AN IMPROVED SPECTRAL AND PROSODIC TRANSFORMATION METHOD IN STRAIGHT-BASED VOICE CONVERSION

Long Qin, Gao-Peng Chen, Zhen-Hua Ling, Li-Rong Dai

iFLYTEK Speech Lab, University of Science and Technology of China, Hefei
qinlong@mail.ustc.edu.cn, {gpchen,zhling}@ustc.edu, lrdai@ustc.edu.cn

ABSTRACT

This paper presents a novel spectral conversion method by considering the glottal effect on STRAIGHT spectrum to improve the performance of former voice conversion system based on codebook mapping. By introducing MoG model into spectral representation, STRAIGHT spectrum is decomposed into excitation-dependent and excitation-independent components, which are transformed separately. Besides, SFC model is adopted to measure the prosodic characteristics of different speakers and realize prosodic conversion. Listening test proves that proposed method can effectively improve the discrimination and speech quality of converted speech at the same time.

1. INTRODUCTION

With the development of corpus-based TTS technique, the intelligibility and naturalness of synthesized speech has been improved a lot. But current corpus-based TTS systems always require large speech databases to synthesize voices of various speakers. So it is significant to realize a high quality voice conversion algorithm to provide current TTS system ability of synthesizing expressive voices of multi-speakers with just a few speech samples of target speakers.

Many techniques have been introduced to solve the problem of voice conversion. For spectral conversion, codebook mapping method [1] and GMM (Gaussian mixture model) method [2] are two main approaches. Considering speech synthesizer, STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weighted spectral contour) [3], as a vocoder-type analysis-synthesis algorithm, has been applied widely in voice conversion field [2][4] for its high quality of reconstructed speech and flexible ability in parameter modification.

Combining STRAIGHT and STASC (Speaker Transformation Algorithm using Segmental Codebooks), we have constructed a voice conversion system [4].Speech signals are firstly decomposed into impulse sequences and smoothed spectral envelopes by STRAIGHT. Then spectral conversion based on codebook mapping and prosodic conversion based on decision tree are implemented separately. By using a phoneme-tied weighting method, the smoothing effects on spectrum, which is caused by superposing quite different spectral code words and would eventually decrease speaker discrimination of converted speech, has been reduced greatly. Through above method, the source speaker's characteristics have been transformed successfully to that of target speaker. However, there still exist two main problems in converted speech, the temporal discontinuity of converted spectrum especially the voice quality presentation and unstable performance of prosodic prediction by decision tree. In order to solve these problems, a voice conversion method by considering the glottal effect on STRAIGHT spectrum during spectral conversion and introducing SFC model into prosodic conversion is presented here.

In the following part of this paper, an introduction to proposed method is presented in section 2. Section 3 introduces the implementation of voice conversion system. Section 4 gives the result of evaluation and section 5 is conclusion.

2. METHOD DESCRIPTION

2.1. Glottal effect on STRAIGHT spectrum

STRAIGHT is a high performance speech analysis-synthesis algorithm. Its basic idea is to decompose the speech signals into excitation and smoothed spectral envelop. However the excitation here is not common glottal waveform but impulse sequences with pitch intervals [5]. Thus the STRAIGHT spectrum consists of not only vocal tract response but also spectral characteristics of glottal waveform. Because the glottal effect presents more acoustic cues dependent on individuals and paralinguistic features instead of phoneme information, it should be converted in a different way from vocal tract features. On the contrary, our former spectral conversion method, which treats the STRAIGHT spectrum as a whole, would introduce inaccuracy and

discontinuity in voice quality of converted speech. So it is necessary to find a reliable method to extract the spectral component corresponding to spectral features of glottal waveforms.

As mentioned in [5], the spectrum of glottal waveform has two main characteristics, F_g (glottal formant) and spectral tilt. The position of F_g can be calculated as :

$$F_g = \frac{1}{2\pi \cdot OQ \cdot T_0} f(RK) \quad (1)$$

where *f(*RK*)* means a function of RK. From Eq.1 we can see that the position of F_g is dependent on T_0 and some source related parameters, such as OQ and RK. When phonation type is fixed, F_g varies linearly with F_0. If F_g of converted speech is predicted incorrectly, the source parameters related with voice quality would be changed.

