THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Glottal Spectral Separation for Speech Synthesis

Citation for published version:
Cabral, JP, Richmond, K, Yamagishi, J & Renals, S 2014, 'Glottal Spectral Separation for Speech Synthesis', *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195-208. https://doi.org/10.1109/JSTSP.2014.2307274

Digital Object Identifier (DOI):
10.1109/JSTSP.2014.2307274

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
IEEE Journal of Selected Topics in Signal Processing

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 25. Aug. 2022

Glottal Spectral Separation for Speech Synthesis

João P. Cabral, Korin Richmond, *Member, IEEE,* Junichi Yamagishi, *Member, IEEE,* and Steve Renals, *Fellow, IEEE*

***Abstract*—This paper proposes an analysis method to separate the glottal source and vocal tract components of speech that is called Glottal Spectral Separation (GSS). This method can produce high-quality synthetic speech using an acoustic glottal source model. In the source-filter models commonly used in speech technology applications it is assumed the source is a spectrally flat excitation signal and the vocal tract filter can be represented by the spectral envelope of speech. Although this model can produce high-quality speech, it has limitations for voice transformation because it does not allow control over glottal parameters which are correlated with voice quality. The main problem with using a speech model that better represents the glottal source and the vocal tract filter is that current analysis methods for separating these components are not robust enough to produce the same speech quality as using a model based on the spectral envelope of speech. The proposed GSS method is an attempt to overcome this problem, and consists of the following three steps. Initially, the glottal source signal is estimated from the speech signal. Then, the speech spectrum is divided by the spectral envelope of the glottal source signal in order to remove the glottal source effects from the speech signal. Finally, the vocal tract transfer function is obtained by computing the spectral envelope of the resulting signal. In this work, the glottal source signal is represented using the Liljencrants-Fant model (LF-model). The experiments we present here show that the analysis-synthesis technique based on GSS can produce speech comparable to that of a high-quality vocoder that is based on the spectral envelope representation. However, it also permit control over voice qualities, namely to transform a modal voice into breathy and tense, by modifying the glottal parameters.**

***Index Terms*—Glottal Spectral Separation, LF-model, parametric speech synthesis, voice quality transformation.**

I. Introduction

SOURCE-filter modeling of speech is based on exciting a vocal tract filter with a glottal source signal. In this framework, the vocal tract is approximated using the spectral envelope of the speech signal, and a simple excitation model is used to approximate the glottal source. This can lead to inaccuracies arising from the vocal tract representation incorporating characteristics of the glottal signal. However, using a simpler excitation has the advantage of avoiding the complex problem of separating the glottal source and vocal tract components from the speech signal. Moreover, there are robust methods to extract the spectral envelope of speech, such as that used by the STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) vocoder [1]. An important problem of this source-filter model is that it does not permit to easily control aspects of the glottal source which are important for voice transformation applications, whatever that be its time-domain shape characteristics or spectral properties such as the glottal formant and spectral tilt [2]. Similar problems arise with entirely spectral representations of the speech signal, such as the Harmonic-plus-Noise Model [3]. Furthermore, important characteristics of the glottal source are lost when this signal is assumed to be incorporated into the vocal tract representation. For example, the mixed phase characteristics of the glottal source are lost, because the vocal tract filter is generally represented as a minimum-phase filter. Such glottal properties are expected to be important for the synthetic speech to sound natural.

The vocal tract filter can be estimated using linear predictive coding (LPC) [4], which assumes speech can be represented by an all-pole model that can be calculated from the speech signal using techniques such as the autocorrelation and covariance methods [5]. However, LPC cannot model voiced sounds that contain zeros in the speech model correctly, such as nasals and voiced fricatives, and may not produce a sufficiently smooth spectrogram due to the difficulty in predicting the correct number of poles.

Mel-cepstral analysis estimates the vocal tract filter as an approximation of the spectral envelope of speech, and is commonly used in speech recognition [6]. This method has several properties which are attractive for speech processing applications, such as its robustness and that it takes into account the perceptual characteristics of the human auditory system. Another common way to obtain the spectral envelope of speech is by computing the spectrum, from which the envelope can be obtained by interpolating the peaks of the amplitude spectrum or by using special analysis windows, such as in the STRAIGHT vocoder.

The excitation used to synthesize speech using either the LPC filter or the spectral envelope has an approximately flat spectrum. It is usually modeled by the residual signal $E(w)$, which is obtained from the speech signal using inverse filtering. Basically, it consists of removing the vocal tract effects from the speech signal $S(w)$ by spectral division, $E(w) = |S(w)|/V(w)$, where $|S(w)|$ is the amplitude spectrum of speech and $V(w)$ the vocal tract transfer function. For voiced speech, the excitation model may vary from a simple

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org

Manuscript received April 6, 2013; revised September 12, 2013 and December 19, 2013; accepted February 2, 2014. The first author is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at the University College Dublin. This paper is based on his PhD work supported by Marie Curie Early Stage Training Site *EdSST* (MEST-CT-2005-020568). This work is also supported by EPSRC through Programme Grant EP/I031022/1, *Natural Speech Technology*, and Healthcare Partnerships Grant EP/I027696/1, *Ultrax*, and by the JST CREST project *uDialogue*.

J. P. Cabral is with the School of Computer Science & Informatics, University College Dublin, Belfield, Dublin 4, Ireland (e-mail: cabralj@scss.tcd.ie).

K. Richmond, J. Yamagishi, and S. Renals are with the Centre for Speech Technology Research, University of Edinburgh, Edinburgh EH8 9AB, U.K. (e-mail: korin@cstr.ed.ac.uk; jyamagis@inf.ed.ac.uk; s.renals@ed.ac.uk).

impulse train (only uses the F0 parameter of the source) to a more complex model, such as the mixture of a periodic signal with noise. However, an increase in the number of source parameters is usually required to permit a better representation of the residual signal and improve the quality of the synthetic speech. Vocoders, such as STRAIGHT, which can extract a smooth spectrogram and use a mixed excitation model can produce high-quality speech.

In order to represent the relevant characteristics of the glottal source on the excitation, it is necessary to separate them from the vocal tract component. A simple way to obtain a better approximation to the vocal tract filter and the glottal source signal compared to traditional LPC inverse filtering consists of performing pre-emphasis of the speech signal prior to LPC analysis. The effect of the pre-emphasis filter is to increase the relative energy of the speech spectrum at higher frequencies. Consequently, the residual has a decaying spectrum which resembles the spectral tilt characteristic of the glottal source signal, although this does not yield an accurate approximation of the tilt. Another method to obtain a more accurate estimate of the vocal tract is to perform LPC analysis on the closed phase of the glottal source (when the glottis is closed) and inverse filtering the speech signal [7]. However, it is often difficult to estimate the closed phase and it may be too short for analysis or the glottis may not even close completely. In the iterative adaptive inverse filtering (IAIF) method [8] the glottal source and the vocal tract are estimated iteratively using inverse filtering. Unlike closed-phase inverse filtering, IAIF performs the analysis on the whole pitch period, which may result in formant frequency and bandwidth errors in the estimated excitation due to the source-tract interaction that occurs when the glottis is opened. Another approach for estimating the glottal and vocal tract parameters consists of simultaneously estimating these parameters using an optimization algorithm to minimize an error measure. This method explicitly represents the glottal source signal as an exogenous input (not known); for example the glottal LPC technique [9] uses the autoregressive with exogenous input (ARX) model of speech production. This method is often employed using an acoustic glottal source model in order to avoid the problem of determining which poles and zeros model the glottal source excitation [10], [11]. The main disadvantages of this approach are the increased computational complexity and convergence problems of the iterative optimization algorithms. The source and vocal tract filter can also be estimated by separating the causal (maximum-phase) and anticausal (minimum-phase) components of speech, for example using the zeros of the z-transform (ZZT) representation [12]. These components are assumed to represent the source and vocal tract components respectively. The main limitations of this technique are its computational complexity and the incomplete separation of the source component from the vocal tract, in other words, the minimum-phase contribution of the voice source (related to the spectral tilt) is not separated.

The GSS method proposed in this work can be divided into three parts. In the first part, the glottal source signal is estimated from the speech signal. For example, this can be done using one of the techniques described in the previous paragraph or by directly estimating parameters of an acoustic glottal source model from the speech signal [13]. Next, the glottal source effects are removed from the speech signal by dividing the amplitude spectrum of the speech signal by the spectral envelope of the source. Finally, the vocal tract transfer function is estimated by computing the spectral envelope of the resulting signal. The GSS method can use a robust spectral envelope analysis technique to estimate the vocal tract, in order to obtain a smooth spectral representation of the vocal tract. Moreover, GSS enables the combination of the robustness of spectral envelope analysis with the flexibility to control and model relevant aspects of the glottal source.

A method called Separation of the Vocal-tract with the LF-model plus noise (SVLN) [14] has been recently proposed that uses a similar idea to the GSS method. This method also divides the speech spectrum by the spectral envelope of the glottal source excitation in order to remove the source effects from the speech signal. However, the excitation signal is represented by periodic and stochastic signals at lower and higher frequency bands. Consequently, the vocal tract filter is computed differently in the two bands.

The GSS method was initially presented in [15]. Here, a more complete and detailed description of this method and its application to voice transformation is presented. After reviewing the LF-model of the glottal source, the GSS method is described in Section III. We have carried out experimental evaluations of the GSS method using the LF-model excitation of voiced speech in copy-synthesis and voice quality transformation (Section IV). In other prior work [16] we proposed a mixed excitation model that combines the LF-model and the aperiodicity component of STRAIGHT for synthesizing speech using the GSS method for HMM-based speech synthesis. In this speech synthesis system, the LF-model and spectral parameters estimated by GSS are used to train the HMMs and to generate the speech waveform during synthesis. Though we do not aim to address speech synthesis here, this paper goes further than [16] by describing the GSS method using a mixed excitation in more detail (Section V) and including new results of a recent experiment conducted to evaluate the quality of speech synthesized with this method (Section VI).

