ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2004)
Hong Kong
December 15-18, 2004

MODELING GLOTTAL EFFECT ON THE SPECTRAL ENVELOP OF STRAIGHT USING MIXTURE OF GAUSSIANS

Zhen-Hua Ling, Yu-Ping Wang, Yu Hu, Ren-Hua Wang

iFlytek Speech Laboratory
University of Science and Technology of China, Hefei
{zhling, ypwang2}@ustc.edu, yuhu@iflytek.com, rhw@ustc.edu.cn

ABSTRACT

This paper presents a method to model the influence of glottal excitation on STRAIGHT spectrum by fitting the spectral envelop with mixture of Gaussians (MOG). The first Gaussian component is used as estimation to glottal formant in STRAIGHT spectrum because analysis result shows that it has an obviously stronger correlation with fundamental frequency than other spectral components and has similar characteristics with glottal formant. Then linear regression is carried out to measure the relationship between F_0 and the parameters of the first Gaussian component. This model is applied to STRAIGHT synthesis process and proved to be effective in compensating the voice quality variation caused by pitch modification.

1. INTRODUCTION

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum) [1], as a high-quality VOCODER-type analysis-synthesis method, has been presented in recent years. This analysis process is based on fundamental frequency extraction using TEMPO (Time-Domain Excitation extractor using Minimum Perturbation Operator)[2] and pitch adaptive spectral smoothing in both time and frequency domains. The synthesis process is implemented by passing a serial of impulse excitations with pitch period intervals through a time varying filter which is calculated from the smoothed spectral envelop. By manipulating the pulse positions in excitation, flexible prosody modification is realized [1].

In the framework of STRAIGHT analysis and synthesis, the excitation consists of only pitch information. Therefore the smoothed spectral envelop is the integration of both spectral presentation of glottal waveform and vocal tract transfer function according to general speech production hypothesis. Here a method for measuring the effect of glottal excitation on STRAIGHT spectral envelop is presented for the following two purposes, while the work introduced in this paper focuses on the first one:

1. It has been proved that some spectral characteristics of glottal waveform are dependent on not only the parameters that define phonation type or voice quality, such as OQ, SQ, but also the fundamental frequency of glottal source [3]. By modeling the effect of F_0 on STRAIGHT spectrum, the pitch modification during STRAIGHT synthesis can be improved.
2. By decomposing STRAIGHT spectrum into source-dependent components and source-independent components, we can provide an alternative for voice quality modification under STRAIGHT framework.

The mixture of Gaussians (MOG) model [4][5] is a speech spectral modeling method. Compared with linear predictive or cepstral coefficients, its parameters have more obvious physics meanings in fitting spectral peaks and are more independent from each other. So it is introduced here to estimate the glottal formant in STRAIGHT spectrum.

In the following part of this paper, an introduction to the method is presented in section 2. Section 3 gives experiment results and related analysis. Section 4 and 5 are discussions and conclusions.

2. METHOD

2.1. The spectral representation of glottal waveform

The spectral characteristics of glottal waveform are studied based on LF model [6], which describes the shape of the differentiated glottal airflow using the following five parameters: T_0, EE, RA, RG, RK. The open quotient of glottal source is related to both RG and RK: OQ = (1+RK)/(2RG). As mentioned in [3], the spectrum of LF model has two main characteristics:

1. Glottal formant (F_g). Assuming an abrupt closure of glottal waveform (when RA=0), the spectrum of differentiated glottal waveform has a (+6, -6)dB/oct asymptotic behavior. So there exists a maximum at

frequency F$_{g,}$, called glottal formant. The position of the glottal formant can be calculated as:

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

where *f*(RK*)* means a function of RK. From Eq.1 we can see that the position of glottal formant is dependent on T$_0$, OQ and RK. When phonation type is fixed, F$_g$ varies linearly with F$_0$.

2. Spectral tilt. When the vocal fold is closed smoothly with positive RA, an additional -6 or -12dB/oct spectral tilt will be added to the spectrum of glottal flow above a specific cut-off frequency.

For STRAIGHT, the effects of both glottal formant and spectral tilt are embodied in the 2-dimentional smoothed spectral envelop.

