An HMM-based speech synthesiser using Glottal Post-Filtering

João P. Cabral[1,2], *Steve Renals*[2], *Korin Richmond*[2], *Junichi Yamagishi*[2]

[1]School of Computer Science and Informatics, University College Dublin, Ireland
[2]The Centre for Speech Technology Research, University of Edinburgh,UK

joao.cabral@ucd.ie, s.renals@ed.ac.uk, jyamagis@inf.ed.ac.uk, korin@cstr.ed.ac.uk

Abstract

Control over voice quality, e.g. breathy and tense voice, is important for speech synthesis applications. For example, transformations can be used to modify aspects of the voice related to speaker's identity and to improve expressiveness. However, it is hard to modify voice characteristics of the synthetic speech, without degrading speech quality. State-of-the-art statistical speech synthesisers, in particular, do not typically allow control over parameters of the glottal source, which are strongly correlated with voice quality. Consequently, the control of voice characteristics in these systems is limited. In contrast, the HMM-based speech synthesiser proposed in this paper uses an acoustic glottal source model. The system passes the glottal signal through a whitening filter to obtain the excitation of voiced sounds. This technique, called glottal post-filtering, allows to transform voice characteristics of the synthetic speech by modifying the source model parameters.

We evaluated the proposed synthesiser in a perceptual experiment, in terms of speech naturalness, intelligibility, and similarity to the original speaker's voice. The results show that it performed as well as a HMM-based synthesiser, which generates the speech signal with a commonly used high-quality speech vocoder.

Index Terms: HMM-based speech synthesis, voice quality, glottal post-filter

1. Introduction

Concatenation-based speech synthesis provide very low parametric flexibility to transform voice quality, because speech is synthesised joining recorded units. In contrast, HMM-based speech synthesisers use a parametric model of speech. The typical model of these systems consists of passing a spectrally flat excitation through a synthesis filter which represents the spectral envelope. In this method, the excitation of unvoiced speech is typically modelled by white noise, while the voiced excitation is modelled by a periodic impulse train. The significant advantage of using this speech representation is that the spectral envelope can be efficiently calculated, e.g. by linear prediction or cepstral analysis. The counterpart is the poor representation of the glottal source, which limits voice quality modelling and might produce unnatural speech quality. Better excitation models than the impulse train have been proposed to improve speech naturalness in HMM-based speech synthesis, e.g. [1, 2]. Such models have more details of the source, such as aperiodicity aspects, but they do not represent the glottal pulse characteristics.

According with the theory of speech production, voiced speech can be obtained by passing a *glottal source model* through a synthesis filter, which represents the vocal tract system. This speech model is different from that of the impulse response that represents the spectral envelope. The main problem with the source-tract model is that the methods to estimate the glottal source and the vocal tract filter are typically less robust than those to estimate the spectral envelope. Nevertheless, this type of speech model has been successfully used in HMM-based synthesis. For example, the system in [3] models the glottal source and the vocal tract filter using LPC parameters, obtained by iterative adaptive inverse filtering [4]. In the synthesis part, the excitation is obtained by transforming a real glottal pulse to have the desired duration and spectral characteristics, using F_0 and the glottal parameters, respectively. In this system, voice transformations could be performed using a library of glottal pulse shapes for different voice qualities.

In previous work [5], we proposed to represent the excitation of the HMM-based speech synthesiser with a spectrally flat signal which is obtained by passing an acoustic glottal source model, the Liljencrants-Fant (LF) model [6], through a post-filter. We call this operation glottal post-filtering. Results of a perceptual test showed that speech synthesised with the post-filtered LF-model sounded more natural than using the impulse train.

In this paper, we propose another HMM-based speech synthesiser that uses a synthesis method with glottal post-filtering. The results showed that this system performs similarly to a different version of the synthesiser that uses the high-quality speech vocoder STRAIGHT [7]. The proposed system has the advantage that allows to modify parameters of the source model, to transform the voice quality of the synthetic speech. The technique to control the pitch of the output speech, in the glottal post-filtering technique, is also improved, in this work.

