Text and speaker independent voice conversion

Elias Azarov, Alexander Petrovsky
Belarusian State University of Informatics and Radioelectronics
P. Brovky 6, 220027, Minsk, Belarus, palex@bsuir.by, www.bsuir.by

Abstract: *This paper describes an approach to the challenging problem of text and speaker independent voice conversion. The approach is based on target speaker's speech production process parameterization using harmonic analysis. Unified model allows processing of any input speech regardless of its content and source speaker. The method provides subjective quality of conversion that is comparable with text/speaker depended methods though requires a drastically simplified teaching routine and a small codebook. The experimental results are given in the paper.*

Keywords**:** Voice conversion, harmonic analysis, speech analysis, vocal tract parameters, spectral envelope, pitch.

1. INTRODUCTION

The voice conversion problem arose with multimedia systems development and has received much attention from different researchers worldwide. The solutions become more and more sophisticated providing better and better conversion quality, however this investigation seems to be at the initial stage. The main reasons for it are listed below:

- complexity of speech processing;
- training set preparation;
- speaker variability;

It is said that speech is one of the most difficult signals to deal with when it comes to synthesis and conversion. Human speech is a product of complex physiological processes that have no effective mathematical description so far. Speech signal has a variable structure combining both periodic and aperiodic components that makes its analysis especially specific. It forces researchers to use special techniques to separate components of different nature from the signal and process them differently.

Majority of voice conversion systems are based on the speaker/text depended approach [1]. It means that the user should have a long, phonetic balanced records prepared for source and target speaker. It also means that records should be in correspondence with each other i.e. synchronized in time domain. It provides the best possible conversion quality but at the same time requires many efforts from the user to make training sets for every pair (source-target) of speakers. If training sets are not synchronized well enough the quality of conversion degrades. The same speaker can utter the same words with different pronunciation which depends on expression, mood and intonation. This fact complicates making strict correspondence between source and target speakers.

An alternative way is text/speaker independent conversion that allows teaching with arbitrary training sets and is free from source-target synchronization. This approach is rarely used because of lower conversion quality in comparison with text/speaker dependent methods. However, it is much more usable and can significantly extend field of voice conversion application.

The text/speaker independent approach described in this paper is based on the assumption that all human speakers of the same language have something in common. In other words the conversion process can be performed through unified set of parameters. The described method has the following benefits:

- simplified training procedure;
- only one (target) speaker is needed for training;
- conversion is possible for any source speaker without retraining;
- small codebook.

In training procedure (Fig.1) an analysis of target speaker's speech is performed in order to build a vocal tract model. The analyzer estimates harmonic and unified parameters, builds spectral envelope, separates periodic components from noise. The parameters of the model are stored in a codebook that is used later in conversion process.

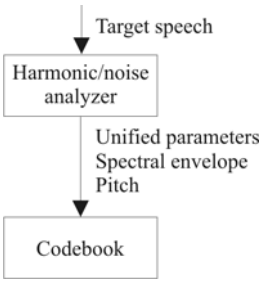


Fig.1 – General training scheme

The conversion procedure is illustrated in Fig.2. Source speech is analyzed by harmonic/noise analyzer and parameters are transmitted into converter where target speaker's frequency envelopes and pitch are estimated. These data are used by synthesizer to make the converted speech.

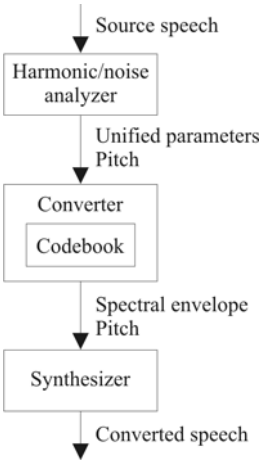


Fig.2 – General conversion scheme

Experiments have shown that proposed method provides a good quality and can be used in multimedia applications.

