Voice Conversion Based on the HNT Model of Speech and Separate VQ Learning

Alexander Pavlovets, Michael Livshitz, Denis Lichachov and Alexander Petrovsky

Computer Engineering Department
Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus
palex@bsuir.by

Abstract

In this paper, a text-dependent voice conversion method based on the mapping codebook approach is proposed. One of the critical tasks in voice conversion framework is speaker parameter estimation. In the given report the method based on the Harmonic-Noise-Transient (HNT) decomposition of speech is offered with an idea to separately process each of the components and further to separately convert them. Informal listening tests have shown the superiority of the presented system over the ACELP-based system.

1. Introduction

Voice conversion problem became very popular in the world. It has applications in many fields, for example in systems that make use of prerecorded speech, such as voice mailboxes or text-to-speech synthesizers based on acoustic unit concatenation. In such cases, voice modification would be a simple and efficient way to create a desired variety of voices while avoiding recording of different speakers [1]. Another field of application is old movies restoration, where the aim is to reconstruct corrupted voices of old actors. Voice modification techniques attempt to transform the speech signals uttered by a given speaker so as to alter the characteristics of his or her voice. This problem – how to modify the speech of one speaker so that it sounds as if it was uttered by another speaker – is known as voice conversion [1]. Voice conversion flowchart is illustrated in common in Figure 1.

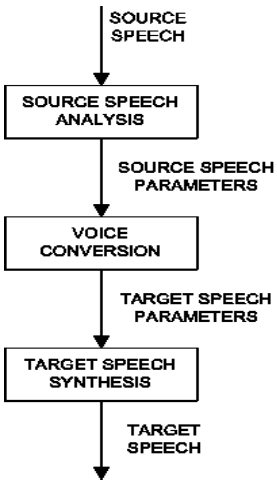

Figure 1. Typical structure of voice conversion process.

An important problem in voice conversion systems is source speaker voice parameters estimation. In our previous work [2] we improved an approach mentioned in [3], that is based on the harmonic plus noise model, by using of speech decomposition into voiced and noise-like components. This approach can be considered as an improvement of the harmonic plus noise model used in [3], and is successfully applied in speech coding framework [4, 5].

Our further investigations have shown a necessity of extending model used in [2], as transient frames of speech can not be handled correctly with it. That's why we have introduced the second mode of speech processing, that deals with transitions.

Identification of voiced regions of speech is quite a difficult task. Decision of voicing can be performed for whole analysis frame [6], searching for maximum voicing frequency [3] or voicing decision for each frequency band [7]. 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 [8] 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 [9]. Speech signal is windowed and window length is chosen to contain small integer multiple of pitch periods. PSHF algorithm performs decomposition in frequency domain by selecting only STFT bins which are aligned with pitch harmonics.

As the methods presented in [8],[9] our method considers voiced and noise-like components present in the whole band. Pitch-Tracking modification is applied to standard DFT (PTDFT) in order to provide spectral analysis in harmonic domain rather than in frequency domain. Model parameters, i.e. pitch frequency and harmonic amplitudes and phases are accurately estimated using analysis-by-synthesis method. Voiced component is defined here 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.

The problem of voice conversion has focused a lot of research effort. For instance, an approach to this problem was speech transformation algorithm using segment codebook (STASC) [10]. The method finds accurate alignments between source and target speaker utterances. Using the alignments,

source speaker acoustic characteristics are mapped to target speaker acoustic characteristics.

A method that improves the quality of the voice conversion output at higher sampling rates is proposed in [11]. It combines the STASC method with Discrete Wavelet Transform (DWT) to estimate the speech spectrum better with higher resolution. Both these researches are combined in [12]. Several works that suggest a possible way to improve the quality of the converted speech consider modification of some specific aspects of the spectral envelope [13] or of the location of the formants [14, 15].

In this paper we present a modification of mapping codebook method [16]. Different codebooks are used depending on the frame type.

The aim of the work is to introduce 2-mode voice conversion system hence providing better performance.

2. Voice conversion system implementation

In our previous works [2, 17] we tried to solve voice conversion problem with such kinds of spectral envelope conversion method as harmonic amplitudes codebook mapping method [17] and line spectral frequencies (LSF) conversion based on the Gaussian Mixture model (GMM) [2]. Taking into consideration the fact that we have implemented 2-mode speech analyser, it would be reasonable to use separate conversion functions for spectral envelopes we get in each mode. So we have applied mapping codebook method with enhancements mentioned in [18] for each mode of analysis, but training of voice conversion system has been carried out separately.

