Robust HNR-based Closed-loop Pitch and Harmonic Parameters Estimation

Alexander Pavlovets, Alexander Petrovsky

Department of Computer Engineering, Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

palex@bsuir.by

Abstract

An important problem in speech coding framework is model parameters estimation. In most cases parametric speech coding methods do not preserve shape of speech waveform. This fact implies straightforward parameters estimation and analysis-by-synthesis method is hardly used.

A novel analysis-by-synthesis parameters estimation method in speech coders based on harmonic models presented. We introduce improved speech model based on robust harmonic and noise components separation. The separation is performed with usage of Pitch Tracking Modified DFT (PTDFT). Harmonic parameters and pitch frequency are estimated simultaneously in a closed-loop manner based on Harmonic-to-Noise Ratio (HNR).

Index Terms: speech coding, pitch, harmonic analysis, speech decomposition.

1. Introduction

Identification of voiced regions of speech is a main task in parametric speech coders. Decision of voicing can be performed searching for maximum voicing frequency [1] or voicing decision for each frequency band [2]. The problem is that the decision is always binary one i.e. the region is declared voiced or unvoiced. From the speech production point of view more accurate assumption is to consider voiced speech as a sum of voiced and noise-like components without identification of voiced/unvoiced regions. In [3] speech decomposition method was proposed which considers voiced and noise-like components of the excitation signal present in the whole speech band. The idea of the work is to use an iterative algorithm based on Discrete Fourier Transform (DFT)/Inverse Discrete Fourier Transform (IDFT) pairs for noise component estimation. Another method of speech decomposition with use of Pitch Scaled Harmonic Filter (PSHF) is presented in [4].

As the methods described in [3,4] our method considers voiced and noise components present in whole band. Pitch-Tracking modification is applied to standard DFT in order to provide spectral analysis in harmonic domain rather than frequency domain. Pitch frequency and harmonic parameters are accurately estimated using analysis-by-synthesis technique. We found this new feature advantageous in comparison to existing harmonic speech coding methods. In our case voiced component is defined as a sum of harmonically related sinusoids with time-varying amplitudes and phases. Decomposition is performed in time domain and noise-like component is defined as a difference between original speech and synthetic voiced component.

2. Harmonic analysis of speech

Speech signal can be presented as a sum of harmonic and residual (noise-like) parts:

$$s(i) = h(i) + r(i)\,, \tag{1}$$

where $h(i)$ – voiced component and $r(i)$ is residual signal.

Harmonic component of speech signal can be defined as:

$$h(i) = \sum_{k=1}^{K} A_k \cos(k \sum_{i=0}^{N-1} \frac{F_0(i)}{F_s} + \theta_k)\,, \tag{2}$$

where: A_k – amplitude of k-th harmonic, K – number of all harmonics, $F_0(i)$ – instantaneous pitch, θ_k – initial phase of k-th harmonic, F_s – sampling frequency, N – frame length.

Identifying frequency-modulated sinusoids in a signal is known to be a tough challenge for classical linear analysis. One way of solution of this problem is suggested in Fan-Chirp Transform-based method [5]. The transform used can be described as the Fourier transform of a warped-time version of the analysis signal.

In our case the core of the harmonic analysis is PTDFT procedure [6]. Modified DFT transform for analysis in harmonic domain is given by:

$$H_n(k) = \sum_{i=0}^{K-1} s_n(i) \exp\left(j \frac{2\pi k i}{F_s} \left(F_0 + \frac{\Delta F_0 i}{2N} \right) \right) w_n(i), \quad j = \sqrt{-1}, \tag{3}$$

where $s_n(i)$ – i-th sample of the n-th frame, F_0 – fundamental frequency, ΔF_0 – fundamental frequency change, $w_n(i)$ – time window.

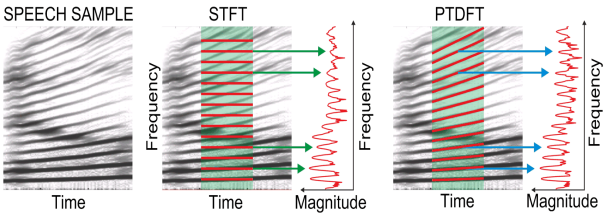

Figure 1: *PTDFT vs. STFT.*

The essence and advances of proposed PTDFT-based method of speech analysis comparing to STFT-based methods can be clearly seen from the Figure 1. Analysing selected region of some speech sample on the spectrogram using PTDFT one can take into consideration prosodical pitch fluctuations and hence provide higher accuracy in short-time harmonic parameters determination.

