Glottal Flow Synthesis for Whisper-to-Speech Conversion

Olivier Perrotin, Ian V. Mcloughlin

Glottal Flow Synthesis for Whisper-to-Speech Conversion

Olivier Perrotin, *Member, IEEE* and Ian V. McLoughlin, *Senior member, IEEE*

***Abstract*—Whisper-to-speech conversion is motivated by laryngeal disorders, in which malfunction of the vocal folds leads to loss of voicing. Many patients with laryngeal disorders can still produce functional whispers, since these are characterised by the absence of vocal fold vibration. Whispers therefore constitute a common ground for speech rehabilitation across many kinds of laryngeal disorder. Whisper-to-speech conversion involves recreating natural-sounding speech from recorded whispers, and is a non-invasive and non-surgical rehabilitation that can maintain a natural method of speaking, unlike the existing methods of rehabilitation. This paper proposes a new rule-based method for whisper-to-speech conversion that replaces the noisy whisper sound source with a synthesised speech-like harmonic source, while maintaining the vocal tract component unaltered. In particular, a novel glottal source generator is developed in which whisper information is used to parameterise the excitation through a high-quality glottis model. Evaluation of the system against the standard pulse train excitation method reveals significantly improved performance. Since our method is glottis-based, it is potentially compatible with the many existing vocal tract component adaptation systems.**

***Index Terms*—Whisper-to-speech conversion, Speech synthesis, Vocal source excitation, Glottal Flow Model, Laryngectomy.**

I. Introduction

VOICE production relies on sound excitation generated in the glottis being transformed by the time-varying shape of the vocal tract (VT) to articulate phonemes. Excitation either comes from vocal fold vibration (phonated or voiced) or from turbulent airflow through an open glottis (unvoiced). The former situation creates sounds that are harmonic while the latter leads to noisy/breathy sounds. Speech consists of phonated vowels interspersed with phonated and unphonated consonants, whereas true whispers are unphonated [1].

Among many disabilities affecting speech production, laryngeal disorders describe malfunction of the vocal folds that affect the quality of phonation. This ranges from hoarse voices, e.g., cases of laryngitis or presence of nodules, to total loss of phonation, for instance following laryngectomy. Recent studies show that laryngeal disorders affect a large number of people, e.g., 7.6% reported laryngeal issues in 2012 in the US, yet few of those are diagnosed and treated appropriately [2]. Current rehabilitative solutions for voice restoration involve the use of alternate vibration sources. The most common is tracheoesophageal speech, which requires surgery to re-direct airflow from the trachea to the oesophagus, using the pharyngo-oesophageal sphincter muscles as a vibration source [3]. The need for surgery and intensive training to master speaking – many patients are unable to produce intelligible speech after surgery [3] – motivates the development of alternative methods, particularly those that are non-invasive and non-surgical. The electrolarynx is an external vibrator that, when held against the neck or placed into the mouth [4], excites the vocal tract [5]. Although less invasive than tracheoesophageal speech, electrolarynx speech has been reported less intelligible and results in a strongly mechanical voice [6].

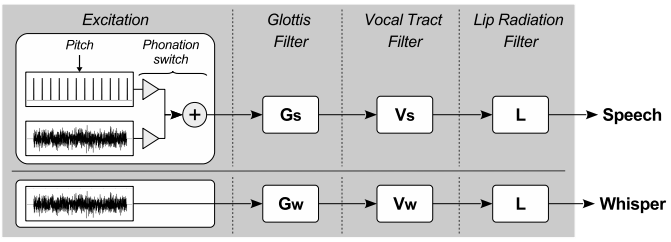

Fig. 1. Linear source filter models for speech (top) and whisper (bottom).

With advances in speech technology, research has recently focused on computational transformation of impaired-to-normal speech. Moreover, since unphonated speech is not affected by loss of laryngeal excitation, whispers are seen as a valid starting point for speech reconstruction – effectively reducing the task to whisper-to-speech conversion (WSC). This paradigm relies on linear models of speech and whisper production [7], shown in Figure 1, featuring an excitation followed by a series of linear filters to model the glottis, the VT and lip radiation, respectively. The glottis filter shapes the excitation into a glottal waveform [8] during phonation, or adds colouration to the noise excitation in the case of whispers, mostly attenuating high frequencies. The glottal signal then passes through the VT, which contributes zeros and resonances (formants) that allow the perception of vowels and consonants. Finally, the radiation filter models lip radiation by a simple derivative component. Distinction between speech and whisper mainly relies on the harmonic or noisy nature of the glottal source, yet differences have been observed in the VT [9], [10].