Further we shall consider the training process of conversion functions and the algorithm of voice conversion system based on HNT model of speech.

2.1. Separate learning process

For training of conversion system in each mode it is necessary to prepare two different sets of conversion parameters of source and target speaker. After the harmonic analysis realized with the PTDFT the following parameters are taken for conversion: spectral envelope expressed by LSFs and fundamental frequency F_0. Analysis of transient frames made with ACELP approach [19] gives us for modification source-filter coeficients expressed by LSFs, pitch lag T_0, adaptive and fixed codebok gains G_a and G_f respectively.

During training phase of the voice conversion system two sets of parameters are obtained for source and target dictor depending on the analyser mode. Then for each set conversion paramerers are computed.

To handle the transformation of such parameters as fundamental frequency F_0, pitch lag T_0, adaptive and fixed codebok gains G_a and G_f the mean-variance linear conversion method is used, with the assumption that average values of these parameters already carry a great deal of the speaker specific information. The underlying assumption is that each speaker's parameter values belong to a Gaussian distribution with a specific mean and variance.

Defining the parameter to be modified as P^t, a linear transformation can then be defined as follows:

$$P^t = \frac{\delta_t}{\delta_s}(P^s - \mu_s) + \mu_t \quad (1)$$

where P^t, P^s – one of the parameters of target and source speakers respectively, δ_s, μ_s, δ_t, μ_t are standard deviation and mean of corresponding parameter of source and target speakers respectively.

The available data for spectral envelope conversion consists of two sets of paired spectral vectors (LSF) obtained by the approach that will be described in Section 3. During training in order to ensure a mapping one to one between source and target spectral vectors, Dynamic Time Warping algorithm (DTW) is used for time alignment. Figure 2 illustrates spectral vectors codebook training phase for any of modes.

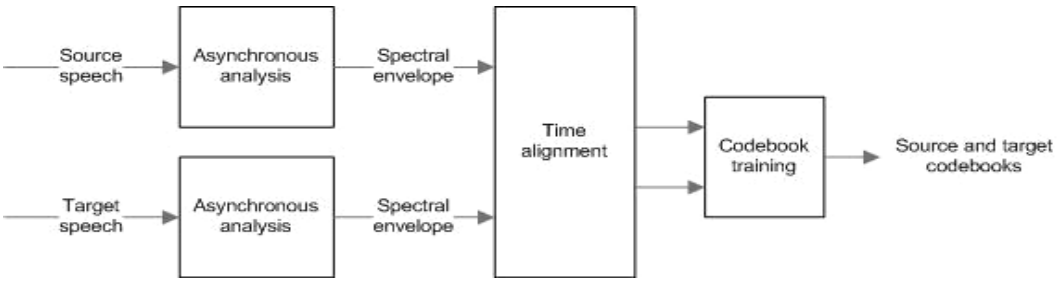

Figure 2. Block diagram of spectral conversion training phase

2.2. Voice conversion flowchart

Once the parameters of the conversion are estimated, the voice conversion is performed as shown in Figure 3. The input of system consists of speech signals sampled at 8 kHz.

First of all voice activity detector (VAD) realized as in [20] discriminates frames of source signal with silence and/or background noise. Such frames are not exposed to processing and are passed directly to output.

The pitch and transient detector block determines whether a frame will be passed to harmonic analysis module based on PTDFT or transient analysis module based on G729 Recommendation. Then parameters extracted with one of the analysis modules are transformed according to the mode and frame of target speaker's speech is synthesized.