2.2. Effect of glottal formant in STRAIGHT spectrum

Glottal formant is the emphasis of our research work because of its linear relationship with fundamental frequency as described in Eq.1. In the general STRAIGHT synthesis process, the spectral envelops at pulse positions are extracted without adjustment to construct filter and synthesize the output speech no matter whether pitch information is modified. That means the glottal formant would not vary with the modification of F$_0$. However, it can be concluded from Eq.1 that a fixed F$_g$ with modified F$_0$ would cause the variation of some important source parameters, such as OQ, RK, which determine the phonation type of synthesized speech. Therefore, it is important to find a method to measure glottal formant in STRAIGHT spectrum and adjust relative spectral parameters when F$_0$ is modified.

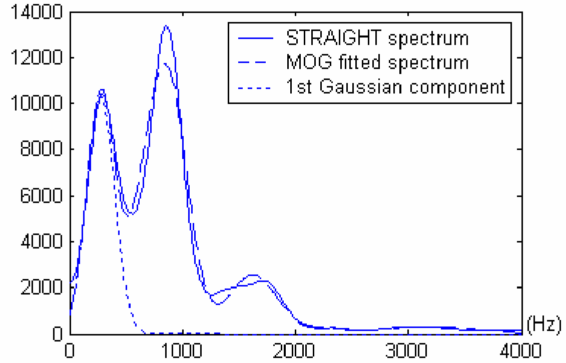

Figure 1: An example of glottal formant in STRAIGHT spectrum and the fitting result of MOG model. The frame is extracted from a vowel /a/ in Mandarin Chinese

Because F$_g$ is always located close to the first harmonic [3], its presentation in the smoothed spectrum of STRAIGHT analysis would be a spectral peak at lower frequency than the first formant. 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. When analyzed vowel has low first formant, such as /i/, /u/, or high F$_0$, the peak presenting glottal formant in STRAIGHT spectrum will be merged with the first formant. So it is necessary to find a reliable method to extract the spectral component corresponding to glottal formant.

2.3. Model the glottal effect using MOG

Mixture of Gaussians (MOG) [4][5] is introduced here to model the glottal formant in STRAIGHT spectrum. MOG is a spectral modeling method and fits the histograms representation of speech spectrum using the following equation:

$$S(\omega) = \sum_{i=1}^{k} w_i \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(\omega - u_i)^2}{2\sigma_i^2}\right] \qquad (2)$$

where $S(\omega)$ is the modeling result for normalized spectral envelop, *k* is the number of Gaussian mixtures. w_i, u_i, σ_i^2 present the weight, mean and variance of each Gaussian component. By considering the spectral envelop as a probability density function, these model parameters can be solved using a form of the Expectation Maximization (EM) algorithm [4].

Figure 1 shows the result of modeling one spectral envelop with MOG, where the dashed line presents the modeled spectrum and the dotted line presents the first Gaussian component, which is used to fit glottal formant. In the next section, the result of our experiment proves that the first Gaussian component is able to capture the characteristics of glottal formant no matter whether the spectral peaks merge or not.

3. EXPERIMENTS AND ANALYSIS

3.1. The correlation analysis between F$_0$ and MOG parameters

In order to prove the representability of the first Gaussian component of MOG model in presenting glottal formant, a correlation analysis is conducted based on a Mandarin Chinese speech corpus pronounced by a female speaker. This corpus is recorded for speech synthesis and contains more than 16,000 sentences with controlled modal voice quality. 30 sentences consisting of 974 syllables are selected from the corpus covering all vowels in Chinese. For each syllable, the pitch contour and spectral envelop are analyzed by STRAIGHT and 3 continuous frames of spectrum with 10 ms interval are extracted from the middle part of each vowel. Then for each frame of spectral envelop, the MOG parameters are calculated.

Here, the speech waveforms are resampled to 8kHz and the number of mixtures is set to 8. Table 1 is the result of correlation analysis between F_0 and the mean and standard deviation of each Gaussian component for all 2922 frames. The result of correlation analysis between F_0 and the first Gaussian component parameters for 5 simple vowels is shown in Table 2.