2. HARMONIC/NOISE ANALYZER

The main speech components, that form speaker's identity (spectral envelope and pitch) can be estimated from speech signal by means of harmonic+noise (H+N) model. The H+N model describes a signal in terms of harmonic (periodic) and noise (aperiodic) parts. The periodic part of the signal is assumed to be a sum of sinusoids with different frequencies, amplitudes and phases. Sinusoidal signal representation is known since 70th years [2] and has been highly developed recently [3]. The H+N model can be described in the following way [3]:

$$s(n) = \sum_{k=1}^{K} A_k(n)\cos\varphi_k(n) + h(n),$$

where $s(n)$ is the source signal $A_k(n)$ - the instantaneous magnitude of the k-th harmonic component, K is the number of the harmonic components, $h(n)$ is the noise component and $\varphi_k(n)$ is the instantaneous phase of the k-th harmonic component. $\varphi_k(n)$ can be derived from the initial phase $\varphi_k(0)$ and instantaneous frequency f_k:

$$\varphi_k(n) = \sum_{i=0}^{n} \frac{2\pi f_k(i)}{F_s} + \varphi_k(0),$$

where F_s is the sampling frequency. The harmonic representation assumes that the next expression is true:

$$f_k(n) = kf_0(n),$$

where $f_0(n)$ is the fundamental frequency.

The H+N model was chosen in this work as the main tool for analysis and synthesis because of its powerful features:

- good energy localization in time and frequency domain;
- periodic/aperiodic separation;
- efficient speech representation;
- low signal distortion;
- main speech components, that indicate speaker's identity (spectral envelope and pitch) can be accurately estimated at every instant of time.

There are a number of techniques that can provide accurate harmonic parameters estimation [4]. In this work the instantaneous harmonic parameters estimation technique [5] is used. The system of filters is applied to the signal, providing instantaneous harmonic parameters ($MAG(n)$ - amplitude, $f(n)$ - frequency, $\varphi(n)$ - phase):

$$MAG(n) = \sqrt{A^2(n) + B^2(n)},$$

$$f(n) = \frac{\alpha(n+1) - \alpha(n)}{2\pi} F_s + f_0(n)\cdot k, \qquad (1)$$

$$\varphi(n) = 2\pi f_0(n)kn + \alpha(n)$$

where

$$A(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{\pi}{F_s} F_\Delta(n-i)\right) \cos\left(\frac{2\pi}{F_s}\varphi_k(n)\right)$$

$$B(n) = \sum_{i=0}^{N-1} \frac{s(i)F_s}{(n-i)\pi} \sin\left(\frac{\pi}{F_s} F_\Delta(n-i)\right) \sin\left(\frac{2\pi}{F_s}\varphi_k(n)\right),$$

$$\alpha(n) = \arctan\left(-\frac{B(n)}{A(n)}\right),$$

$$\varphi_k(n) = \left(\sum_{i=0}^{n} F_0(n) - \sum_{i=0}^{N/2} F_0(n)\right)k$$

The general scheme of the harmonic/noise analyzer is presented in Fig.3.