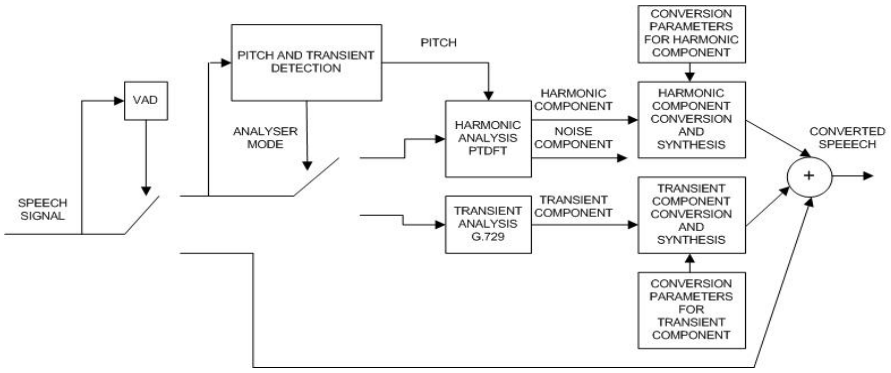

Figure 3. Voice conversion system flowchart

3. HNT model of speech

Multimode approach has already been successfully represented in fields of speech and audio coding. Among others one can mention works [21, 22]. First of them describes hybrid speech coder which combines a frequency-domain parametric coder (for stationary voiced and stationary unvoiced speech) with a time-domain waveform coder (for transition speech). Another one utilizes segmentation of audio into three separate signals: a signal which models all sinusoidal content with a sum of time-varying sinusoids, a signal which models all attack transients present using transform coding, and a noise signal which models all of the high frequency input signal not modeled by the transients.

Speech in our model is considered as:

$$s(i) = h(i) + r(i) \tag{2}$$

where *h(i)* – voiced (harmonic) component and *r(i)* is residual signal (noise-like component or transient).

3.1. Mode selector

The speech analyser mode in our case is fully determined by pitch presence or absence (Fig. 4).

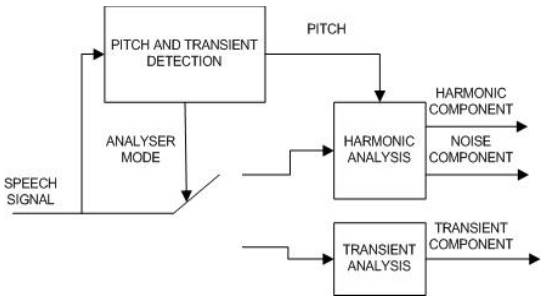

Figure 4. Speech decomposition scheme.

Detailed structure of pitch determination algorithm will be considered further.

3.2. Harmonic analysis

Harmonic component of speech signal can be defined as:

$$h(i) = \sum_{k=1}^{K} A_k \cos(k \sum_{i=0}^{N-1} \frac{F_0(i)}{F_s} + \theta_k) \tag{3}$$

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.

The core of the harmonic analysis in our case is PTDFT procedure [5].

Modified DFT transform for analysis in harmonic domain is given by:

$$H_n(k) = \sum_{i=0}^{K-1} s_n(i) \exp\left(j \frac{2\pi k i}{F_s} \left(F_0 + \frac{\Delta F_0 i}{2N} \right) \right) w_n(i), \quad j = \sqrt{-1}, \tag{4}$$

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.

Hence harmonic amplitudes and phases can be found:

$$A_n(k) = \frac{\sqrt{\mathrm{Re}^2(H_n(k)) + \mathrm{Im}^2(H_n(k))}}{\sum_{i=0}^{L-1} w_n(i)}, \tag{5}$$

$$\theta_n(k) = -arctg \frac{\mathrm{Im}(H_n(k))}{\mathrm{Re}(H_n(k))}. \tag{6}$$

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 this solution is to design a spectral window, which follows fundamental frequency changes. Good results can be achieved using Kaiser window as a prototype [23]:

$$w_n(i) = \frac{I_0\left(\beta \sqrt{1 - \left[(2x - L_n + 1)/(L_n - 1) \right]^2} \right)}{I_0(\beta)}, \tag{7}$$

where $i=0\ldots N-1$, 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:

$$x = \frac{a_{2,n}(N-1-i)^2 + a_{1,n}(N-1-i)}{a_{2,n}(N-1)+a_{1,n}}, \tag{8}$$

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

$$a_{2,n} = \frac{2\pi\Delta F_0}{F_s N}, \tag{9}$$

$$a_{1,n} = \frac{2\pi F_0}{F_s}. \tag{10}$$

3.2.1. Closed-loop parameters determination algorithm

Harmonic analysis using PTDFT and hence speech decomposition has been shown [4,5] to be quiet 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.

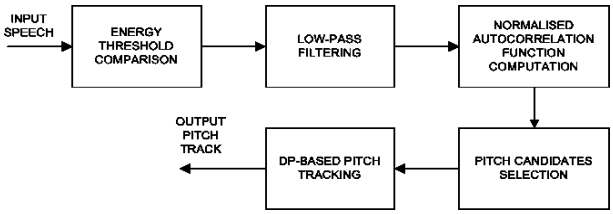

Figure 5. Preliminary pitch track estimation.