WSC research to date tends to reduce the problem to a source-filter model where the source is the excitation component, and the filter is a combination of the glottis, vocal tract and lip radiation filters. During conversion, the whisper

O. Perrotin is with Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, F-38000 Grenoble, France, e-mail: olivier.perrotin@gipsa-lab.grenoble-inp.fr

I.V. McLoughlin is with the University of Science and Technology of China, Hefei, P.R. China, e-mail: ivm@ustc.edu.cn

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes audio examples of whisper-to-speech conversion used for the method evaluation. Contact olivier.perrotin@gipsa-lab.grenoble-inp.fr for further questions about this work.

Manuscript received July 23, 2019; accepted January 28, 2020.

TABLE I
MACHINE LEARNING METHODS FOR WSC. TOP AND BOTTOM TABLES SHOW THE PREDICTION AND SYNTHESIS METHODS, RESPECTIVELY

	Toda *et al.* [11]	Tran *et al.* [12]	Li *et al.* [13]	McL. *et al.* [14]	Janke *et al.* [15]		Meenakshi *et al.* [16]
Speech feature	***Machine learning methods for mapping with whisper spectral envelope***						
Spectral envelope	GMM (MFCC)	GMM (MFCC)	RBM (full)	DNN (full)	DNN (MFCC)		Bi-LSTM (MFCC)
Aperiodicity	GMM	GMM					Bi-LSTM
f_0 + *V/UV*	GMM				ANN		
V/UV		DNN	SVM	SVM			Bi-LSTM
f_0		GMM	SVR	GMM			Bi-LSTM
	Synthesis methods						
Source	STRAIGHT	STRAIGHT	Pulse train	Pulse train	Pulse train	Unit-selection	STRAIGHT
Filter	MLSA	MLSA	Spectral envelope	Spectral envelope	MLSA		Spectral envelope

excitation is extracted, to be replaced by a synthetic voiced source, while the filter part is adapted to compensate for VT differences between whisper and speech. The main challenges in WSC are (i) generating a high-quality phonated source; (ii) predicting phonated source parameters such as pitch contour and voiced/unvoiced decisions; (iii) adapting the glottis and vocal tract filters. Most previous research has focused on steps (ii) and (iii), namely pitch prediction [1], [11]–[22] or vocal tract adaptation through filter modification [11]–[14], [16], [21]–[25] while using simple pulse trains for the replacement phonated excitation. Nevertheless, we strongly believe that overall reconstruction quality is severely hampered by lack of naturalness in the excitation source. Hence this paper presents a complementary approach by focusing on step (i): a new method of phonated source synthesis, focusing on timbre naturalness. It proposes a new excitation source driven by natural whisper information, along with a spectral glottis model, to produce high quality phonated excitation.

After reviewing WSC methods, the remainder of this paper details the proposed whisper decomposition process, excitation synthesis and glottis filter design in Sections II to IV. Evaluation in Section V is followed by a conclusion in Section VI.

A. Whisper-to-speech systems

As mentioned, existing WSC research tends to focus on pitch prediction and VT adaptation. There are two approaches;

1) Rule-based: Standard frameworks for rule-based WSC consist of: 1) decomposing whisper, using linear prediction (LP) to estimate a linear filter that gives a parametric representation of the speech spectral envelope [26], then applying inverse filtering to get the excitation; 2) applying rules to modify filter parameters, and predict the replacement phonated excitation source parameters; 3) generating speech by filtering the new excitation source with the modified filter [1], [20]–[25]. Filter modification starts with formant extraction from the poles of the LP coefficients [1], [20], [23]–[25] or from line spectral pairs [21], [22]. Then formants can be shifted to account for the formant position difference between speech and whisper [1], [21], [24], [25]; enhanced, i.e., make them more prominent [21]–[25]; and their frequency trajectories are sometimes smoothed as the noisy nature of whispers leads to higher variability than in speech [1], [21], [23]–[25]. The simplest source for replacement excitation is a pulse train that is fully harmonic [20]. To add more variability, the Mixed Excitation Linear Prediction (MELP) vocoder [27] mixes pulse train and noise spectra on different frequency bands. The mixing coefficient per frequency band is often called an aperiodicity component. Morris et al. used MELP with a fixed aperiodicity component, where frequency bands below and above 3 kHz were pulse train harmonics and noise, respectively [21]. An alternative is using codebooks of excitation with Code-Excited Linear Prediction (CELP) [23]–[25]. The same authors later presented a different approach using sine wave synthesis [1]. To control these sources, only one study predicted voiced/unvoiced decisions, using phoneme classification [25]. Several approaches predicted pitch from whisper intensity [21], or formant frequencies and bandwidths [1], [22]. The rules used for such prediction were fully empirical, yet more recent work attempted to derive pitch from whisper spectral information [20] by learning rules from a pitch-annotated database of whispers, i.e., judges were asked to attribute a perceived pitch to whispers in the database.