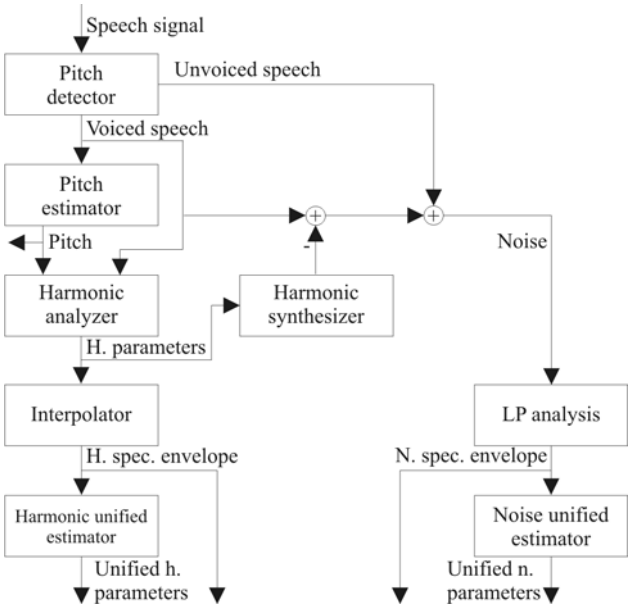


Fig.3 – Harmonic/noise analyzer structure

The harmonic/noise analyzer separates speech frames into harmonic and noise parts and then estimates spectral envelopes and unified parameters from both of them. Pitch detector classifies type of a frame. If the frame is unvoiced it is considered as a noise, if the frame is voiced it is analyzed by harmonic analysis. The pitch is estimated as described in [5] by harmonic analysis of the low frequency band of the speech (80-420Hz). Then harmonic analyzer estimates amplitude, frequency and phase for each harmonic using (1). After estimation of the harmonic parameters, the periodic part of the signal is synthesized and subtracted from the original signal in order to separate the noise one. The harmonic spectral envelope is interpolated for predefined set of frequency points to ensure insensibility to pitch. The spectral interpolation of the noise part is made by LP analysis. The unified parameters are evaluated from spectral envelopes of each part. In this work is made the assumption that the spectral bandwidth can be divided into subbands where relative energy in a band will be close for any speaker, uttering the same sound. Thus the unified parameters are simply calculated as a sum of energy values in every specified band.

3. TRAINING PROCEDURE

As has been stated above, the energy subband values can identify the sound being uttered and therefore can be used as unified parameters. The conversion function should allow converting them to spectral envelope of a

target speaker. After series of experiments had been made was concluded that spectral envelope can be estimated as a linear combination of the energy values. The conversion function can be written in the following matrix form:

$$\overline{E}_T = \overline{U} \cdot K_T , \qquad (2)$$

where $\overline{E}_T$ - target spectral envelope vector, $\overline{U}$ -unified parameters vector, K_T - conversion matrix. Thus the conversion function is a coefficient matrix that provides minimum error of target spectral envelopes estimation. The training procedure solves a system of linear equations that minimizes the error in the least square sense:

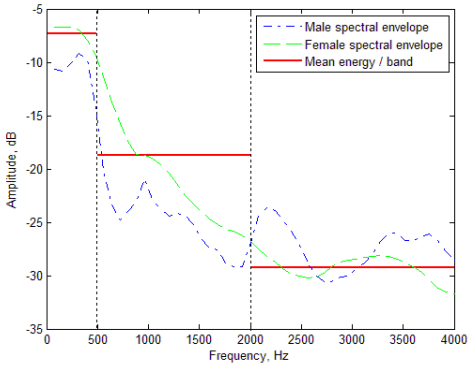
$$\min_{K_T} \left| \overline{E}_T - \overline{U}_T \cdot K_T \right| ,$$

The training set (both $\overline{E}_T$ and $\overline{U}_T$) are estimated from speech of the target speaker. Experimentally has been determined that the optimal number of spectrum subbands for unified parameters estimation is three.

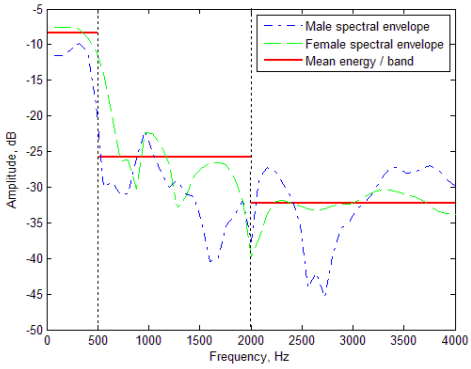
The codebook should provide information about the pitch of the target speaker as well. In this paper the statistical approach is used for pitch conversion [6]. The codebook contains conversion matrix K_T, average pitch and standard deviation of pitch of the source speaker.

3. CONVERSION

First the source speech is analyzed by harmonic/noise analyzer. Then using obtained unified parameters the target spectral envelope is estimated from (2). Fig.3 illustrates estimating target's speaker envelopes from different sets of unified parameters.



