Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis

João P. Cabral, Steve Renals, Korin Richmond and Junichi Yamagishi

The Centre for Speech Technology Research
University of Edinburgh,UK
jscabral@inf.ed.ac.uk, s.renals@ed.ac.uk, korin@cstr.ed.ac.uk, jyamagis@inf.ed.ac.uk

Abstract

This paper proposes the use of the Liljencrants-Fant model (LF-model) to represent the glottal source signal in HMM-based speech synthesis systems. These systems generally use a pulse train to model the periodicity of the excitation signal of voiced speech. However, this model produces a strong and uniform harmonic structure throughout the spectrum of the excitation which makes the synthetic speech sound buzzy. The use of a mixed band excitation and phase manipulation reduces this effect but it can result in degradation of the speech quality if the noise component is not weighted carefully. In turn, the LF-waveform has a decaying spectrum at higher frequencies, which is more similar to the real glottal source excitation signal.

We conducted a perceptual experiment to test the hypothesis that the LF-model can perform as well as or better than the pulse train in a HMM-based speech synthesizer. In the synthesis, we used the mean values of the LF-parameters, calculated by measurements of the recorded speech. The result of this study is important not only regarding the improvement in speech quality of these type of systems, but also because the LF-model can be used to model many characteristics of the glottal source, such as voice quality, which are important for voice transformation and generation of expressive speech.

Index Terms: LF-model, Statistical parametric speech synthesis, HMM-based speech synthesis

1. Introduction

Glottal source modeling has been commonly used in rule-based speech synthesizers, since they are fully parametric. For example, the formant synthesizer proposed by Klatt and Klatt uses the KLGLOTT88 [1] model, which permits the control of several glottal parameters such as the open quotient, breathiness and spectral tilt. Concatenative synthesizers model the glottal source by inverse filtering, but unit-selection synthesizers typically aim to avoid signal processing and simply concatenate the speech units in order to obtain better speech quality. Thus, this type of system does not permit flexibility to control any glottal parameters besides the fundamental frequency.

Methods for speech transformation usually use inverse filtering to separate the speech signal into vocal tract and excitation components. Typically, voice quality transformations are performed on the spectrum of the vocal tract but a model of the glottal source can also be used to simulate different aspects of voice quality by controlling the glottal parameters, such as in [2].

Emerging applications, such as dialogue systems or virtual characters, demand expressive speech which is difficult to obtain with unit-selection synthesis. To generate expressive speech unit-selection requires larger speech databases, e.g. a speech corpus recorded with different emotions. However, the recordings are costly and demanding to conduct.

Statistical speech synthesis generates high-quality speech and is fully parametric, e.g. [3] and [4]. The high degree of parametric flexibility can overcome the limitations of concatenation synthesizers to generate variable speech. Compared with formant speech synthesizers, one great advantage of HMM-based synthesizers is that the parameters are automatically obtained from training data. Typically, the features used are the spectrum and F_0, which is controlled with a binary pulse. However, a drawback of this approach is that the synthetic speech is characterized by a buzzy quality. This is explained by the strong harmonic structure of the pulse signal at higher frequencies when compared with the true glottal source signal. To reduce this effect, more recent versions of this approach, such as [5], use the high-quality STRAIGHT [6] method for analysis and synthesis, and a multi-band mixed excitation with phase manipulation of the periodic pulse component.

In the work described here, we employ a more parametric model of the excitation (described in Section 2) than the traditional pulse train used in the HMM-based speech synthesis. The goal of doing this is to improve the naturalness of the synthetic speech and to enhance the parametrization of the glottal source. The glottal parameters may be used to better model and transform effects related to voice quality and the speech characteristics of the speaker. For example, in [7], the authors showed that F_0 is strongly correlated with the glottal parameters. Thus, the control and modeling of these parameters could also improve the naturalness of the synthetic speech.