2) Machine learning based: Using machine learning to predict elements of the reconstructed speech, systems are trained from parallel databases of natural speech and whispers. On the one hand, speech is decomposed into various features (e.g., spectral envelope, aperiodic components, voiced/unvoiced (V/UV) decisions, pitch (f_0)). On the other hand, whispers are described by spectral envelope only. For each speech feature, a machine learning model is trained to map to the whisper spectral envelope. Various implementations follow this framework, summarised in Table I. Speech and whisper spectral envelopes can be mapped via Restricted Boltzmann Array (RBM) [13], or converted to Mel Frequency Cepstrum Coefficients (MFCC) for regression with Gaussian Mixture Models (GMM) [11], [12]. Deep Neural Networks (DNN) [14], [15] and Bidirectional Long Short-Term Memory Networks (Bi-LSTM) [16] have also been used. The f_0 and V/UV decisions are sometimes combined (where $f_0 = 0$ means 'unvoiced') [11], [15], although performance improves when they are predicted separately using DNN [12], support vector machine (SVM), support vector regression (SVR) [13], or Bi-LSTM [16]. Finally, the STRAIGHT vocoder [28] has been used to generate mixed-excitation when aperiodicity components are available [11], [12], [16], but pulse trains are used when no aperiodicity components exist. Excitation is then either filtered by a Mel Log Spectrum Approximation (MLSA) filter if MFCC is used [11], [12], [15], or by a full spectral

TABLE II
PARTITIONING MODELS FOR SOURCE AND FILTER.

	Source	**Filter**
DSP-based	Excitation	Glottis + Vocal Tract + Lip
Voice-based	Excitation + Glottis + Lip	Vocal tract

envelope if available [13], [14], [16]. Alternatively, an end-to-end WSC system using Generative Adversarial Networks that does not use any parametrisation of speech and whisper has also been implemented [29]. Finally, recent research has proposed electrolarynx-to-speech [30] and oesophageal speech-to-normal speech conversion [31], using statistical methods [11].

In summary, both rule-based and machine learning-based methods focus mainly on whisper feature adaptation via filter modification, and on speech-related feature prediction (e.g., pitch contour generation or V/UV detection). Most use simple excitation sources such as alternating pulse trains and noise, leading to buzzy and robotic speech [13]–[15], [20], [21]. Only those which can predict an aperiodic component per frequency band with machine learning used STRAIGHT mixed-excitation [11], [12], [16].

B. Speech decomposition

All excitation sources used in current WSC systems and described above have a flat spectral envelope, and excite a filter that includes the glottis, vocal tract and lip contributions. For shorthand we call this type *DSP-based* since such a decomposition is widely used in audio and speech DSP. The 'Source' only includes an excitation, and all filters reside in the 'Filter' part. However, it is possible to group these linear model components differently. For example, *Voice-based* source-filter decomposition, shown in Table II, follows the natural production of speech by defining the 'Source' as a combination of excitation and glottis filter, and the 'Filter' as the vocal tract contribution. It is also common to group the lip radiation filter (a simple derivative) with the glottis filter to output the glottal flow derivative (GFD) instead of the glottal flow source. A time-domain parametric model of the GFD is often used, such as the widespread LF model [32]. Solutions to jointly estimate the glottal source and the vocal tract filter from speech have been proposed [33], [34]. Other methods first estimate a glottal model from the speech signal, deconvolve the model from the signal and apply linear prediction to estimate the vocal tract [35], [36]. An alternative is to make use of a natural glottal pulse library [37]–[39] or perform principal component analysis on the library [40]. In each case, the source consists of the temporal GFD waveforms.