II. Liljencrants-Fant Model

A. Waveform

The Liljencrants-Fant model (LF-model) [17] is an acoustic glottal source model, widely used for speech processing applications such as the study of voice characteristics, speech synthesis and voice transformation. For the reader's convenience, we shall briefly summarize this model here. This model represents the flow derivative waveform during one pitch cycle, with duration equal to the fundamental period T_0, according to the following equations:

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

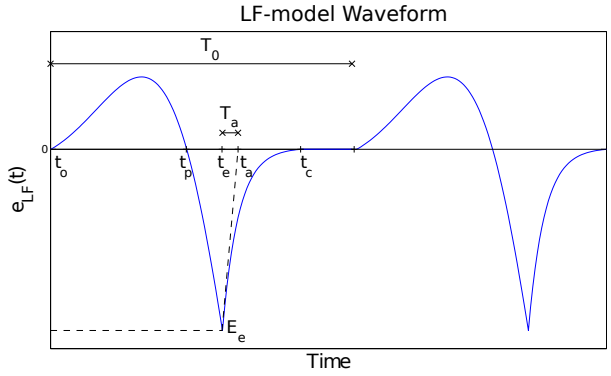

Fig. 1. Segment of the LF-model waveform and representation of the glottal parameters during one fundamental period of the model.

$$\int_0^{T_0} e_{LF}(t)dt = 0 \tag{2}$$

$$e_{LF}(t_e^-) = e_{LF}(t_e^+) = -E_e, \tag{3}$$

where $w_g = \pi/t_p$. (2) and (3) represent the zero energy balance and amplitude continuity constraints, respectively. The value of the parameter t_o is arbitrary, as it represents the start of the LF-model waveform. In this work, t_o is assumed to be zero and it is omitted in the formulas that describe the LF-model. In general, the parameters α, E_0 and ϵ are derived from (2) and (3). Therefore, the LF-model given by (1) can be defined by the six parameters: t_p, t_e, T_a, t_c, T_0, and E_e. Figure 1 represents these parameters for a cycle of the LF-model $e_{LF}(t)$.

The region between the start of the glottal pulse and the instant of maximum airflow t_p, is called the *opening phase*. At t_p, the vocal folds start to close and the flow amplitude decreases until the abrupt closure of the glottis (discontinuity in the LF-model waveform) at the *instant of maximum excitation*, t_e. The time interval which corresponds to the duration when the vocal folds are opened and there is airflow through the glottis (duration equal to $t_e + T_a$) is called the *open phase*. The next part represents the transition between the open phase and the closed phase, which is called *return phase* (the return phase is sometimes considered as part of the open phase). The duration of the return phase is given by $T_a = t_a - t_e$ and it measures the abruptness of the closure. t_a is defined as the point where the tangent to the decaying exponential at $t = t_e$ hits the time axis. Finally, the *closed phase* is the region of the glottal cycle when the vocal folds are completely closed that starts at the glottal closure instant t_c. A simplified version of the LF-model is often used which consists of setting t_c equal to the fundamental period ($t_c = T_0$). This 5 parameter version is used in this work.

B. Dimensionless Parameters

The LF-model can also be described by dimensionless parameters which are correlated with voice quality and spectral properties [18]. One is the open quotient $OQ = (t_e + T_a)/T_0$, which measures the relative duration of the open phase. The second is the speed quotient $SQ = t_p/(t_e - t_p)$, which measures the asymmetry of the glottal pulse. The third is called return quotient $RQ = (t_a - t_e)/T_0$ and it measures the relative duration of the return phase. These are the main dimensionless parameters and the ones used in this work.

Several studies have shown that these parameters are strongly correlated with voice quality, as in [18]–[20]. For example, breathy voice is usually characterized by a lack of tension of the vocal folds and incomplete closure of the folds, which results in high OQ and RQ values compared to modal voice (neutral voice quality). In contrast, tense voice is associated with increased vocal folds tension when compared with modal voice, which has the effect of producing higher SQ values. For this voice type, the glottal open intervals are also shorter, which results in lower OQ and RQ values.

C. Spectrum

The amplitude spectrum of the LF-model is characterized by a spectral peak at the lower frequencies, often called the "*glottal formant*", and the *spectral tilt* (attenuation at higher frequencies). A detailed description of the spectral representation of the LF-model is given in [21]. The glottal formant depends mainly on the open quotient and asymmetry coefficient [22]. On the other hand, the spectral tilt effect is mainly dependent on the return phase of the LF-model which acts as a low-pass filter of order one at cut-off frequency $F_c = 1/(2\pi T_a)$.

The LF-model can be described as a mixed-phase model, because it has both causal and anti-causal properties [23], [24]. In general, source-filter models which do not represent the glottal source characteristics of the excitation are minimum-phase. For example, when the excitation of voiced speech is an impulse train, the speech signal is the impulse response of a minimum-phase filter which represents the vocal tract. Recent work suggests that a mixed-phase model of voiced speech is more appropriate than the minimum-phase model due to the maximum-phase characteristic (anti-causality) of the source signal [12], [25]. Thus, the convolution of the LF-model with the vocal tract filter is expected to give a better representation of the phase spectrum of speech than the traditional impulse response of the minimum-phase filter (which represents the spectral envelope).

III. Glottal Spectral Separation

A. Speech Model

The Glottal Spectral Separation (GSS) method assumes that voiced speech is the convolution of a glottal source signal with the vocal tract filter. In the frequency domain, this speech production model can be represented by

$$S(w) = P(w)U(w)V(w)R(w), \tag{4}$$

where $P(w)$ is the Fourier transform (FT) of an impulse train, $U(w)$ is the FT of a glottal pulse, $V(w)$ is the vocal tract transfer function and $R(w)$ is the radiation characteristic, which can be modeled by a differentiating filter. In the experiment of Section IV, the LF-model which is a model of the glottal

source derivative is used to represent $G(w) = U(w)R(w)$. Meanwhile, in Section V an extension to this excitation model of $G(w)$ is proposed which consists of mixing the LF-model and a noise signal.

This model (4) is different from the traditional model used by the LPC vocoder [26], which is given by:

$$S(w) = P(w)H(w) \tag{5}$$

In this representation, the input excitation is represented by the impulse train and $H(w)$ represents the spectral envelope of $S(w)$. The vocal tract, the lip radiation and the glottal source effects are all incorporated into $H(w)$. There are other vocoders based on the same model which employ a more complex excitation model, such as STRAIGHT [27] and MELP [28].

B. Analysis

The block diagram of the GSS analysis method is illustrated in Figure 2. The glottal source signal $v(t)$ is estimated from the speech signal $s(t)$ and the glottal parameters are extracted from $v(t)$. This analysis step can be achieved using any glottal source estimation method. In Section IV, LPC inverse filtering with pre-emphasis is used due to its simplicity. But in Section VI, the more sophisticated IAIF method is used. Both techniques are based on the LPC model, which assume an all-pole model of speech. Although this model has limitations in terms of representing some sounds, such as nasals, this does not affect GSS analysis significantly because the LPC model is only used to estimate the glottal source component of speech. By using a better speech model, such as Discrete All-pole Modeling (DAP), it could be possible to obtain more accurate estimates of the glottal source signal [29], at the cost of increasing the complexity of the analysis. A *post-processing operation* on the glottal parameters can be employed in order to reduce possible estimation errors, for example using a smoothing technique.

For separating the spectral properties of the glottal source from the speech, the speech spectrum is divided by the spectral envelope of the glottal source derivative, $E_p(w)$. In this work, this envelope is obtained by computing the amplitude spectrum of one period of the LF-model signal. Note that this amplitude spectrum does not have harmonics and thus corresponds to the spectral envelope of a periodic glottal source signal $E(w)$. The FT of the resulting signal can be represented by $S(w)/E_p(w)$. From (4), this signal can be described by

$$\frac{S(w)}{E_p(w)} = P(w)V(w)\frac{U(w)R(w)}{E_p(w)} \tag{6}$$

Assuming that $R(w)$ is modeled by the derivative function and that the estimated $E_p(w)$ is a good approximation of the glottal source derivative, then $E_p(w) \simeq U(w)R(w) = G(w)$. Under this approximation, (6) can be rewritten as

$$\frac{S(w)}{E_p(w)} \simeq P(w)V(w) \tag{7}$$

This equation shows that the vocal tract filter $V(w)$ can be estimated as the spectral envelope of $S(w)/E_p(w)$, by

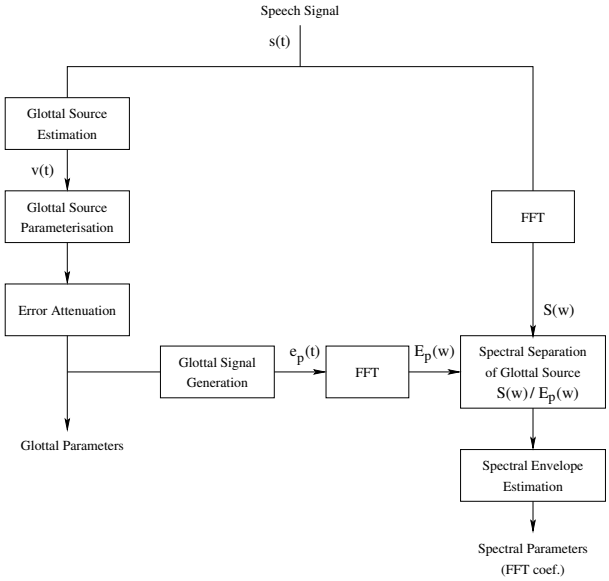

Fig. 2. Block diagram of speech analysis in the GSS method.

comparison with (5). In this work, an implementation of STRAIGHT (Matlab version V40_006b) was modified in order to divide the amplitude of the short-time speech spectrum $S(w)$ by the amplitude spectrum of one period of the LF-model $E_p(w)$. The spectral envelope of the resulting signal $S(w)/E_p(w)$ is then computed instead of using $S(w)$, yielding the spectral parameters.