2. Liljencrants-Fant model

2.1. Waveform

The Liljencrants-Fant (LF) model [6] is an acoustic model of the glottal source derivative, which is shown in Figure 1. It can be represented by the following equation:

$$e_{LF}(t) = \tag{1}$$
$$\begin{cases} E_0 e^{\alpha t}\sin(w_g t), & t_o \leq t \leq t_e \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \leq t_c \\ 0, & t_c < t \leq T_0 \end{cases}$$

where $w_g = \pi/t_p$. The LF-model is defined by six shape parameters: t_c, t_p, t_e, T_a, T_0, and E_e. The remaining parameters (E_0, ϵ and α) can be calculated by using the energy and continuity constraints, which are given by $\int_0^{T_0} e_{LF}(t)dt = 0$ and $e_{LF}(t_e) = e_{LF}(t_e^+) = -E_e$, respectively. The first branch of equation (1) starts at the instant of glottal opening, $t_o = 0$,

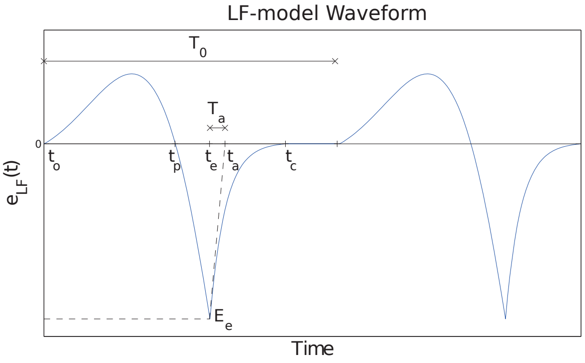

Figure 1: *Segment of the LF-model waveform.*

and ends at the instant of abrupt glottal closure, t_e. The amplitude of maximum excitation, E_e, occurs at this discontinuity point. The second part is called the return phase and it represents the transition between the abrupt closure and the closed phase (when the vocal folds are completely closed), which has zero value.

The LF-model is often represented by the first two branches of equation (1), for simplification. In this case, the instant of complete closure, t_c, is set equal to the period T_0 in the second branch. In this work, this simplified LF-model version, which is defined by five parameters, is used.

2.2. Voice quality parameters

The parameters of the LF-model can also be expressed as dimensionless quotients. The main dimensionless parameters are the open quotient (OQ), speed quotient (SQ), and return quotient (RQ). OQ measures the relative duration of the open phase (with duration equal to $t_e + T_a$), SQ is related to the asymmetry of the glottal pulse, and RQ measures the relative duration of the return phase. They are given by the following equations:

$$OQ = \frac{t_e + T_a}{T_0} \tag{2}$$

$$SQ = \frac{t_p}{t_e - t_p} \tag{3}$$

$$RQ = \frac{T_a}{T_0} \tag{4}$$

These dimensionless parameters are strongly correlated with voice quality, e.g. [8]. For example, breathy voice is typically characterised by high OQ, high RQ, and low SQ. A tense voice has the opposite pattern, that is, low OQ, low RQ, and high SQ.

2.3. Spectral representation

The spectrum of the LF-model is characterised by a spectral peak at low frequency, often called the “glottal formant”, and the spectral tilt. Figure 2 shows the stylised spectrum of the LF-model [9]. The LF-model transfer function has a low-pass characteristic. For example, the spectral tilt is equivalent to a first order low-pass filter which contributes with -6dB/oct attenuation for frequencies above the cut-off frequency F_c.

In [9], the authors derived formulae which relate shape parameters of the LF-model waveform with the spectral parameters. They showed that the frequency of the spectral peak, F_g, can be given by:

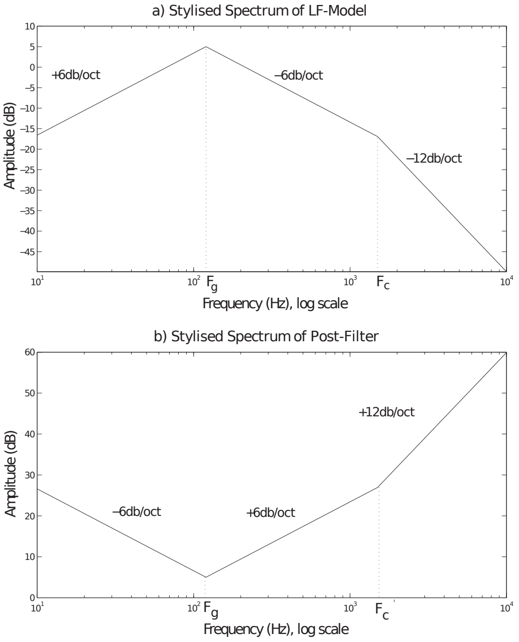

Figure 2: *Stylised spectrum of the LF-model (a) and its corresponding Post-filter spectrum (b).*