Figure 1 shows one frame STRAIGHT spectrum (real line) of vowel /a/ in Mandarin Chinese pronounced by a female speaker and sampled in 8kHz. For female speakers, the first formant of /a/ is generally above 1000Hz, so the first spectral peak in Figure 1 demonstrated the existence of glottal formant in STRAIGHT spectrum.

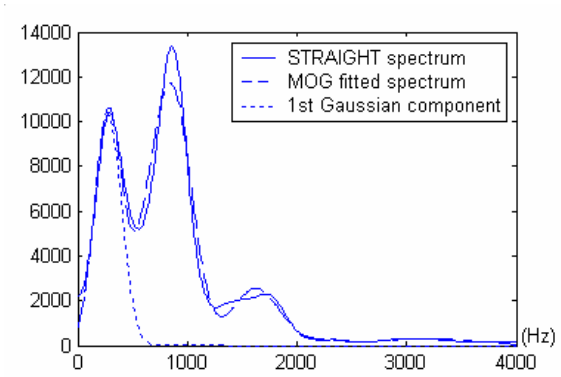

Figure 1: An example of glottal formant in STRAIGHT spectrum and the fitting result of MOG model. The frame is extracted from a vowel /a/ in Mandarin Chinese

2.2. Measuring glottal formant using MoG

As we proved the glottal formant affects the STRAIGHT spectrum greatly, we consider decompose the spectrum into excitation-dependent and excitation-independent components, so that the two parts can be modified separately.

Mixture of Gaussians (MoG) is introduced here to model the glottal formant in STRAIGHT spectrum [5]. MOG is a spectral modeling method and fits the histograms representation of speech spectrum using the mixture of Gaussians equation. Figure 1 shows the result of modeling one spectrum contour with MoG, where the dashed line presents the first Gaussian component, which is used to fit glottal formant. The result of following experiment proves that the parameters of the first Gaussian component has well linear relationship with F0 as Eq.1 and is able to capture the characteristics of glottal formant.

2.3. Spectral conversion by decomposing STRAIGHT spectrum

For spectral conversion, the first step is to remove the first Gaussian component of MoG model from both source and target speaker's STRAIGHT spectrum in the training part. Here the parameters of MoG model are estimated by Expectation Maximization (EM) algorithm. And the rest spectral envelop, treated as excitation-independent component, is modeled by an all-pole model and transformed into 20-order LSF coefficients. Then the source speaker's LSF coefficients are converted by codebook mapping. Based on the linear relationship between F_0 and F_g, we can directly rebuild the first Gaussian component of target speaker's spectrum from the predicted F_0 by SFC model. At last, the converted excitation-independent spectrum together with the predicted first Gaussian component can reconstruct the target speaker's spectral envelop.

2.4. Prosody conversion by SFC model

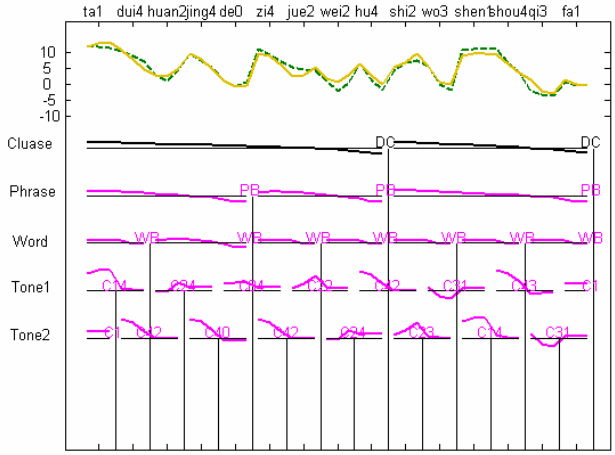

Figure 2: Decompose a melodic contour as the superposition of the contributions of four layers. The dashed line is the observed F0. The solid line is the synthesized F0.

A trainable SFC prosodic model[6] is applied to the prosody conversion which considers prosodic parameters (F0, syllabic lengthening) are interpreted as the superposition of overlapping multi-parametric contours. These contours are associated with high-level prosodic features operating at different scopes, such as tones, stress, prosodic boundary, etc. Each feature label corresponds to

a metalinguistic function (morphological, lexical, syntactic, attitudinal…) which is represented by a neural network. The observed contour is the sum of the outputs of the corresponding neural networks (Figure 2). An analysis-by-synthesis scheme is implemented for automatic learning.