A



B

Fig.4 – Target spectral envelope estimation
(A – $\overline{U}$ = (-7,-18,-28) **dB, B –** $\overline{U}$ = (-7,-25,-31) **dB)**

As can be seen the conversion method provides smooth and different envelopes, depending on the codebook and the set of unified parameters.

The pitch is estimated by means of the Gaussian normalization algorithm. The method is based on matching the average pitch and the standard deviation of pitch of a given source speaker to those of a target speaker [6]. The transformed pitch value $p_t^{S->T}$ is estimated as:

$$p_t^{S->T} = \frac{p_t^S - \mu^S}{\sigma^S} \sigma^T + \mu^T ,$$

where μ^S and σ^S are the average pitch and standard deviation of pitch of the source speaker respectively, μ^T and σ^T are the average pitch and standard deviation of pitch of the target speaker respectively, p_t^S is a given pitch value of the source speaker. The parameters μ^S and σ^S are stored in the codebook.

Transformed speech is synthesized using obtained spectral envelopes and pitch contour. The general scheme of the synthesizer is presented in Fig.5.

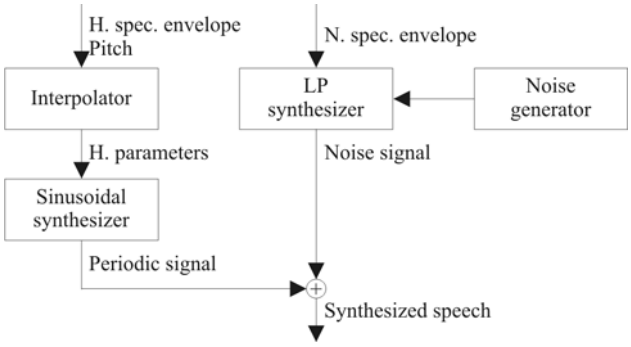


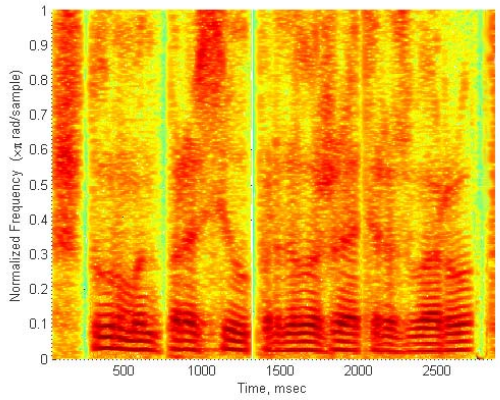
Fig.5 – Speech synthesizer structure

The interpolator makes inverse interpolation of the spectral envelope. Harmonic parameters are recalculated in pitch-defined frequency points $f_k(n)$. The sinusoidal synthesizer forms the periodic signal, using harmonic parameters evaluated by the interpolator.

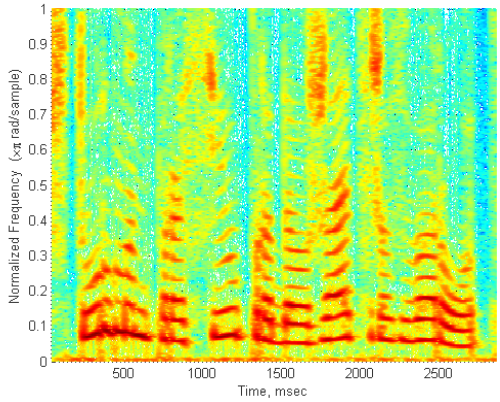
Noise spectral envelope is processed by LP synthesizer in order to generate noise part of the target speaker's speech.

