTRACTION SCHEME

3.1 New Definition of the Normalized Short-time Autocorrelation Function

For a signal $f(t)$ at time t_0 with lag T, we define the normalized short-time autocorrelation function as follows.

<Definition-1> (Figure 2-a))

$$R(t_0,T) \equiv \frac{\int_{t_0-\frac{I(T)}{2}}^{t_0+\frac{I(T)}{2}} f(t-\frac{T}{2})\, f(t+\frac{T}{2})\,dt}{\int_{t_0-\frac{I(T)}{2}}^{t_0+\frac{I(T)}{2}} \left[\{f(t-\frac{T}{2})\}^2+\{f(t+\frac{T}{2})\}^2\right] dt \Big/ 2}, \quad (1)$$

where $I(T)=T$. The frame length is set equal to twice of the time lag T and short-time power of the signal segmented by the frame is used for normalization. If the signal $f(t)$ is perfectly periodic with periodicity of T_0, the following equality holds at $T=nT_0$ (n is an in-

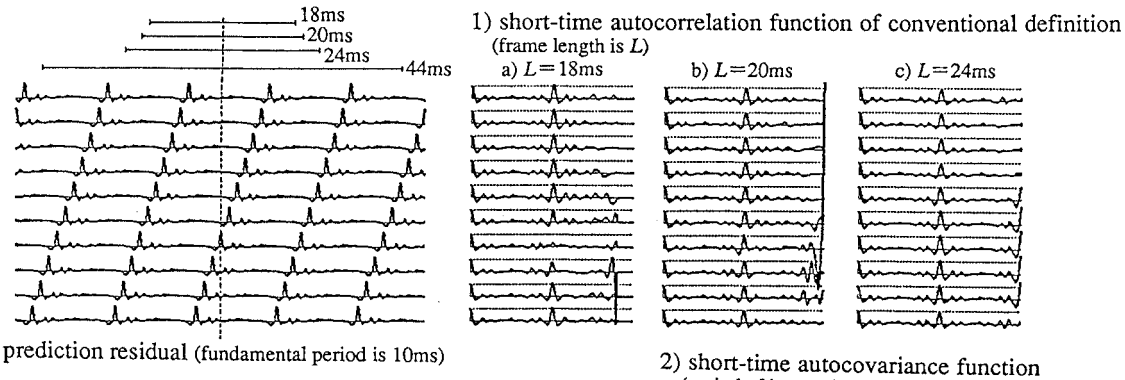

Fig. 1. Normalized autocorrelation functions of the prediction residual obtained by a running PARCOR analysis for synthesized vowel /a/ with 10ms fundamental period. Panels 1) and 2) respectively show the short-time autocorrelation function of conventional definition, and the short-time autocovariance function.

teger) irrespective of the frame position:

$$\frac{\int_{t_0-\frac{I(T)}{2}}^{t_0+\frac{I(T)}{2}} f(t-\frac{T}{2})\,f(t+\frac{T}{2})dt}{\int_{t_0-\frac{I(T)}{2}}^{t_0+\frac{I(T)}{2}} \left[\{f(t-\frac{T}{2})\}^2+\{f(t+\frac{T}{2})\}^2\right] dt \Big/ 2} = 1. \qquad (2)$$

Figure 3 shows the normalized autocorrelation function of the above definition for the prediction residual obtained through a running PARCOR analysis of the synthesized vowel /a/ with fundamental period of 10ms.

3.2 Conventional Definition for the Normalized Short-time Autocorrelation Function

In conventional schemes with fixed frame length L, a normalized short-time autocorrelation function for the signal $f(t)$ at time t_0 with lag T is defined by

<Definition-2> (Figure 2-b))

$$R(t_0,T) \equiv \frac{\int_{t_0-\frac{I(T)}{2}}^{t_0+\frac{I(T)}{2}} f(t-\frac{T}{2})\,f(t+\frac{T}{2})dt}{\int_{t_0-\frac{I(0)}{2}}^{t_0+\frac{I(0)}{2}} \{f(t)\}^2 dt}, \qquad (3)$$

where $I(T)=L-T$.

In order to detect the fundamental frequency of a signal, it is possible to make use of the short-time autocovariance instead of the short-time autocorrelation function defined above. A normalized short-time autocovariance of the signal $f(t)$ at time t_0 with lag T can be defined by

<Definition-3> (Figure 2-c))

$$R(t_0,T) \equiv \frac{\int_{t_0-\frac{I}{2}}^{t_0+\frac{I}{2}} f(t-\frac{T}{2})\,f(t+\frac{T}{2})dt}{\int_{t_0-\frac{I}{2}}^{t_0+\frac{I}{2}} \{f(t)\}^2 dt}, \qquad (4)$$

where I is the fixed period of integration.