Another possible implementation of the GSS method is to first compute the spectral envelope of the short-time speech signal using the STRAIGHT vocoder and then divide the resulting spectral envelope by $E_p(w)$, in order to obtain the vocal tract transfer function $V(w) = H(w)/E_p(w)$. This approach avoids the need to modify STRAIGHT. The two implementations of the GSS method yield similar results, because the spectral division is a linear operation. Initially, the first implementation was used in the experiment of Section IV. However, the second implementation was chosen in the latest experiment of Section VI for the practical reason of separating the spectral envelope computation by STRAIGHT from the spectral division performed by GSS.

The GSS analysis could also be performed using a model of the glottal flow instead of its derivative. In this case, the glottal flow pulse generated from this model does not include the radiation effect, unlike $E_p(w)$. Then, the spectrum obtained using GSS is the combination of the vocal tract and the radiation effect, that is $V(w)R(w)$.

Figure 3 shows the spectral envelope of the signal $S(w)/E_{LF}(w)$, which was calculated by removing the spectral effects of the LF-model signal $E_{LF}(w)$ from $S(w)$ and using STRAIGHT to compute the spectral envelope of the resulting signal. STRAIGHT uses a pitch-adaptive Short-term Fourier Transform (SFT) to compute a smooth spectrogram that has minimal periodicity interference. This spectral analysis method is described in more detail in [24] and [27]. The estimated vocal tract transfer function is flatter than the spectral envelope of the original speech signal $S(w)$, due to

the removal of the tilt characteristic of the LF-model. The frequency of the first maximum peak is also different between the two spectra because of the removal of the glottal peak characteristic of the LF-model by GSS. In general, the signal $S(w)/E_{LF}(w)$ has a high *DC component* due to the very low amplitude of $E_{LF}(w)$ near the zero frequency. The high DC component could affect the estimation of the spectral envelope. However, this problem is not relevant when using STRAIGHT to compute the spectral envelope, because it removes the DC component from the input signal before computing its spectral envelope.

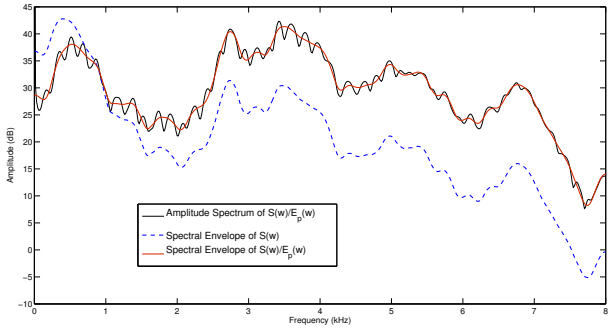

Fig. 3. Spectral envelope of a 40 ms short-time speech signal calculated by the GSS method, using the LF-model and STRAIGHT. The spectral envelope calculated only using STRAIGHT is also represented, for comparison.

The great advantage of GSS compared with typical source-tract separation techniques is that it is possible to obtain approximately smooth parameter contours for glottal and vocal tract parameters. This is very important for speech synthesis and voice transformation applications, in which parameter discontinuities are a major cause of speech distortion. There are two operations steps during GSS which contribute to the parameter smoothing. One is that errors in the glottal parameter estimation can be alleviated before separating the glottal source aspects from the speech signal, for example by using a median filter for performing the smoothing. The other is that the vocal tract spectrum can be computed using a robust spectral envelope estimation method, as the glottal source and the vocal tract parameters are estimated independently. These two operations are generally not performed by other techniques for estimating the glottal source and the vocal tract because they use a unique speech model that only enables these two components to be calculated jointly or iteratively.

The LF-model is usually able to represent the glottal source derivative well. However, this model may not fit more irregular source pulse shapes correctly. These may occur due to natural glottal source effects (such as diplophony[1]) or due to analysis errors. These errors may be caused by inaccurate estimation of the glottal source signal, epoch error detection and problems when fitting the LF-model to the glottal source signal. However, in the experiments here, the LF-model appeared to generally fit the glottal source derivative signal well (by visual comparison of the estimated LF-model and glottal source derivative signals).

[1]The diplophony effect, in which two different pulses appear to occur within one glottal pulse cycle, is common with creaky voice

C. Synthesis

According to (4), voiced speech can be synthesized from the GSS parameters by convolving the periodic excitation with the vocal tract filter. In the frequency domain, this is given by:

$$Y(w) = P(w)G(w)V(w), \tag{8}$$

where $Y(w)$ is the FT of the synthetic speech. $G(w)$ is calculated using the excitation parameters, whereas the vocal tract filter is defined by the spectral parameters.

The GSS method can be used with different types of glottal source model. However, the model used for synthesis is expected to be the same as that used in the analysis.

IV. Perceptual Evaluation of GSS Method using the LF-model

An experiment was conducted to evaluate the GSS method using the LF-model. This method was compared against a method which used the spectral envelope of speech and the impulse train excitation to synthesize speech. This experiment permitted to test the hypothesis that the LF-model can produce better speech quality than using an impulse train and to confirm the parametric flexibility of the glottal source model for voice transformations. The GSS method permits a reliable comparison of these two excitation models because the spectral parameters used to synthesize speech can be calculated using the same spectral envelope estimation technique.

A. Recorded Speech

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, which were obtained from the following University of Stuttgart webpage:
http://www.ims.uni-stuttgart.de/phonetik/EGG/page10.htm.
The sentences contained only sonorant sounds, as the study concerned 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 incidence of errors in the estimated LF-parameters.

B. GSS Analysis

The fundamental frequency F_0 and the glottal epochs were estimated in the first stage of the GSS method, since they were used to estimate the glottal source derivative signal pitch-synchronously. The glottal epoch parameter corresponds to the maximal amplitude peak of the glottal flow derivative cycle, so it was also used as an estimate of the instant of maximum excitation of the LF-model, t_e. The F_0 and the epoch parameters were estimated using the F_0 and epoch detectors [30], [31] of the ESPS tools. F_0 values were calculated using the *get_f0* function, while the epochs were calculated using the *epochs* function and the estimated F_0 values. In this way, the extracted epochs were consistent with the F_0 values, that is, epochs were only estimated for voiced speech ($F_0 > 0$).

The inverse filtering technique with pre-emphasis was used for estimation of the glottal source derivative signal, $v(t)$,

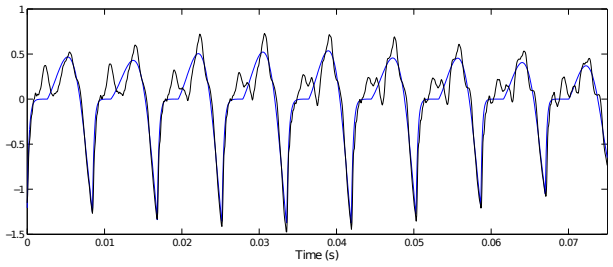

Fig. 4. Example of several LF-model cycles (in blue) obtained from the fitting technique for a segment of the glottal source derivative signal (in black).

because this is a straightforward and popular method for estimation of the glottal source signal. The coefficients of the inverse filter were calculated pitch-synchronously (analysis window centered at the glottal epoch i) from the pre-emphasized speech signal (α=0.97), using the autocorrelation method (order 18) and a Hanning window with duration equal to twice the fundamental period or a minimum of 20 ms long. Then, the *derivative of the glottal volume velocity* (DGVV), $v^i(t)$, was estimated by inverse filtering the short-time signal $x^i(t)$ sampled at 16 kHz.

Initial estimates of the LF-model parameters, with the exception of t_e, were obtained by performing direct measurements on the estimated $v^i(t)$, as described in [15]. This short-time signal was one period long and delimited by two consecutive glottal epochs, which were indexed as $i-1$ and i, respectively.

The E_e parameter was directly estimated from the residual signal as the absolute value of the amplitude of $v^i(t)$ at the glottal closure instant (glottal epoch $i-1$). In order to obtain more accurate estimates of the parameters t_o, t_p and T_a, a non-linear optimization algorithm was used that fitted each period of the LF-model signal to a low-pass filtered version of the DGVV signal. The initial estimates of these parameters were used in the iterative process. This is a common time-domain approach for estimating the LF-model parameters. However, our technique differed from standard methods in that we fitted the LF-model waveform for each pitch cycle starting at the instant of maximum excitation (epoch), t_e, instead of starting at the glottal opening instant.

The fitting method consisted of minimizing the mean-squared error between one period of the LF-model signal and the short-time signal, $v^i(t)$. In this work, the *Levenberg-Marquardt algorithm* [32] was used to solve this optimization problem (a *non-linear least squares* problem), which was implemented using the MATLAB function *lsqnonlin*. Figure 4 shows an example of the estimated LF-model signals for a segment of the DGVV signal. After the fitting procedure, t_e was calculated as $t_e = T_0 - t_o$ (t_e is equal to the duration from the glottal opening instant to the instant of maximum excitation).

The LF-parameter trajectories obtained for each utterance were smoothed using the median function. This operation reduces trajectory discontinuities caused by estimation errors. Figure 5 shows the trajectories of the time-domain parameters of the LF-model estimated for an utterance.

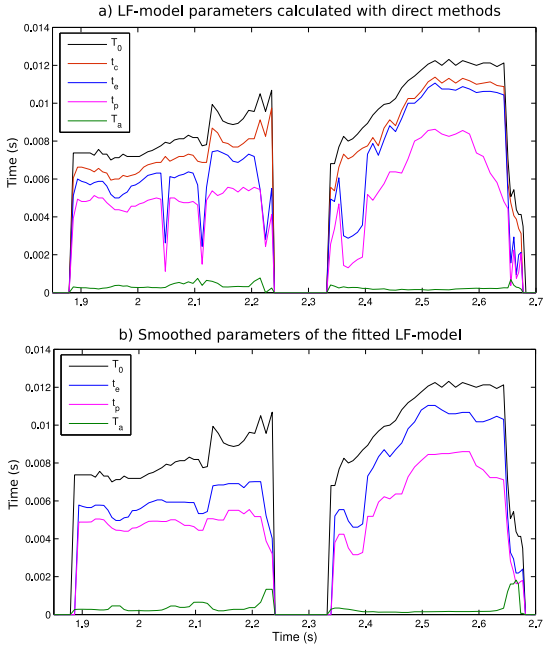

Fig. 5. Trajectories of the LF-model parameters estimated for a segment of a recorded utterance. This segment corresponds to the words "danger trail". a) Trajectories estimated based on amplitude measurements; b) Smoothed trajectories of the parameters estimated by the fitting method.