$$F_g = \frac{1}{2\pi}\sqrt{\frac{E_e}{I}}, \tag{5}$$

where E_e is the amplitude of maximum excitation of the LF-model and I is the integral of the glottal flow pulse.

The cut-off frequency F_c mainly depends on the return phase parameter T_a, e.g. [6], and it can be estimated by:

$$F_c = \frac{1}{2\pi T_a} \tag{6}$$

3. Speech Synthesis with Glottal Post-filtering

3.1. Overview

The speech production model used to synthesise speech with glottal post-filtering consists of shaping the spectrally flat excitation, $X(w)$, with the spectral envelope, $H(w)$. Glottal post-filtering is used to generate the excitation of voiced speech by passing the LF-model signal through the glottal-post filter (GPF), $F(w)$. This filter transforms the input LF-model signal into the spectrally flat excitation. Speech synthesised with this excitation model can be represented by:

$$Y(w) = E_{LF}(w)F(w)H(w), \tag{7}$$

where $E_{LF}(w)$ is the Fourier Transform (FT) of the LF-model.

3.2. Glottal Post-Filter Calculation

The stylised spectrum of the GPF is described by three linear segments, whose slopes are symmetric to the slopes of the LF-model spectrum. The stylised spectrum that corresponds to the filter transfer function is shown in Figure 2 b).

The parameters of the GPF, the frequencies F_c and F_g, are calculated from a set of LF-model parameters. The LF-model signal obtained from these parameters is called the *reference LF-model*. This set of parameters can be estimated from the recorded speech of the speaker, e.g. estimating the mean values of the LF-parameters for that speaker. However, it is not necessary for the LF-model to accurately represent the glottal source characteristics of the speaker, because it is transformed into a spectrally flat signal by the GPF. An important issue in the selection of the reference LF-model is the duration of the open phase. We suggest to choose an open phase duration ($T_o = t_e + T_a$) close to the minimum pitch period of the speaker. This avoids problems in synthesising speech with high F_0 values, as explained in the next section.

The parameter F_g of the GPF can be calculated using (5). For this, the integral I of the LF-model has to be calculated. First, the LF-model waveform is obtained from (1), by setting $T_a = 0$. The return phase is not considered, because the authors in [9] assume that F_g does not depend on T_a in (5). Next, the resulting LF-model waveform is integrated to obtain the glottal flow pulse, $u_{LF}(n)$, associated with the LF-model. I is calculated as the integral of $u_{LF}(n)$. Finally, the frequency F_g is calculated as $F_g = 1/(2\pi)\sqrt{E_e * F_s/I}$, where F_s is the sampling frequency.

The other parameter used to obtain the GPF is the frequency F_c, which is calculated from the LF-parameter T_a, using (6).

3.3. Pitch Control

The fundamental period, $T_0 = 1/F_0$, determines the duration of the LF-model and it is used to model the pitch of the synthetic speech. The glottal post-filtering technique proposed in our past work [5], time-scales the reference LF-model waveform to obtain a signal with the desired duration. This operation is used to control F_0 without modifying the voice quality parameters (RQ, SQ, OQ) of the reference LF-model waveform. However, time-scaling the LF-model signal changes its spectrum. The effect of a positive and negative scale factors is to compress and expand the spectrum, respectively. It is also possible to demonstrate that the spectrum of the LF-model changes with time-scaling using equations (5) and (6).

