Real-time pitch modification system for speech and singing voice

Elias Azarov, Maxim Vashkevich, Denis Likhachov, Alexander Petrovsky

Computer engineering department, Belarusian State University of Informatics and Radioelectronics, 6, P.Brovky str., 220013, Minsk, Belarus

azarov@bsuir.by, vashkevich@bsuir.by, likhachov@bsuir.by, palex@bsuir.by

Abstract

A real-time pitch modification system has been developed. The implemented processing scheme is based on hybrid deterministic/stochastic decomposition of the signal and includes extraction of instantaneous pitch, pitch-synchronous time-frequency analysis, parametrical morphing and synthesis. The scheme provides high quality output with considerably high naturalness. The aim of the presentation is to show capabilities of the designed real-time signal processing framework. The system implements speech-specific intonation change routines such as lowering, uplifting, tremolo etc. In order to make the presentation more expressive we designed a special singing mode in which the system automatically corrects wrong notes. The target melody and voice effects are specified using musical instruments digital interface (MIDI).

Index Terms: pitch estimation, pitch correction, harmonic modeling.

1. Introduction

The most convenient approach to pitch manipulation is harmonic modeling which implies decomposition of the signal into periodical components of multiple frequencies and extraction of their parameters [1,2]. Separation of speech into deterministic and stochastic components is required for reducing audible artifacts. The present work is aimed at improving analysis/synthesis techniques of harmonic modeling focusing on their practical aspects. We introduce a real-time processing system that implements instantaneous harmonic parameterization, parameters modification and synthesis.

The system provides accurate instantaneous pitch and harmonic parameters extraction, fine voiced/unvoiced classification for each harmonic and low-aliasing synthesis that altogether results in noticeable improvement in quality of reconstructed speech.

The system is implemented on a personal computer (PC). It captures the users voice via microphone and plays the processing result through loudspeakers (headphones). The user can hear their own processed voice with fixed latency (around 100–150ms).

The system can operate in two modes: speech and singing voice processing. In speech mode the signal is processed according to a chosen pitch correction pattern, performing change of key and intonation. In singing voice mode the system acts as a karaoke player. It plays backing (and optionally lead tune) of a chosen song and processes user's voice fitting it to the tune. Synthesis of a personalized singing voice is a difficult task [3]. So singing as an extreme phonation mode is very indicative of capabilities of the model.

2. Implementation

The system is implemented as a frame-based processing routine and includes the following functional blocks: pitch estimation, time-warping, harmonic parameters extraction, voiced/unvoiced detection, parameters modification and synthesis as shown in figure 1.

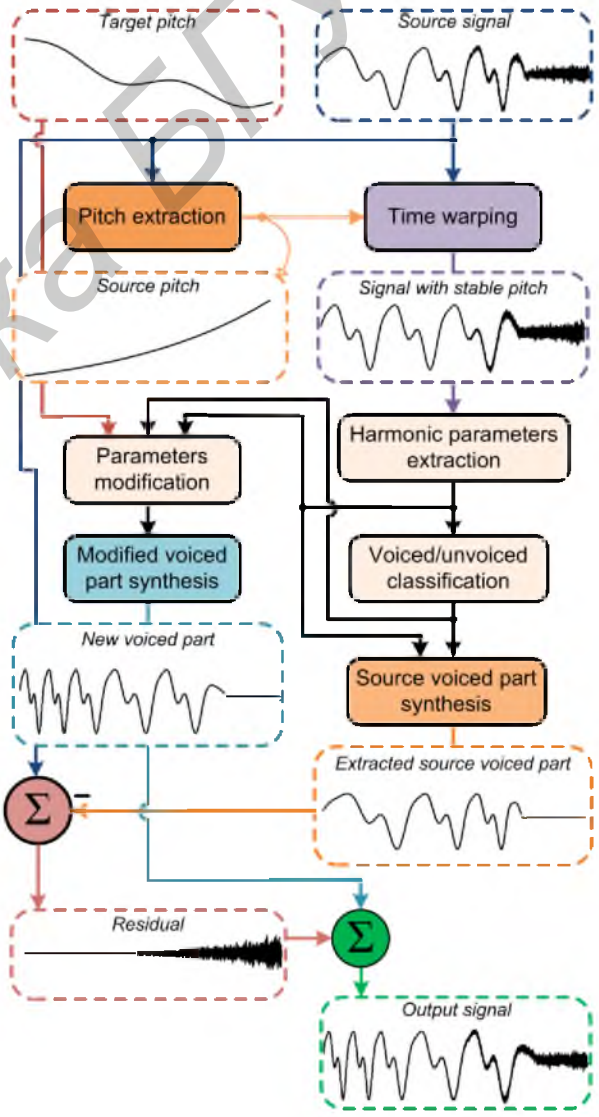

Figure 1: *Voice processing scheme.*

The scheme operates on frames of variable length (which depends on pitch value) with 5ms offset. Parameters of underlying harmonic model are updated each 5ms. The sampling rate is 44.1 kHz.