For the vocal tract filter estimation, the speech analysis was not performed pitch-synchronously. The speech signal was segmented at 5 ms frame rate into 40 ms long frames, $s^j(t)$, instead. These values were chosen to be the same as the default values of STRAIGHT analysis. However, it was necessary to map each speech frame (using its center point), $s^j(t)$, to the closest glottal epoch i, because the LF-model parameters were calculated for speech frames delimited by two contiguous glottal epochs. The set of LF-model parameter values associated with each selected epoch i was used to generate one period of the LF-model signal, $e^i_{LF}(t)$, starting at the glottal opening instant t_o. Then, each speech frame $s^j(t)$ was multiplied by a Hamming window and zero-padded to have the length of 1024 samples, for the short-term Fourier transform (SFT) analysis. The LF-model signal $e^i_{LF}(t)$ was also zero-padded to 1024 sample points. Next, the speech spectrum, $S^j(w)$, was divided by the amplitude spectrum of the LF-model signal, $\left|E^i_{LF}(w)\right|$. Finally, the STRAIGHT vocoder was used to calculate the spectral envelope of the resulting signal $V^j(w)$. For unvoiced speech, the spectral parameters were estimated by computing the spectral envelope of $S^j(w)$ using STRAIGHT.

C. Copy-synthesis

Each utterance was synthesized with the modal voice, by copy-synthesis, using the parameters estimated during GSS analysis. Each voiced frame i of the excitation signal was generated by concatenating two periods of the LF-model waveform. They started at t_e and had durations T^i_0 and T^{i+1}_0, respectively. The first LF-model cycle was generated from the glottal parameters estimated for the frame i: t^i_e, t^i_p, T^i_a, and E^i_e.

The t_e and t_p parameters of the second cycle were calculated under the assumption that the dimensionless parameters of the LF-model (OQ, SQ and RQ) were the same as the first cycle. That is, the glottal parameters are assumed to vary linearly with the fundamental period. For example, the t_p estimate for the second cycle was $\hat{t}_p = t_p^i T_0^{i+1}/T_0^i$. This linear approximation for the variation of certain LF-model parameters is considered to be good because the variation of the dimensionless parameters between contiguous frames is generally not significant. The T_a and E_e parameters of the second cycle were set equal to the values of the first cycle respectively, as they did not show significant variation with T_0 from the analysis measurements. The spectrum of the synthetic speech frame, $Y^i(w)$, was calculated by multiplying the amplitude spectrum of the LF-model waveform by the vocal tract transfer function, which is given by the spectral parameters (FFT coefficients). In this process, the LF-model spectrum was calculated by performing the 1024 point FFT, using a Hamming window. The speech waveform was generated by computing the inverse fast Fourier transform (IFFT) of $Y^i(w)$ and removing the Hamming window effect from the resulting signal (simply dividing the signal by the window). Finally, the speech frames were concatenated using a pitch-synchronous overlap-and-add (PSOLA) technique [33] with windows centered at the instants of maximum excitation, which is a standard technique for synthesizing speech pitch-synchronously. The overlap windows were asymmetric, in order to obtain perfect overlap-and-add (they add to one), as in the pitch-synchronous time-scaling method [34]. Each overlap window was obtained by concatenating the first half of a Hanning window, which had duration T_0^{i-1}, with the second half of a Hanning window, which had duration T_0^i.

The modal voice utterances were also synthesized using the impulse train instead of the LF-model. The speech synthesis method using the impulse train was similar to the GSS method using the LF-model, with the exception that the LF-model waveform was replaced by a delta pulse and the spectral parameters represented the spectral envelope of speech (computed by STRAIGHT) instead of the vocal tract filter estimated by the GSS method. The delta pulse was placed at the instant of maximum excitation t_e (approximately at the center of the excitation), and had amplitude equal to $\sqrt{T_0}$. The F_0 values were the same as those estimated during GSS analysis. Note that STRAIGHT uses F_0 in the computation of the spectral envelope of speech. For this reason, STRAIGHT was modified to use the F_0 values estimated during GSS analysis instead of its default F_0 estimation method TEMPO [35]. Although the STRAIGHT vocoder also uses FFT processing for generating the speech waveform, its technique is slightly different from the synthesis technique used in this experiment. In particular, the STRAIGHT method processes the phase of the delta pulse to add some randomness to the phase of the excitation signal. This phase processing is explained in more detail in Section V-A. Another difference is that the STRAIGHT method obtains the impulse response by calculating the complex cepstrum of the smooth spectral envelope. In other words, STRAIGHT synthesizes speech by passing the mixed excitation through the minimum-phase filter, which represents the spectral envelope of the speech signal. In contrast, the technique used in this experiment uses IFFT and PSOLA, as explained in the previous paragraph. However, several utterances synthesized from the impulse train in this experiment were compared against the same utterances synthesized by STRAIGHT (using the delta pulse as the voiced excitation without phase processing) and no significant differences in speech quality were perceived between the two methods. The advantage of using the same signal processing technique for producing speech using the impulse and the LF-model excitation signals is that it permitted a closer comparison between them.

D. Voice Quality Transformation

Five utterances spoken with modal voice were transformed into breathy and tense voices by modifying the mean values of the OQ, SQ, and RQ parameters of the LF-model. Our approach for glottal parameter transformation differs from standard approaches which usually calculate scale factors of the glottal parameters at the frame level instead of scale factors of the mean parameter values at the utterance level.

During synthesis, the F_0 and spectral parameters remained the same. In order to obtain the new trajectories, the voice quality parameters of the LF-model (OQ^i, SQ^i and RQ^i) were calculated for each frame i of an utterance, using the formulas given in Section II-B. Then, for each utterance, the variations of the mean values of the dimensionless 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 $\Delta\overline{OQ}_{breathy} = E[OQ_{breathy}] - E[OQ_{modal}]$, where $E[x]$ represents the mean computed over the total number of speech frames of an utterance.

The transformed trajectories of the LF-model parameters were obtained by multiplying the measurements of the glottal parameters of the modal voice by scale factors, so as to reproduce the target variation of the voice quality parameters (mean values of OQ, SQ, and RQ). The formulas used to calculate the scale factors were derived from the formulas of the voice quality parameters, given in Section II-B, and from the deltas of the mean values of the voice quality parameters. For example, for transforming the voice quality of the speech frame i, from modal to breathy, the scale factors are given by

$$k_{T_a}^i = 1 + \frac{\Delta\overline{OQ}_{breathy}}{RQ^i} \tag{9}$$

$$k_{t_p}^i = \frac{t_e^i}{t_p^i} \frac{\Delta\overline{SQ}_{breathy} + SQ^i}{1 + \Delta\overline{SQ}_{breathy} + SQ^i} \tag{10}$$

$$k_{t_e}^i = \frac{T_0^i}{t_e^i}(\Delta\overline{OQ}_{breathy} + OQ^i) - \frac{k_{T_a}^i T_a^i}{t_e^i} \tag{11}$$

The scale factors used to transform a modal voice into a tense voice were also calculated the same way as for breathy voice. Figure 6 shows the estimated trajectories of the LF-parameters for a segment of speech spoken with modal voice and the transformed trajectories for synthesizing that speech

segment with breathy voice. The main effect of scaling the LF-parameters using (9) to (11) is to change the mean component of the LF-parameter trajectories, while the dynamic component of the LF-parameter trajectories remains approximately unchanged. Thus, the local aspects of voice quality which are correlated with prosody are preserved, such as voice quality variations in stressed syllables. On the other hand, the mean values of the LF-model parameter trajectories which are expected to be related to the overall voice quality of the utterance are modified by the scaling operations. This voice transformation is based on the assumption that the characteristics related to the voice quality type are approximately the same throughout the utterance.

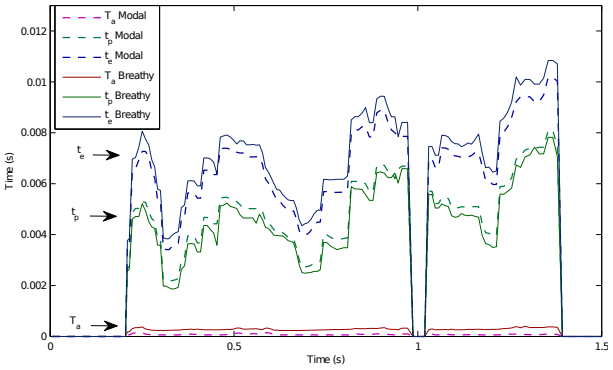

Fig. 6. Estimated trajectories of the LF-parameters for an utterance spoken with a modal voice and the respective transformed trajectories which were calculated to synthesize speech with a breathy voice.

E. Experiment under Laboratory Conditions

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

The listening test was divided into five parts. In the first, subjects were presented with 20 pairs of stimuli (10 utterances, randomly chosen and repeated twice with the order of the samples alternated). Each pair consisted of a sentence synthesized using the LF-model and the same sentence synthesized using the impulse train. For each pair, they had to select the version that sounded more natural. Each synthetic utterance had been previously scaled in amplitude to have the absolute value of the maximal amplitude equal to that of the recorded utterance.

The second and third parts of the test were similar to the first, but the recorded speech was compared to speech synthesized using the impulse train and speech synthesized using the LF-model, respectively.

In the fourth part, listeners were first presented with two pairs of recorded utterances in order to demonstrate the difference between modal and tense voices. This test consisted of 10 pairs, corresponding to 5 different sentences. Each pair contained a sentence synthesized with modal voice (by copy-synthesis) and the same sentence synthesized with the transformed trajectories of the LF-parameters which were calculated for the tense voice. Subjects had to select the speech sample that sounded most similar to the tense voice. Finally, the fifth part was similar to the fourth, with the difference that sentences synthesized with breathy voice were used instead of sentences synthesized with tense voice. In this part, listeners were asked to select the speech sample that sounded most similar to breathy voice.

