GUSLAR: A FRAMEWORK FOR AUTOMATED SINGING VOICE CORRECTION

Elias Azarov, Maxim Vashkevich and Alexander Petrovsky

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

ABSTRACT

The paper presents a solution for singing voice processing that is used in a karaoke application with automated voice correction[1]. The intended purpose of the application is to automatically improve user's performance towards performance of a professional singer by implementation of voice effects such as pitch correction, artificial polyphony, time stretching and other. The proposed framework incorporates a complete processing workflow including analysis, morphing and synthesis. The framework uses an original model of voiced speech which represents each harmonic as a multicomponent function and provides high quality processing in conditions of partial glottalization.

Index Terms— speech analysis, speech synthesis, singing voice processing

1. INTRODUCTION

Singing voice modification has been subject of great research and consumer interest in the last decade. The main field of its application is voice processing systems of various kinds. The goal of such systems is to give a nonprofessional musician an opportunity for creating professional singing voices. The well-known singing synthesis system Vocaloid [1] can create singing voices just from typed lyrics and melody score. Functionality of the system has been recently extended by VocaListener [2] – a plug-in for personalized synthesis which can extract some characteristics of user's original voice and apply them to synthesized singing. There are some other systems that utilize melody and lyrics read by the user [3]. The synthesis application that we address in this paper is different in a way. It takes as inputs the melody of the song and original singing of the user. The user tries to sing as good as they can and the system corrects singing according to the given melody. The goal is to preserve original voice and subtle nuances of user's performance as accurately as possible.

The proposed solution is based on the speech modeling system GUSLY [4]. The model is experimentally applied to singing voice and its performance is experimentally evaluated in the paper.

1.1. Relation to prior work

The present study focuses on further development of GUSLY in order to make it suitable for singing voice correction. Currently the most popular and respected system for voice processing is TANDEM-STRAIGHT [5] that has been successfully applied to singing [6]. Both systems are similar on conceptual level (i.e. modeling consists in extracting and manipulating pitch, temporary spectral envelopes and excitation) though use different methods for implementing all their components: analysis, morphing and synthesis. Due to its paradigm of warped-time processing and multicomponent harmonic excitation GUSLY has an ability to capture and render fine subharmonic components that is potentially beneficial for modeling various phonation phenomena such as glottalization, creaky voice, diplophonic phonation etc. This ability might be valuable for singing voice processing since these effects are typical in singing. The introduced extensions of GUSLY can be summarized as follows: 1) the estimator of instantaneous pitch [7] is now capable of processing extreme low and high voices; 2) a robust voiced/unvoiced detection scheme is proposed that utilize an artificial neural network (ANN); ANNs have been already applied for speech detection [8,9] and specifically for voiced/unvoiced speech classification [10,11]; our solution is different as far as applied for singing voice and has an advantage of time-warping which to the knowledge of the authors has not been explored in this context; 3) estimation of harmonic parameters is adapted to partial glottalization.

2. GUSLY OUTLINE

2.1. Voiced and mixed speech modeling

In GUSLY voiced and mixed excitations are considered as a quasi-periodic process with constant pitch. The pitch period determines how many harmonics are distinguished by the model. The model considers each of them as a multicomponent function and represents signal $s(n)$ as

[1] The presented framework is used in an automated karaoke application by IT ForYou company

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

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

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 $k = 1, \dots, K$. According to (1) 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.

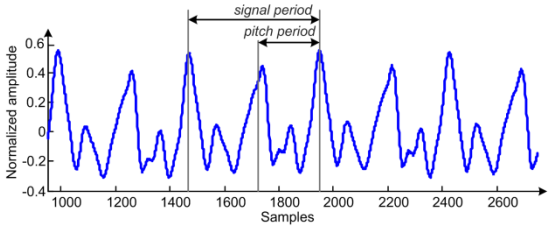

Figure 1 - Glottalization pattern

We use term 'glottalization' in the same sense as it is defined in [12] i.e. as a speech production phenomenon which is characterized by cycles of normal quasi-periodic phonations demarcated by a significantly lower vibration period of lower amplitude as shown in figure 1. In frequency domain glottalization is characterized by emergence of a regular subharmonic structure.