2. Glottal source model

We have used the LF-model [8] because, in general, it gives a good approximation of the differentiated glottal volume velocity (DGVV) which we intend to model. It has also been extensively studied and used in speech research so we could find and compare different techniques to extract the glottal features.

2.1. LF-model

The model we use is divided into three parts and is given by Equation 1. Figure 1 shows the LF-waveform and the glottal parameters. The first part of the model is described by an exponentially increasing sine wave that starts at the opening instant of the vocal folds, t_o, and ends at the instant of maximum negative amplitude, t_e. The second branch is given by a decaying exponential function that models the closure after the abrupt flow termination. The time constant t_a is the duration from t_e to the point where a tangent to the exponential at $t = t_e$ hits

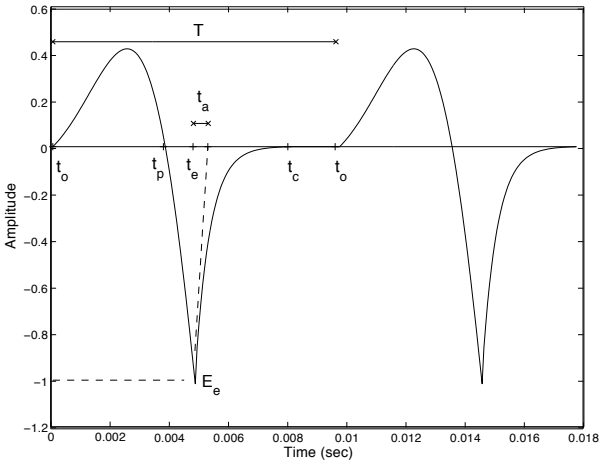


Figure 1: Segment of the LF-model waveform with the representation of the glottal parameters of Equation 1 during one period.

the time axis and measures the abruptness of the closure. The exponential part ends in the zero t_c. For simplification, it is usually assumed that $t_c - t_0 = T$, i.e., the fundamental period. Instead, we consider the glottal folds can be totally closed for a longer duration, from t_c until the end of the period. The other two parameters are the instant of maximum airflow t_p and the excitation amplitude E_e.

$$e(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t), & 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \leq t_c \\ 0, & t_c < t \leq T \end{cases} \quad (1)$$

where $w_g = \frac{\pi}{t_p}$.

The parameters ϵ and α can be calculated from Equation 1 by imposing $e(t_e) = E_e$ and the energy balance $\int_0^T e(t) = 0$.

The LF-model can also be described by other parameters which are related with properties of the glottal flow in the frequency domain. The most relevant parameters are the open quotient (OQ), speed quotient (SQ), and the return quotient (RQ), which can be calculated from the basic time domain parameters as follows [9]:

$$OQ = \frac{t_e + t_a}{T} \quad (2)$$

$$SQ = \frac{t_p}{t_e - t_p} \quad (3)$$

$$RQ = \frac{t_a}{T} \quad (4)$$

In the spectral domain, the LF-model can be stylized by three asymptotic lines with +6dB/oct, -6dB/oct and -12db/oct slopes [10]. Figure 2 shows this spectral representation. The crossing point of the first two lines corresponds to a peak (called glottal spectral peak) at the frequency F_g. The last line is due to the spectral tilt which contributes with an additional -6dB/oct above the frequency F_c. The frequency F_g can be calculated as in [11]:

$$F_g = \frac{1}{2\pi O_q T}\sqrt{\frac{e(\alpha_m)}{j(\alpha_m)}} \quad (5)$$

where $j(\alpha_m)$ and $e(\alpha_m)$ are functions of the asymmetry coefficient $\alpha_m = SQ/(1+SQ)$. F_c depends on several glottal parameters and it can be computed as described in [12]. However, it mostly depends on the glottal parameter t_a and it can be approximated by a simpler expression given in [8]:

$$F_c = \frac{1}{t_a 2\pi} \quad (6)$$

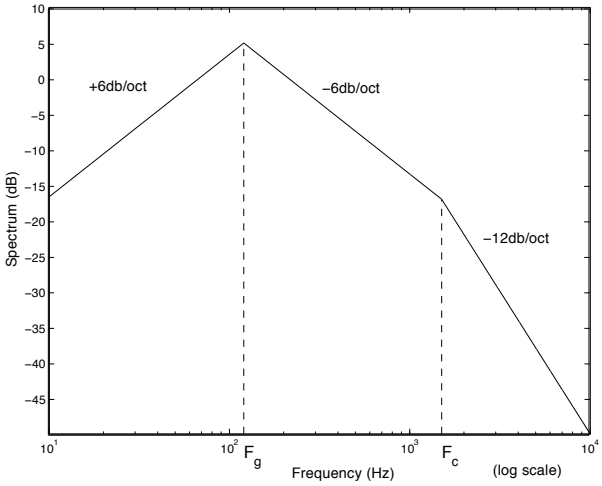


Figure 2: Linear stylization of the LF-Model spectrum.

2.2. Feature extraction

We measured the LF-parameters in ten utterances of the male speaker, which were selected from the speech corpus that was used to train the statistical models of the speech synthesizer. We calculated the mean values of the glottal parameters to generate the excitation signal in the speech synthesis.

Each utterance, sampled at 16 kHz, was analyzed pitch-synchronously using the epochs (instants of maximum excitation) calculated with the Entropic Signal Processing System (ESPS) tools. The algorithm to calculate the epochs is described in [13] and [14]. We obtained an estimate of the DGVV waveform by inverse filtering the speech signal. The resulting signal was high-pass filtered with a pre-emphasis filter ($\alpha = 0.97$) to eliminate the effect of the lip radiation. The LPC coefficients were calculated for each frame using a Hanning window, centered at the glottal epochs and with duration of 20 ms. Then, the residual was low-pass filtered at 4 kHz to reduce the high-frequency rumble effect on the energy envelope of the residual and permit a more accurate estimation of the glottal parameters.

The parameter t_e was estimated from the pitch-marks and E_e was the value of the waveform at that time instant. The other LF-parameters were estimated for each pitch cycle in the voiced regions.

The time instants t_c, t_o, and t_p can be extracted from the estimated glottal flow waveform. For example, t_p and t_o were calculated from the the electroglottographic (EGG) signal in [15]. We obtained an estimation of the glottal flow waveform by

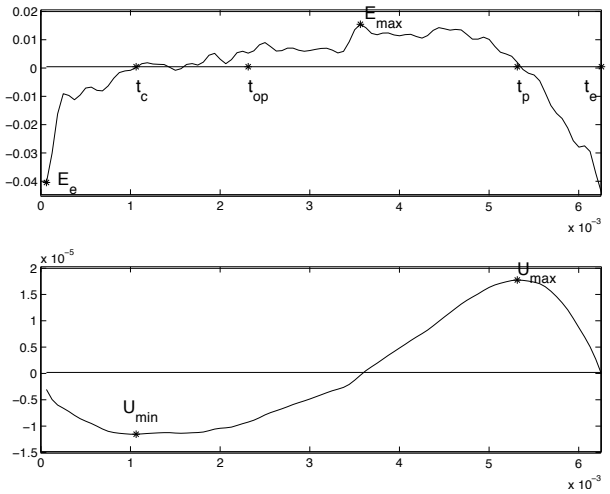


Figure 3: Estimation of t_c, t_o, and t_p. Top: a pitch cycle of the LPC residual; Bottom: integration of the residual signal (estimation of the glottal flow).