When the duration of the reference LF-model signal is adjusted, it is important to preserve its shape in order to obtain a spectrally flat excitation. The method to control the pitch in [5] was improved to avoid the problem related with time-scaling, as follows. When the desired F_0 is higher than the F_0 of the reference LF-model, its closed phase is truncated by the required number of samples. Conversely, for a lower F_0 than that of the reference LF-model, this signal is padded by the required number of zeros. This operation allows to control the pitch period without affecting the spectrum of the LF-model signal, unless the truncation region is longer than the closed phase of the LF-model. If the length of the closed phase is not long enough to perform the truncation, the open phase of the glottal signal can be truncated or decimated, but this alters the shape of the LF-model signal and its spectrum. This effect can be avoided by choosing a reference LF-model with a sufficiently short open phase. For example, an LF-model with open phase equal to the minimum T_0 characteristic of the speaker is a good solution.

3.4. Voice Quality Transformation

The characteristics of the glottal source signal can be modified using a set of LF-model parameters different from that of the reference LF-model to synthesise speech. For example, if the return phase parameter T_a is lower than that of the reference model, the spectral tilt decreases (lower attenuation at the higher frequencies). The variations in the spectrum of the LF-model signal produce similar changes in the spectrum of the synthetic speech, because the post-filter remains the same. Therefore, by modifying the input LF-model parameters is possible to modify the voice characteristics of the synthetic speech. For example, voice quality can be modified by controlling the glottal parameters correlated with voice quality: OQ, SQ, and RQ.

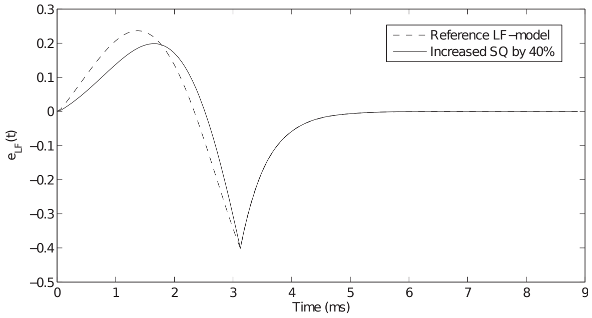

Figure 3: *Waveforms of the reference LF-model and the LF-model obtained by increasing the SQ of the reference LF-model by 40%.*

The limitation of voice quality transformation with glottal post-filtering is that speech cannot be synthesised directly from the glottal source signal. Instead, the method produces variations of the glottal characteristics relative to the speech which is synthesised with the reference LF-model. For example, if we take a reference LF-model with OQ=0.6, then by using a LF-model with lower OQ in synthesis, e.g. OQ=0.3, the resulting synthetic speech has the spectral effects of decreasing the OQ.

Figure 4 shows an example of the reference LF-model waveform and the signal obtained by increasing the SQ of the reference LF-model by 40%. Figure 4 a) shows the difference between the spectrum of these two signals. The effect of increasing the SQ of the reference LF-model signal is to decrease the spectral tilt (increase of energy at the higher frequencies) and to change the frequency and amplitude of the glottal formant. The excitation is affected by the same variation, because the glottal post-filter does not change. Figure 4 b) shows the spectrum of the two filtered signals. When the input of the filter is the reference LF-model signal, the excitation is spectrally flat. Instead, when the SQ of the reference LF-model is increased, the spectrum of the excitation is no longer flat. This variation in the spectrum of the excitation has the same effect on the spectrum of the synthetic speech. As result, by changing the SQ of the reference LF-model signal, the synthetic speech will exhibit different voice quality.

4. Application to HMM-based Speech Synthesis

4.1. Baseline System

In this work, the GPF is integrated into a HMM-based speech synthesiser based on the Nitech-HTS 2005 system [10]. This system uses the Matlab version of STRAIGHT for analysis and synthesis. In the analysis, STRAIGHT is used to extract the FFT parameters of the spectral envelope and aperiodicity parameters. F_0 is also estimated by using the Entropic Signal Processing System (ESPS) tools [11]. For synthesising speech,