2.2. Analysis procedure

In GUSLY parameters of the model are estimated in warped-time domain that requires prior estimation of instantaneous pitch. Time-warping implies resampling of the signal using a constant number of samples per pitch period.

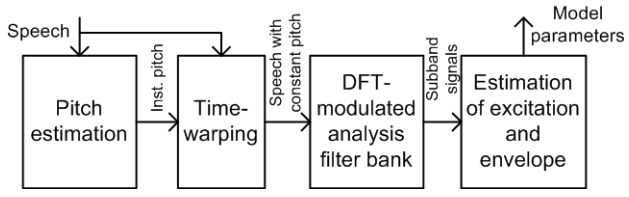

Figure 2 - Analysis scheme

Analysis workflow is shown in figure 2 and can be briefly summarized in the following way: 1) estimation of instantaneous pitch is made; 2) time warping is applied that results in a speech signal with constant pitch [13]; 3) the signal is separated into individual harmonics using a DFT analysis filter bank; 4) subband analytical signals are decomposed into instantaneous harmonic parameters using modified Prony's method.

2.3. Synthesis procedure

The output signal is synthesized using the same functional blocks applied in reverse order as shown in figure 3: 1) subband signals are synthesized using (1); the sample rate of the signals varies with target (modified) instantaneous pitch; 2) DFT synthesis filter bank is applied which performs antialiasing filtering of each harmonic component (the subband signals are decimated in order to reduce overall computational cost); 3) inverse time warping is applied to form target pitch contour.

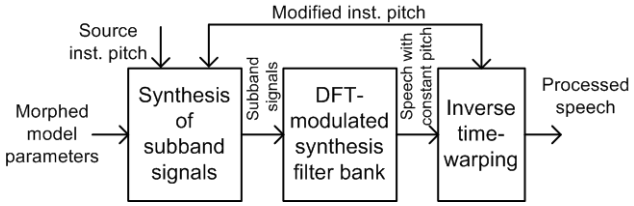

Figure 3 - Synthesis scheme

3. GUSLAR: EXTENDING GUSLY FOR SINGING VOICE PROCESSING

3.1. Pitch estimator extension

A characteristic feature of singing voice in contrast to speech is a much wider range of possible pitch values. So it is required to extend working range of the instantaneous robust algorithm for pitch tracking IRAPT [7] that is used in GUSLY. On the other hand the system is very sensitive to pitch errors so it is important to preserve accuracy and robustness of the original estimator. The key idea of IRAPT is to use instantaneous parameters of sinusoidal modeling for calculation of so-called instantaneous normalized cross-correlation function (INCCF) that is used as a period candidate generation function. To estimate parameters of the model speech is split into analytical subband signals. The bandwidth of the subbands should separate individual harmonics and at the same time be wide enough to process rapid frequency variations. Thus for a bass voice very narrow channels (about 25-30 Hz) should be used, however, temporal resolution for soprano in this case would be very poor.

To solve the problem we introduce segmented calculation of INCCF. The idea is to upsample and downsample the signal (by factors 4 and 2 respectively) and apply the same DFT filter bank as in IRAPT (figure 4). This produces three different sets of instantaneous parameters each of them corresponding to a specific segments of

allowed lag range. For upsampled version instantaneous amplitudes are filtered by a low-pass filter in order to attenuate multicomponent subband signals. Segmented INCCFs are calculated from corresponding parameter sets and then combined into resulting period candidate generation function.

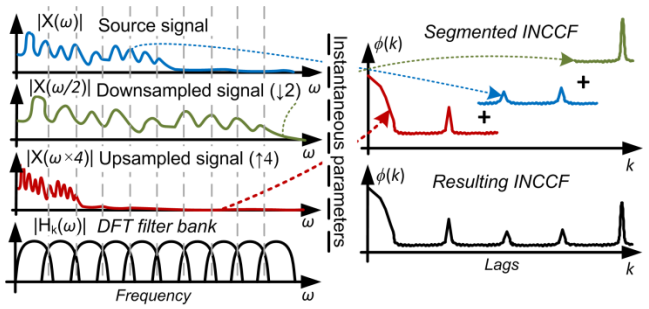

Figure 4 - INCCF calculation

Remaining steps of pitch estimation process are performed as in IRAPT-1 [7].