3.3 Description of Proposed Scheme (Figure 4)

The proposed scheme consists of three stages: pre-processing, main processing and post-processing. The pre-processing is a stage for separating the source signal from the spectral envelope characteristics. The main processing is for converting the source signal into a time function appropriate for the detection of the fundamental period, and to detect candidates for the fundamental

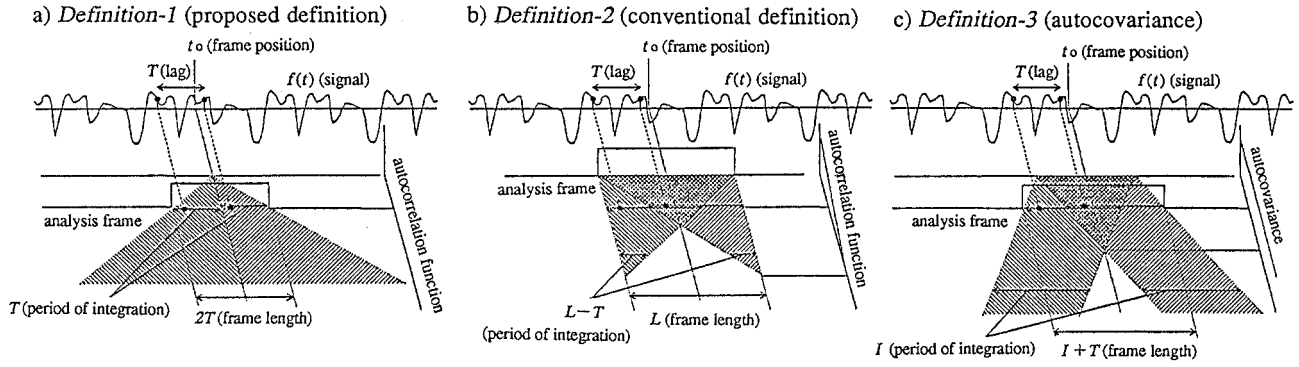

Fig. 2. Three kinds of definition for the short-time autocorrelation function.

prediction residual (fundamental period is 10ms) autocorrelation function

Fig.3. Newly defined normalized autocorrelation function of the prediction residual obtained by a running PARCOR analysis for synthesized vowel /a/ with 10ms fundamental period.

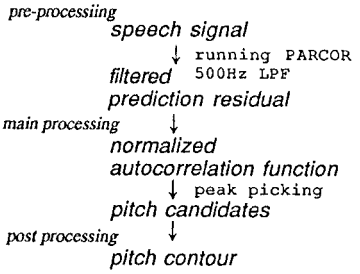

Fig. 4. Algorithm of the proposed scheme for pitch extraction.

period. The post-processing is for selecting the most probable fundamental period.

The major function of the pre-processing stage is a PARCOR analysis of the input speech. Since, in the proposed scheme, the prediction residual should be calculated at each sampling point and fed to the main processing stage, the pre-processing stage performs the running PARCOR analysis using the lattice algorithm, where the partial autocorrelation coefficients are calculated by a running algorithm[4]. The prediction residual is then low-pass filtered at 500 Hz by a linear-phase filter.

The main processing stage is an autocorrelation analysis (*Definition-1*) of the prediction residual, followed by interpolation and peak-picking of the resulting autocorrelation function to select several candidates for the fundamental period at each frame position. The positions of maxima of $R(t_0,T)$ are selected as candidates. In order to increase the accuracy of pitch period measurements, the locations of such maxima are determined after parabolic interpolation of the sampled autocorrelation function.

The post-processing stage performs voiced/unvoiced decision and selection of the most probable fundamental period in the case of voiced decision. These processes are conducted on the basis of peak values of $R(t_0,T)$ taking into account the quasi-continuity of the fundamental frequency.

4. EXPERIMENTAL EVALUATION USING NATURAL SPEECH

Experiments using natural speech were conducted to examine differences between the proposed scheme based on equation 1(*Definition-1*) and schemes based on equation 3(*Definition-2*) and 4(*Definition-3*). For the fixed frame lengths L's in *Definition-2*, 16ms, 24ms, 32ms are selected, and, for the periods of integration I's in *Definition-3*, $L/2$ are selected.