F. Web Experiment

The same experiment was also conducted on the web, after the lab evaluation. Twelve listeners participated in the test, using headphones. The listening panel consisted of students and staff from the University of Edinburgh, including seven speech synthesis experts and ten native speakers. No payment was offered to the participants in this experiment.

For the web experiment, each synthesized utterance was multiplied by a scale factor so that the total speech power of the utterance was equal to the total power of the respective recorded utterance. This amplitude scaling was different from the one used in the lab test. The reason for this adjustment was to reduce the difference in loudness between the synthetic and the recorded utterances of each pair, which was found in the stimuli after the lab test had finished.

G. Results

The results obtained from the lab and web listening tests are shown in Figure 7. All the results are statistically significant with $p\text{-}value \leqslant 0.01$.

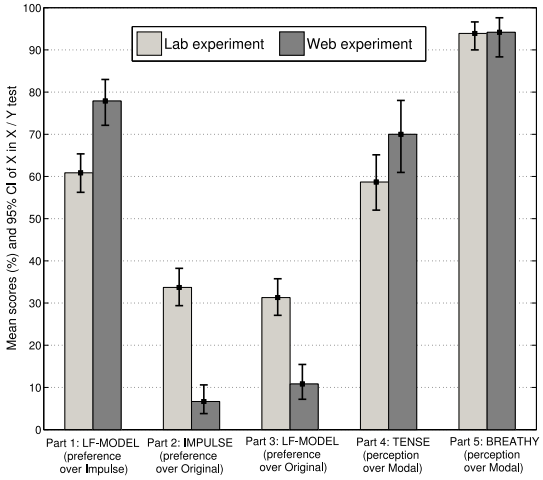

Fig. 7. Preference rates and 95% confidence intervals obtained for each part of the forced-choice test.

In general, speech synthesized using the LF-model sounded more natural than speech synthesized using the impulse train. The preference for the LF-model was significantly higher in the web test than in the lab evaluation. In the web test, the participation of speech synthesis experts and the power normalization of the speech samples are possible causes of the difference in results to the lab test. The results obtained in the two experiments were expected because the impulse train produces a buzzy speech quality, whereas that effect is attenuated by using the LF-model to represent the excitation. One important factor of the LF-model which could contribute

to this improvement in speech quality is that it is a mixed-phase signal, as described in Section II-C, whereas the phase spectrum of the delta pulse is constant.

Synthetic speech obtained higher scores than expected when compared to recorded speech, especially in the lab test. This result was unexpected, since the LF-model does not represent all the details of the true glottal source signal. For example, the LF-model cannot model certain voice effects such as aspiration noise, which is often perceived in voiced speech. 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 using the impulse excitation to the LF-model. An explanation might be that a small number of listeners (six out of ten) preferred speech spoken with a more buzzy voice quality over the natural voice of the speaker. Another explanation might be that the differences in loudness, which were observed between speech samples used in the lab test, influenced the perception of speech naturalness for some listeners. However, the differences between the results of the lab and web tests were not further investigated because in both experiments the results showed a significant improvement of the speech quality by using the LF-model instead of the impulse train. The unexpectedly good results obtained by synthetic speech in the comparisons against natural speech also indicate that the GSS synthesis method can produce high-quality speech by copy-synthesis, either using the impulse train or the LF-model.

Speech synthesized using the transformed LF-parameter trajectories to reproduce a breathy voice quality almost always sounded more breathy than speech synthesized using the estimated trajectories for modal voice. The results obtained for speech synthesized using the transformed LF-parameter trajectories to reproduce a tense voice quality were not as decisive as those for breathy voice. A possible reason to explain this result is that speech features other than the LF-parameters are important to correctly model this voice quality (e.g. the F_0 parameter).

V. Mixed Excitation Model for Synthesis using GSS Parameters

We have extended the GSS method to use a mixed excitation model, in which an acoustic glottal source model is combined with the noise excitation of the STRAIGHT vocoder.

A. Mixed Excitation Model of STRAIGHT

The mixed excitation used by the STRAIGHT vocoder (version V40_006b) is the sum of the periodic and noise components, which is given by:

$$X(w) = \sqrt{1/F_0} D(w)\Phi(w)W_p(w) + N(w)W_a(w), \quad (12)$$

where $D(w)$ is the FT of the delta pulse, $N(w)$ is the FT of white noise, and $\Phi(w)$ represents an all-pass filter function. Finally, $W_p(w)$ and $W_a(w)$ are the weighting functions of the periodic and noise components, respectively. The noise is modeled by a random sequence with zero mean and unit variance. For the impulse train to have the same energy as the noise signal, the pulse is multiplied by $\sqrt{1/F_0}$.

The all-pass filter $\Phi(w)$ is used to reduce the degradation in speech quality associated with the strong periodicity of the pulse train, $P(w)$. It introduces randomness in the phase of this signal by manipulating the group delay at higher frequencies [24], [27].

STRAIGHT measures the aperiodic component of the spectrum, $P_{AP}(w)$, using the ratio between the *lower* and *upper smoothed spectral envelopes* of the short-time signal as explained in [27]. The upper envelope, $|S_U(w)|^2$, is calculated from the speech spectrum by connecting *spectral peaks* and the lower envelope, $|S_L(w)|^2$, is calculated by connecting *spectral valleys*. A more detailed description of the aperiodicity measurements in STRAIGHT can also be found at [24]. Figure 8 shows an example of the aperiodicity spectrum calculated for a voiced speech frame.

During synthesis, the periodic and noise components of the excitation are added together to yield the mixed excitation, which is approximately flat. The delta pulse spectrum, $D(w)$, and the noise spectrum, $N(w)$, are also approximately flat and have the same energy. In this process, $W_p(w)$ and $W_a(w)$ are calculated from $P_{AP}(w)$. The plots a) and b) of Figure 9 show an example of the amplitude spectra of the two excitation components before and after the weighting, respectively.

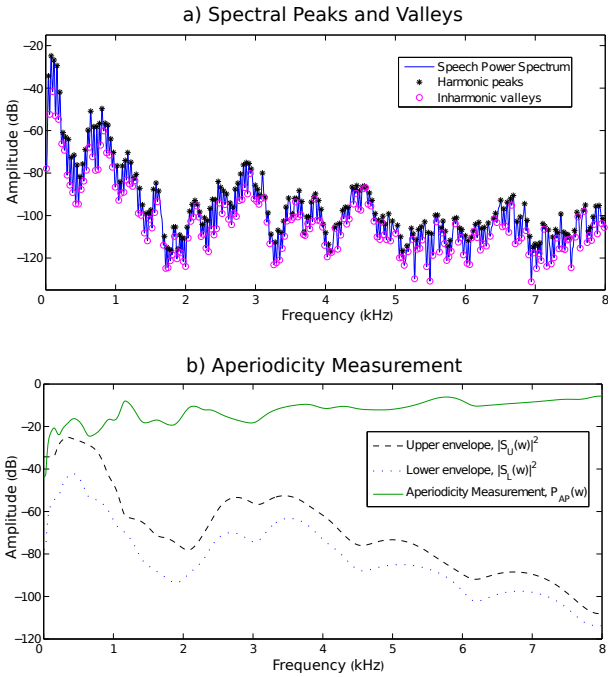

Fig. 8. Example of the aperiodicity spectrum. Top: amplitudes of the spectral peaks and valleys obtained from the amplitude spectrum of the speech signal, by STRAIGHT. Bottom: lower and upper spectral envelopes calculated by STRAIGHT and the resulting aperiodicity spectrum.

B. Mixed Excitation Model Adapted to GSS Parameters

The mixed excitation model of STRAIGHT was adapted to synthesize speech using the GSS parameters as follows:

$$G(w) = E(w)W_p(w) + K_n N(w)E_p(w)W_a(w), \quad (13)$$

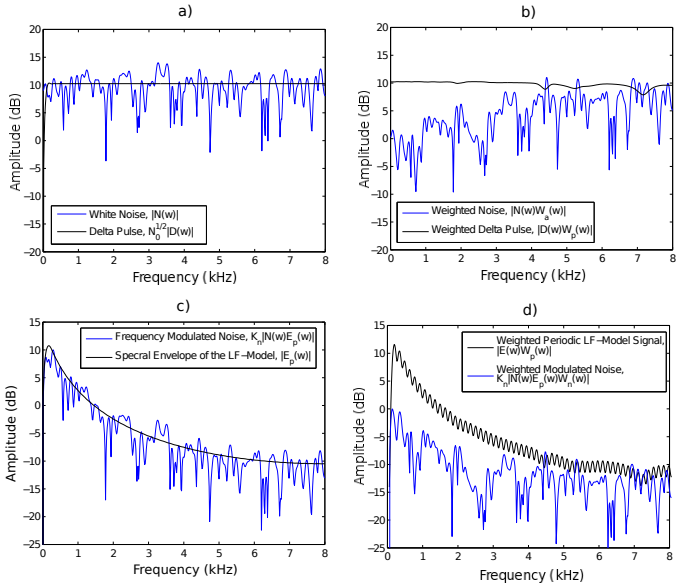

Fig. 9. Weighting effect on the mixed excitation components using the STRAIGHT aperiodicity measurements and two types of periodic signal. In a) and b), the periodic component is represented by the delta pulse. In c) and d) the mixed excitation is generated using the LF-model: c) amplitude spectrum of white noise shaped by the spectral envelope of the LF-model, and d) effect of weighting on the modulated noise and LF-model periodic signal.