Non-orthogonal transformation kernel can cause energy leakage to neighboring spectral lines. Time-varying windows are usually used in order to deal with leakage phenomenon. The idea of proposed solution is to design a spectral window with a shape that follows fundamental frequency changes, in fact spectral matching is introduced. Good results can be achieved using Kaiser window as a prototype [7] (Figure 2):

$$w_n(i) = \frac{I_0\left(\beta\sqrt{1-\left[(2x-L_n+1)/(L_n-1)\right]^2}\right)}{I_0(\beta)}, \quad (4)$$

where i=0…N-1, N is window length, β is window parameter, $I_0(\,)$ is zeroth order Bessel function, x is a function enabling time-varying feature, given as:

$$x = \frac{a_{2,n}(N-1-i)^2 + a_{1,n}(N-1-i)}{a_{2,n}(N-1)+a_{1,n}}, \quad (5)$$

where $a_{2,n}$ and $a_{1,n}$ – parameters that provide linear pitch change:

$$a_{1,n} = \frac{2\pi F_0}{F_s},\ a_{2,n} = \frac{2\pi\Delta F_0}{2F_s N}. \quad (6)$$

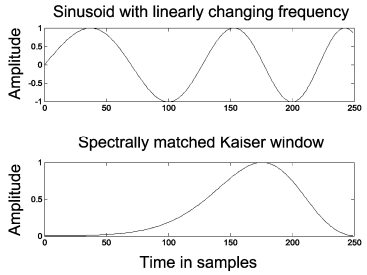

Figure 2: *An example of spectrally matched Kaiser window.*

Accuracy of the harmonic parameters estimation can be measured by *HNR*:

$$HNR = 10\lg\frac{E_H}{E_N}, \quad (7)$$

where E_H – energy of synthesized according to (2) harmonic component, E_N – energy of noise-like component. The noise-like component is defined as a difference between original speech and synthetic harmonic component:

$$r(i) = s(i) - \hat{h}(i). \quad (8)$$

In PTDFT-based analysis of speech a problem of crucial importance, that mainly influences on the accuracy of harmonic parameters determination, is pitch estimation accuracy. Section 3 is devoted to the possible solution of this task.

3. Closed-loop pitch and harmonic parameters estimation algorithm

Harmonic analysis using PTDFT and hence speech decomposition has been shown [6] to be quite robust and accurate if pitch value is accurately determined. The solution proposed here contains simultaneous determination of pitch and harmonic parameters (amplitudes and phases) in a closed-loop way.

First of all robust and reliable method for rough (preliminary) pitch estimation must be used. In our case time-domain pitch determination algorithm (PDA) based on normalized autocorrelation function (NACF) computation in combination with postprocessing based on dynamic programming (DP) method was applied.

During preliminary pitch track estimation firstly the comparison of input speech frame energy with the defined threshold is done. Then after low-pass filtering with cutoff frequency of 1 kHz information about pitch track is obtained by peak-picking of normalized autocorrelation function:

$$\psi(k) = \frac{\sum_{j=1}^{N} s_j s_{j+k}}{\sqrt{\sum_{j=1}^{N} s_j^2 \sum_{j=1}^{N} s_{j+k}^2}}, \quad (9)$$

where k – pitch lag. Peaks located within range corresponding to possible pitch lags (in our case 16 to 160) are considered as pitch candidates. In order to discard spurious peaks all candidates with correlation value below 30% of the maximum one are discarded. Next step of algorithm is pitch tracking based on DP [8]. For each candidate cost function is computed with respect to past information about the pitch track:

$$D_{i,j} = d_{i,j} + \min_{k\in I_{i-1}}\left\{D_{i-1,k} + \delta_{i,j,k}\right\}, \quad (10)$$

where: d_{ij} – local cost of j-th candidate in the time instant i, $\delta_{i,j,k}$ – cost between k-th candidate in the time instant i-1 and j-th candidate in the time instant i, $1 \le j \le I_i$; I – number of candidates. During the cost function calculation information about correlation coefficients and distance between candidates of current and previous frames is used. The aim is to find maximally smooth pitch track.

After the DP procedure the candidate j with minimal cost $D_{i,j}$ is selected as a preliminary pitch frequency estimate for current frame.

Let's consider some speech sample with pitch varying within 300 – 400 Hz and the dependency of *HNR* from the possible fundamental frequency. To do that we have selected one of the voiced frames of the speech sample and analyzed it with the method proposed in Section 2 for the typical pitch values from 50 to 500 Hz. The result of this procedure can be seen at the Figure 3.