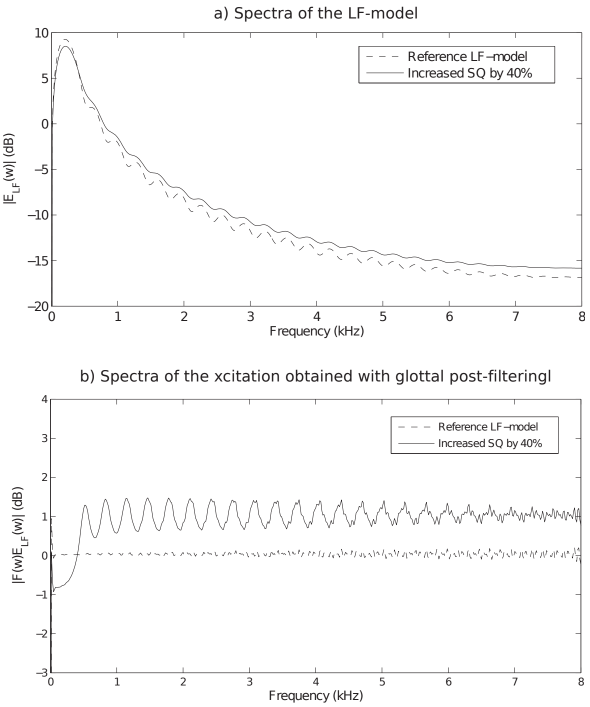

Figure 4: *a) Spectra of the reference LF-model and its modified version with higher SQ. b) Spectra of the two glottal post-filtered LF-model signals.*

STRAIGHT shapes the excitation with the spectral envelope using a minimum-phase filter. The excitation of unvoiced speech is represented by white noise, whereas the excitation of voiced speech is modelled by mixing noise with an impulse train. The mixing operation consists of weighting the two components in the frequency domain, by using the aperiodicity parameters, and adding them together. STRAIGHT also uses an all-pass filter function that modifies the phase of the impulse, to reduce the buzziness effect of this signal. The voiced excitation is obtained by mixing the processed impulse train, $P(w)$, with the noise $N(w)$. That is,

$$X(w) = P(w)W_p(w) + N(w)W_a(w), \tag{8}$$

where $W_p(w)$ and $W_a(w)$, are the weighting functions of the periodic and noise components, respectively.

For statistical modelling, the system uses a 5 state context-dependent HMM. Each feature vector contains the static and dynamic features (Δ and Δ^2) of the spectrum and excitation: 39^{th} order mel-cepstral coefficients (obtained from the FFT coefficients), five aperiodicity parameters (mean values of the aperiodicity measurement in five frequency bands), and $\log F_0$. The state output probability distribution used to model each speech parameter is a Gaussian function.

4.2. Implementation of Glottal Post-filtering

4.2.1. Calculation of the Glottal Post-Filter

The parameters of the reference LF-model were obtained from measurements of the LF-parameters, which were performed on some utterances of the speech corpus used to build the voice of the HMM-based speech synthesiser. The method to estimate the LF-model parameters is similar to the method we used in [12]. Basically, it consists of using a non-linear optimisation algorithm to fit pitch-synchronously the LF-model to the glottal source derivative. In this analysis method, the glottal source derivative signal was estimated from the speech signal using the iterative adaptive inverse filtering method [4].

The LF-model measurements were used to calculate the mean values of the dimensionless parameters: OQ, SQ, and RQ. An estimate of the maximum F_0 of the speaker was also calculated. The parameter t_e of the reference LF-model was set approximately equal to the minimum T_0 of the speaker. Then, the other time parameters of the reference LF-model were calculated by using the mean values of the dimensionless parameters and equations (2) to (4). In this way, the reference LF-model was short enough to avoid the problem of synthesising high-pitched speech and the dimensionless parameters were equal to the mean values obtained from the measurements.

The GPF was implemented as a linear phase FIR filter, to preserve the phase information of the LF-model.

4.2.2. Synthesis

The STRAIGHT synthesis method used by the HMM-based speech synthesiser was replaced by the synthesis method with glottal post-filtering. Speech is synthesised as shown in the block diagram of Figure 5. This method also uses a multi-band mixed excitation model, which is represented by

$$X(w) = K_e E_{LF}(w)F(w)W_p(w) + N(w)W_a(w), \tag{9}$$

where $E_{LF}(w)$ is the FT of a periodic LF-model signal, $F(w)$ represents the transfer function of the GPF, $N(w)$ is the FT of the noise signal, and K_e is a scale factor to match the energy of the periodic excitation to the energy of the noise. $W_p(w)$ and $W_a(w)$ represent the weighting functions of STRAIGHT, which are obtained from the aperiodicity parameters.