The WSC methods of section I-A are DSP-based vocoding techniques. Yet voice-based decomposition using GFD seems more appealing for high quality WSC since the decoupling of glottis and VT filter increases the flexibility of source parameterisation without affecting the VT components. In particular, voice quality parameters such as tension of vocal force that are typical of phonated speech and essential for expressivity can be generated by modifying the glottis filter [41]. Furthermore, as we have stated, WSC is primarily an excitation+glottis filter reconstruction task rather than a VT filter modification task. Hence, we propose synthesising a speech glottal source model, including an excitation, glottis filter and lip radiation filter, directly from whisper parameters.

C. Proposed system

The proposed WSC system aims to replace the whisper excitation and glottis filter by generating a new voiced excitation, and creating a new glottis filter adapted for speech. To increase naturalness, both the excitation and glottis filter are dynamically generated from the whisper decomposition. The following sections describe, in turn, the whisper decomposition, replacement excitation and glottis filter syntheses.

A full diagram of the proposed WSC system is shown in Figure 2. This begins with whisper input at the top right. The GFM-IAIF method [42] (Glottal Flow Model-based Iterative Adaptive Inverse Filtering, explained in section II-B) extracts the VT and glottis filters to be used in the whisper decomposition process in the top left panel. Successive inverse filtering of the vocal tract and glottis provides the GFD and the excitation, respectively. The whisper excitation is input to the speech excitation synthesis block in the middle left panel (detailed in section III). Similarly, the whisper glottis filter and GFD drive the glottis filter synthesis in the middle right panel (described in section IV). The bottom panel shows the speech reconstruction process. The synthesised excitation passes through the synthesised glottis filter, providing a voiced GFD. In order to control the alternation of voiced and unvoiced phones in speech, the synthesised voiced GFD and the whisper unvoiced GFD are mixed as detailed in section IV-B. Finally, the whisper VT filter is applied to the synthesised GFD to obtain the reconstructed speech. Previous studies have shown the importance of VT modification [11], [21], [25] for WSC; however, *no VT adjustment* is used in this paper, so the method is compatible with existing WSC systems that modify the VT; all of our performance gains are achieved by improved glottal/excitation modelling alone. Analysis uses a sliding window of size W_{win} advancing by W_{shift} each frame with corresponding overlap-add at the output.

II. PROPOSED WHISPER DECOMPOSITION PROCESS

The challenge in whisper decomposition and speech reconstruction is to find coherent models for whisper and speech VT filters, and whisper and speech glottis filters. We first discuss models for speech and whispers, then present a common framework for their analysis-resynthesis.

A. Speech and whisper modelling

1) Speech production model: In a frequency domain glottal flow model [8], [43], the opening and closing phases of the vocal fold vibration are described by a resonance close to the fundamental frequency, called 'glottal formant' as an analogy to the VT resonances [8], and a high-frequency attenuation, respectively. The glottal formant is modelled by a second order resonant filter with frequency F_g and bandwidth B_g, that are directly correlated to vocal fold oscillation asymmetry and open duration relative to the fundamental period. F_g is