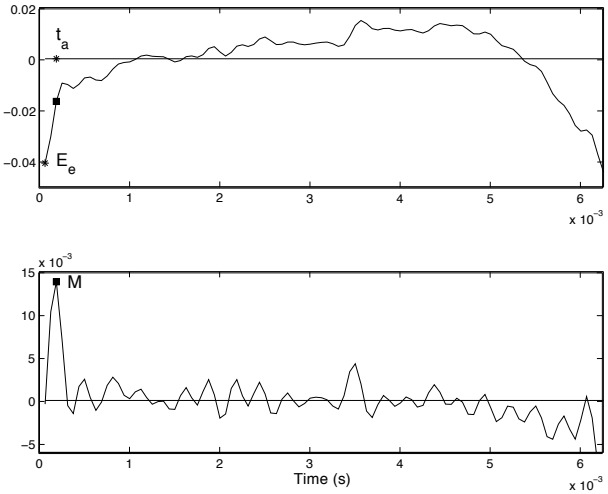


Figure 4: Estimation of t_a. Top: a pitch cycle of the LPC residual; Bottom: derivative of the residual signal.

taking the integration of the LPC residual signal. The residual was first high-pass filtered by a linear phase FIR filter with cut-off frequency of 80 Hz to reduce the low frequency amplitude fluctuation that results from the integration operation. Figure 3 helps to explain the method to estimate these parameters. In the figure, the point of maximal flow amplitude U_{max} gives the instant t_p and the point of minimum flow amplitude U_{min} is the estimation of t_c. From [16], the point t_o can be approximated by:

$$t_o = \frac{2(U_{max} - U_{min})}{\pi E_{max}} \tag{7}$$

where E_{max} is the maximal value of the residual in the period. There are methods that measure these parameters using amplitude thresholds or zero crossings, e.g. [17], but they do not necessarily give precise results because they are sensitive to rumble noise and it is difficult to set the appropriate thresholds.

The estimation of t_a is typically more difficult. It is usually obtained by fitting a model to the inverse filtered signal, which requires the use of an optimization algorithm and more complex calculations. We use a simple and effective method which consists of calculating the derivative of each pitch cycle of the residual and then detecting the peak of maximal amplitude in the return phase (starts at t_e and has duration equal to t_a). This peak is represented by M in Figure 4. Figure 1 shows that the tangent to the exponential decaying curve in the LF-model has the maximum slope at the instant t_e. Thus, we calculate $t_a = E_e/(MF_s)$, where F_s is the sampling rate, by assuming that the amplitude of the peak, M, is equal to the slope of the tangent at $t = t_e$.

Figure 5 a) shows the contours of the measured parameters for two voiced regions of an utterance spoken by a male speaker. In general, the parameters appear to increase linearly with T, with exception of t_a which is approximately constant. In [7], the measurements of the parameters for 3 vowels spoken with different pitch show similar relations, except for deviations at high values of T. In that study, each vowel followed its own raising path and the parameters were influenced by the preceding phone. Our results also show variation of the glottal parameters trajectories between different phonetic segments of the utterances.

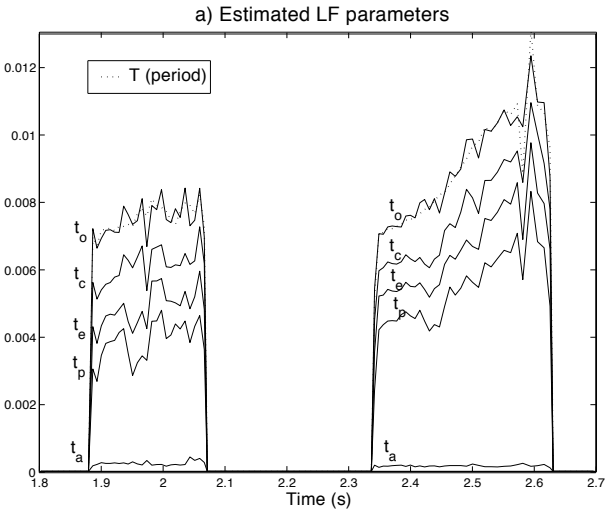
The values of the parameters, with the exception of t_a and E_e, are normalized by the pitch period. We use the median function to obtain smoother variations in the curves. Figure 5 b) shows the curves of the LF-parameters after the normalization and smoothing operations. Finally, the mean values of the glottal parameters are calculated.

