interspeech 2008 incorporating SST 08

Glottal Spectral Separation for 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

The great advantage of using a glottal source model in parametric speech synthesis is the degree of parametric flexibility it gives to transform and model aspects of voice quality and speaker identity. However, few studies have addressed how the glottal source affects the quality of synthetic speech.

Here, we have developed the Glottal Spectral Separation (GSS) method which consists of separating the glottal source effects from the spectral envelope of the speech. It enables us to compare the LF-model with the simple impulse excitation, using the same spectral envelope to synthesize speech. The results of a perceptual evaluation showed that the LF-model clearly outperformed the impulse. The GSS method was also used to successfully transform a modal voice into a breathy or tense voice, by modifying the LF-parameters.

The proposed technique could be used to improve the speech quality and source parametrization of HMM-based speech synthesizers, which use an impulse excitation.

Index Terms: Glottal Spectral Separation, HMM-based speech synthesis, LF-model.

1. Introduction

Parametric speech synthesis has received greater attention in recent years with the development of statistical HMM-based speech synthesizers. This type of system can produce speech of comparable quality to the unit-selection method, although it does not sound as natural [1]. State-of-the-art HMM-based speech synthesizers use the STRAIGHT vocoder [2] to shape a multi-band mixed excitation with the spectral envelope e.g. [1]. However, the periodic component of the excitation is modelled by a delta pulse that controls only F_0. There are other glottal parameters which are important for voice quality and speech naturalness, such as the open quotient, the speed quotient and the return quotient [3]. Also, the impulse train has a strong harmonic structure which produces a buzzy speech quality.

In general, rule-based formant synthesizers produce speech by passing a glottal source model through the vocal tract filter. For example, the KLSYN88 system [4] uses a modified version of the Liljencrants-Fant (LF) model [5]. However, the speech sounds unnatural due to the limitation of the acoustic models.

A source-filter method can also be used to transform acoustic features of speech signals. For example, Linear Prediction Coding (LPC) based methods [6] are often employed in concatenative speech synthesis to reduce discontinuities at concatenation points or to modify voice quality. Although LPC is simple to implement, it does not optimally separate the effects of the source from the vocal tract and degrades speech quality. There are other techniques which more accurately estimate the source and vocal tract filter than traditional LPC e.g. glottal inverse filtering using a more complete model of speech [7] and Adaptive Iterative Inverse Filtering [8]. However, these methods are more complex and typically depend on a good estimation of the poles and zeros of a speech production model.

In previous work [9], we integrated a glottal source model into the HMM-based speech synthesizer by using the LF-model [5] instead of an impulse signal. The spectrum of the excitation signal was flattened using a post-filter. This operation was necessary because the STRAIGHT spectrum expects a spectrally flat excitation while the LF-model presents a decaying spectrum type that depends on the source parameters. The limitation of this implementation was that the post-filter was difficult to derive from the LF-model to obtain a spectrally flat excitation. Consequently, in the synthesis we used constant parameters for the LF-model in order not to modify the post-filter. Nevertheless, the results indicated that speech synthesized with the LF-model sounded more natural than using the impulse signal.

In this paper, we propose an alternative to post-filtering the excitation signal, called Glottal Spectrum Separation (GSS). The GSS method enables us to produce speech automatically from the trajectories of the LF-model parameters and spectral features. This method, like the post-filtering technique, avoids the difficult task of calculating the zeros and poles of a parametric model of speech to estimate the glottal source signal and the vocal tract transfer function.

2. Method

In the frequency domain, the speech production model can be represented by

$$S(w) = D(w)G(w)V(w)R(w) \quad (1)$$

where $D(w)$ is the Fourier Transform (FT) of an impulse train, $G(w)$ is the FT of a glottal pulse, $V(w)$ is the vocal tract transfer function and $R(w)$ is the radiation characteristic. $R(w)$ can be modelled by a differentiating filter but $G(w)$ and $V(w)$ are more difficult to calculate accurately from the speech signal.

This model can be simplified to $S(w) = D(w)V(w)$. In this case, the input excitation is represented by the impulse train and the vocal tract filter is the spectral envelope of the speech signal: $V(w) = \hat{H}(w)$, as in the LPC vocoder [6]. The effects of the source and lip radiation are incorporated in the spectrum.