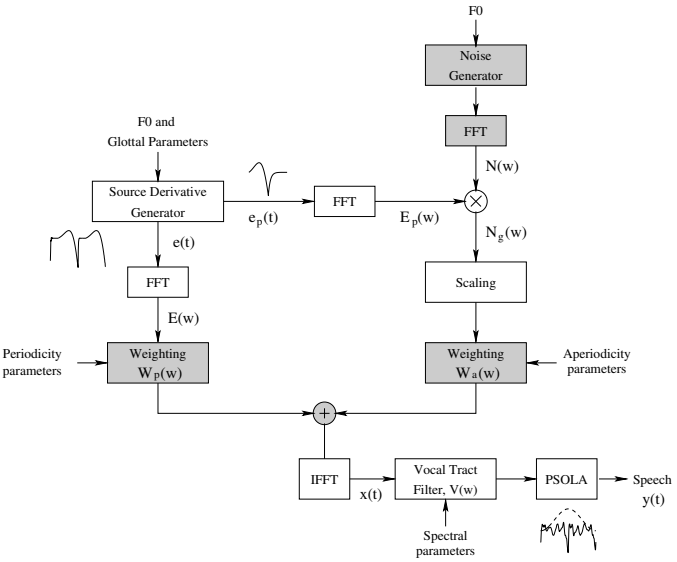

Fig. 10. Block diagram of the speech synthesis method using the parameters estimated by the GSS method. The glottal source derivative waveform represented in this figure was obtained using the LF-model, as an example.

where $E(w)$ and $N(w)$ represent the FT of the periodic component of the glottal source derivative and white noise, respectively. $E_p(w)$ represents the spectral envelope of the glottal signal $E(w)$ and K_n is a scale factor to normalize the energy of the noise relative to the source signal. Finally, $W_p(w)$ and $W_a(w)$ are the weighting functions of the periodic and aperiodic components of the excitation, respectively, which are computed by STRAIGHT. Note that this model could also be valid with other types of weighting functions with similar properties (mainly that produce a spectrally flat excitation). Figure 10 shows the flowchart of the speech synthesis method using this excitation model. In this figure the blocks which the STRAIGHT and GSS methods have in common between are shaded. They differ in the periodic component of the excitation (processed delta pulse for STRAIGHT and LF-model signal for GSS) and the synthesis filter (spectral envelope of STRAIGHT against vocal tract representation of GSS), as explained in the previous sections. They also differ in the signal processing technique for producing speech, because STRAIGHT uses different FFT processing and does not perform PSOLA. This difference was explained in Section IV-C and the reader can find more details about the STRAIGHT technique in [35] and [24].

Both $E(w)$ and $E_p(w)$ are calculated using the glottal parameters and F_0. In this work, the LF-model is used to represent the glottal source derivative signal.

There are two main differences between the method described in Figure 10 and that used in Section IV-C to synthesize voiced speech using the LF-model excitation. The first is that the signal $E_p(w)$ is used to weight the glottal source and noise components of the excitation. The other is that amplitude scaling of the noise is performed using the factor K_n. The remainder of this section expands on these two differences.

In contrast to the delta pulse, the glottal source signal $E(w)$ is not spectrally flat, and its energy does not depend on the fundamental period alone. In general, the shape of the glottal source waveform depends on all the glottal parameters, and its energy varies with these parameters too. For this reason, either the glottal source signal or the white noise have to be transformed so that the weighting operation is performed correctly for synthesizing speech using the GSS parameters. The solution proposed in this paper is to shape the spectral envelope of the source derivative on the white noise before the weighting operation. The spectral envelope of the source can be described as the impulse response $D(w)E_p(w)$, in which $E_p(w)$ is the transfer function of one period of the glottal source signal. This operation can be represented by $N_g(w) = |E_p(w)|N(w)$, where $N_g(w)$ is the *frequency modulated noise*. Figure 9 c) shows an example of $N_g(w)$, which was obtained using the LF-model signal as the *modulating signal*. Figure 9 d) shows the weighting effect on both the LF-model signal and the modulated noise. In this example, the amplitude spectrum of the LF-model component of the excitation, $E(w)$, has harmonics because it consists of two cycles of the LF-model waveform.

Besides adjusting the amplitude spectrum of the noise for matching its shape with that of the glottal source signal, an energy scaling operation also needs to be performed for the two signals to have the same power. The reason for this operation is that white noise $N(w)$ has power equal to one, whereas the delta pulse train $P(w)$ has power $1/T_0$. This scaling operation is similar to that applied to the periodic component of the excitation of STRAIGHT by the factor $\sqrt{1/F_0}$ in (12). However, here the scaling is performed on the noise signal $N_g(w)$ using the scale factor $K_n = 1/\sqrt{T_0}$, where $N_g(w)$ has the same duration as the periodic excitation. It is important that the amplitude scaling is performed on the

noise instead of the periodic component, in order to avoid the variation of amplitude parameters of the glottal source model. For example, if the LF-model waveform is scaled in amplitude so that it matches the unit power, then E_e is altered.

Finally, the synthetic speech frames are concatenated using the PSOLA technique with asymmetric windows that was described in Section IV-C.

VI. Perceptual Evaluation of GSS Method Using Mixed Excitation Model

A forced-choice perceptual experiment was conducted to test whether the mixed excitation model of the GSS method described in the previous section improves speech quality compared with excitation using the LF-model only. In this experiment, the STRAIGHT vocoder (Matlab version V40_006b) was used as a baseline.

A. GSS Analysis

The glottal source derivative was estimated from the speech signal using the IAIF method [8]. This technique was implemented to obtain more accurate estimates of the LF-model parameters than the inverse filtering with pre-emphasis used in Section IV. It consists of estimating the glottal source and the vocal tract components iteratively using the inverse filtering technique. The glottal flow is first modeled as a low-order all-pole signal (2 poles). This model is estimated by LPC analysis and its spectral effects are removed from the speech signal. Then, the resulting signal is used to obtain the initial estimate of the vocal tract using linear prediction. The glottal source waveform is also estimated by inverse filtering the speech signal using the estimated all-pole model. Next, a second estimate of the vocal tract and glottal source is performed similarly using a higher order parametric model of the glottal flow.

The epochs were estimated using the ESPS tools, similar to the previous experiment. However, in this experiment they were additionally verified and corrected by visual inspection of the waveforms of the residual and by comparison with the epochs detected on the electroglottograph (EGG) signal, which is available together with the recorded speech in the corpora used for this experiment. This process enabled the effect of epoch estimation errors to be avoided in the speech quality produced using GSS, as the main goal of this experiment was to compare the excitation models of the different methods. The LF-model parameters were estimated pitch-synchronously using a waveform fitting method as described in Section IV-B.

Finally, the spectral envelope of speech and aperiodicity parameters were obtained using the STRAIGHT vocoder. The F_0 values estimated using the ESPS tools were used in this spectral analysis, meaning the same F_0 values were used for both the GSS and STRAIGHT methods. The spectral envelope was used to calculate the vocal tract filter by GSS, whereas the aperiodicity parameters were used to represent the mixed excitation model. These parameters were also used to synthesize speech with the STRAIGHT method.

B. Copy-Synthesis Methods

Speech was synthesized using the GSS parameters and the mixed excitation model as described in Section V-B. The GSS parameters were also used to synthesize speech using the LF-model only, without performing any spectral weighting operation with the aperiodicity parameters.

For synthesis using STRAIGHT, the F0 contour was obtained using the ESPS tools, in order to obtain an F0 contour similar to that derived from the epochs in the GSS method. In addition, the manually corrected epochs estimated for the GSS method were used to automatically adjust the time intervals of the unvoiced parts of the F0 contour (when F0 is equal to zero), in order to obtain similar durations for the unvoiced/voiced regions of speech with the two methods. This post-processing of the F0 contour enabled differences in speech quality between the two methods caused by voiced/unvoiced classification errors to be avoided. A modified version of STRAIGHT was also used in the experiment to synthesize speech without using the aperiodicity parameters and the noise component of the mixed excitation. In this case, the excitation is only represented by the impulse train with phase processing.

C. Stimuli

Recorded speech from three speakers (two male and one female) was used in this experiment. One set of utterances consisted of the ten sentences spoken by a male speaker with a modal voice, used in Section IV. The other two sets consisted of ten utterances from the US English BDL (male) and US English SLT (female) speech corpora of the CMU ARCTIC speech database [36].

All utterances were also generated by copy-synthesis using the GSS and STRAIGHT methods with mixed excitation, labeled "GSS-MIX" and "STR-MIX" respectively. Speech was also synthesized using the GSS method with the voiced excitation represented by the LF-model only ("GSS-LF") and the STRAIGHT method with the simple impulse excitation ("STR-IMP"). These two versions were used to evaluate the effect of the noise component of the mixed excitation on the speech quality.

Pilot work using the BDL voice indicated that speech synthesized by STRAIGHT with simple excitation (STR-IMP) was preferred on average over the version with mixed excitation (STR-MIX). This result indicated that the relative energy of the noise component was over-estimated for this voice. For this reason, we tested different scaling factors of the noise component for the three voices (ranging from -10 dB to +5 dB). From our own perceptual evaluation of the speech quality we chose the attenuation by 5 dB for the male voices and no scaling for the female voice.

D. Pairwise Preference Experiment

The GSS-MIX method was compared against the equivalent with simpler excitation (GSS-LF), the STR-MIX method and recorded speech. Meanwhile, the STR-MIX method was also compared against its equivalent with simple excitation STR-IMP.

In total, the experiment consisted of 120 pairs of utterances. For each sentence, all stimuli were normalized in energy and amplitude.

The evaluation was conducted in a supervised perceptual testing lab at the University of Edinburgh, using a web interface and headphones. The approximate duration of the evaluation was 30 minutes.

Subjects were asked to listen to the pairs of stimuli and for each pair they had to select the version that sounded better (A or B). They were able to listen to the files in any order, as many times as they wished. 43 participants took part in the experiment, students and staff of the University of Edinburgh. They were all native speakers of UK English and were paid to participate.

E. Results and Discussion

Figure 11 shows the preference rates obtained from the pairwise comparisons between the different methods and the 95% confidence intervals. The preference rates are statistically significant ($p - value \leq 0.01$), with the exception of the comparison of STR-MIX versus STR-IMP for the male voices.