We design four prosodic layers, a tone layer (two syllables' concatenative tones), a word layer, a phrase layer and a clause layer. They are the contributions to F0 in different scopes. The F0 contours of speakers are decomposed into different layers respectively. Then the mapping of the corresponding features' function in the corresponding layer between source speaker and target speaker is constructed.

3. SYSTEM IMPLEMENTATION

3.1.System framework

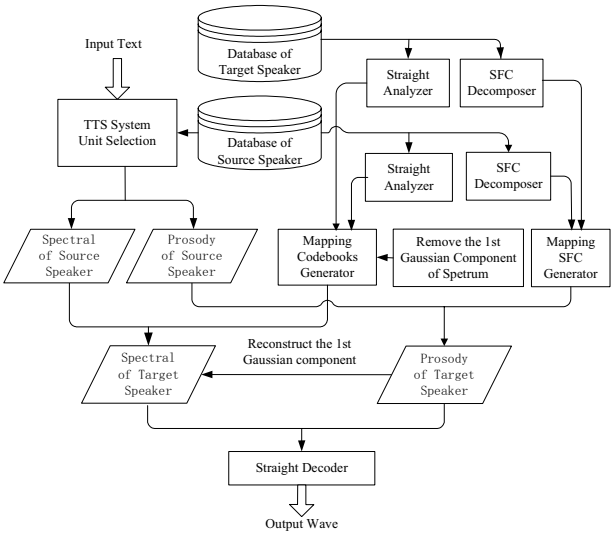

Figure 3: Flowchart of TTS system with voice conversion

The framework of our voice conversion system is illustrated in Figure 3. Our TTS system guarantees the optimal candidate units for source speaker. The units are represented by impulse sequences and smoothed spectrum by STRAIGHT. Then spectral conversion and prosodic conversion are implemented separately. The STRAIGHT spectrum is decomposed into excitation-dependent and excitation-independent component to avoid the effect of glottal formant. The excitation-dependent component is reconstructed according to the relationship between F0 and MoG parameters. While the excitation-independent component of spectrum is converted by the codebook mapping method. And the prosodic conversion is performed by SFC model. Finally, we combine the converted spectrum and prosody into the STRAIGHT decoder to synthesize target speaker's speech.

3.2.The correlation analysis between F0 and MoG model parameters

In order to prove the relationship between the first Gaussian component of MoG model and glottal formant, a correlation analysis is conducted based on 30 sentences which consist of 974 syllables and cover all vowels in Chinese of a male source speaker and a female target speaker. For each syllable, the pitch contour and spectral envelop are decomposed by STRAIGHT and 3 continuous frames of spectrum with 10 ms interval are extracted from the middle part of each vowel. Then the MoG parameters of every frame spectral envelop are calculated. Here, the speech waveforms are resampled to 8kHz and the number of mixtures is set to 6. As Table 1 and 2 illustrate, that the mean and the standard deviation of the first Gaussian component have a close relationship with F_0, which has been listed in Eq.1.

No.	Mean	Std. Deviation
1	**0.853**	**0.701**
2	-0.109	0.024
3	0.036	0.056
4	0.048	0.019
5	0.071	-0.068
6	0.048	-0.041

Table 1: The result of correlation analysis between F0 and MoG model parameters of a mal source speaker

No.	Mean	Std. Deviation
1	**0.935**	**0.836**
2	0.092	0.153
3	-0.022	0.309
4	-0.051	0.047
5	-0.052	0.076
6	0.009	0.094

Table 2: The result of correlation analysis between F_0 and MoG model parameters of a female target speaker

3.3. Rebuild the glottal formant

According to the correlation between F_0 and MoG parameters, we construct a LR (linear regression) model using 2922 frames of a female target speaker to predict the mean and standard deviation of the first Gaussian component with the converted F_0 as independent factor. The linear equation is shown as follows:

$$M = 0.5448P + 130.548 \quad (2)$$

$$V = 0.0843P + 83.6891 \quad (3)$$

Where M stands for the mean and V denotes the standard deviation of the first Gaussian component. P is the converted F_0 of target speaker. The energy of the first Gaussian component of target speaker can be calculated by the ratio of the energy of the first Gaussian component to a single unit.