The GSS method calculates the FT of a model of the glottal flow derivative, $E(w)$, from the speech signal. Then, it removes the source effects from the spectral envelope of the signal: $V(w) = \hat{H}(w)/E(w)$. The speech can be recovered using the same source model as the input into the vocal tract filter:

$$S(w) = D(w)E(w)\frac{\hat{H}(w)}{E(w)} = D(w)\hat{H}(w) \quad (2)$$

Supported by Marie Curie Early Stage Training Site EdSST (MEST-CT-2005-020568)

Copyright © 2008 ISCA

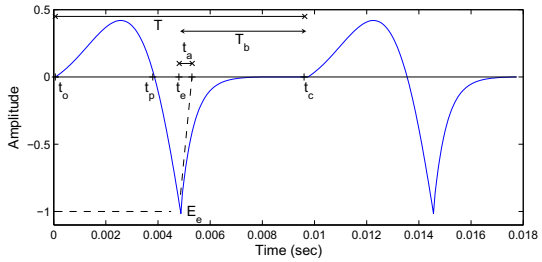

Figure 1: Segment of the LF-model waveform with the representation of the glottal parameters during one period.

2.1. Glottal Source Model

We used the LF-model of the glottal flow derivative signal:

$$e_{LF}(t) = \begin{cases} E_0 e^{\alpha t}\sin(w_g t), & 0 \le t \le t_e \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \le t_c \end{cases} \quad (3)$$

where $w_g = \pi/t_p$. The parameters E_0, ϵ and α can be calculated from equations 3. Figure 1 shows a segment of the LF-model and the five glottal parameters. The cycle of the LF-model starts at the opening instant of the vocal folds, t_o. The maximum glottal flow is represented by t_p, which is a zero of the glottal flow derivative. The abrupt glottal closure is given by the discontinuity point t_e and E_e is the excitation amplitude. T_a is the effective duration of the return phase, which measures the abruptness of the transition to the complete closure, t_c. For convenience, t_c is set equal to the fundamental period ($t_c = T$).

The LF-model can also be described by parameters related to voice quality and spectral properties [3]: open quotient $OQ = (t_e + T_a)/T$, speed quotient $SQ = t_p/(t_e - t_p)$, and the return quotient $RQ = T_a/T$.

The spectrum of the LF-model has a decaying characteristic, which is more accentuated at higher frequencies (spectral tilt), and a glottal spectral peak at lower frequencies [9].

2.2. Analysis: Glottal Spectral Separation

The method for extracting the glottal and spectral features for voiced speech is illustrated in Figure 2. Speech frames $s^i(t)$ are sampled at 16 kHz, 40 ms long and centered in the glottal epochs t_g^i which are calculated as in [9].

The glottal flow derivative waveform $v_g^i(t)$ is estimated by inverse filtering the short-time signal, using a conventional LPC analysis method. The coefficients of the inverse filter are calculated pitch-synchronously from the pre-emphasized speech signal (α=0.97), using the autocorrelation method (order 18) and a Hanning window.

The parameters of the LF-model are calculated for the pitch cycle of $v_g^i(t)$ which starts at t_g^{i-1} and has duration $T^i = t_g^i - t_g^{i-1}$. The point of maximum glottal closure t_e is set to coincide with the glottal epoch t_g. Thus, the parameter E_e^i is estimated by the amplitude of $v_g^i(t)$ at t_g^i and the parameter t_e^i is calculated as $t_e^i = T^i - T_b^i$, where T_b^i is the duration from t_g^{i-1} to the next glottal opening instant (see Figure 1). The parameters T_b^i, t_p^i, and T_a^i are obtained by fitting the LF-model to a low-pass and linear-phase filtered version (0 to $4kHz$) of $v_g^i(t)$, to reduce the effect of the high-frequency noise. The initial estimates of the parameters for the optimization algorithm are calculated using the direct methods in [9]. Then, the values of the parameters are varied for a maximum number of iterations to minimize the mean-squared error between the LF-model and the short-time signal using the Levenberg-Marquardt algorithm [10].

The next step is to estimate the spectral parameters. First, the STRAIGHT analysis method accurately extracts the spectral envelope, $\hat{H}^i(w)$, for each frame. STRAIGHT estimates the spectral envelope more accurately than other vocoders, such as the LPC vocoder [6], because it removes almost perfectly the periodicity interference. Then, a cycle of the estimated LF-model, with the duration T^i and starting from the glottal opening t_o, is calculated. This segment is zero-padded to 1024 sample points for the short-time Fourier analysis. Finally, the spectral envelope is divided by the amplitude spectrum of the LF-model $\left|E_{LF}^i(w)\right|$ to remove the source model effects.

Figure 3 shows the spectral envelope of a speech signal after the separation of the LF-model spectrum. The spectral envelope is flattened due to the removal of the tilt characteristic of the LF-model. The glottal source separation produces a high DC component which will be eliminated in the synthesis part.

