AN INTELLIGENT PITCH TRACKER BASED ON FORMAL LANGUAGE THEORY AND PHONETIC KNOWLEDGE

M. Stamenkovic, J. Bakran

Abstract: This work presents a new technique for construction of an accurate and efficient pitch tracker. The proposed model is based on formal language theory and phonetic knowledge. The phonetic knowledge consists of: acoustic description of speech segments and defined properties of speech signal.

INTRODUCTION

For both speech synthesis and analysis (recognition) the pitch period (fundamental frequency - F0) is one of the essential features. Unfortunately, the accurate and exact estimation of F0 is extremely difficult to achieve because of considerable influences of different kind of noise (variable spectral shape of segments) that mask inherent periodicity of voiced periods [1,2]. So far many pitch tracking algorithms have been published [1] and different speech processing techniques have been proposed. We can classify them into three main categories:

1) The algorithms which utilize the time-domain features of speech signals;
2) The algorithms which utilize the frequency-domain properties of speech signals;
3) The algorithms which utilize both the time and frequency domain properties of speech.

The methods in the time domain relate to pick-valley measurements, zero-crossing, autocorrelation function, short-time energy etc. The assumption for these methods is that formant structures were previously removed. The algorithms utilizing frequency-domain properties of speech extract F0 as a spectral component of FFT analysis. The third group of methods benefits the features of both, time and frequency domain analysis. All three categories above include different postprocessing techniques for smoothing and averaging F0 contour.

In this paper we adopt a hybrid approach for estimating F0 and proposed a simple way how to incorporate the different kind of phonetic knowledge in the model. The implemented phonetic knowledge considerably improves detection of voiced/unvoiced segment at the end of the utterances and in the noisy conditions.

THE ALGORITHM

Speech signal is low pass filtered at
4.5 kHz (4-th order Chebyshev's filter) and sampled at $F_s=10$ kHz through 12 bits A/D converter. The digitized speech is then uniformly divided into 10 ms frames (100 samples per frame). The pitch determination is realized into 4 steps:

1. Calculation of U/V string

Let $A=\{U, S, V\}$ be an alphabet and let $s_{t+n,N}$ denotes speech frame that start at $t$-th sample and occupies next $N$ samples of speech signal $s(t)$. Each speech segment is assigned an element $x \in A$ by calculated probabilities:

$$p_t(U) = \begin{cases} 1 & \text{for } (R) \land (E) \\ C_1 \exp[-R/R_0] & \text{otherwise} \end{cases}$$

(1)

$$p_t(S) = \begin{cases} 1 & \text{for } E \land (E) \\ C_2 \exp[-E/E_c] & \text{otherwise} \end{cases}$$

(2)

$$p_t(V) = \begin{cases} 1 & \text{for } (R) \land (E) \\ C_3 \exp[-R/R_0] & \text{otherwise} \end{cases}$$

(3)

where $E_c$ represents mean energy of silences frames and $R_0$ is mean ratio of zero crossing and energy of frame representing voiceless sounds. The value for $E$ and $R$ are obtained:

$$E = \frac{1}{100} \sum_{j=1}^{100} |s(t+j)|$$

(4)

$$R = ZC / E$$

(5)

where $ZC$ is number of the zero crossing in the frame. The constants $C_1$, $C_2$ and $C_3$ are determined empirically. After the segmentation, we obtain string

$$\alpha = x_1 \ldots x_j \ldots x_m, \quad x \in A$$

(6)

2. Calculation of raw FO

After determination of voiced segments, the pitch period is determined by short time autocorrelation function of central clipped signal [4]:

$$ACF(k) = \sum_{n=0}^{N-1} y(t+n) y(t+n+k)$$

(7)

where $y(j)$ is clipping transformation by [4]. The autocorrelation of voiced speech has salient picks that indicates the periodicity of the signal. The pitch period is calculated as a lag from $ACF(0)$ of the first $ACF(i)$ ($i=1..N$) that exceeds predefined threshold. We used adaptive thresholding that lowers the thresholds around $\pm 1$ ms of previous pitch period for a factor of 1.5. The autocorrelation frame length was chosen to be $N=200$ in order to analyze low voices as 50 Hz. The pitch frequency defined in such manner is:

$$F_0 = \frac{F_s}{T}$$

(8)

4. Postprocessing

In order to compensate false U/V decision the context condition of natural U/V sequences an FST translation schemata of $\omega A^* \cup B$ is defined as follows:

$$\text{FST}=(Q, \Sigma, \Delta, \xi, \rho, F)$$

$Q = F = \{1, 2, 3, 4, 5\}$

$$\Sigma = \Delta^* = A \cup B; \quad q_0 = 1$$

$$\delta(q, x) = (r, y), \quad q, r \in Q, \quad x \in \Sigma, \quad y \in \Delta^*$$

$$q \rightarrow y$$

1 5 4 V 2 V 3 \ldots V 1 3 2 S 1 V S 4 8 1 6 5 S 1 US

EUROSPeECH '89, Paris, France, September 1989

1471
If two successive frames $s_{(l+N)}$ and $s_{(l+2N)}$ are voiced and step of increment/decrement is greater than 20 Hz, then a nonlinear smoothing technique is applied: a search for first voiced candidate that satisfies the condition:

$$|s_{(t,t+N)} - s_{(t+kN,t+(k+1)N)}| \leq 20 \quad (9)$$

is made and a mean of k-th next frame and the first frame is assigned to the second one.

RESULTS

The accuracy of the frequency determination depends on sampling frequency $F_s$ [Hz] and estimated $F_0$ [Hz]:

$$\Delta F [Hz] = \frac{F_0^2}{F_s + F_0} \quad (10)$$

Calculated pitch contour for "Kako ste vi?" ("How are you?") is presented in fig. 1.

Note: This algorithm is implemented on PC XT/AT computer in PASCAL and FORTRAN and the object program for performance comparison is available on request.

REFERENCES


