Singing Voice Correction System with Smartphone Interface

Elias Azarov, Maxim Vashkevich, Denis Likhachov and Alexander Petrovsky
Belarusian State University of Informatics and Radioelectronics,
6, P.Brovky str., 220013, Minsk, Belarus
{azarov, vashkevich, likhachov, palex}@bsuir.by

Abstract

A singing voice processing system has been developed with interactive smartphone interface. The system performs automated singing voice correction according to the target melody of the song and specified audio effects. The aim of the presentation is to show capabilities of the designed signal processing framework which can be potentially used in music production and entertainment services[1].

1. Introduction and motivations

An automated voice correction system modifies user's singing to be perceived as 'professional' i.e. to be in tune with the melody of the song. There are many sophisticated solutions for singing voice processing including Vocaloid [1] and VocaListener [2] that produce really impressive results of voice morphing.

The crucial part of a voice morphing system is underlying signal model which interprets the signal in parametric domain. The present work focuses on finding a specific model for high quality voice morphing and was inspired by recent development of speech phonation models [3] and their application to singing voice processing [4]. The proposed modeling framework GUSLAR [5] is designed specifically for singing voice processing and can change pitch/tempo of the signal and add artificial polyphony. Processing of voiced speech is made in warped-time domain where it is possible to use narrow-band filters and extract harmonic and subharmonic components. Due to warped-time processing GUSLAR can be potentially beneficial for modeling various phonation phenomena such as glottalization, creaky voice, diplophonic phonation etc. This might be valuable for singing voice processing since these effects are typical in singing.

The demo presented here performs processing of user's singing on the spot. Voice recording is done using an interactive karaoke application implemented on a smartphone.

2. Scientific and technical description

2.1. Modeling of voiced and mixed sounds

Harmonic model implies manipulating of each harmonic of the signal separately. Here, in singing voice processing, GUSLAR tries to extract and process subharmonic components as well. In order to make bands of analysis filters narrow enough we utilize very long analysis frames (up to 16 pitch periods that corresponds to 35–320 ms for pitch range 450–50 Hz). It is possible to use such large windows without frequency smoothing due to time-warping which results in a signal with stable pitch. An example of harmonic analysis is shown in figure 1.

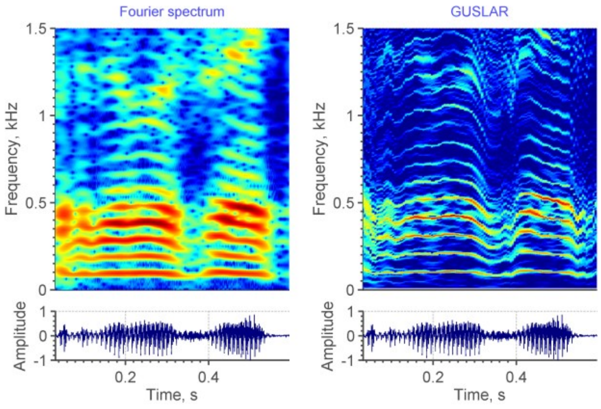


Figure 1. Time-frequency representation of voiced speech

The model considers each harmonic as a multicomponent periodic function and represents voiced speech signal $s(n)$ as

[1] The presented framework is used in an automated karaoke application developed by IT Mobile company (Moscow)

$$s(n) = \sum_{k=1}^{K} G_k(n) \sum_{c=1}^{C} A_k^c(n) \cos\left(f_k^c n + \varphi_k^c(0)\right)$$
$$= \sum_{k=1}^{K} G_k(n)\, e_k(n),$$

where $G_k(n)$ is a gain factor specified by the spectral envelope, C – number of sinusoidal components for each harmonic, f_k^c and $\varphi_k^c(0)$ – frequency and initial phase of c-th component of k-th harmonic respectively, $e_k(n)$ - excitation signal of k-th harmonic. Amplitudes $A_k^c(n)$ are normalized in order to set the unit energy to each harmonic's excitation: $\frac{1}{2}\sum_{c=1}^{C}[A_c^k(n)]^2 = 1$ for $k = 1, \ldots, K$.

According to the model the actual period of excitation can be longer than the period of pitch. That makes the model suitable for processing speech fragments with partial glottalization.

2.2. Generating target pitch contour

The target pitch contour for corrected singing voice is generated according to a given melody and the pitch estimated from user's singing. Firstly tessitura matching is carried out by shifting octaves of separate melody pieces. Then fine time alignment of source pitch and melody is made by dynamic programming (DP). This reduces audible artifacts occurring at note transitions. Then the source pitch contour heuristically segmented into notes which are drawn up to the melody. The form of original pitch contour is preserved at borders of the segments in order to attenuate the effect of 'computer accent'.