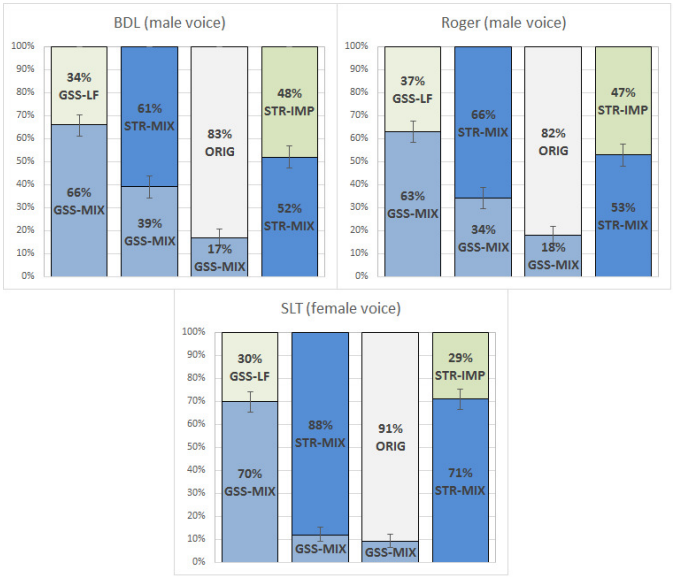

Fig. 11. Preference rates obtained by the different speech synthesis methods.

In contrast to Section IV, where the LF-model was significantly better than the impulse train, the mixed excitation model using the LF-model was outperformed by the mixed excitation of STRAIGHT in this experiment. Although the results of the two experiments cannot be directly compared because the synthesis methods are different, it is suspected that the signal processing technique of STRAIGHT produces significantly better speech quality than the synthesis method used in Section IV, which used the same FFT processing and PSOLA techniques as GSS. In particular, STRAIGHT performs phase processing on the impulse signal to reduce the "buzziness" of the impulse excitation.

The GSS-MIX method obtained higher preference rates than the GSS-LF method for all the voices, which is in accord with the assumption that the mixed excitation model using the LF-model improves the representation of the noise characteristics of voiced speech compared to the LF-model excitation. The mixed excitation model also significantly improved the average speech quality of STRAIGHT compared with the simple excitation for the female voice. However, this improvement was not statistically significant for the male voice. This may be explained by the fact that "buzziness" is expected to be higher for high-pitch voices (like the female SLT), due to stronger periodicity of the spectrum.

The results obtained with the GSS method were somewhat worse than with STRAIGHT. Possible factors to explain this are errors in the LF-model parameter estimation and the differences in signal processing between the two methods. LF-model parameter errors may produce significant variations in the vocal tract filter and shape of the LF-model waveform between contiguous frames, which causes speech distortion. PSOLA is used for reducing this effect by smoothing the waveform in the transition between contiguous frames, but it may not be sufficient in some cases, especially if the estimation errors result in very irregular shapes of the LF-model pulses. The performance of PSOLA is also expected to be worse for higher-pitch voices due to shorter OLA windows, which supports the worse results obtained for the female compared with the male voices.

It is interesting to compare the results of GSS with those obtained by this method in the evaluation of the SVLN method [14]. In [14], the LF-model parameters estimated by SVLN were also used by the GSS method to estimate the vocal tract spectrum and to synthesize speech. Thus, the effect of the LF-model parametrization was excluded from the comparison between the two techniques. The SVLN obtained slightly better results than GSS in terms of speech quality (the difference between the overall mean scores was around 0.5 in the scale of 1 to 5 for both male and female voices). This experiment also indicated that STRAIGHT was better than GSS (difference of about 0.5 on the mean scores) for male voices but they obtained comparable results for the female voices, on average. These results contradict our results. Though different voices and experimental conditions have been used, this contradiction may be explained by the effect of LF-model parameter estimation on the speech quality. We verified experimentally that irregular shapes of the LF-model waveform synthesized by GSS produced distortion in the synthetic speech for some utterances of the female voice, as shown in Figure 12. In this example, there is a rapid variation of the shape of the glottal waveform between contiguous pulses and the third and fourth pulses clearly have an irregular shape.

We expect that the weaker results of the GSS method for the female voice in our experiment are related to problems in the LF-parameter estimation for this voice. This assumption is in agreement with the poor performance of LPC analysis in formant estimation of high-pitch voices as indicated in [37], which could result in more inaccurate estimates of the glottal source signal by IAIF and consequently more errors in LF-model parameter estimation. In contrast, the SVLN method uses a phase minimization technique to estimate a

shape parameter of the LF-model, R_d, from the speech signal (the LF-model waveform is synthesized from this parameter only). In addition, this method does not require glottal source signal estimation and in [38] it is shown to be more robust than IAIF for estimation of R_d. This paper also shows that the estimation of the LF-parameter by IAIF is significantly more robust for the BDL (male) than the SLT (female) voices, supporting our view that the female voice is more affected by glottal parameter errors than the male voice in our experiment.

In balance, the results of GSS are encouraging because they indicate this method can produce very high-quality speech in some cases, especially for the male voices. For example, on average the quality of this method was at least comparable to the high-quality STRAIGHT vocoder for a significant part of the male utterances (preference for GSS close to 40%). The relatively high preference rates obtained by GSS against recorded speech (9% to 18%) are also an indicator that it can produce very high-quality speech for some utterances.

As future work the robustness of the LF-model parameter estimation in GSS will be more extensively evaluated for different male and female voices, and its effect on synthetic speech distortion will be investigated.

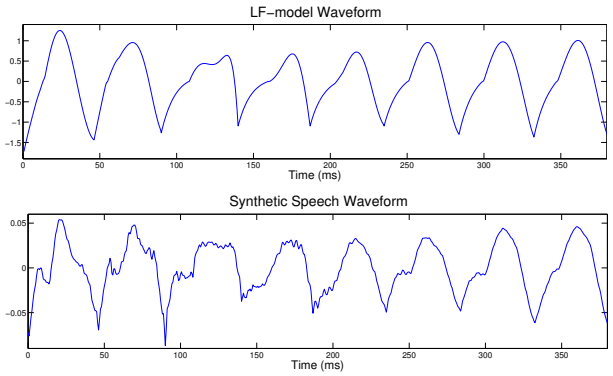

Fig. 12. Example of a segment of the LF-model waveform with distortion (top figure) which was synthesized using the GSS method for the female voice. The effect of this distortion on the speech segment synthesized using the GSS-LF method is shown at the bottom.

VII. Conclusion

This paper proposes the Glottal Spectral Separation (GSS) method for estimating the glottal source and vocal tract components of speech. GSS can be divided into three main parts. First, glottal source parameters are estimated from the speech signal, for example using an inverse filtering technique for calculating the glottal source signal and fitting this signal to an acoustic glottal source model. Next, the spectral effects of the glottal source signal are removed pitch-synchronously from the speech signal by dividing the amplitude spectrum of the speech by the amplitude spectrum of the glottal signal. Finally, the spectral envelope of the resulting signal is computed in order to estimate the vocal tract spectrum.

An experiment was conducted to evaluate the GSS method for copy-synthesis and voice quality transformation, where the LF-model of the derivative of glottal volume velocity was used. Results showed that the quality of speech synthesized by convolving the LF-model and the vocal tract filter estimated by GSS analysis was significantly better than convolving the traditional impulse train with the spectral envelope of speech. The explanation is that the LF-model signal contains more phase information than the impulse train, which reduces the buzziness of the synthetic speech. The experiment also showed that by using the GSS method and transforming the LF-model parameters estimated for utterances spoken with a modal voice, it was possible to modify the voice quality, namely to the target voice qualities breathy and tense. This is a great advantage of using GSS for analysis-synthesis compared with other methods which do not use an acoustic glottal source model, such as the STRAIGHT vocoder.

Finally, a method for synthesizing speech using GSS and a mixed excitation model was also proposed. This model consists of mixing the LF-model signal with a noise signal defined by the aperiodicity measurements extracted by the STRAIGHT vocoder. A second perceptual experiment was conducted to evaluate this model. Results showed that the improved excitation model produced better speech quality by copy-synthesis than the LF-model excitation, as expected. However, the STRAIGHT vocoder was preferred on average over the GSS method with mixed excitation in this experiment. The main reason to explain this result is that GSS depends on both the robustness of glottal source and spectral parameter estimation, whereas STRAIGHT requires only a robust spectral envelope estimation. The development of accurate and robust glottal source estimation techniques is an ongoing problem of research and future advances in this area could be used to further improve the GSS analysis-synthesis method. There is also room for further improvement in the mixed excitation model of the GSS method. For example, the spectral model of the noise excitation for voiced speech proposed in this paper cannot adequately represent important effects in the time-domain such as noise bursts or aspiration noise, which are important for speech quality and to better reproduce certain aspects of voice quality (e.g. breathiness). Currently, a mixed excitation model which combines the LF-model signal with a time-domain model of the noise component is being developed, in order to overcome this limitation.

References

[1] H. Kawahara, I. Masuda-Katsuse, and A. 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 sounds," *Speech Commun.*, vol. 27, no. 3–4, pp. 187–207, 1999.

[2] B. Doval and C. d'Alessandro, "The spectrum of glottal flow models," National Centre for Scientific Research, Stockholm, Sweden, Notes et Documents LIMSI-CNRS (Notes and Documents of the Laboratory for Mechanics and Engineering Sciences), 1999.

[3] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans Speech Audio Process*, vol. 9, no. 1, pp. 21–29, 2001.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1976.

[5] J. R. Deller, J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. New York, USA: Macmillan, 1993.

[6] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," in *Pattern Recognition and Artificial Intelligence*, C. H. Chen, Ed. Academic, 1976, pp. 374–388.