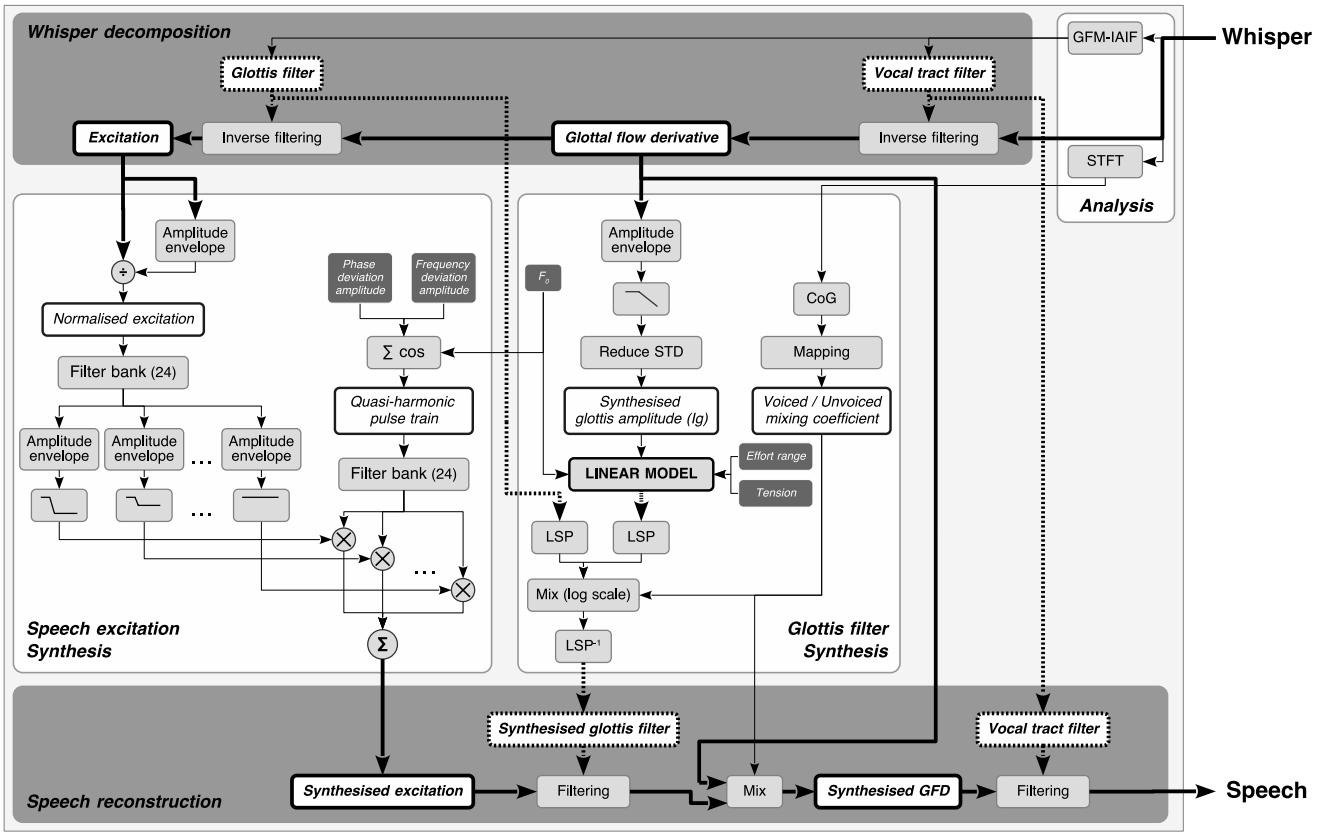

Fig. 2. Flow chart of the proposed WSC system. Thick plain arrows and thick-contoured boxes represent time-domain signals. Thick dashed arrows and dashed-contoured boxes represent filter frequency responses. Dark boxes with white text are the parameter inputs of the system. Grey boxes are functions.

usually located between the 1st and 2nd harmonic [8]. The high-frequency attenuation, called spectral tilt [8], is modelled by a first order low-pass filter with a cut-off frequency F_{st}, and is linked to the abruptness of vocal fold closure. The more tensed the voice, the quicker the folds close, increasing high-frequency amplitude through a positive shift of F_{st}, and vice versa. In summary, the glottis spectral envelope is modelled by a third order filter with a pair of conjugate poles $\{a, a^*\}$ (glottal formant) and a real pole b (spectral tilt):

$$G(z) = \{(1 - az^{-1})(1 - a^*z^{-1})(1 - bz^{-1})\}^{-1} \quad (1)$$

Varying section areas in the pharynx and oral cavity caused by the moving articulators (tongue, jaw, lips) introduce cascade resonances in the glottis spectrum [7]. Although the nasal cavity attenuates some frequencies, these are commonly approximated with additional resonances [44]. An all-pole model of N_v pairs of complex conjugate poles describes the VT:

$$V(z) = \left\{\prod_{i=1}^{N_v}(1 - c_i z^{-1})(1 - {c_i}^* z^{-1})\right\}^{-1} \quad (2)$$

Finally, the lip radiation is modelled with a simple derivative filter with a coefficient d close to 1 [7].

$$L(z) = 1 - dz^{-1} \quad (3)$$

In summary, the speech glottis filter encompasses a low-frequency resonance and a spectral tilt. Conversely, the VT response is globally flat across the frequency range, with generally higher frequency resonances than the glottal formant.

