REAL-TIME VOICE CONVERSION BASED ON INSTANTANEOUS HARMONIC PARAMETERS

Elias Azarov, Alexander Petrovsky

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

ABSTRACT

The paper presents a voice conversion framework that can be used in real-time applications. The conversion technique is based on hybrid (deterministic/stochastic) parametric speech representation. The conversion approach has been tested in two modifications for narrow-band and wide-band speech signals. Though the real-time requirement adds some significant limitations (frame by frame processing) the approach provides high quality of the reconstructed speech and recognizability of the target speaker's identity due to improved feature mapping. The proposed solution is embedded in a mobile communication system as an entertainment service.

Index Terms— Real-time voice conversion, parametric speech representation.

1. INTRODUCTION

Voice conversion can be considered as a feature transformation process that provides mapping of features from different domains. The most common parameterization techniques used in voice conversion are based on short time spectrum estimates. Mel-frequency cepstral coefficients (MFCC), mel-generalized cepstral coefficients (MGC) and line spectral frequencies (LSF) are commonly used for efficient spectral envelopes representation [1]. The choice of a specific parameterization technique is very important in voice conversion systems design.

A conventional system involves two main phases: training and conversion itself. The most common approach to voice conversion implies training on parallel sentences uttered by source and target speakers. The result of the training is a conversion function that contains the rules of the transformation. The most sophisticated voice conversion systems implement mapping between entire phonetic speech units and prosodic features in order to model identity of the target speaker. However in real-time voice conversion applicable methods are confined to frame-based transformations. In most cases only short-time spectral envelope and fundamental frequency are mapped. The problem of spectrum envelopes mapping has a number of solutions among which are codebook mapping, space modeling with Gaussian mixture models (GMM) [2] and frequency warping (FW) [3]. It is known that a single linear transformation function is not as effective as a set of local transformation functions each for one Gaussian [1] on the other hand acoustic space clusterization leads to known overfitting and discontinuity problems. Moreover, estimating target envelopes as a weighted sum of linear regression models leads to oversmoothing.

The solution to voice conversion problem, proposed in the present work, is based on instantaneous parametric speech representation. The speech signal is decomposed into deterministic/stochastic parts by means of instantaneous harmonic analysis [4]. Localized parametric description of the signal enables to perform accurate dynamic time warping (DTW) that is combined in the proposed system with iterative estimation of the conversion function. The results, obtained in this work show that due to aforementioned features it is possible to perform high quality real-time voice conversion using single linear transformation function and short training sequences.

2. OVERVIEW OF THE CONVERSION SYSTEM

In order to estimate the best possible conversion function for envelopes transformation the system uses a speech database with parallel utterances of the source and target speakers. The method of envelopes conversion is based on linear regression. The aim of the training procedure is to estimate values of a conversion matrix (regression coefficients) that fits source and target training sequences. The training process illustrated in Figure 1 consists of the following steps:

1. Parameterization of parallel speech phrases. The speech is represented as sets of harmonic and noise envelopes and pitch.
2. Initial regression coefficients set.
3. Iterative dynamic time warping of converted and target sequences and updating regression coefficients.

Speech is represented as a set of harmonic/residual envelopes and pitch contours. Then the iterative time warping procedure along with coefficients estimation is carried out. Recalculation of DTW for converted and target envelopes provides accurate synchronization of the training sequences eliminating influence of the difference between

vocal tract lengths. The conversion process illustrated in Figure 2 can be divided into following steps:

1. Parameterization of the source speech signal.
2. Transformation of the harmonic and residual envelopes using regression coefficients estimated at training phase.
3. Transformation of the pitch contour using statistical rules.
4. Synthesis of the waveform of the output signal.

Because of real-time condition the system processes speech signal frame by frame without modification of tempo of the speech.