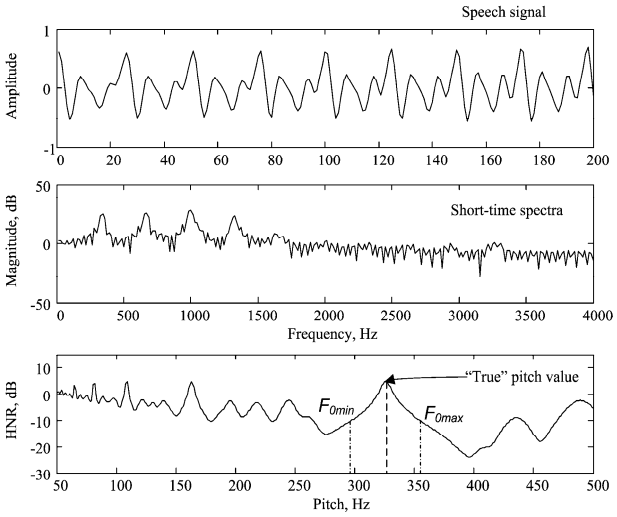

Figure 3: *Speech frame in time (**top**) and frequency (**middle**) domain, and HNR(F0) function (**bottom**).*

It can be seen from the Figure 3 that relation $HNR(F_0)$ has its local maxima at the point, corresponding to the pitch of this speech frame, and this maxima is unimodal in the vicinity of that point (frequency region of 280 – 390 Hz). The experiments have shown that the character of this dependency

is common for the voiced frames. Hence, harmonic parameters will have their optimal values when:

$$F_0^{opt} = \arg\max(HNR(F_0)),\ F_{0min} \le F_0 \le F_{0max}. \qquad (11)$$

So, the following method of harmonic parameters determination is advisable: first of all rough pitch estimation and pitch tracking is held and then in the vicinity of the initial pitch value closed-loop analysis-by-synthesis maxima searching takes place resulting in simultaneous pitch and harmonic parameters refinement.

A detailed structure of closed-loop parameters estimation and pitch refinement algorithm used is shown in Figure 4.

Closed-loop procedure for pitch track refinement (stage 1) is performed after the preliminary pitch track estimation step. The goal of closed-loop estimation is to find optimal pitch value for the current frame which maximizes energy of voiced component. PTDFT is the core of this process. Closed-loop procedure works as follows: first, the boundaries for pitch search are selected based upon information about initial pitch value in the frame and they are referred to as F_{0min} and F_{0max}. PTDFT is performed then and voiced component is generated according to (2, 3) and *HNR* values are computed according to (7). The next step involves golden section method for pitch search which maximizes *HNR* value within given boundaries. In order to ensure that we have one maximum of the *HNR* function within given boundaries PTDFT analysis is done with four leading harmonics. This is because larger number of harmonics used for pitch estimation can introduce distortions and thus it is harder to find optimal pitch value. The boundaries of the search are chosen as ±15% of the initial pitch value.

A problem of ΔF_0 computation is solved in the same way (stage 2). Closed-loop procedure here is performed after the stage 1 completion.

An example of speech decomposition using proposed approach is presented at the Figure 5.

4. Experimental results

Edinburgh Database [9] was used to evaluate the proposed method. It consists of 50 utterances by one English male and female speakers each. The total duration of utterances is 7 min at 20-kHz sampling rate. Laryngograph data was recorded simultaneously with speech and was used by the creators of the database to produce fundamental frequency estimates. They also identified regions where the fundamental frequency was inexistent.

We have tested our HNR-based method on the database described above and then compared a performance of the proposed method to classical PDA's mentioned in [9]. Along with the accuracy of the HNR-based PDA for the clean speech, performance under noisy condition (additive Gaussian noise) was also evaluated at five different SNR levels: 15, 10, 5, 2 and 0 dB. Pitch was determined by the proposed method every 12.5 ms. Two error parameters were computed during evaluation: one of them was F_0 frame error (*FFE*), that takes into consideration both Gross Pitch Errors (GPE) and Voicing Decision Errors (VDE) [10], and mean of the fine pitch errors (*MFPE*). The results of testing for clean and noisy male and female speech are summarized in Tables 1 and 2.

The results of the tests clearly show that the method proposed outperforms classical approaches mentioned in [9] and is quite robust under noisy conditions.