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
7	-0.045	0.019
8	0.224	0.012

Table 1: The result of correlation analysis between F_0 and MOG model parameters

Vowel	Frame Num	1st Mean	1st Std. Dev
/a/	93	0.975	0.942
/e/	147	0.985	0.970
/o/	30	0.979	0.974
/i/	307	0.860	0.485
/u/	239	0.898	0.810

Table 2: The result of correlation analysis between F_0 and the parameters of the first Gaussian component for 5 simple vowels

The following conclusions can be drawn from Table 1 and 2:

1. There explicitly exists spectral component that is strong correlated with F_0 and should be adjusted during pitch modification
2. The relationship between pitch information and STRAIGHT spectrum is mainly presented by the parameters describing the first Gaussian component in MOG model. Because the first Gaussian component is always used to fit the first spectral peak, it can be treated as a measurement of glottal formant on speech spectrum, which also varies linearly with F_0 according to Eq.1.
3. For the vowels with high first formant, such as /a/, /e/, the correlation between F_0 and parameters of the first Gaussian component is quite high (>0.9). That means the glottal formant can be measured accurately from STRAIGHT spectrum for the convenience of fitting it with single Gaussian function. While for the vowels with low first formant, such as /i/, /u/, the correlation is not so high but still obvious enough, especially for the mean. This proves the ability of MOG in estimating glottal formant position even peak merging happens.

3.2. Linear regression model for predicting the first Gaussian component parameters

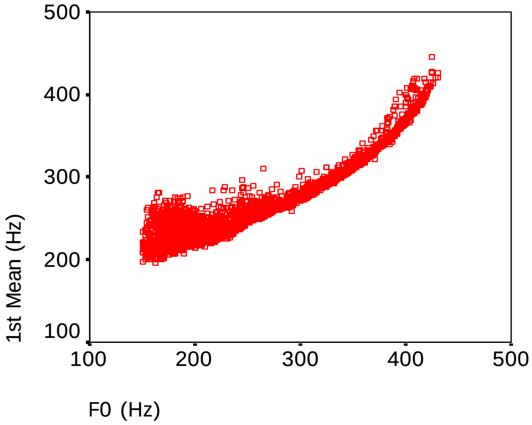

Figure 2: The plot of F_0 and the mean of the first Gaussian component for all 2922 frames

Figures 2 plot the relationship between F_0 and the mean of the first Gaussian component for 2922 frames. According to the linear relationship between F_0 and F_g as Eq.1, a linear regression (LR) model is constructed here to predict the mean and standard deviation of the first Gaussian component with F_0 as independent factor. The model summary is listed in Table 3. This model can realize spectral adjustment during F_0 modification and can also be regarded as description to a given voice quality as Eq.1.

Dependent factor	R	R Square	Std. Err of the estimation (Hz)
1st Mean	0.935	0.873	16.93
1st Std. Dev	0.836	0.699	11.07

Table 3: The summary of linear regression model for predicting the mean and standard deviation of the first Gaussian component based on F_0

3.3. The application of LR model on pitch modification of STRAIGHT

In order to test the effects LR model on spectral adjustment during F_0 modification, an experiment is conducted upon a waveform segment of vowel /o/ selected from the speech database mentioned in section 3.1. This segment is sampled at 8kHz and the pitch is modified to half of the original value using STRAGHT. The iterative adaptive inverse filtering (IAIF) [7] method is used here to estimate the glottal source from synthesized and modified speech. The estimated glottal waveforms and their spectrums are presented in Figure 3. Because STRAIGHT discards original phase information during analysis-synthesis process, the shape of estimated glottal waveforms in Figure 3 is not like the general glottal pulse described by LF model any more. However, it still can be observed from Figure 3 that for the speech synthesized by STRAIGHT with half F_0 reduction and no spectral adjustment, labeled with (B), there is longer closed quotient in estimated glottal waveform than that of speech without F_0 modification (A) and the open quotient reduces visibly, that would introduce a change in voice quality. Besides, in the spectral representation of glottal waveform, the difference between the amplitude of the first and second harmonic is always considered as a measurement for phonation type [8]. For (B), the first harmonic decreases greatly which lowers H1-H2 compared with (A). Both the reduction of OQ and H1-H2 indicate that the voice quality of reconstructed speech would be transformed from modal voice to creaky one, which has been proved by perceptual test. On the other hand, if the STRAIGHT spectral envelops are modeled by MOG and the mean and variation of the first Gaussian component, which is considered as a measurement of glottal formant, are adjusted according to the LR model constructed in section 3.2 when F_0 is modified (C), both the time domain estimated glottal waveform and its spectrum can capture more characteristics of original one in pulse shape and H1-H2. Listening test proves that (C) has closer voice quality to (A) and sounds more natural.