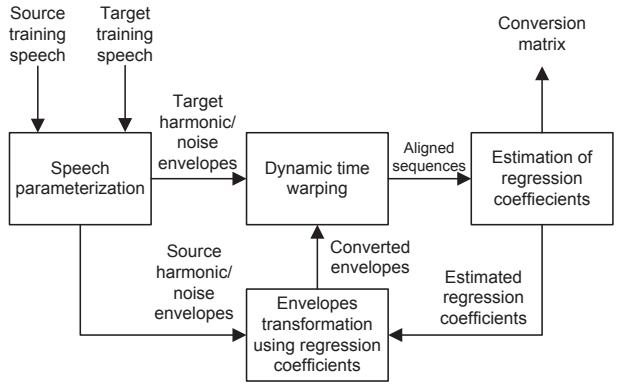


Figure 1 - Training scheme

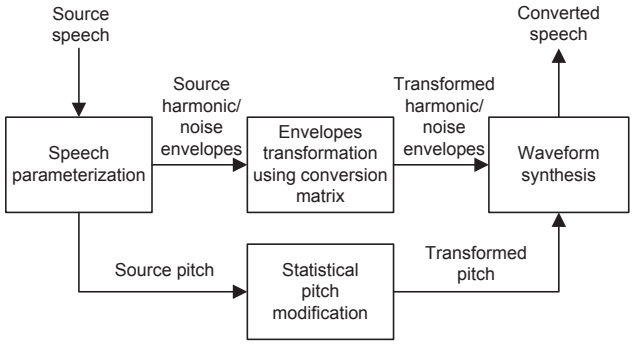


Figure 2 - Conversion scheme

3. SPEECH PARAMETERIZATION BASED ON INSTANTANEOUS HARMONIC PARAMETERS

The hybrid deterministic/stochastic model assumes that the signal $s(n)$ can be expressed as the sum of its periodic and noise parts:

$$s(n) = \sum_{k=1}^{K} \mathrm{MAG}_k(n)\cos\varphi_k(n) + r(n),$$

where $\mathrm{MAG}_k(n)$ - the instantaneous magnitude of the k-th sinusoidal component, K is the number of components, $\varphi_k(n)$ is the instantaneous phase of the k-th component and $r(n)$ is the stochastic part of the signal. Instantaneous phase $\varphi_k(n)$ and instantaneous frequency $f_k(n)$ are related as follows:

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

where F_s is the sampling frequency and $\varphi_k(0)$ is the initial phase of the k-th component. The harmonic model states that frequencies $f_k(n)$ are integer multiples of the fundamental frequency $f_0(n)$ and can be calculated as:

$$f_k(n) = k f_0(n).$$

The harmonic model is often used in speech coding since the instantaneous harmonic parameters $\mathrm{MAG}_k(n)$, $f_k(n)$ and $\varphi_k(0)$ represent voiced speech in a highly efficient way.

Instantaneous harmonic parameters are calculated through the technique based on analysis filters. Filter bands are recalculated for each frame of the signal using estimated pitch values. Given a speech frame, multiplied by a window function, $s(n)\ \ 0 \le n \le N-1$ and a filter passband specified by center frequency contour $F_c(n)$ and bandwidth $2F_\Delta$, instantaneous magnitude $\mathrm{MAG}(n)$, phase $\varphi(n)$ and frequency $f(n)$ are calculated as [4]:

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

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

$$f(n) = \frac{\varphi(n+1) - \varphi(n)}{2\pi} F_s.$$

where

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

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

$$\varphi_c(n,i) = \begin{cases} \sum_{j=n}^{i} F_c(j), & n < i \\ -\sum_{j=i}^{n} F_c(j), & n > i. \\ 0, & n = i \end{cases}$$

Central frequencies of the filter bands are calculated as the instantaneous fundamental frequency multiplied by the number k of the respective harmonic $F_C^k(n) = k f_0(n)$.

The procedure goes from the first harmonic to the last, adjusting fundamental frequency at every step. The fundamental frequency recalculation formula can be written as follows:

$$f_0(n) = \sum_{i=0}^{k} \frac{f_i(n)\mathrm{MAG}_i(n)}{(i+1)\sum_{j=0}^{k}\mathrm{MAG}_j(n)}.$$

The fundamental frequency values become more precise while moving up the frequency range. It allows making proper analysis of high order harmonics with significant frequency modulations.

Harmonic envelopes are calculated from instantaneous harmonic parameters using linear interpolation. The deterministic part of the signal is synthesized using the estimated harmonic parameters and subtracted from the source signal frame in order to obtain the residual. The residual (stochastic) part of the signal $r(n)$ is parameterized as a bark-band noise. The noise envelopes are calculated as energies of the signal in bark subbands.