3. System

3.1. General description

We integrated the LF-model into the speaker-dependent HMM-based speech synthesizer called Nitech-HTS 2005 [5]. This system uses the high-quality STRAIGHT method [6] to extract F_0, to compute the mel-cepstrum and to estimate spectral aperiodicity. The Nitech-HTS 2005 system uses a mixed multi-band excitation signal with phase manipulation, but it also has the option to use only the pulse train. The F_0 and aperiodicity parameters are used to generate the mixed-excitation signal. Speech is synthesized from the mixed-excitation and the mel-cepstral coefficients using an Mel Log Spectrum Approximation (MLSA) filter.

The system generates the excitation signals of voiced speech using a pulse (centered within a 512 sample length frame) which is processed to obtain phase randomization at the higher frequencies and summed with white Gaussian noise, in the spectral domain. The noise component is estimated on five frequency bands: 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. For unvoiced speech, the glottal source component is modeled only by white Gaussian noise. The resulting short-time signals are multiplied by asymmetric windows and added using the Pitch-Synchronously Overlap-and-Add (PSOLA) algorithm [18]. The weighting windows are centered in the pulse and are composed of two half-hanning windows: the first half of a hanning window lasts the duration of the previous period and the second (decaying half) lasts the duration of the current period. In case of the unvoiced frames the durations of the windows are set to a constant.



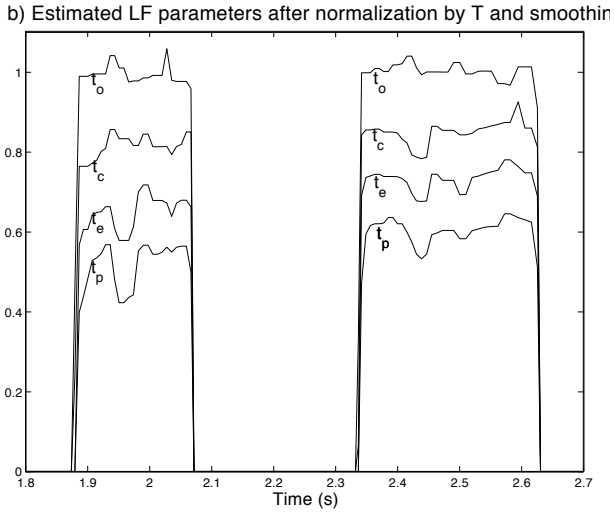


Figure 5: Curves of the estimated glottal parameters.

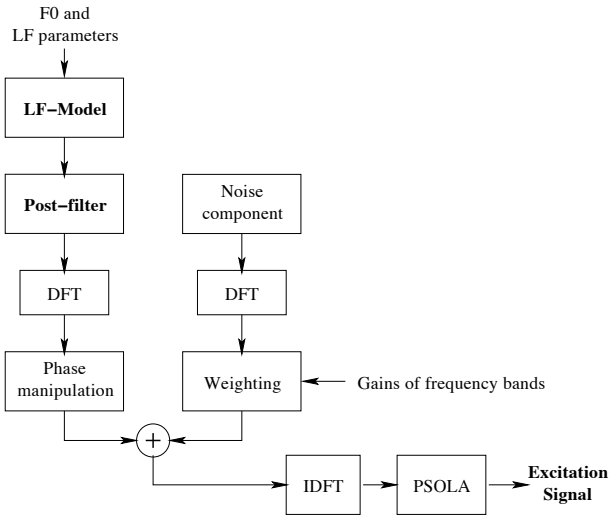


Figure 6: Block diagram of the excitation generation part using the LF-model.