3.2. Voiced/unvoiced classification

Inaccurate voiced/unvoiced classification leads to annoying artifacts, especially when pitch modifications are significant. As far as the proposed analysis scheme extracts instantaneous harmonic parameters it seems reasonable to use them for voiced/unvoiced detection. A harmonicity measure based on instantaneous frequency was given in [14] that can be used as a possible solution. In investigations [15,16] was shown that detector's capabilities can be further improved by time-warping. However the problem that emerges when combining these ideas is that voiced/unvoiced detector should be robust to typical pitch estimation errors like halving or doubling. To reduce impact of the pitch errors it is necessary to combine measurements of adjacent frames. We found that in the current conditions a good alternative is to use an ANN. Advantage of ANN is that the performance of the estimator can be improved by adding inaccurately classified cases in the training dataset. The structure of the proposed detector is shown in figure 5.

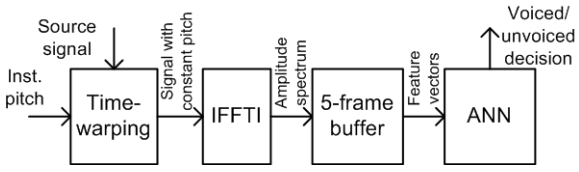

Figure 5 - Vocied/unvoiced decision scheme

Time-warped signal is divided into overlapping frames and log spectrums are calculated using the Discrete Fourier Transform (DFT). Spectrums of five adjacent frames are concatenated into feature vectors which are given to ANN's input. Time warping removes pitch variations and due to external estimation of instantaneous pitch each harmonic is placed on a predefined set of bins. This significantly narrows parameter space and provides better generalization of the ANN as will be shown in the experimental section.

3.3. Excitation parameters estimation

Applying parametric morphing to singing voices in practice, we found that non professional performers often introduce partial or even pronounced glottalization. It is true especially for male singers.

Assuming that pitch is constant it is possible to perform individual separation of subharmonic components by applying a filter bank with double or triple number of channels and correspondent lengthening of the filter prototype (we use 8 pitch periods in original GUSLY and 16 in its current modification). In GUSLY excitation parameters are extracted using modified instantaneous Prony's method. Each subband signal of the analysis filter bank is represented as a sum of sinusoids with close frequencies using derivatives of the signal. The method performs quite well but its computational cost is rather high. We found that when the number of channels is increased further decomposition of subband signals by Prony's method does not introduce noticeable quality improvements and therefore can be skipped.

3.4. Excitation synthesis

During synthesis of excitation signals it is important to preserve phase synchronization between harmonics in order to reduce phasiness (reverberation) artifacts [17]. Investigation of this effect and some notable solutions were presented in [18,19]. According to (1) in the current modeling framework we can directly synchronize periodical components of excitation signals using relative phase parameter $\Delta\varphi_k^c(n)$

$$e_k(n) = \sum_{c=1}^{C} A_k^c(n)\cos\left(f_k^c n + \Delta\varphi_k^c(n)\right) \qquad (2)$$

which is calculated as $\Delta\varphi_k^c(n) = \varphi_k^c(n) - \frac{\varphi_0(n) f_k^c}{f_0}$. The relative phase parameter $\Delta\varphi_k^c(n)$ is unwrapped along sample indices n. Using (2) for excitation synthesis preserves shape of the waveform and results in natural synchronization of voice pulses to correct relative positions.

When pitch is changed frequencies of excitation components are changed as well by direct and inverse time-warping. Their amplitudes $A_k^c(n)$ and relative phases $\Delta\varphi_k^c(n)$ are interpolated in polar coordinates according to source and target instantaneous pitch values.

Output waveform of the signal can be synthesized from modified model parameters either using original GUSLY's scheme (see figure 3) or the overlap-add method (OLA), both producing close subjective quality. Considering that our target application operates in non real-time mode we used OLA for the reason of implementation simplicity,

though GUSLY's scheme benefits from antialiasing filtering (introduced by DFT synthesis filter bank) and much lower computational cost.