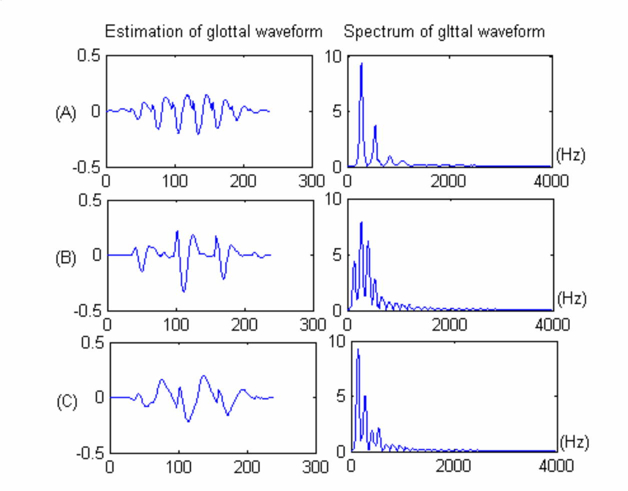

Figure 3: The comparison of the estimated glottal waveform using IAIF and its spectrum among speech synthesized by STRAIGHT without F_0 modification (A), with half reduction of original F_0 (B) and with half reduction of original F_0 combining spectral adjustment (C)

4. DISCUSSION

In section 3.3, the first Gaussian component of MOG model is extracted to measure the glottal formant caused by excitation source. The position of glottal formant is adjusted according to pitch modification during synthesis. Furthermore, the glottal formant is influenced by not only F_0 but also phonation type. Here the phonation type of analyzed speech is partly presented by specific LR model mentioned in section 3.2. If the LR model is adjusted, the glottal formant can be modified to synthesize different voice quality. Besides, the spectral tilt of glottal spectrum can also be modified by adjusting the weight of each Gaussian component in MOG. To realize the modification of voice quality under STRAIGHT framework is the goal of our further research.

5. CONCLUSION

MOG model is introduced into modeling STRAIGHT spectral envelop to measure the effect of glottal source on speech spectrum. Experiments show that the mean and standard deviation of the first Gaussian component have high correlation with F_0 and can be used to estimate glottal formant. After a LR model is constructed to predict these MOG parameters based on F_0, a pitch modification experiment is conducted. The result demonstrated that by adjusting spectral parameters properly, the STRAIGHT synthesizer can obtain better performance in maintain the phonation type of reconstructed speech during pitch modification.

6. REFERENCES

[1] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, “Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in Sound”, Speech Communication, vol.27, pp. 187-207, 1999.

[2] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, “Fixed Point Analysis of Frequency to Instan-taneous Frequency Mapping for Acurrate Estimation of F0 and Periodicity”, Proc. Eurospeech, pp. 2781-2784, 1999.

[3] C. d'Alessandro, B. Doval, “Voice quality modification for emotional speech synthesis”, Proc. Eurospeech, pp. 1653-1656, 2003.

[4] M. Lee, M. J. T. Smith, “Spectral modeling of the singing voice using asymmetric generalized Gaussian functions”, Proc. Stockholm Music Acoustics Conference, pp. 483-486, 2003.

[5] P. Zolfaghari, S. Watanabe,A. Nakamura and S. Katagiri, "Bayesian Modelling of the Speech Spectrum Using Mixture of Gaussians," Proc. ICASSP, pp. 553--556, 2004.

[6] G. Fant, J. Liljencrants, and Q. Lin, “A four-parameter model of glottal flow”, STL-QPSR, Speech, Music and Hearing, Royal Institute of Technology, Stockholm, 4, 1-13, 1985.

[7] P. Alku. “Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering”, Speech Communication, vol.11, pp. 109-118, 1992.

[8] C. Gobl, and A. Ní Chasaide, “Acoustic characteristics of voice quality”, Speech Communication, vol. 11, pp. 481-490,1992.