The periodic component of the excitation is the concatenation of two LF-model signals, which start at the instant of maximum excitation t_e. These signals are obtained by adjusting the length of the reference LF-model (by truncating/padding with zeros) to the target $T_0 = 1/F_0$. That is, for synthesising the speech frame i, the first LF-model has the duration T_0^{i-1} (equal to the period of the previous frame) and the second has the duration T_0^i. The resulting LF-model waveform is approximately centered at the instant of maximum excitation, t_e.

The speech signal, $Y(w)$, is synthesised in the frequency domain as $Y(w) = X(w)H(w)$, where $H(w)$ is the FFT spectrum obtained from the mel-cepstral coefficients generated by the synthesiser. As in STRAIGHT, the FFT coefficients are a representation of the spectral envelope. The speech frames are concatenated by using the overlap-and-add technique. The overlap windows are asymmetric, to obtain perfect overlap-and-add (they add to one), as in the pitch-synchronous time-scaling method [13]. Each overlap window is obtained by concatenating the first half of a Hanning window with the second half of a Hanning window, which may have different durations. The first part has duration T_0^{i-1}, whereas the second has duration T_0^i.

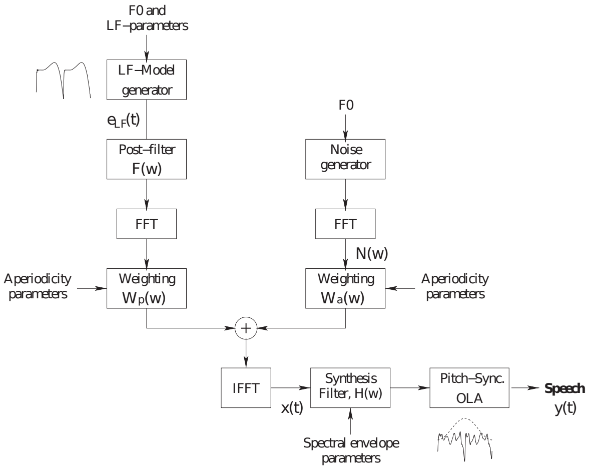

Figure 5: *Block diagram of the speech synthesis method using glottal post-filtering.*

5. Experiments

5.1. Perceptual Evaluation of Speech Quality

5.1.1. Speech Databases

Three synthetic voices were built for each system using the following speech databases:

- Voice A: UK English full voice from the male database released for the Blizzard Challenge 2009 ($\approx$ 8 h).
- Voice B: ARCTIC subset of the male database ($\approx$ 1 h).
- Voice C: UK English female speech database ($\approx$ 6 h).

5.1.2. Systems

Seven systems were evaluated in the perceptual experiment, including the HMM-based speech synthesiser with STRAIGHT vocoder and the system with glottal post-filtering, which were described in section 4. The natural speech of the original speaker was also included. The other systems are omitted in this paper, because they are not related with this work.

5.1.3. Listening Test Design

A perceptual evaluation was conducted to evaluate the HMM-based speech synthesisers, in terms of speech naturalness, speech intelligibility, and similarity of the synthetic voice to the original speaker's voice. Each participant conducted the evaluation in a supervised perceptual lab at the University of Edinburgh, by following the instructions given by a computer program and using headphones. The Blizzard listening test set-up [14] was used to perform this evaluation. However, some adjustments were performed to the original listening evaluation design. The test was similar for the three voices and it was divided into different sections Each section contained several parts and corresponded to one of the following listener tasks:

- Similarity (SIM) task: listeners heard an utterance and chose how similar the synthetic voice sounded to the voice of reference samples of the original speaker on a scale from 1 [Sounds like a totally different person] to 5 [Sounds like exactly the same person].
- ABX task: listeners heard one utterance from each of two systems (A and B) and chose one of the three possible responses: [A sounds more natural than B], [B sounds more natural than A], and [A and B sound equally natural].
- Mean Opinion Score (MOS) task: listeners heard one utterance and chose a score which represented how natural or unnatural the sentence sounded on a scale of 1 [Completely Unnatural] to 5 [Completely Natural].
- Intelligibility (WER) task: listeners heard one utterance that corresponded to a semantically unpredictable sentence and they had to type what they heard.

5.1.4. Listeners

Ninety six undergraduate students from the University of Edinburgh were recruited to participate in the evaluation. Subjects were equally distributed among the three evaluations associated with each voice (A, B and C). They were all native speakers of UK English, aged 18-25 and were paid for their participation.