4. EXPERIMENTAL EVALUATION

4.1. Pitch tracking algorithm

In order to evaluate performance of the proposed pitch detection algorithm the PTDB-TUG speech database [20] is used. To emulate signing voices we artificially scale pitch by upsampling and sampling rate adjustments. The proposed algorithm (denoted as '$IRAPT_3$') is compared to RAPT [21], YIN [22], SWIPE' [23] in terms of gross pitch error (GPE) and mean fine pitch error (MFPE) [7] – table 1.

	Male speech		Female speech	
	GPE	MFPE	GPE	MFPE
RAPT	2.24	1.72	6.99	1.19
YIN	1.78	1.26	**5.95**	**0.83**
SWIPE'	**0.94**	1.32	6.77	1.11
$IRAPT_3$	1.76	**1.19**	6.77	0.91

Table 1 - Pitch detection performance evaluation

The experiment shows that the proposed technique has a close performance to other robust pitch detection algorithms.

4.2. Voiced/unvoiced detector

In this subsection we experimentally compare performance of the ANN-based voiced/unvoiced detector when operating in source and warped time domains.

A small database of speech and singing voice samples was manually classified. Then the database was converted to feature vectors in two modes: without time-warping and with time-warping. Conversion conditions are listed in table 2. For both cases each feature vector was obtained as concatenation of five successive short-time log spectrums.

Sampling rate, kHz	Window length, samples / ms	Offset, samples / ms	Available number of harmonics	Feature vector dimension
log spectrum without time-warping				
16	640 / 40	80 / 5	6->200	1600
log spectrum after time-warping				
0.63-27.3	128 / 4.7-208	80 / 3-127	10	320

Table 2 - Conditions of feature vector extraction for voiced/unvoiced classification

A neural network with logistic units and one hidden layer (100 hidden units) is used. In order to compare how good the net generalizes we use training sets of different durations as shown in figure 6. Despite the fact that after time-warping the number of features is five times smaller, the ANN generalizes much faster providing smaller error rate on the test set (3.5% versus 5.2% for two-minute training set).

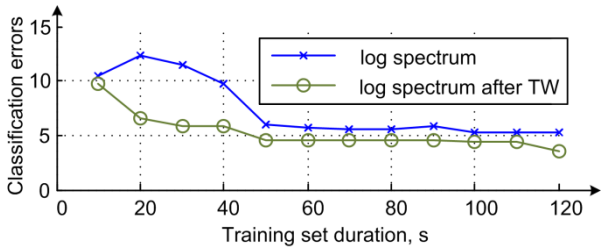

Figure 6 - Voiced/unvoiced classification errors on the test data set

4.3. Subjective modeling evaluation

Performance of the whole framework is evaluated using subjective mean opinion score (MOS) measures. We use recordings of three different Russian songs each performed by three nonprofessional singers. The melodies of the songs are written using a MIDI (Musical Instruments Digital Interface) sequencer or extracted from professional singing. For MIDI-based melodies we add smoothing and vibrato as described in [24] to make target pitch contours more natural. The voices were processed by the modeling framework resulting in pitches correction and adding artificial polyphony. Polyphonic effect is achieved by mixing several outputs with different target pitch contours. Some samples are available for download at http://dsp.tut.su/guslar_demo.zip.

Four volunteers who have a good ear for music were asked to rate naturalness and harmony of the source and corrected singings in 1-to-5 scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). Then fragments with glottalization were separated from the processed voices and the volunteers were asked to rate each of them individually. Average results of the listening tests are presented in table 3.

Overall singing		Glottalization fragments	
Naturalness	Harmony	Naturalness	Harmony
Source singing			
5	2.9	-	-
Corrected singing			
4.5	5	4.2	4.5

Table 3 - Subjective quality evaluation (mean opinion scores)

The results show that the model can effectively correct singing voices with glottalization, providing high subjective quality.

5. CONCLUSIONS

A framework for singing voice correction has been proposed. The framework is based on speech processing system GUSLY that has been extended for creating high quality singing effects. The framework's performance has been evaluated using subjective measures.

6. REFERENCES

[1] H. Kenmochi, and H. Ohshita, “Vocaloid – commercial singing synthesizer based on sample concatenation,” in *Proc. Interspeech 2007*, 2007, 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] T. Saitou, M. Goto, M. Unoki, and M. Akagi “Speech-to-singing synthesis: converting speaking voices to singing voices by controlling acoustic features unique to singing voices,” in *Proc. Interspeech 2007*, 2007, pp. 4011–4010.