First of all robust and reliable method for rough pitch estimation must be used. In our case time-domain pitch determination algorithm based on normalized autocorrelation function (NACF) computation in combination with postprocessing based on dinamic programming (DP) method was applied. The sequence of preliminary pitch track estimation steps is shown in Figure 5.

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:

$$\psi(k) = \frac{\sum_{j=1}^{N} s_j s_{j+k}}{\sqrt{\sum_{j=1}^{N} s_j^2 \sum_{j=1}^{N} s_{j+k}^2}}, \tag{11}$$

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 dynamic programming (DP) [24]. For each candidate cost function is computed with respect to past information about the pitch track:

$$D_{i,j} = d_{i,j} + \min_{k \in I_{i-1}} \left\{ D_{i-1,k} + \delta_{i,j,k} \right\}, \tag{12}$$

where: $d_{i,j}$ – 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.

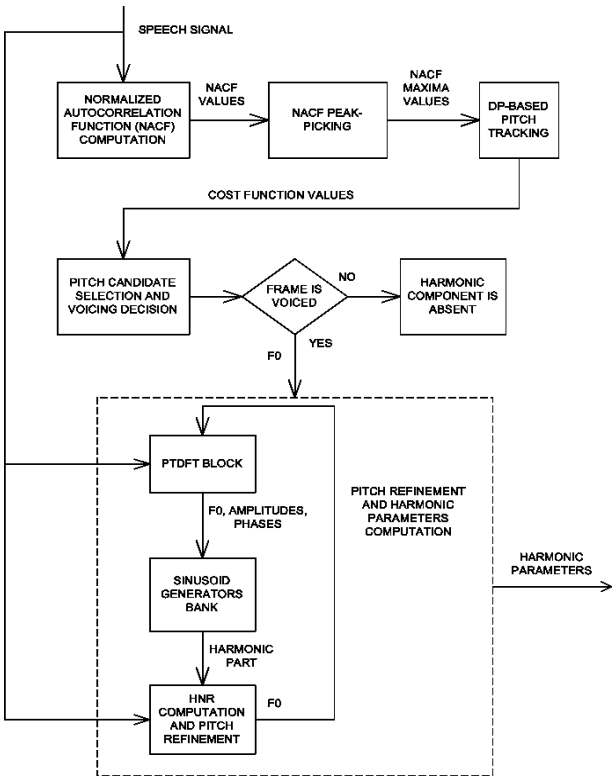

Figure 6. Closed-loop harmonic parameters estimation and pitch refinement algorithm.

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

The detailed structure of closed-loop parameters determination and pitch refinement algorithm used is shown in Figure 6.

Accuracy of the parameters estimation can be measured by Harmonic-to-Noise Ratio (*HNR*):

$$HNR = 10 \lg \frac{E_H}{E_N}, \tag{13}$$

where E_H – energy of synthesized according to (3) 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:

$$r(i) = s(i) - h(i)\,. \qquad (14)$$

Closed-loop procedure for pitch track refinement 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, voiced component is generated according to (3) and *HNR* coefficients are computed. The next step involves golden section method for pitch search which maximizes *HNR* coefficient within given boundaries:

$$F_0^{opt} = \arg\max(HNR(F_0)),\ F_{0\min} \le F_0 \le F_{0max}. \qquad (15)$$

An example of speech decomposition is illustrated in Figure 7.

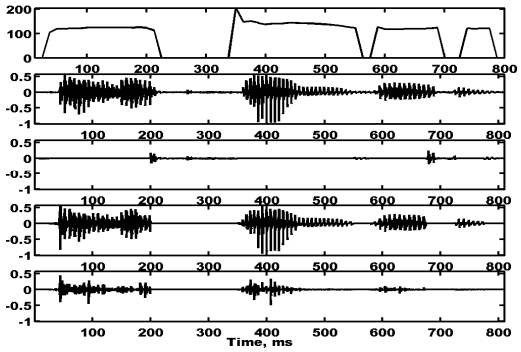

Figure 7. Example of speech signal decomposition. From top to bottom: pitch track, original speech waveform, residual waveform, harmonic component waveform, noise component waveform.

3.3. Transient analysis

For analysis of transient frames we have used ACELP approach standardized as ITU-T G.729 [19].