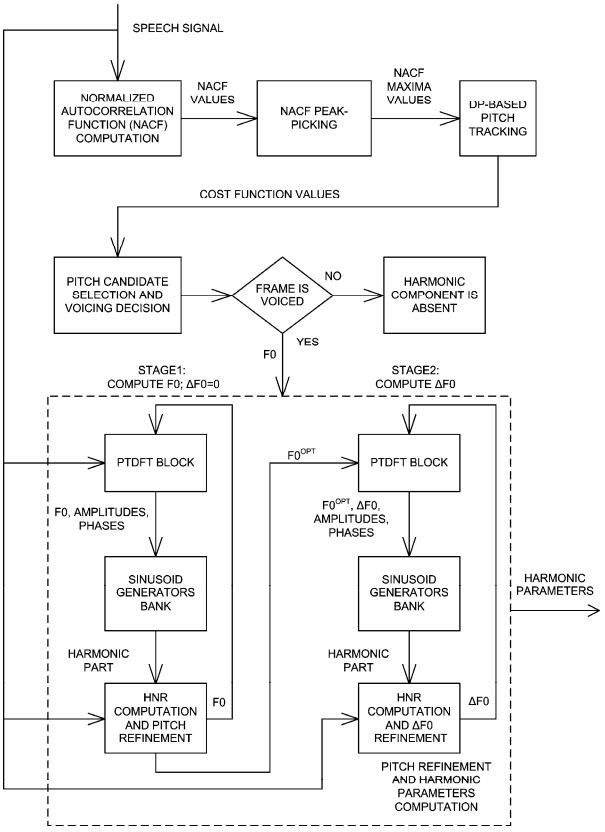

Figure 4: *Closed-loop harmonic parameters estimation and pitch refinement algorithm.*

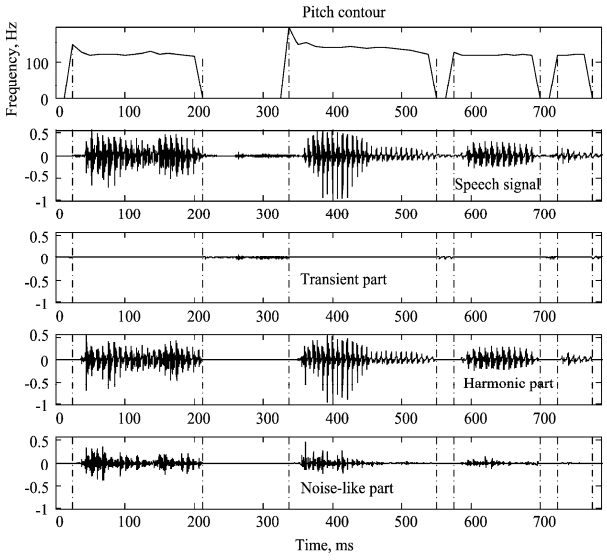

Figure 5: *Speech signal decomposition example.*

5. Real-time implementation

The presented method of pitch and harmonic parameters estimation was embedded into low bit-rate harmonic coders (2.4 and 4.8 kbps), that were realized on DSP TMS320C6713B.

Real-time computational process is organized in the following way. Speech frame (25 ms) is received using McBSP port and EDMA logic of DSP. Analysis is performed twice per frame (on every 100 samples) according to methods

described in Sections 2 and 3. It was experimentally determined that good trade-off between computational complexity and accuracy is achieved when closed-loop procedure consists of 10 iterations for joint fundamental frequency refinement and harmonic parameters computation and of 5 iterations for fundamental frequency change and further harmonic parameters refinement. Computational complexity of proposed analysis methods and algorithms is about 110 MIPS.

Table 1. *Results of testing 7 PDA's under clean speech conditions for male and female speech*

PDA	Male speech	*FFE*, %	*MFPE*, Hz	Female speech	*FFE*, %	*MFPE*, Hz
CPD		40,93	2,94		55,87	6,39
FBPT		19,20	1,86		19,27	5,40
HPS		47,58	3,25		41,40	4,59
IPTA		28,85	2,67		24,49	4,38
PP		25,01	2,64		21,96	6,11
eSRPD		17,92	1,40		12,44	4,17
HNR-based		8,14	1,43		5,66	3,65

Table 2. *Performance of HNR-based PDA under noisy speech conditions for male and female speech*

SNR, dB	Male speech	*FFE*, %	*MFPE*, Hz	Female speech	*FFE*, %	*MFPE*, Hz
15		12,28	1,62		12,47	4,17
10		12,08	1,63		11,27	4,22
5		13,23	1,64		10,17	4,39
2		15,23	1,72		11,36	4,41
0		15,01	1,66		11,85	4,47

Several comparisons based on the Modified Bark Spectral Distortion (MBSD) value were made between the proposed PTDFT-based speech coding systems and some standardized codecs (Figure 6 and 7). Speech samples from 5 dictors (3 male and 2 female) with a common length of about 30 seconds were used during evaluation process. First of the proposed coders (2.4 kbps) outperforms such speech codecs as MELP (2.4 kbps), ITU-T G.723.1 (5.3 kbps), ITU-T G.726 (16 kbps), another one (4.8 kbps) has showed better performance than ITU-T G.729 (8 kbps) (Figure 6).