5.2. Results

Figure 6 shows the results obtained in the perceptual evaluation of the full male and female voices, for the MOS, SIM and WER tasks. The results are analysed as in [14], i.e., in terms of the median for the first two tasks and the mean for the WER task. In the MOS and SIM plots, the median is represented by a solid bar across a box, showing the quartiles.

Natural speech was rated always significantly higher than the synthetic speech, in all parts of the evaluation. In general, the system with STRAIGHT synthesis and the system with glottal post-filtering obtained similar results, for all voices. Also, the comparisons between the two systems were not statistically significant ($p - value \geqslant 1$), for all tasks. From the results of the evaluation, they are equally natural, intelligible, and similar to the original speaker. Also, the results of the ABX task were consistent with the MOS, since the preference rates obtained by the two systems were similar and not statistically significant. This result indicates that the use of a flattened LF-model signal for the excitation does not affect significantly the speech quality of the synthesiser, when compared with the impulse signal.

In our previous work [5], the results showed that the system with the post-filtered LF-model produced more natural speech than the system with impulse train. We expected the same result in the experiment of this paper, although the systems are different between the two experiments. Possibly, the difference between the two types of excitation was perceptually less significant in this experiment, because they were mixed with noise (unlike in [5]). On the other hand, the similarity and intelligibility results were expected, because the main difference between the impulse train and the excitation obtained with GPF is that the first has stronger harmonics. That is, they are both spectrally flat signals, which do not contain the source characteristics associated with the identity of the speaker's voice (they are incorporated into the spectral envelope). Note that glottal post-filtering does not produce an excitation which represents the glottal source characteristics, but it allows to transform them.

5.3. Voice Quality Transformations

We also conducted experiments to investigate the effect of modifying the parameters of the reference LF-model on the voice quality of the synthetic speech. A small set of sentences were synthesised with the HMM-based speech synthe-

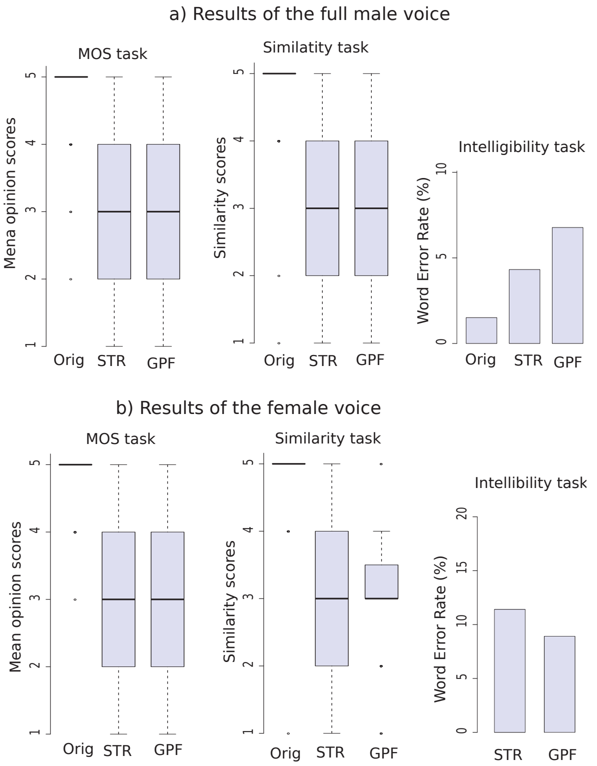

Figure 6: *Results of the perceptual evaluation obtained by the original speech (Orig), the system with STRAIGHT synthesis (STR) and the system with glottal-post-filtering (GPF).*

siser that uses the GPF, for different shapes of the input LF-model. Speech synthesised with the reference LF-model, was considered to have neutral voice quality. This voice quality was transformed by varying one of the dimensionless parameters: OQ, SQ and RQ. Each parameter was decreased and increased by different degrees. For example, the OQ was multiplied by scale factors, which ranged from 0.2 to 1.8 Examples of the synthetic speech samples can be found at `http://muster.ucd.ie./~joao/web/hts-gpf`.