3.2. Integration of the LF-model

We modified the synthesis part of the Nitech-HTS 2005 system to give the option to use the LF-model instead of the pulse train. By using the same approach based on STRAIGHT for the analysis and synthesis we can compare the effect on the speech quality of the two excitation source models. Figure 6 shows the schematic diagram of the excitation generation using the LF-model. When the system uses the new option, each voiced frame contains two pitch cycles of the LF-waveform, centered at the instant of maximum excitation, t_e. The LF-parameters, with the exception of t_a and E_e, are obtained by multiplying the normalized mean values of the glottal parameters by the synthesis period $T = 1/F_0$. In these calculations we assume that the variation of these parameters with T is approximately linear and constant. The parameters T_a and E_e are set to the constant mean value (not normalized) because they showed to have no correlation with T.

The STRAIGHT method estimates the spectrum envelope of the speech signal which is not a good estimation of the vocal tract because it only eliminates the F_0 effects of the glottal source from the speech. Thus, all the other aspects of the glottal source, such as the spectral tilt and the differences of amplitude of the first harmonics in the excitation spectrum, are described by the mel-cepstrum.

The pulse signal is used to model the periodicity of the excitation in the STRAIGHT speech synthesis method. The advantage of using this signal is that is spectrally flat. However, a drawback of using this model is that it has a strong harmonic structure at the higher frequencies when compared with the excitation of real speech, which has the effect of making the synthetic speech sound buzzy. Figure 7 shows the spectrum of a segment of the pulse train.

Glottal source models fit well in the source-filter theory which separates the speech signal in three independent processes: glottal excitation, vocal tract filter, and lip radiation. In this case, speech can be generated by feeding the glottal excitation through the vocal tract filter and performing a simple differentiation operation to model the lip radiation effect. However, source models are not appropriate for the synthesis with

STRAIGHT because this method uses a MLSA filter obtained from the mel-centrum instead of a vocal tract filter. We adapt the LF-model to the STRAIGHT synthesis method by using a post-filter that transforms the spectrum of the LF-signal into an approximately flat spectrum. This is equivalent to remove the spectral properties of the glottal spectral peak and the spectral tilt from the LF-waveform since they are described by the mel-spectrum. The post-filter is a linear phase FIR filter described by three linear segments which are symmetric to the slopes of the LF-model spectrum represented in Figure 2: -6dB/oct, +6dB/oct and +12dB/oct, respectively. We calculated the frequencies F_g and F_c from the equations 5 and 6, respectively, and using the mean values of the glottal parameters. Figure 8 shows the spectrum of a segment of the LF-model and the same segment after post-filtering.

If the HMM-based speech synthesizer was used to model the glottal parameters and generate them as it does with F_0, it would be necessary to use a time-varying post-filter which could be a limitation of approach.

The advantages of using the LF-model within this system are that it produces a less harmonic structure at the high-frequencies of the spectrum than the pulse train and permits flexibility to transform voice quality by modifying the glottal parameters.

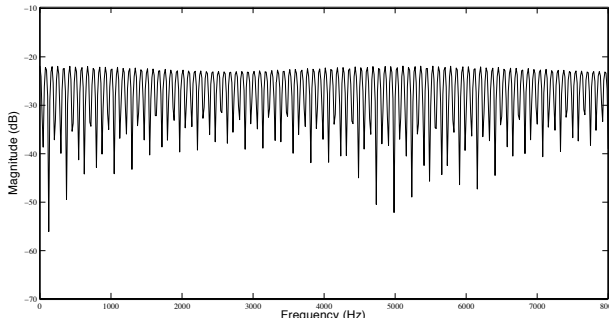


Figure 7: Spectrum of a segment of the pulse train (with duration 25 ms).