After applying the parameterization technique the speech signal is represented as a set of instantaneous harmonic envelopes, short-time noise envelopes and a pitch contour. All these parameters are transformed during voice conversion technique and then the resulting waveform is synthesized.

4. PARAMETRIC VOICE CONVERSION

4.1. Training phase

As was said above the spectral envelopes are converted using linear transformation. In order to estimate needed regression coefficients the training procedure uses source and target sequences of spectral envelopes. The required conversion function can be written in the following matrix form:

$$E_t = E_s^T \cdot K,$$

where E_t - target spectral envelope vector, E_s - source spectral envelope vector, K - conversion matrix. The conversion function is a coefficient matrix that provides estimation of target spectral envelopes with minimum estimation error. The training procedure involves solution of a system of linear equations that minimizes the error in the least square sense (T denotes transposition):

$$\min_K |E_t - E_s^T \cdot K|,$$

The training sets (sequences of E_t and E_s) are estimated from speech of the source and target speakers using the technique described above. The envelope vectors consist of concatenated harmonic and noise parts. The training procedure is iterative and implies dynamic time warping of the training sequences at every step. The sequences are warped in order to minimize the conversion error. The spectral-distortion measure minimized by means of dynamic programming is mean log spectral distance between converted and target sequences.

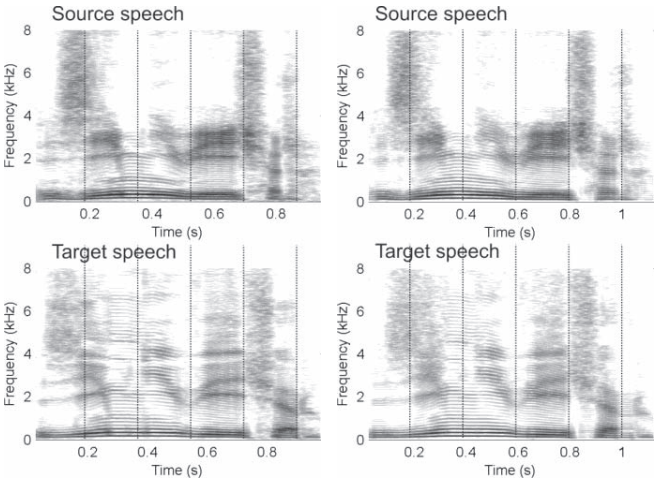


Figure 3 - An example of iterative DTW based on linear regression (a) – before DTW, (b) – after DTW

Using converted sequence instead of source reduces influence of the speaker-specific features on time warping result. The method works surprisingly well and gives accurate estimates of the warping function. A result of iterative DTW based on linear regression is shown in figure 3. The training procedure has a fast convergence and requires just a few iterations.

Using only one linear function it is possible to transform instantaneous spectral envelopes directly providing continuous converted values.

4.2. Conversion phase

Speech signal is split into 50ms frames with 10ms offset. Pitch values are estimated using analysis filters described above as was presented in [4]. The converted 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 [5]. The transformed pitch value $p_t^{S\to T}$ is estimated as:

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

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 estimated during training phase. Harmonic and noise envelopes are transformed using conversion function and then output speech signal is synthesized.

The conversion technique was implemented in real-time. The system was originally designed for mobile communication application. In order to communicate with Global System for Mobile Communications (GSM) the

Session Initiation Protocol (SIP) was used. Overall 30ms latency of the analysis-conversion-synthesis sequence was achieved on a personal computer due to optimized C++ code.

5. EVALUATION

5.1. Conversion systems

The proposed conversion system was implemented in two versions: for narrow-band ($F_s = 8$kHz) and wide-band ($F_s = 44.1$kHz) speech. The narrow-band version was tested with both uncompressed speech signals and speech signal, passed through GSM.