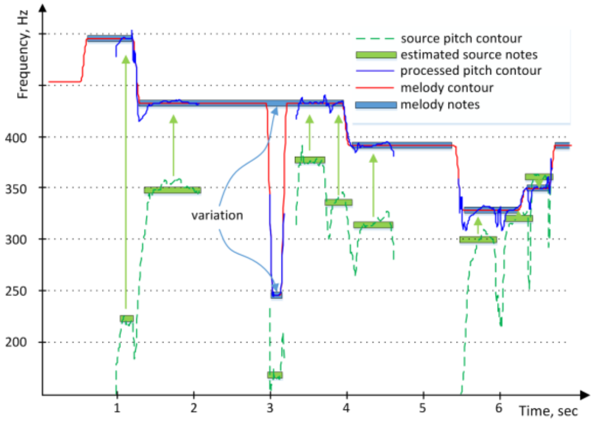


Figure 2. Changing pitch of singing voice

To reduce gaps between source and target pitch contours some melody variations are allowed. Variations are predefined in the melody by simultaneous notes and resolved during processing with DP. An example of pitch contour generation is shown in figure 2.

3. Implementation and use

The demo is implemented as an interactive internet service. Using MATLAB implementation of GUSLAR a remote server processes incoming sound files according to a given melody score. The general scheme of the voice correction system is presented in figure 3. In order to record user's voice and communicate with the server a dedicated client application is implemented on a smartphone. A typical demo session involves two steps.

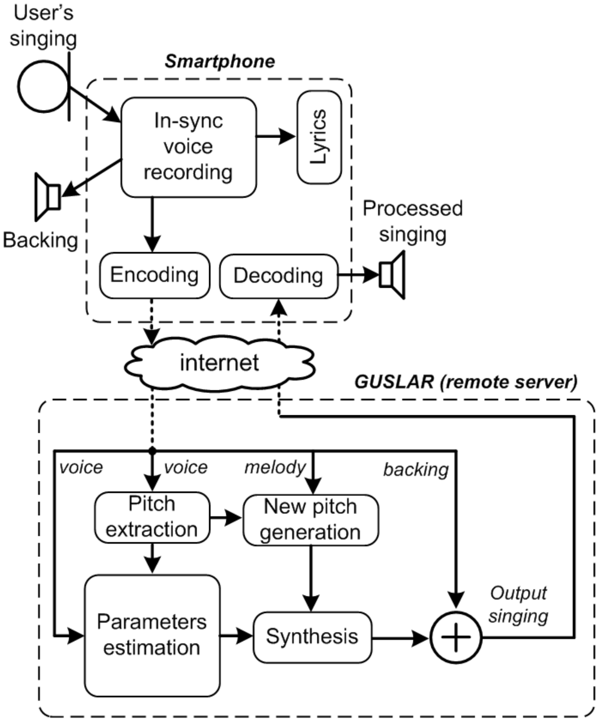
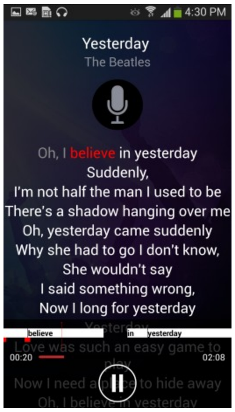


Figure 3. Singing voice correction system

At the first step the user sings while listening to the backing in earphones and seeing lyrics on the screen as shown in figure 4. When recording session is finished the data are encoded and transmitted to the server.

The second step is voice processing. The pitch contour is extracted from user's singing and then the target contour is generated using the melody of the song. Other model parameters are estimated from the signal and morphing is applied. The synthesized signal is mixed with the backing and the result is encoded and returned to the user. The user can listen to the result with the demo smartphone, or alternatively the result can be sent to a specified e-mail.



Oh, I believe in yesterday
Suddenly,
I'm not half the man I used to be
There's a shadow hanging over me
Oh, yesterday came suddenly
Why she had to go I don't know,
She wouldn't say
I said something wrong,
Now I long for yesterday
Yesterday
Love was such an easy game to play
Now I need a place to hide away
Oh, I believe in yesterday

Figure 4. User interface for recording session

Some audio samples and screenshots of the application can be found at http://dsp.tut.su/guslar_demo.html.

4. Conclusions and future developments

An automated voice correction system is presented. The interactive demo includes voice processing framework GUSLAR implemented as a server and a client smartphone application.

Future developments are aimed to further improvement of the model and real-time GUSLAR implementation.

5. Acknowledgements

The authors are grateful to IT Mobile company for support and particularly for help with implementation of the demo.

6. References

[1] H. Kenmochi, and H. Ohshita, “Vocaloid – commercial singing synthesizer based on sample concatenation,” in *Proc. Interspeech*, 2007, Antwerp, Belgium, pp. 4011–4010.

[2] 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 *Proc. IEEE ICASSP'2011*, Prague, Czech Republic, May 2011, pp. 453-456.

[3] H. Kawahara, T. Takahashi, M. Morise and H. Banno “Development of exploratory research tools based on TANDEM-STRAIGHT,” *Proc. APSIPA*, Japan Sapporo, Oct. 2009.

[4] H. Kawahara, and M. Morise "Analysis and synthesis of strong vocal expressions: extension and application of audio texture features to singing voice," *Proc. ICASSP'2012*, Kyoto, Japan, March 2012, pp.5389-5392.

[5] E. Azarov, M. Vashkevich, and A. Petrovsky “GUSLAR: a framework for automated singing voice correction,” *Proc. IEEE ICASSP'14*, Florence, Italy, May 2014. To be published.