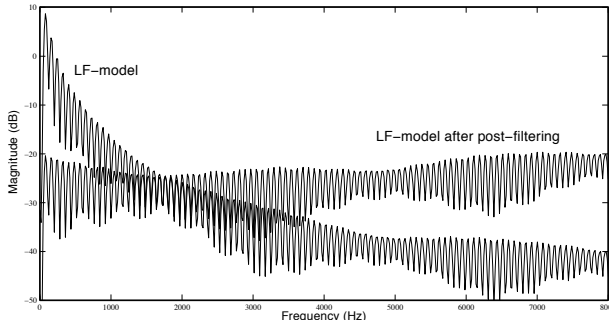


Figure 8: Spectrums of a segment (with duration 25 ms) of the LF-model signal and this signal after the post-filtering to obtain an approximately flat spectrum.

4. Perceptual evaluation

A forced-choice perceptual test was conducted to evaluate the performance of the LF-model when compared with the traditional pulse train used to model the excitation signal in the HMM-based synthesizer described in the previous section.

4.1. Stimuli

Speech was synthesized using the simple excitation (without multi-band noise or phase manipulation). The aperiodic components could also be used with the LF-model but they would have the same effect on the synthetic speech as when using the pulse train because the periodic and noise components are assumed to be independent. The US-English voice EM001 (male speaker) was built from the speech database released for the Blizzard Challenge 2007 (a total of approximately 8 hours of speech data). In the training part, the HMMs were modeled with the 39 order mel-cepstral coefficients obtained by the STRAIGHT analysis, the $\log F_0$ and the aperiodicity measurements.

The mean values of the LF-parameters were calculated from the measures of the parameters obtained for eight speech utterances of the speech database of the speaker EM001.

The stimuli consisted of ten different utterances. For each utterance two speech signals were synthesized, using the LF-model and the pulse model for the excitation. The duration of the speech signals varied from 2.6 to 7.2 sec.

4.2. Experiment

The instructions presented to the subjects were simply to listen the two synthetic speech samples for each utterance and select the one that sounded most natural. At the end, they had to indicate if they used headphones or speakers, and if they were native speakers of English (U.K./U.S.) or not.

4.3. Listeners

Students and staff of Edinburgh University were asked to perform the test which was presented via a web interface browser. Eighteen listeners participated in the test, from which seven were native speakers of English.

4.4. Results

The results of the perceptual experiment are presented in Table 1. In general, subjects preferred the speech generated with the LF-model than with the pulse train. This result was expected because the source model presents a less harmonic structure at the higher frequencies when compared to the pulse signal, which reduces the buzzy effect of the synthetic speech. Although there is a clear improvement with the LF-model, the rate of 64% indicates that this model does not overcome completely the limitations of the pulse train. Thus, additional properties of the glottal source need to be modeled, such as the noise, to obtain more natural speech.

5. Conclusions and future work

The LF-model of the glottal source was implemented in a HMM-based speech synthesis system which originally used the pulse signal to model the excitation. The glottal source model increases the parametric flexibility of the system and permits to transform voice characteristics of the speech by modifying the glottal parameters.

A perceptual experiment was conducted to evaluate the per-

	Excitation	
	LF-Model	Pulse Train
Non-native speakers	61%	39%
Native speakers	68,6%	31,4%
Total scores and 95% CI	64% $\pm$ 6.7%	36% $\pm$ 6.7%

Table 1: Scores, in percentage, obtained by each excitation model in the evaluation of the naturalness of the synthetic speech.

formance of the LF-model when compared with the pulse train in the quality of the synthetic speech. The results indicate that the speech synthesized with the LF-model sounds more natural. Although the difference in speech quality in comparison with the pulse model is not large, the LF-model can be used with the multi-band mixed excitation to obtain further improvements.

In this work, the statistical parametric synthesizer used the STRAIGHT for the analysis and synthesis. This method uses the pulse train to model the periodicity of the voiced excitation. The LF-model is not compatible with STRAIGHT for the excitation generation because it models more characteristics of the source besides the period and presents a decaying spectrum in contrast to the spectrally flat spectrum of the pulse. Thus, a post-filter was used to adapt the glottal source model to the STRAIGHT spectrum. The LF-model was used with the mean values of the glottal parameters, which were estimated from recorded utterances of the speech database.