5.2. Database

Four different voices were used in the experiments – two male and two female voices (labeled as m1, m2, f1, f2 respectively). The database contained 100 Russian sentences per speaker. The overall duration of the speech database was about 3 minutes per speaker. The sentences were uttered with different intonations, natural for the speakers; however, the text of the sentences was the same in order to create parallel training sequences. All the sentences were sampled at 44.1kHz, then resampled at 8kHz and transmitted through GSM. For training 10 parallel sentences were used for each conversion method.

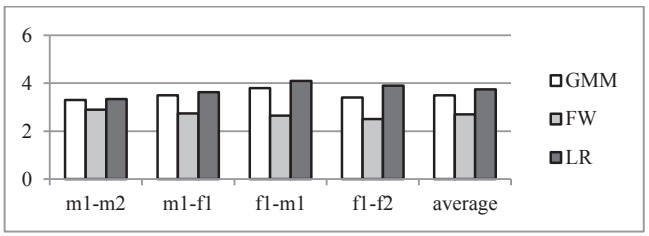
5.3. Subjective tests

Mean opinion scores tests were carried out in order to evaluate quality and identity of the reconstructed speech. The subjective test compared three different methods: the GMM-based approach (labeled as GMM), frequency warping approach (labeled as FW) and the proposed technique (labeled as LR).

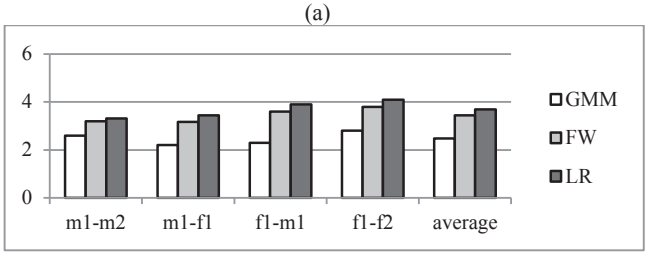
Ten listeners participated in evaluation experiments. They were asked to listen converted-target sentence pairs corresponding to three different bases (uncompressed 44.1kHz, uncompressed 8kHz and GSM 8kHz) and four conversion directions. First the listeners rated the sentence pairs on a scale from 1="different person" to 5="the same person" regardless to the speech quality. Then the listeners rated the quality of the reconstructed speech on the scale 1="bad – synthetic speech" to 5="excellent – natural speech" using the same sentence pairs. The listening results for wide-band speech are presented in figure 4. The average scores for narrow-band uncompressed speech are listed in table 1.

Table 1 - Average mean opinion scores for narrow-band speech

	Identity			Quality		
	GMM	FW	LR	GMM	FW	LR
Uncompressed	**3,1**	2,1	**3,1**	2,5	3,1	**3,4**
GSM	2,8	2,2	**2,9**	2,2	2,9	**3,0**



(a)



(b)

Figure 4 - Mean opinion scores for wide-band speech
(a) – identity, (b) – quality

6. CONCLUSIONS

A system for real-time voice conversion has been proposed. The implemented speech parameterization technique performs deterministic/stochastic decomposition by means of instantaneous harmonic analysis. An iterative dynamic time warping scheme has been proposed in the training phase that provides accurate alignment of the training sequences subject to the speaker-depended features of the spectral envelopes. The conversion scheme uses only one linear transformation function, however the method showed good results in comparison with GMM and FW-based methods.

7. ACKNOWLEDGMENTS

This work was supported in part by the IT Mobile Company (Moscow, Russia).

8. REFERENCES

[1] E. Helander, et. al. "Voice Conversion Using Partial Least Squares Regression," IEEE Transcations on Audio, Speech, and Language Processing, Vol. 18, No. 5, pp. 912-921, July 5 2010.

[2] Y. Stylianou, O. Capp.e, and E. Moulines. Continuous probabilistic transform for voice conversion. IEEE Trans. Speech and Audio Processing, Vol. 6, No. 2, pp. 131–142, 1998.

[3] D. Erro, A. Moreno and A. Bonafonte "Voice Conversion Based on Weighted Freqeuncy Warping," IEEE Transcations on Audio, Speech, and Language Processing, Vol. 18, No. 5, pp. 922-931, July 5 2010.

[4] E. Azarov and A. Petrovsky, "Instantaneous harmonic analysis for vocal processing" in *Proc. DAFx-09*, Como, Italy, September 1-4, 2009.

[5] 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.