We clearly perceived the variation of the voice quality of the synthetic speech, with the degree of transformation of each LF-model parameter. Moreover, each parameter appears to have a different effect on the voice quality. This result is expected, because the variation of each parameter has a different effect on the spectrum of the LF-model [9]. The voice quality transformations also seemed to not produce speech artefacts, even for relatively large degrees of transformation of the LF-parameters.

6. Conclusion and Future Work

This paper proposes a HMM-based speech synthesiser that allows to transform relevant glottal source parameters. This is achieved by using the glottal post-filtering method for synthesis. It consists of modifying the LF-model, which is the input to the time-invariant glottal post-filter. This system was compared against a state-of-the-art speech synthesiser, which uses the STRAIGHT vocoder to synthesise speech. The results of a perceptual evaluation showed that the two systems performed equally in terms of speech naturalness, intelligibility, and similarity of the synthetic voice to the speaker's original voice. The great advantage of the proposed system is to allow voice transformation, by controlling the LF-model parameters.

An efficient technique to control the pitch of the synthetic speech by the glottal post-filtering method was also proposed, in this paper. It consists of truncating or padding with zeros the closed phase of the LF-model to obtain the desired pitch, without modifying the LF-model in the open phase.

Formal evaluations of voice quality transformation by the system with LF-model are going to be conducted. We also plan to use more source parameters than the LF-parameters, such as jitter and aspiration noise, to improve voice quality modelling.

7. Acknowledgements

The first author is now supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at University College Dublin, Ireland. This paper is based on his PhD research undertaken at the University of Edinburgh, supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568).

8. References

[1] Maia, R., Toda, T., Zen, H., Nankaku, Y. and Tokuda, K., "A Trainable Excitation Model for HMM-based Speech Synthesis", in Proc. of INTERSPEECH, Belgium, 2007.

[2] Drugman, T., Wilfart, G. and Dutoit, T., "A Deterministic plus Stochastic Model of the Residual Signal for Improved Parametric Speech Synthesis", in Proc. of INTERSPEECH, UK, 2009.

[3] Raitio, T., Suni, A., Pulakka, H., Vainio, M. and Alku, P, "HMM-Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering", in Proc. of the INTERSPEECH, Australia, 2008.

[4] Alku, P., Vilkman, E. and Laine, U. K., "Analysis of Glottal Waveform in Different Phonation Types Using the New IAIF Method", in Proc. of the ICPhS, 4, pp. 362–365, France, 1991.

[5] Cabral, J., Renals, S., Richmond, K. and Yamagishi, J., "Towards an improved modeling of the glottal source in statistical parametric speech synthesis", in Proc. of the SSW6, Germany, 2007.

[6] Doval, B. and d'Alessandro, C., "Fant, G. and Liljencrants and J. and Lin, Q.", STL-QPSR Tech. Report, 26(4):1–13, KTH, 1985.

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

[8] Childers, D. G. and Ahn, Chieteuk, "Modeling the glottal volume-velocity waveform for three voice types", J. Acoust. Soc. Amer., 97(1):505–519, 1995.

[9] Doval, B. and d'Alessandro, C., "The spectrum of glottal flow models", in Notes et Documents LIMSI-CNRS, Sweden, 1999.

[10] Zen, H., Toda, T., Nakamura, M. and Tokuda, K., "Details of the Nitech HMM-based Speech Synthesis System for the Blizzard Challenge 2005", IEICE Trans. Inf. and Syst., E90-D:(1):325–333, 2007.

[11] Secrest, B. G. and Doddington, G. R., "An integrated pitch tracking algorithm for speech systems", in Proc. of the ICASSP, pp. 1352–1355, USA, 1983.

[12] Cabral, J. and Renals, S. and Richmond, K. and Yamagishi, J., "Glottal spectral separation for parametric speech synthesis", in Proc. of INTERSPEECH, pp. 1829–1832, Australia 2008.

[13] Cabral, J. P. and Oliveira, L. C., "Pitch-Synchronous Time-Scaling for Prosodic and Voice Quality Transformations", in Proc. of the INTERSPEECH, pp. 1137–1140, Portugal, 2005.

[14] King, S. and Karaiskos, V., "The Blizzard Challenge 2009", in Proc. of the Blizzard Challenge workshop, UK, 2009.