4.1 Speech Materials

The speech materials used for the experimental evaluation were recordings of general weather forecast, read by three male and three female radio announcers. The recorded speech was low-pass filtered at 4.8 kHz and digitized at 10 kHz with 12 bit accuracy for further processing.

4.2 Experimental Evaluation

In order to provide references for evaluation of the pitch extraction schemes, F_0 values for individual periods of the voiced portions are measured by visual inspection of the speech waveform low-pass filtered at 500 Hz. The F_0 value at an arbitrary frame position is obtained by interpolation. The performance of each scheme is evaluated by comparing its results with the reference values. The performance of each scheme is evaluated only for the portions of the speech signal judged as voiced by visual inspection.

In order to reduce the influence of frame positioning, the analysis frame is shifted with intervals of 1ms. Among several candidates for the current analysis frame position, the candidate corresponding to the value closest to that obtained by visual inspection is selected as the most probable candidate.

The group of candidates with the maximum peak values of the autocorrelation functions for every frame position shall be called *the first candidate group* and that with the second maximum peaks called *the second candidate group* and so on. Reliable pitch extraction is possible without complicated processing at the post-processing stage, if the most probable candidate for each frame position belongs to the first candidate group. Figure 5 shows the percentage of the most probable candidates in each group. In the case of *Definition-1*, more than 90% of the first candidates are taken as the most probable candidates. In the case of *Definition-2*, similar results were obtained if the frame length is appropriately selected, but if the frame length is inappropriately selected, a large percentage of the most probable candidates fails to belong to the first candidate group.

The peak value of the normalized autocorrelation function is often used as a measure for periodicity of a signal to perform voiced/unvoiced decision. It is desirable that the normalized autocorrelation function of the most probable candidate takes a value close to 1, but never exceeds it. Figure 6 shows the distributions of the peak values of the normalized autocorrelation for the most probable candidates. Panels a), b), c) are respectively for the schemes based on *Definition-1*, *Definition-2*, *Definition-3*. In the case of *Definition-2*, the autocorrelation function decreases as the lag being larger. This effect is compensated in the figure by a appropriate weighting in order to have a good comparison. In the case of *Definition-2* and *Definition-3* with short frame lengths, the value of the normalized autocorrelation function sometimes exceeds 1.

5. CONCLUDING REMARKS

In order to cope with errors due to frame position, a new scheme of pitch extraction of speech has been proposed. Experiments using natural speech are conducted between the proposed method and the conventional methods to compare and examine the differences. It was found that

1) the percentage of the most probable candidate in the first candidate groups obtained by the proposed scheme is as high as those obtained by the conventional schemes with appropriate frame lengths, and

2) the peak values of the normalized autocorrelation function for the most probable candidate never exceeds 1 and are usually close to 1 for the proposed scheme.

These experimental results show that the proposed scheme has an advantage in the automatic extraction of the fundamental frequency.

REFERENCES

[1] 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 Processing, Vol. ASSP-24, pp. 399-418, 1976.

[2] W. Hess, *Pitch Determination of Speech Signals*, Springer-Verlag, Berlin, 1983.

[3] H. Fujisaki, "Automatic Extraction of fundamental Period of Speech by Auto-Correlation Analysis and Peak Detection," J. Acoust. Soc. Am., Vol. 32(A), p. 1518, 1960.

[4] H. Fujisaki, K. Hirose and K. Shimizu, "A New System for Reliable Pitch Extraction of Speech," Proc. ICASSP 87, 34.16.4, 1987.

[5] H. Fujisaki, K. Hirose and S. Seto, "A New Pitch Extraction Method of Speech with Minimal Influence of Analysis Frame Position, Record Fall Meeting of Acoustical Soc. Japan, pp. 249- 250, 1989.

[6] H. Fujisaki, K. Hirose and S. Seto, "A Method for Pitch Extraction of Speech with Reduced Errors Due to Analysis Frame Position," Trans. of the Committee on Speech Research, Acoust. Soc. Japan, SP89-69, 1989.