4. EXPERIMENTAL RESULTS

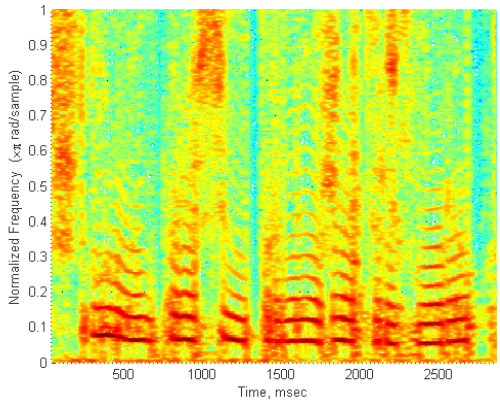
Five codebooks were trained for five different speakers (2 - male, 3 - female). All sound material was sampled at 8kHz. Training sets were one minute length each. A preliminary experiment was carried out in order to estimate accuracy features of the proposed conversion model. Speech records were converted to the original speaker (using source speaker's codebook). This test showed that three unified parameters are enough to estimate spectral envelopes of a speaker. Experience has shown that the optimal segmentation of the spectrum into subbands is 0-500Hz, 500-2000Hz and 2000-4000Hz. Then cross conversions were performed between those speakers. In Fig.6 some results are presented as source-target-transformed spectrograms.



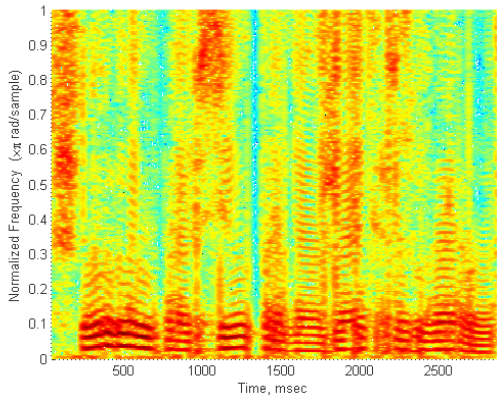
A



B



C



D

Fig.6 – Spectrograms of source-target-transformed signals (A – male source speech, B- female source speech, C – male->female conversion, D – female->male conversion)

Test records were processed through the text/speaker independent conversion technique described in this paper. As can be seen in Fig.6 spectrograms of converted speech perceive target speaker's formant structure and pitch level. The test records were different from those used for training. One can hear that the converted voice is completely different from the original and has pitch and timbre of the target speaker.

The converted samples sound almost natural with slight audible artifacts. The proposed method is close to text-speaker dependent methods regarding conversion quality.

5. CONCLUSION

In this paper a text/speaker independent voice conversion method has been proposed. The method is based on using unified spectral parameters for spectral envelopes conversion. The codebook consists of the conversion matrix and statistical parameters of target speaker's speech. Series of experiments have been executed that approved effectiveness of this approach.

6. REFERENCES

[1] M. Abe, S. Nakamura, K Shikano. Voice conversion through vector quantization. *Proceedings of the International Conference "Acoustics, Speech, and Signal Processing. (ICASSP-88)"*, New York, USA, 1-14 April 1988, pp. 655-658.

[2] R. Schafer, L. Rabiner. Design and simulation of a speech Analysis-Synthesis-System based on Short-Time Fourier analysis, *IEEE Trans. AU-21*, No. 3, June 1973, p.165.

[3] P. Zubrycki, A. Petrovsky. Accurate speech decomposition into periodic and aperiodic components based on discrete harmonic transform. *Proc. of the 15th European Signal Process. Conf., (EUSIPCO-2007)*, Poznan, 2007, pp.2336-2340.

[4] L.B. Almeida, J.M. Tribolet *Nonstationary spectral modeling of voiced speech, IEEE Trans. on Acoust., Speech and Sig. Proc.* ,Vol. ASSP-31. no. 3. pp. 664 – 678, 1983.

[5] E. Azarov, A. Petrovsky, M. Parfieniuk. Estimation of the instantaneous harmonic parameters of speech. *Proc. of the 16th European Signal Process. Conf., (EUSIPCO-2008)*, Lausanne, 2008, CD-ROM.

[6] KY Lee, Y Zhao, Statistical Conversion Algorithms of Pitch Contours Based on Prosodic Phrases. *Proceedings of the International Conference* "Speech Prosody 2004". *(SP 2004)"*, Nara, Japan March 23-26 2004, CD-ROM.