In the near future, we will implement the statistical parametric approach with the glottal source model and a good method to estimate the vocal tract filter. Another interesting topic for future work is to model the glottal parameters with the HMMs.

6. References

[1] Klatt, D.H. and Klatt, L.C., "Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers", J. Acoust. Soc. Amer., Vol. 87(2):820–857, 1990.

[2] Childers, D. G., "Glottal Source Modelling for Voice Conversion", Speech Communication, 7(6):697–708, 1995.

[3] Tokuda, K., Zen, H. and Black, A.W., "An HMM-based Speech Synthesis System Applied to English.", Proc. of the 2002 IEEE SSW, pp.227230, USA, 2002.

[4] Black, A.W., Zen, H. and Toda, T., "Statistical Parametric Speech Synthesis", Proc. of the IEEE ICASSP, pp.1229–1232, Hawaii, 2007.

[5] Zen, H., Toda, T., Nakamura, M. and Tokuda, K., "Details of Nitech HMM-based Speech Synthesis System for the Blizzard Challenge 2005", IEICE Trans. Inf. and Syst., Vol.E90-D, No.1, pp. 325–333, 2007.

[6] Kawahara, H., Masuda-Katsuse, I. and Cheveigné, A., "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f_0 extraction: Possible role of a repetitive structure in sounds", Speech Communication, Vol. 27, pp. 187–207, 1999.

[7] Tooher, M. and McKenna, J.G., "Variation of the glottal LF parameters across F0, vowels, and phonetic environment", Proc. of the ITRW (VOQUAL'03), Geneva, Switzerland, August 2003.

[8] Fant, G., "The voice source in connected speech", Speech Communication, Vol. 22, pp. 125–139, 1997.

[9] Fant, G. and Lin, Q., "Frequency domain interpretation and derivation of glottal flow parameters", STL-QPSR, 29(2-3), pp. 1–21, 1988.

[10] Doval, B. and d'Alessandro, C., "The spectrum of glottal flow models." Notes et document LIMSI, num. 99–07, 1999.

[11] d'Alessandro, C. and Doval, B., "Voice quality modification for emotional speech synthesis", Proc. of the Eurospeech 2003, pp. 1653-1656, Geneva, Switzerland, 2003.

[12] Doval, B. and d'Alessandro, C., "Spectral Correlates of Glottal Waveform Models: An Analytical Study", Proc. of the Int. Conf. on Acoust., Speech and Signal Proc., Germany, Vol. 2, pp. 1295–1298, 1997.

[13] Talkin, D., "Voicing epoch determination with dynamic programming", J. Acoust. Soc. Amer., 85, Supplement 1, 1989.

[14] Talkin, D. and Rowley, J., "Pitch-Synchronous analysis and synthesis for TTS systems", Proc. of the ESCA Workshop on Speech Synthesis, C. Benoit, Ed., Imprimerie des Ecureuils, Gieres, France, 1990.

[15] Krishnamurthy, A.K. and Childers, D.G., "Two-channel speech analysis", IEEE Trans. Signal Process., Vol. 34, no. 4, pp. 730-743, 1986.

[16] Gobl, C. and Ní Chasaide, A., "Amplitude-based source parameters for measuring voice quality", Proc. of the ITRW (VOQUAL'03), pp. 151–156, Geneva, Switzerland, August 2003.

[17] Arroabarren, I. and Carlosena, A., "Glottal source parameterization: a comparative study", Proc. of the ITRW (VOQUAL'03), pp. 29–34, Geneva, Switzerland, August 2003.

[18] Moulines, E. and Charpentier, F., "Pitch-synchronous waveform processing techniques for text to speech synthesis using diphones", Speech Communications, Vol. 9, pp. 453–476, December 1990.