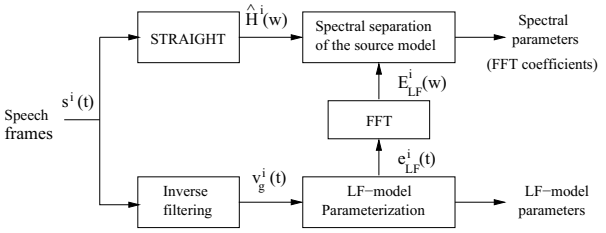

Figure 2: Block diagram of the analysis part of the GSS method.

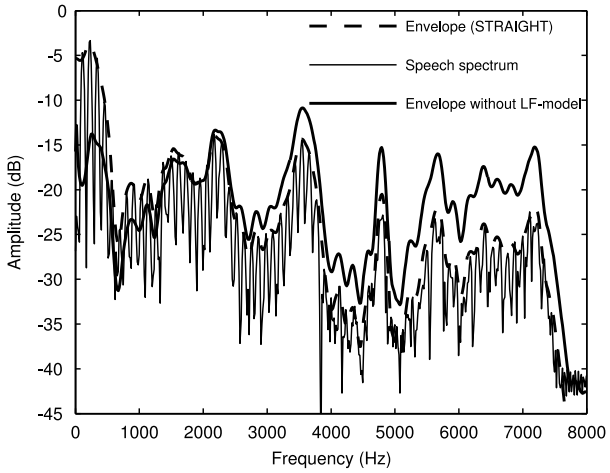

Figure 3: Separation of the LF-model spectrum from the spectral envelope of the speech signal.

2.3. Synthesis

Each voiced frame i of the excitation signal contains two pitch cycles of the LF-waveform, which start from t_e and have durations T^i and T^{i+1}, respectively. The first LF-model cycle is generated from the glottal parameters estimated for the frame i: t_e^i, t_p^i, T_a^i and E_e^i. The parameters t_e and t_p of the second cycle are calculated under the assumption that the voice quality parameters (OQ, SQ and RQ) are the same along the LF-waveform. According to this assumption, the glottal parameters vary linearly with the period. For example, the estimated t_p in the sec-

ond cycle is $\hat{t}_p = t_p^i T^{i+1}/T^i$. In general, this approximation gives good results because the variation of the pitch period between contiguous frames is small. The parameters T_a and E_e of the second cycle are set the same as the first cycle because they did not show significant variation with T in our measurements.

The resulting LF-signal is multiplied by a Hamming window and zero-padded to 1024 samples to calculate the FFT. The amplitude spectrum of the excitation is multiplied by the amplitude spectrum $\hat{H}^i(w)$ which is given by the spectral parameters. Then, the time-domain waveform is obtained by calculating the IFFT from the resulting amplitude spectrum and phase of the excitation signal. Next, we removed the effect of the Hamming window and the DC offset from the synthesized speech.

Finally, the resulting short-time signals are multiplied by asymmetric windows and added using the Pitch-Synchronous Overlap-and-Add (PSOLA) algorithm [11].

3. Perceptual Experiment

A forced-choice test was conducted to compare the LF-model with the impulse train in relation to speech naturalness and parametric flexibility for voice quality transformations.

3.1. Recordings

A male English speaker was asked to read ten sentences with a modal voice and two different voice qualities: breathy and tense. He had listened to examples of tense and breathy speech beforehand. The sentences contained only sonorant sounds, as we are only interested here in the study of voiced speech. The use of other sounds, such as voiced fricatives and unvoiced speech could decrease the performance of the epochs detector and increase the errors in the estimation of the LF-parameters.

3.2. Synthetic Speech

First, the trajectories of the LF-parameters calculated for each recorded utterance were smoothed using the median function to alleviate estimation errors. Each utterance was synthesized with the modal voice, by copy-synthesis, using the GSS method.

The utterances spoken with a modal voice were also synthesized with the impulse excitation. In this case, we used the spectral parameters estimated by STRAIGHT (without removing the source model effects). The excitation signal was generated by replacing the LF-waveform with a delta pulse placed at the instant of maximum excitation t_e (approximately at the center of the excitation), with amplitude equal to E_e.