[7] D. Y. Wong, J. Markel, and J. A. H. Gray, "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans Acoust Speech Signal Process*, vol. 27, no. 4, pp. 350–355, Aug. 1979.
[8] P. Alku, E. Vilkman, and U. K. Laine, "Analysis of glottal waveform in different phonation types using the new IAIF method," in *Proc. of International Congress of Phonetic Sciences (ICPhS)*, France, 1991.
[9] P. Hedelin, "A glottal LPC-vocoder," in *Proc. of ICASSP*, USA, 1984.
[10] A. Krishnamurthy, "Glottal source estimation using a sum-of-exponentials model," *IEEE Trans Signal Process*, vol. 40, no. 3, pp. 682–686, Mar. 1992.
[11] Q. Fu and P. J. Murphy, "Robust glottal source estimation based on joint source-filter model optimization," *IEEE Trans Audio Speech Lang Process*, vol. 14, no. 2, pp. 492–501, Mar. 2006.
[12] B. Bozkurt, "Zeros of the z-transform (ZZT) representation and chirp group delay processing for the analysis of source and filter characteristics of speech signals," Ph.D. thesis, Faculté Polytechnique De Mons, Belgium, 2005.
[13] J. P. Cabral, S. Renals, K. Richmond, and J. Yamagishi, "Towards an Improved Modeling of the Glottal Source in Statistical Parametric Speech Synthesis," in *Proc. of the 6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Germany, August 2007, pp. 113–118.
[14] G. Degottex, P. Lachantin, A. Roebel, and X. Rodet, "Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis," *Speech Communication*, vol. 55, no. 2, pp. 278–294, Feb. 2013.
[15] J. P. Cabral, S. Renals, J. Yamagishi, and K. Richmond, "Glottal spectral separation for parametric speech synthesis," in *Proc. of INTERSPEECH*, Brisbane, Australia, September 2008.
[16] ——, "HMM-based speech synthesiser using the LF-model of the glottal source," in *Proc. of the ICASSP*, Prague, May 2011, pp. 4704–4707.
[17] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," Royal Institute of Technology, KTH, Stockholm, Sweden, STL-QPSR, 1985.
[18] D. Childers, "Glottal source modeling for voice conversion," *Speech Communication*, vol. 16, no. 2, pp. 127–138, 1995.
[19] C. Gobl, "A preliminary study of acoustic voice quality correlates," Royal Institute of Technology, KTH, Stockholm, STL-QPSR, 1989.
[20] G. Fant, "The LF-model revisited. Transformations and frequency domain analysis," Royal Institute of Technology, KTH, Stockholm, STL-QPSR, 1995.
[21] B. Doval and C. d'Alessandro, "Spectral correlates of glottal waveform models: an Analytic study," in *Proc. of ICASSP*, Munich, Germany, 1997, pp. 1295–1298.
[22] C. d'Alessandro, N. D'Alessandro, S. Le Beux, and B. Doval, "Comparing time-domain and spectral-domain voice source models for gesture controlled vocal instruments," in *Proc. of 5th International Conference on Voice Physiology and Biomechanics*, Tokyo, Japan, July 2006.
[23] B. Doval, C. d'Alessandro, and N. Henrich, "The voice source as a causal/anticausal linear filter," in *Proc. of ITRW (VOQUAL'03)*, Geneva, Switzerland, August 2003.
[24] J. P. Cabral, "HMM-based Speech Synthesis Using an Acoustic Glottal Source Model," Ph.D. Thesis, University of Edinburgh, archived at https://www.era.lib.ed.ac.uk/handle/1842/4877, UK, 2010.
[25] W. R. Gardner, "Modeling and Quantization Techniques for Speech Compression Systems," Ph.D. thesis, University of California, San Diego, USA, 1994.
[26] J. G. Proakis and D. G. Manolakis, *Digital Signal Processing: Principles, Algorithms, and Applications*. Upper Saddle River, Englewood Cliffs, NJ: Prentice-Hall, 1996.
[27] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. of MAVEBA*, Firenze, Italy, September 2001.
[28] A. McCree and T. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans Speech Audio Process*, vol. 3, no. 4, pp. 242–250, Jul. 1995.
[29] P. Alku and E. Vilkman, "Estimation of the glottal pulseform based on discrete all-pole modeling," in *Proc. of ICSLP*, Yokohama, Japan, September 1994.
[30] D. Talkin and J. Rowley, "Pitch-synchronous analysis and synthesis for TTS systems," in *Proc. of ESCA Workshop on Speech Synthesis*, Autrans, France, September 1990.
[31] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Elsevier Science, 1995, pp. 495–518.
[32] D. Marquardt, "An algorithm for least-squares estimation of nonlinear parameters," *J. Soc. Ind. Appl. Math.*, vol. 11, pp. 431–441, 1963.
[33] H. Valbret, E. Moulines, and J. P. Tubach, "Voice transformation using PSOLA technique," *Speech Commun.*, vol. 11, no. 2–3, pp. 175–187, 1992.
[34] J. P. Cabral and L. C. Oliveira, "Pitch-synchronous time-scaling for prosodic and voice quality transformations," in *Proc. of INTERSPEECH*, Lisbon, Portugal, September 2005, pp. 1137–1140.
[35] H. Kawahara, "STRAIGHT - TEMPO: A universal tool to manipulate linguistic and para-linguistic speech information," in *Proc. of IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Florida, USA, October 1997.
[36] J. Kominek and A. Black, "The CMU Arctic speech databases," in *Proc. of 5th ISCA Speech Synthesis Workshop (SSW5)*, Pittsburgh, USA, 2004.
[37] P. Alku, "An automatic method to estimate the time-based parameters of the glottal pulseform," in *Proc. of ICASSP*, Orlando, USA, March 1992.
[38] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans Audio Speech Lang Process*, vol. 19, no. 5, pp. 1080–1090, Jul. 2011.

João P. Cabral is a postdoctoral research fellow at Trinity College Dublin, in the School of Computer Science and Statistics, as part of the Centre for Global Intelligent Content (CNGL). He received B.Sc. and M.Sc. degrees from Instituto Superior Técnico (IST), Lisbon, Portugal, in Electrical and Computer Engineering, in 2003 and 2006 respectively. He spent the final year of his B.Sc. at the Royal Institute of Technology (KTH), Sweden, under the programme Socrates-Erasmus, where he started working in speech signal processing funded by the Department of Signals, Sensors and Systems. His M.Sc. thesis was also in this area ("Transforming Prosody and Voice Quality to Generate Emotions in Speech"). He was awarded a Ph.D. degree in Computer Science and Informatics from the University of Edinburgh, U.K., in 2010, funded by a European Commission Marie Curie Fellowship. His Ph.D. thesis contributed with the novel integration of an acoustic glottal source model in HMM-based speech synthesis, for improvement of speech quality and control over voice characteristics. Before joining Trinity College Dublin in 2013, he also worked as a postdoctoral research fellow at the University College Dublin, as part of CNGL, from 2010. His main areas of expertise are text-to-speech synthesis, glottal source modelling and voice transformation. However, his current research interests also include machine learning, automatic speech recognition, spoken dialogue systems and Computer-Assisted Language Learning (CALL).

Korin Richmond has been involved with human language and speech technology since 1991. This began with an M.A. degree at Edinburgh University, reading Linguistics and Russian (1991-1995). He was subsequently awarded an M.Sc. degree in Cognitive Science and Natural Language Processing from Edinburgh University in 1997, and a Ph.D. degree at the Centre for Speech Technology Research (CSTR) in 2002. This Ph.D. thesis ("Estimating Articulatory Parameters from the Acoustic Speech Signal"), applied a flexible machine-learning framework to corpora of acoustic-articulatory data, giving an inversion mapping method that surpasses all other methods to date.

As a research fellow at CSTR for twelve years, his research has broadened to multiple areas, though often with emphasis on exploiting articulation, including: statistical parametric synthesis (e.g. Researcher Co-Investigator of EPSRC-funded "ProbTTS" project); unit selection synthesis (e.g. implemented the "MULTISYN" module for the FESTIVAL 2.0 TTS system); and lexicography (e.g. jointly produced "COMBILEX", an advanced multi-accent lexicon, licensed by leading companies and universities worldwide). He has also contributed as a core developer of CSTR/CMU's Festival and Edinburgh Speech Tools C/C++ library since 2002. Dr. Richmond's current work aims to develop ultrasound as a tool for child speech therapy.

Dr. Richmond is a member of ISCA and IEEE, and serves on the Speech and Language Processing Technical Committee of the IEEE Signal Processing Society.

Junichi Yamagishi is a senior research fellow and holds an EPSRC Career Acceleration Fellowship in the Centre for Speech Technology Research (CSTR) at the University of Edinburgh. He is also an associate professor of National Institute of Informatics (NII) in Japan. He was awarded a Ph.D. by Tokyo Institute of Technology in 2006 for a thesis that pioneered speaker-adaptive speech synthesis and was awarded the Tejima Prize as the best Ph.D. thesis of Tokyo Institute of Technology in 2007. Since 2006, he has been in CSTR and has authored and co-authored about 100 refereed papers in international journals and conferences. His work has led directly to three large-scale EC FP7 projects and two collaborations based around clinical applications of this technology. A recent coauthored paper was awarded the 2010 IEEE Signal Processing Society Best Student Paper Award. He was awarded the Itakura Prize (Innovative Young Researchers Prize) from the Acoustic Society of Japan for his achievements in adaptive speech synthesis and the 2012 Kiyasu Special Industrial Achievement Award from the Information Processing Society of Japan. In 2012 he was an area chair for the Interspeech conference and elected to membership of the IEEE Signal Processing Society Speech & Language Technical Committee. He is an external member of the Euan MacDonald Centre for Motor Neurone Disease Research.

Steve Renals is Professor of Speech Technology at the University of Edinburgh, where he is the director of the Institute of Language, Cognition, and Communication (ILCC) in the School of Informatics. He received a BSc (1986) from the University of Sheffield and an MSc (1987) and PhD (1991) from the University of Edinburgh. He has held teaching and research positions at the Universities of Cambridge and Sheffield, and at the International Computer Science Institute. He has over 200 publications in speech technology and spoken language processing and has coordinated a number of large collaborative projects. He is an IEEE fellow, senior area editor of the IEEE/ACM Transactions on Audio, Speech, and Language Processing, and a member of the ISCA Advisory Council.