[4] Azarov, E., Vashkevich, M., and Petrovsky A., “Instantaneous harmonic representation of speech using multicomponent sinusoidal excitation,” *Proc. Interspeech'13*, Lyon, France, Aug. 2013, pp. 1697–1701.

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

[6] Kawahara H., and Morise M. "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.

[7] Azarov, E., Vashkevich, M., and Petrovsky A., “Instantaneous pitch estimation based on RAPT framework,” *Proc. EUSIPCO'12*, Bucharest, Romania, Aug. 2012, pp. 2787–2791.

[8] Pham, T. V., Tang, C. T., and Stadtschnitzer, M., “Using artificial neural network for robust voice activity detection under adverse conditions”, in *Proc. RIVF'2009*, Danang, Vietnam, July 2009, pp. 1-8.

[9] Hughes, T., and Mierle, K., “ Recurrent neural networks for voice activity detection”, in *Proc. ICASSP'2013*, Vancouver, Canada, May 2013, pp. 7378–7382.

[10] Ghiselli-Crippa, T., and El-Jaroudi, A., "A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech," in *Proc. ICASSP-91.*, Toronto, Canada, April 1991, pp. 441–444.

[11] Kia, S. J., and Coghill, G.G. “A mapping neural network and its application to voiced-unvoiced-silence classification” in *Proc. First New Zealand International Two-Stream Conference on Artificial Neural Networks and Expert Systems*, Dunedin, New Zealand, Nov. 1993, pp. 104–108.

[12] Hedelin, P., and Huber, D. "Pitch period determination of aperiodic speech signals," in *Proc. IEEE ICASSP'1990*, Albuquerque, NM, April 1990, pp. 361–364.

[13] Nilsson M., Resch B., Kim Moo-Young, and Kleijn W.B. "A canonical representation of speech," in *Proc. IEEE ICASSP'2007*, Honolulu, USA, April 2007, volume IV, pp. 849–852.

[14] Artifianto, D., and Kobayashi, T., "Voiced/unvoiced determination of speech signal in noisy environment using harmonicity measure based on instantaneous frequency" in *Proc. IEEE ICASSP'2005*, Philadelphia, USA, March 2005, pp. 877-880.

[15] Malyska, N., and Quatieri, T. F., "A time-warping framework for speech turbulence-noise component estimation during aperiodic phonation" in *Proc. IEEE ICASSP'2011*, Prague, Czech Republic, May 2011, pp. 5404-5407.

[16] Wang, T., and Cuperman, V., "Robust voicing estimation with dynamic time warping" in *Proc. IEEE ICASSP'1998*, Seatle, USA, May 1998, pp. 533-536.

[17] 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.

[18] Yegnanarayana, B., Veldhuis, R., “Extraction of Vocal-Tract System Characteristics from Speech Signal”, *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 4, pp. 313-327, 1998.

[19] Bonada, J., “High Quality Voice Transformations based on Modeling Radiated Voice Pulses in Frequency Domain”, *Proc. of the 7th Int. Conference on Digital Audio Effects*, Naples, Italy, Oct 2004, pp. 291–295.

[20] G. Pirker, M. Wohlmayr, S. Petrik and F. Pernkopf, “A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario”, in *Proc. Interspeech'2011*, Florence, Italy, August 2011, pp. 1509-1512.

[21] D. Talkin, "A Robust Algorithm for Pitch Tracking (RAPT)" in "Speech Coding & Synthesis", W B Kleijn, K K Paliwal eds, Elsevier ISBN 0444821694, 1995.

[22] A. Cheveigné and H. Kawahara "YIN, a fundamental frequency estimator for speech and music", *Journal Acoust. Soc. Am.*, vol. 111, no. 4, pp 1917-1930, Apr. 2002.

[23] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music", *Journal Acoust. Soc. Am.*, vol. 123, no. 4, pp 1638-1652, Sep. 2008.

[24] Dong, M., Chan, P., Cen, L., and Li, H. “Aligning singing voice with MIDI melody using synthesized audio signal” in *Proc. of the 7th Int. Symposium on Chinese Spoken Language Processing (ISCSLP)*, Tainan, Taiwan, Nov. 2010, pp.95-98.