3.4. Predict target speaker's F0 by SFC model

The prediction of target speaker's F0 based on source speaker's F0 is realized by linear mapping. Training set is composed of four groups of 50 sentences respectively from four speakers. Result is showed in Table 3. We can see that the conversion from female to male is much better than from male to female and all the conversions are good enough for voice conversion system.

Source	Target			
RMSE/ Corr	Male1	Male2	Female1	Female2
Male1	—	22.55/0.81	19.74/0.84	24.88/0.82
Male2	18.48 /0.86	—	21.07/0.87	22.46/0.89
Female1	16.50 /0.88	15.86/0.87	—	21.85/0.89
Female2	18.52 /0.84	15.90/0.86	20.14/0.86	—

Table 3: The prediction error of target speakers measured by RMSE (Hz) and correlation.

4. EVALUATION

We collect the same 110 synthesized sentences of a male and a female speaker by a corpus-based TTS system. Then voice conversion is conducted between the two speakers from both female to male and male to female. The first 100 sentences are used for training, and the last 10 sentences are used for evaluation. Five listeners with more than two years' experience in perceptions are asked to give the result of listening test.

For comparing between the proposed method and the non glottal-formant-separated method, we also list the performance of our former system in bracket.

4.1.Evaluation experiment on speaker individuality

The converted speech is compared with the corresponding sentences of the source speaker and the target speaker to get evaluation based on discrimination. We use 5 grades: 5 means very close to the target speaker while 1 is very close to the source speaker. The result is shown in Table 4.

Conversion Type	F to M	M to F
Average Grade	4.6 (4.6)	4.5 (4.4)
MOS	3.7 (3.6)	3.6 (3.4)

Table 4: Subjective evaluation results

4.2.Evaluation experiment on speech quality

In order to evaluate the quality of converted speech, the opinion test is performed, where 5 means excellent and 1 is bad. The result is shown in Table 4. As the MOS (Mean Opinion Score) of original speech synthesized by the TTS system is about 4.0, that there isn't much decrease in the procedure of voice conversion.

5. CONCLUSION

The glottal effect on STRAIGHT spectrum is taken into account in our voice conversion system based on STRAIGHT and codebook mapping. At first the spectral component related with glottal formant is extracted from STRAIGHT spectrum and then rebuilt from F_0, which is predicted by SFC model during spectral conversion. The excitation-independent component is converted in traditional codebook mapping way. Listening test proves that converted speech using proposed method has better discrimination and speech quality. However, current measurement of glottal effect on STRAIGHT spectrum is not accurate enough. To improve its performance is the goal of our further research.

6. ACKNOWLEDGEMENT

This research is supported by the National Natural Science Foundation of China (60475015).

7. REFERENCES

[1] Arslan L.M., "Speaker transformation algorithm using segmental codebooks (STASC)", *Speech Communication 28*, pp. 211-226, 1999.

[2] Toda T., Lu J., Saruwatari H., and Shikano K., "STRAIGHT-based voice conversion algorithm based on Gaussian mixture model", *Proc. ICSLP*, Beijing, China, pp. 279-282, Oct. 2000.

[3] Kawahara H., "Restructuring speech representations using a pitch-adaptive time frequency smoothing and a instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sound", *Speech Communication* 27, pp. 187-207, 1999

[4] Shuang Z.W., Wang Z.X., Hu Y., and Wang R.H., "A novel voice conversion system based on codebook mapping with phoneme-tied weighting", *INTERSPEECH-2004,* pp. 1197-1200

[5] Ling Z.H., Wang Y.P., Hu Y., and Wang R.H., "Modeling glottal effect on the spectral envelop of STRAIGHT using mixture of Gaussians", *Proc. ISCSLP*, pp. 73-76, 2004.

[6] Chen G.P., Bailly G., Liu Q.F., and Wang R.H., "A superposed prosodic model for Chinese text-to-speech synthesis"*,Proc. ISCSLP*, pp. 177-180, 2004.