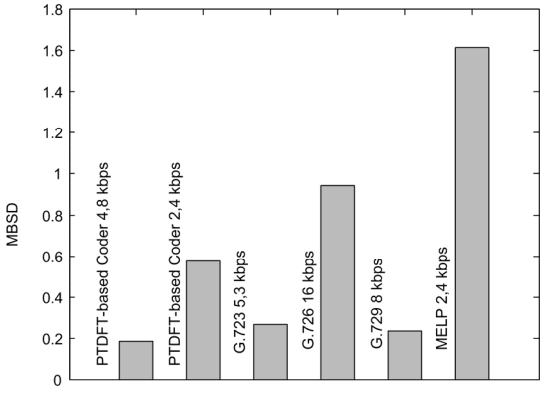

Figure 6: *Speech codecs evaluation results.*

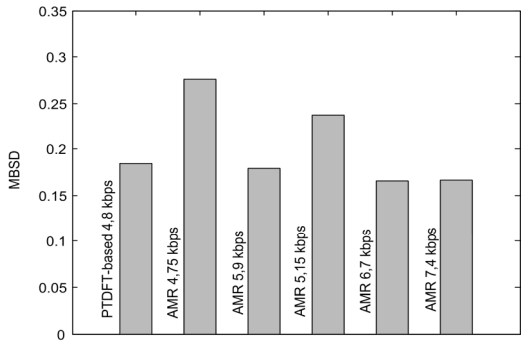

Figure 7: *Comparison of PTDFT-based coder 4,8 kbps with AMR family of speech codecs.*

Figure 7 contains the results of comparison of PTDFT-based coder 4,8 kbps with a set of speech codecs belonging to AMR family. From the Figure 7 it can be seen that proposed PTDFT-based coder's performance is clearly better than AMR codecs 4,75 and 5,15 kbps and is slightly worse than codecs 6,7 and 7,4 kbps.

6. Conclusions

A new method of pitch and harmonic parameters estimation has been developed that uses closed-loop analysis-by-synthesis approach for joint pitch refinement and speech decomposition and based on the *HNR* for optimization purposes. An analysis of errors indicates high accuracy and good robustness of the method under noisy conditions.

7. References

[1] Stylianou, Y., Laroche, J. and Moulines, E., "High-quality speech modification based on a harmonic + noise model", Proc. EUROSPEECH'95, Spain, pp. 451 – 454.

[2] Griffin, D. and Lim, J. "Multiband excitation vocoder", IEEE Trans. Acoust., Speech and Sig. Proc., vol. 36, №8, pp. 1223 – 1235, Aug. 1988.

[3] Yegnanarayana, B., d'Alessandro, C. and Darsinos, V., "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components", IEEE Trans. on Speech and Audio Proc., vol.6, № 1, pp. 1 – 11, Jan. 1998.

[4] Jackson P.J.B., and Shadle, C.H., "Pitch-scaled estimation of simultaneous voiced and turbulence-noise components in speech", IEEE Trans. on Speech and Audio Proc., vol.9, № 7, pp. 713 – 726, Oct. 2001.

[5] Weruaga, L. and Kepesi, M., "The fan-chirp transform for non-stationary harmonic signals", Signal Processing, vol. 87, № 6, pp. 1504 – 1522, June 2007.

[6] Petrovsky, A., Zubricki, P. and Sawicki, A., "Tonal and noise components separation based on a pitch synchronous DFT analyzer as a speech coding method", Proc. ECCTD'03, Poland, vol. 3, pp.169 – 172.

[7] Sercov, V. and Petrovsky, A., "An improved speech model with allowance for time-varying pitch harmonic amplitudes and frequencies in low bit-rate MBE coders", Proc. EUROSPEECH'99, Hungary, pp. 1479 – 1482.

[8] Talkin, D., "Robust algorithm for pitch tracking" in "Speech Coding and Synthesis", Kleijn, W.B. and Palival, K.K. (Eds.) Elsevier, Amsterdam, Netherlands, 1995.

[9] Bagshaw, P.C. et al., "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching", Proc. EUROSPEECH'93, Germany, pp. 1003–1006.

[10] Chu, W. and Alwan, A., "Reducing F0 Frame Error of F0 Tracking Algorithms Under Noisy Conditions with an Unvoiced/Voiced Classification Frontend", Proc. ICASSP'09, pp. 3969 – 3972.