Voice Conversion Systems with Web Interface

Elias Azarov, e-mail: azarov@bsuir.by
Maxim Vashkevich, e-mail: vashkevich@bsuir.by
Denis Likhachov, e-mail: likhachov@bsuir.by
Alexander Petrovsky, e-mail: palex@bsuir.by
Computer engineering department
Belarusian State University of Informatics and Radioelectronics, Minsk, Belarus

Two speech processing systems have been developed for real-time and non-real-time voice conversion. Using the real-time version the user can apply conversion during voice over IP (VoIP) calls imitating identity of a specified target speaker. Non-real-time processing system converts prerecorded audio books read by a professional reader imitating voice of the user. Both systems require some speech samples of the user for training. The training procedures are similar for both systems however the user is considered as a source speaker in the first case and as a target speaker in the second. For parametric representation of speech we use a speech model based on instantaneous harmonic parameters with multicomponent sinusoidal excitation. The voice conversion itself is made using artificial neural networks (ANN) with rectified linear units.

1. Introduction

In this paper we present a voice conversion technique that has been implemented in two versions for real-time (referred to as 'CloneVoice') and non-real-time (referred to as 'CloneAudioBook') speech processing. CloneVoice is intended for VoIP communications and allows the user of the system to speak somebody else's voice. The current implementation of the system establishes VoIP to GSM connection using a voice conversion server as shown in figure 1. In order to get access to the voice conversion server a dedicated application is designed for iPhone.

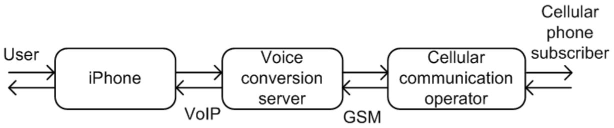

Figure 1: Schematic representation of real-time voice conversion using VoIP

CloneAudioBook is applied to prerecorded audio books which are stored in a database. The audio book chosen by the user is processed by the voice conversion server and then can be downloaded using a web interface as shown in figure 1. The aim of the conversion is to change the voice of the original reader to the voice of the user.

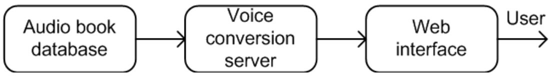

Figure 2: Schematic representation of non-real-time voice conversion for audio books

Before conversion can be performed the user is asked to utter a set of phrases which are used for training of the voice conversion function. The development of these voice conversion applications was inspired by the recent success of neural network applied to voice conversion [1] and recent advancement of contemporary speech morphing models [2].

2. Implementation

The system is divided into two main stages: training and conversion as shown in figure 3. For training parallel utterances of the source and target speakers are used. They are aligned in time and then the conversion function is trained using ANN. The conversion function matches features of the source speaker to those of target speaker. The training core is implemented in MATLAB and compiled into executables using the built-in compiler.

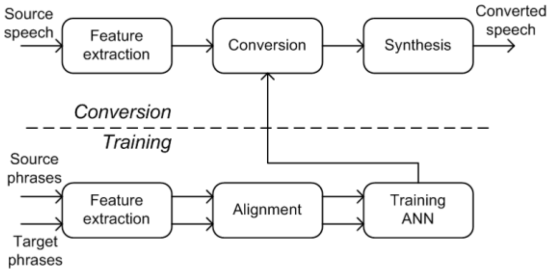

Figure 3: Schematic representation of the voice conversion system

During conversion the conversion function is applied to speech features and then the waveform of output speech is synthesized. Since the conversion stage is time critical it is implemented in C++.

For training and conversion a parametrical representation of speech is used. Instantaneous spectral envelope, pitch and excitation type (voiced, unvoiced or mixed) are extracted for each 5 ms of the signal. Unvoiced speech is modeled as a random process with a specified spectral power density. For spectral mapping we use a feed-forward ANN that consists of four layers. The ANN uses rectified linear units that implement the function $RL(x) = \max(0, x)$. The network performs mapping between the source and target envelopes and uses speakers' state vectors and to reduce oversmoothing. The state vectors are calculated using current normalized pitch values and voiced/unvoiced decisions. Weights and biases of the ANN are trained using the backpropagation algorithm.

3. Evaluation

CloneVoice and CloneAudioBook operate under different conditions considering time constrains and quality of the source speech. The input of CloneVoice is a noisy speech, recorded by iPhone in natural conditions sampled at 16 kHz, the output is sampled at 8 kHz which subsequently processed with GSM encoding/decoding scheme. The input of CloneAudioBook is a clean speech sampled at 44.1 kHz recorded in an audio recording studio and the output is sampled at 44.1 kHz as well. So it is very naturally to expect different performance of the systems regarding perceptual quality of the converted speech. We have performed some subjective evaluations which are summarized in table 1.

Table 1. Subjective evaluations of voice conversion quality (mean opinion scores)

	CloneVoice	CloneVoice (b.p. mode)	CloneAudioBook
Intelligibility	3,1	3,4	4,1
Quality	3,0	3,3	4,4
Similarity	2,9	-	3,9

Listeners were asked to rate perceptual quality of the processed speech in 1-to-5 scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad) in terms of intelligibility (how easy the words are recognized), quality (naturalness) and similarity to the target speaker. In order to evaluate the influence of the communication channel CloneVoice system has been tested in two modes: full processing mode and bypass mode where no voice conversion is applied.

References

[1] S. Desai, A.W. Black, B. Yegnanarayana, and B. Prahallad. “Spectral mapping using artificial neural networks for voice conversion,” IEEE Trans. Audio, Speech and Language Processing, Vol. 18, No. 5, pp. 954-964, 2010.

[2] H. Kawahara, R. Nisimura, T. Irino, M. Morise, T. Takahashi, and B. Banno, “Temporally variable multi-aspect auditory morphing enabling extrapolation without objective and perceptual breakdown,” Proc. ICASSP, Taipei, Taiwan, April 2009.

[3] Azarov, E., Vashkevich, M., and Petrovsky A., “Instantaneous pitch estimation based on RAPT framework,” Proc. EUSIPCO, Bucharest, Romania, Aug. 2012.