4. Experimental results

The performance of proposed voice conversion system was evaluated by informal listening tests. For the comparison voice conversion system entirely based on the ACELP model as analysis approach and on the mapping codebook method (briefly ACELP-based system) was used.

The tests have shown that voice produced by the proposed system is rather natural, its intelligibility is higher in comparison with the ACELP-based system.

Figures 9, 10 contain spectrograms of male-to-female voice conversion samples. A phrase in Polish was taken to be modified: “Lubić czardaszowy pląs” (Figure 8). 15 phrases from [25] were used for training . Obviously, the result of conversion by HNT-based system better matches the harmonic structure of target speaker and contains less noise.

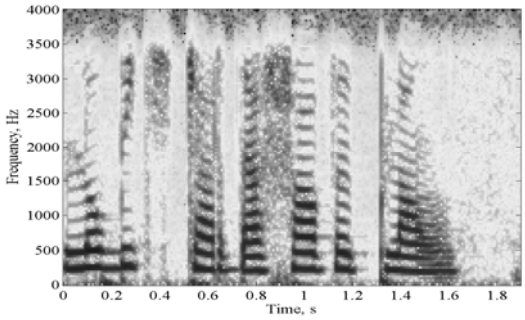

Figure 8. Target speaker phrase spectrogram

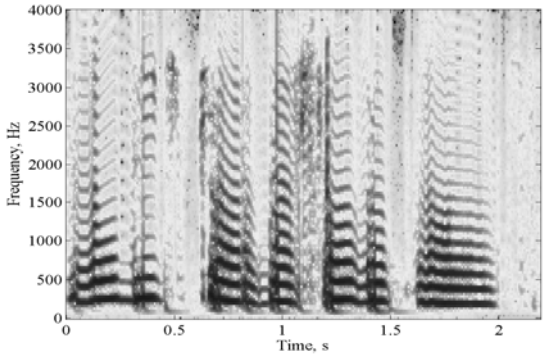

Figure 9. Example of conversion by HNT-based system

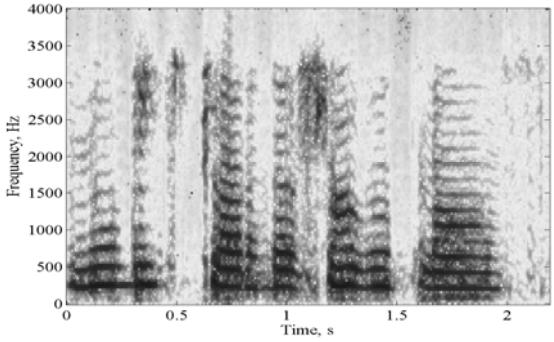

Figure 10. Example of conversion by ACELP-based system

5. Conclusions and future work

In this paper we have presented a voice conversion system based on 2-mode analysis approach with separate conversion parameters for each mode. Informal listening tests have shown the superiority of the presented system over the ACELP-based system.

The advantage of the HNT-based voice conversion system can be realized as easy way to aggregate analysis-synthesis benefits of harmonic model of speech with time-

domain transient analysis and conversion. Thus a part of source speaker's voice in converted speech can be considerably decreased.

The future work is connected with embedded applications (mobile phones) of the presented approach such as combining a test-to-speech (TTS) system and a real-time voice conversion system with limited training base with the purpose of reception of multilingual TTS system and very small memory size of database. As known, a typical structure of TTS system consists of three blocks: text processing, prosody control, and acoustic module. The acoustic module can be realized as a voice conversion system on which entrance a data from the prosody control block move. Thus, the database is presented in harmonic area (it is preliminary processed by means of the HNT analyzer and its compression is carried out [4]).

6. References

[1] Moulines, E. and Sagisaka, Y., Eds. "Voice conversion: state of the art and perspectives". *Speech Communication*, vol. 16, Feb. 1995.

[2] Pavlovets, A., Kien, T., Zubricki, P. and Petrovsky A. "Speech analysis – synthesis based on the PTDFT for voice conversion", in *Proc. of the 2007 Int. Workshop on Spectral Methods and Multirate Sig. Proc., SMMSP*, Moscow, Russia, Sep. 2007, pp. 203 – 210.

[3] Stylianou, Y., Laroche, J. and Moulines, E. "High-quality speech modification based on a harmonic + noise model", in *Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH*, Madrid, Spain, Sep. 1995, pp. 451 – 454.