Five utterances were synthesized with a breathy and tense quality using the glottal epochs and spectral parameters of the modal voice, but using transformed trajectories of the glottal parameters of this voice. To obtain the new trajectories, the voice quality parameters of the LF-model (OQ, SQ and RQ) were calculated from the measured LF-parameters, using the equations in Section 2.1. For each utterance, the variations of the mean values of the glottal parameters between each voice quality and the modal voice were calculated. For example, the variation of the mean value of the OQ for the breathy voice is $\Delta OQ_{breathy} = E[OQ_{breathy}(j)] - E[OQ_{modal}(i)]$. In this equation, $E[x]$ represents the mean function, $OQ_{breathy}(j)$ is the OQ calculated for the frame j of the utterance recorded with a breathy voice and $OQ_{modal}(i)$ is the OQ calculated for the frame i of the same utterance spoken with a modal voice. Then, the scale factors of the LF-parameters, k_{T_a}, k_{t_p} and k_{t_e}, were calculated from the equations in Section 2.1 and the variations of the mean values of the voice quality parameters. For example, the scale factors to transform the parameters of the frame i of the modal voice into the breathy voice values, are calculated as follows:

$$k_{T_a}^i = 1 + \frac{\Delta OQ_{breathy}}{RQ^i} \tag{4}$$

$$k_{t_p}^i = \frac{t_e^i}{t_p^i} \frac{\Delta SQ_{breathy} + SQ^i}{1 + \Delta SQ_{breathy} + SQ^i} \tag{5}$$

$$k_{t_e}^i = \frac{T^i}{t_e^i}(\Delta OQ_{breathy} + OQ^i) - \frac{k_{T_a}^i T_a^i}{t_e^i} \tag{6}$$

Figure 4 shows the trajectories of the LF-parameters for a modal voice and a breathy voice.

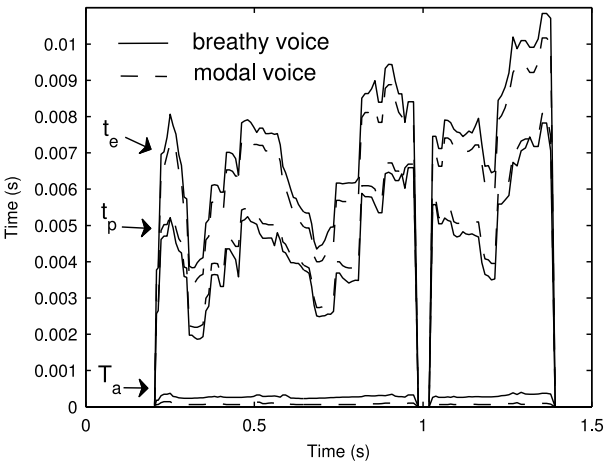

Figure 4: Trajectories of the LF-parameters calculated for an utterance spoken with a modal voice and the transformations of those trajectories to synthesize speech with a breathy voice.

3.3. Lab Experiment

This experiment was conducted in a quiet room using headphones. Twenty three students, who were all English native speakers, were paid to participate in the test.

The listening test was divided into five parts. In part 1, subjects were presented with 20 pairs of stimuli (10 utterances, randomly chosen and repeated twice with the order of the samples alternated). Each pair contained the utterance synthesized with the LF-model and with the impulse excitation. For each pair, they had to select the version that sounded more natural.

Parts 2 and 3 of the test were similar to the first, but the recorded speech was compared to the speech synthesized with the impulse train and with the LF-model, respectively.

In part 4, the listeners were first presented with two pairs of recorded utterances to show the difference between modal and tense voices. This test consisted of 10 pairs, corresponding to 5 utterances. Each pair contained an utterance synthesized with the modal voice (by copy-synthesis) and an utterance synthesized with the trajectories of the glottal parameters calculated for the tense voice. Subjects had to select the sample that sounded more tense. This part was repeated for the breathy voice quality in part 5.

3.4. Web Experiment

After the lab test, we noticed that the recorded speech was louder than the synthesized speech, despite normalizing the am-

plitude. We found the amplitude of the recorded speech was more symmetric in relation to the zero axis y=0 than the synthetic speech. Thus, we instead normalized the power of the speech. After this correction, the test was repeated on the web.

Twelve students and staff from the university participated in the test, using headphones, including 7 speech synthesis experts and 10 native speakers. No payment was offered.

4. Results

The results obtained from the two experiments are shown in Figure 5. All the results are statistically significant with p<=0.01.

In general, speech synthesized with the LF-model sounded more natural than speech synthesized with the impulse train. These results are supported by findings in previous work [9]. The preference for the LF-model was significantly higher in the web than in the lab evaluation. We think that the participation of speech synthesis experts and the normalization of the power of the speech samples, in the web test, influenced this variation.