2.1. Parameters extraction

Since the analysis scheme is pitch-dependent, its performance strongly depends on robustness and time-frequency resolution of pitch estimation. We designed a special pitch extraction algorithm based on multirate sampling, which decomposes the signal into complex subbands and estimate current pitch from their instantaneous frequency values. Pitch is extracted each 5ms and then interpolated between adjacent frames for each sample of the signal.

We apply time-warping to the input signal which implies adaptive resampling of the signal with variable sample rate proportional to the extracted instantaneous pitch. Time-warping stabilizes pitch and reduces smoothing in frequency domain caused by pitch modulations.

The harmonic parameters are extracted from the warped signal using a DFT-modulated analysis filter bank. Since the warped signal has stable pitch the analysis window of the bank always contains fixed number of pitch periods. Using estimates of amplitude, instantaneous frequency and phase of each harmonic we evaluate spectral envelope.

Using instantaneous frequency values each subband signal of the filter bank is classified as periodic or stochastic. Decisions for each harmonic are combined into voiced/unvoiced spectral regions.

Voiced part is synthesized and subtracted from the input signal that gives residual which is not changed during pitch modification.

2.2. Synthesis

Extracted harmonic parameters are modified according to the target pitch, retaining original shape of the amplitude envelope. New phases of the harmonics are generated continuously, approximating original relative offset with respect to the phase of the fundamental frequency component.

Due to processing in the warped time domain where frequencies of harmonics are fixed, it is possible to implement an efficient synthesis scheme with antialiasing filtering. We use a DFT-modulated synthesis filter bank in which each channel corresponds to a harmonic component. Decimated subchannel sequences are generated using modified harmonic parameters and interpolated with the filter bank that results in an output waveform with stable pitch. To get the final output signal inverse time-warping is applied to the waveform according to the target pitch contour.

2.3. Generating target pitch

In speech processing mode the pitch is changed according to one of predefined patterns. The user can choose one from the list and adjust settings of the effect if available. The patterns include constant and adaptive lowering, uplifting, intonation change, tremolo etc. An example of speech processing is given in figure 2.

In singing voice processing mode the target pitch contour is generated using selected tune and input pitch contour of the user. The tunes are synchronized with correspondent backings and stored as MIDI events.

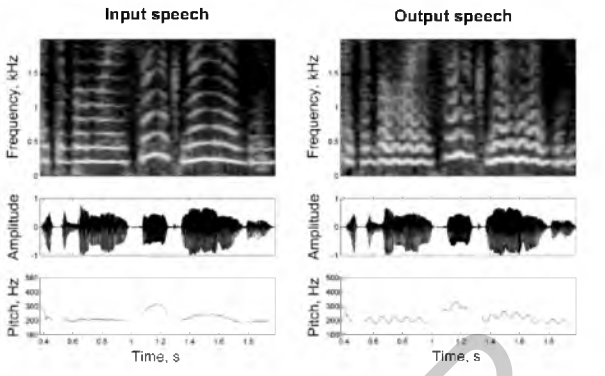

Figure 2: *Speech processing example (tremolo).*

An optimization procedure is applied to the target melody in order to attenuate 'computer accent' effect. The source pitch contour and voiced/unvoiced decisions are analyzed in order to find the best moments for transition between notes. The system can also perform a polyphonic effect by mixing several outputs with different target frequency contours.

3. Demo

The demo includes a fully functional mockup of real-time voice processing system implemented on PC. The system performs pitch correction of the input signal and operates in two modes: speech or singing voice processing.

4. Conclusions

The paper presents a voice processing real-time system based on instantaneous harmonic modeling of speech. Instantaneous pitch is extracted from input voice using newly proposed algorithm based on multirate sampling. Voiced/unvoiced classification as well as harmonic parameters extraction are made in warped time domain that considerably improves overall accuracy of the analysis. An efficient synthesis scheme operates on decimated samples and performs antialiasing filtering of each harmonic. The system produces high quality speech reconstruction and can be also applied to singing voice correction.

5. Acknowledgements

The authors are grateful to the ITForYou company for support. This work was also supported by Belarusian Republican Foundation for Fundamental Research (grant No F14MV-014).

6. References

[1] J. Bonada, and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models" *IEEE Signal Processing Magazine*, vol. 24, issue 2, pp. 67-79, March 2007

[2] J. Laroche, Y. Stylianou and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *IEEE ICASSP 1993 – IEEE International Conference on Acoustic, Speech, and Signal Processing, April 27-30, Minneapolis, USA, Proceedings*, 1993. – pp. 550-553.

[3] T. Nakano, and M. Goto, "VocaListener2: A singing synthesis system able to mimic a user's singing in terms of voice timbre changes as well as pitch and dynamics," *in IEEE ICASSP 2011*, Prague, Czech Republic, May. 2011, pp. 453-456.