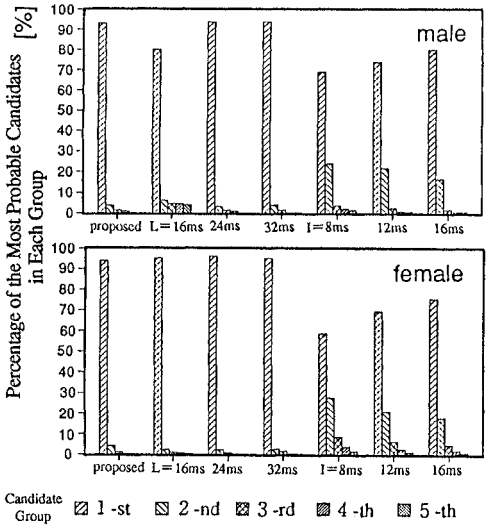

Fig. 5. Percentage of the most probable candidates in each group.

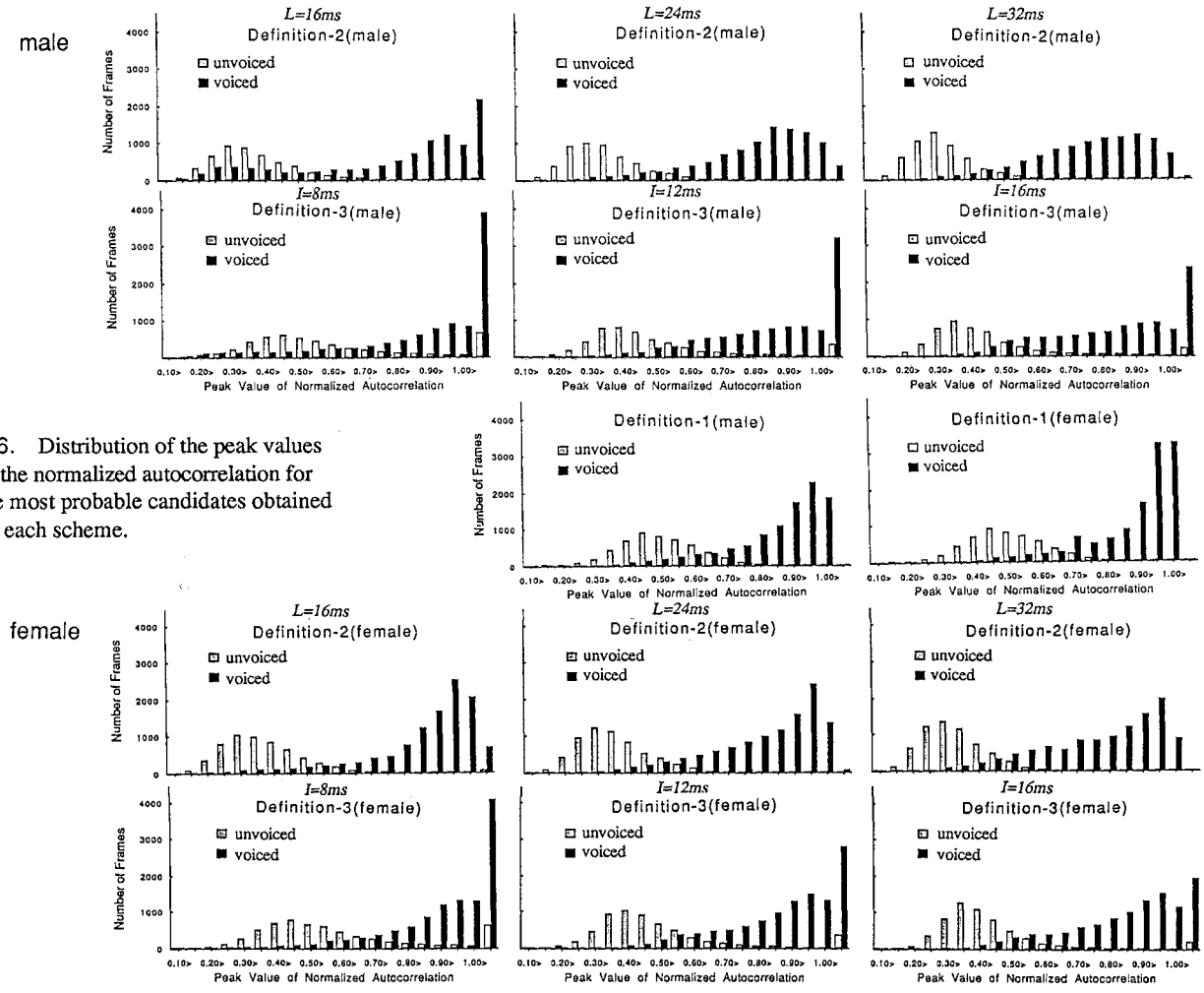

Fig. 6. Distribution of the peak values of the normalized autocorrelation for the most probable candidates obtained by each scheme.