Synthetic speech obtained higher score than expected when compared with the recorded speech, especially in the lab test. This was unexpected because the source models do not represent all the details of the real voice source signal. A detailed analysis of the lab test results showed that six listeners clearly preferred the synthetic speech to the recorded speech. The same listeners also clearly preferred speech synthesized with the impulse excitation to the LF-model. An explanation might be that a small fraction of listeners preferred speech spoken with a more buzzy voice quality than the natural voice of the speaker. Another explanation might be that the differences in loudness between speech samples, detected in the lab experiment, influenced the perception of speech naturalness for some listeners.

The speech synthesized from the modal speech by transforming the LF-parameters to produce a breathy quality almost always sounded more breathy than the speech synthesized with the natural voice. The results obtained for speech synthesized with a tense voice were not as good as those obtained for breathy voice. Possible reasons to explain this result are that a tense voice is more difficult to distinguish from the natural voice and that other speech features than the LF-parameters are important to model this voice quality well e.g. F_0 or duration.

5. Conclusion

In this work, we have developed the GSS method to estimate the voice source and the vocal tract, which avoids some problems of existing popular techniques. In particular, GSS does not have the difficulty of calculating the poles and zeros of a speech model. Instead, it estimates the vocal tract from the spectral envelope of the speech signal by removing the spectral effects of the calculated glottal source model.

The proposed method was used to compare speech synthesized with an impulse excitation and speech synthesized with the LF-model of the source. The results of the perceptual evaluation supported previous results which showed an increase in speech quality by using the LF-model than the delta pulse signal. The perceptual test also demonstrated the parametric flexibility of the LF-model by transforming the modal voice into breathy and tense voice qualities, without modifying the F_0. In contrast, the delta pulse only permits control of F_0.

This work is important for HMM-based speech synthesizers which typically use the impulse signal for the voiced excitation. The proposed GSS method could be integrated into these types of system so that they could use a source model without modifying the speech vocoder currently used to synthesize the speech.

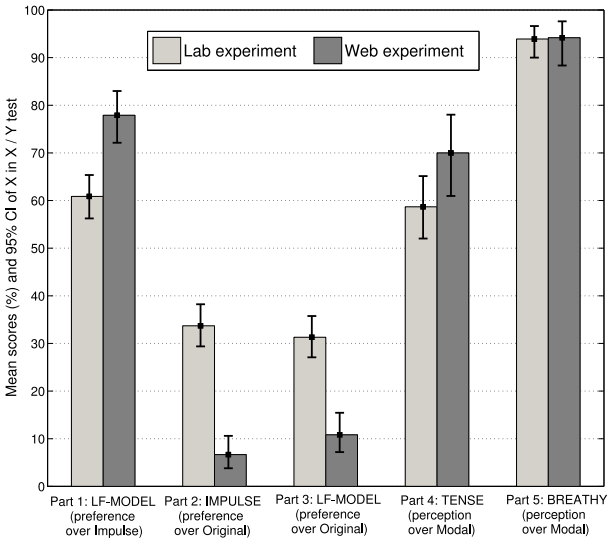

Figure 5: Mean scores and 95% confidence intervals for one type of speech, in each part of the forced-choice test.

Further experiments are required to evaluate the performance of the GSS when integrated into a parametric speech synthesizer.

6. References

[1] Yamagishi, J., Zen, H., Toda, T. and Tokuda, K., "Speaker-Independent HMM-based Speech Synthesis System - HTS-2007 System for the Blizzard Challenge 2007", Proc. of Blizzard Challenge 2007, 2007.

[2] 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, 187–207, 1999.

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

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

[5] Fant, G., Liljencrants, J. and Lin, Q., "A four-parameter model of glottal flow", STL-QPSR, 26(4), 1–13, 1985.

[6] Sproat, R. and Olive, J., "An Approach to Text-to-Speech Synthesis", in W. Kleijn and K. Paliwal (eds.), Speech Coding and Synthesis, 611–633, Amsterdam, Elsevier, 1995.

[7] Funaki, K. and Mitome, Y., "A speech analysis method based on a glottal source model", in ICSLP, 45–48, 1990.

[8] Alku, P., "An Automatic Method to Estimate the Time-Based Parameters of the Glottal Pulseform", in Proc. of the IEEE ICASSP'92, vol. 2, 29–32, 1992.

[9] 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 6th ISCA Workshop on Speech Synthesis, Germany, 2007.

[10] Marquardt, D., "An Algorithm for Least-Squares Estimation of Nonlinear Parameters", SIAM Journal on Applied Mathematics, 11, 431-441, 1963.

[11] Talkin, D. and Rowley, J., "Pitch-Synchronous analysis and synthesis for TTS systems", Proc. of the ESCA Workshop on Speech Synthesis, 55–58, 1990.