[4] Petrowsky, A., Zubricki, P. and Sawicki, A. "Tonal and noise components separation based on a pitch synchronous DFT analyzer as a speech coding method," in *Proc. European Conf. Circuit Theory and Design*, Cracow, Poland, Sep. 2003, vol. 3, pp.169 – 172.

[5] Zubricki, P., Pavlovec, A. and Petrovsky, A. "Analysis-by-synthesis parameters estimation in the harmonic coding framework by pitch tracking modified DFT" in "New trends in audio and video", Dobrucki, A., Petrovsky, A. and Skarbek, W. Eds. Byalystok 2006, pp. 233 – 246.

[6] Tremain, T. "The government standard linear predictive coding algorithm: LPC-10", *Speech Technology Magazine*, vol. 1, №. 2, 1982, pp. 40 – 49.

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

[8] 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.

[9] 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.

[10] Arslan, L. and Talkin, D. "Voice Conversion by Segmental Codebook Mapping of Line Spectral Frequencies and Excitation Spectrum'', in *Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH* Rhodes, Greece, Sep. 1997, vol. 3, pp. 1347 – 1350.

[11] Turk, O. and Arslan, L. "Subband Based Voice Conversion", in *Proc. of the Int. Conf. on Spoken Language Proc. ICSLP*, Denver, USA, Sep. 2002, vol. 1, pp. 289 – 292.

[12] Turk, O. "New methods for voice conversion", PhD. Thesis, Bogaziçi University, Istanbul, Turkey, 2003.

[13] Vích, R. and Vondra, M. "Voice conversion based on Spectral Envelope Transformation", www.iiassvietri.it/school2004/ School_Materials_1/ oral_contributions/Vondra_short_slides.pdf.

[14] Malkin, J., Xiao Li and Bilmes, J. "A Graphical Model for Formant Tracking", in *Proc. of the Int. Conf. on Acoust., Speech and Sig. Proc. ICASSP*, Philadelphia, USA, March 2005, vol. 1, pp. 913 – 916.

[15] En-Najjary, T., Rosec, O. and Chonavel, T. "A voice conversion method based on joint pitch and spectral envelope transformation", in *Proc. of the Int. Conf. on Spoken Language Proc. Interspeech – ICSLP*, Jeju Island, Korea, Oct. 2004, pp.1225 – 1228.

[16] Abe, M., Nakamura, S., Shikano, K. and Kuwabara, H. "Voice conversion through vector quantization", in *Proc. of the Int. Conf. on Acoust., Speech and Sig. Proc. ICASSP*, New York, USA, Apr. 1988, vol. 1, pp. 655 – 658.

[17] Pavlovets, A. and Petrovsky, A. "Voice conversion as a part of the voice analysis/synthesis system based on the periodic-aperiodic decomposition of speech", in *Proc. of the 9th Int. Conf. on Pattern Recognition and Information Proc., PRIP*, Minsk, Belarus, May 2007.

[18] Stylianou Y., Cappe O., Moulines E. "Continuous probabilistic transform for voice conversion", *IEEE Trans. on Speech and Audio Processing*, vol. 6, № 2, pp. 131 – 142, March 1998.

[19] ITU-T Rec. G.729, "Coding of speech at 8 kbit/s using conjugate-structure algebraic-code-excited linear – prediction (CS-ACELP)", Mar. 1996.

[20] ITU-T Rec. G.729, annex B, "A silence compression scheme for G.729 optimized for terminals conforming to Recommendation V.70", Nov. 1996.

[21] Shlomot, E., Cuperman, V. and Gersho, A. "Hybrid coding: combined harmonic and waveform coding of speech at 4 kb/s", *IEEE Trans. on Speech and Audio Proc.*, vol.9, № 6, pp. 632 – 646, Sep. 2001.

[22] Levine, S. and Smith, J.O. "A sines+transients+noise audio representation for data compression and time/pitch scale modifications" in *Proc. 105th Conv. Audio Eng. Soc., preprint #4781, Sep. 1998.*

[23] 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", in *Proc. of the European Conf. on Speech Communication and Technology EUROSPEECH*, Budapest, Hungary, Sep. 1999, pp. 1479 – 1482.

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

[25] Grocholevski, S. "First Database for Spoken Polish", in *Proc. Int. Conf. On Language Resources and Evaluation*, Grenada, 1998, pp. 1059 – 1062.