2) Whisper production model: To our knowledge, previous research did not study the whisper glottis and VT spectral shapes separately. Nevertheless, studies of the overall spectral envelope show that whispers present higher spectral balance and centre of gravity than speech [45]. This translates into an absence of low-frequency resonance, and reduced high-frequency tilt [46]–[48]. Moreover, whisper vocal tract formants have higher positions and bandwidths than speech [9], [10]. To ensure coherence between whisper and speech models, we assume that whispers follow the same speech glottis filter model, with extremely high glottal formant bandwidth and high spectral tilt cut-off frequency. Likewise, we will use a similar model for speech and whisper VT filters, differing in terms of pole angles and magnitudes.

B. Source-filter separation of speech and whisper

The glottis-based voice coding techniques presented in section I-B are able to accurately predict the glottis and VT filters from a speech signal [34]–[36]. Nevertheless, their use of a time-domain model of the glottal pulse makes them unsuitable for whisper decomposition. Conversely, other methods such as Iterative Adaptive Inverse Filtering (IAIF) [49] focus on estimating both glottis and VT filter frequency responses from the signal regardless of its harmonicity. A variant called Glottal Flow Model (GFM)-IAIF [42] has been recently developed

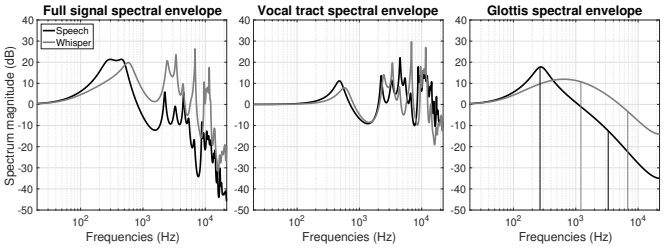

Fig. 3. Source-filter decomposition using GFM-IAIF on speech (black curves) and whisper (grey curves) for the vowel in "*is*". In the right panel the vertical lines mark the glottal formant (left) and spectral tilt cut-off frequency (right).

to ensure that the extracted glottis filter follows the model described in eqn. 1. Compared to IAIF, GFM-IAIF has been shown to more accurately extract glottal formant and spectral tilt features [42], key parameters for distinguishing whisper from speech glottis filters. Figure 3 shows an example of speech and whisper decomposition. From the left panel (spectral envelope), one can see that the whisper spectral envelope tends to be less tilted in high frequencies than for speech. The middle panel shows that whisper and speech VT envelopes appear approximately flat over the full spectrum, meaning that most of the spectral tilt has been removed. Similar formant repartition is observed between both signals, although the first formant in whispers is higher than for speech [9]. Most of the difference between speech and whisper spectral envelopes evidently lies in the glottis spectrum (right panel). The whisper glottal formant is higher in frequency and larger than in speech, and has less spectral tilt [46], [47]. To conclude, GFM-IAIF allows separate extraction of whisper and speech glottis and VT filters, permitting independent modification of each, necessary in the proposed system (see architecture in Fig. 2).

III. Synthesis of replacement excitation

Most previous WSC methods use a simple pulse train or a mixture of pulse train plus noise as excitation source for the phonated part of the signal [13]–[15], [20], [21]. However, the high regularity of the pulse train is unnatural, with a characteristic "buzziness". Yet it is known that preserving the speech excitation phase leads to a major reduction in buzziness [50]. Attempts to add irregularities to the pulse train were conducted in speech synthesis frameworks, either based on rules [28], [51], on natural excitation libraries [52] or by encompassing glottal pulse phase characteristics in statistical parametric synthesis [53]. In our application, we have at our disposal a whisper excitation input which already contains many natural variations. We therefore propose extracting some of those irregularities to increase the naturalness of the phonated excitation source.

1) Additive synthesis: Firstly, a pulse train is generated through additive synthesis, as a sum of N harmonics:

$$e_{\sqcup\!\sqcup}(t) = \sum_{i=1}^{N} \cos\left(2\pi \int_0^{T_{\text{tot}}} f_i(t)dt + \phi_i\right) \tag{4}$$

where T_{tot} is the duration of the signal, f_i and ϕ_i are the instantaneous frequency and phase of harmonic i, respectively. The number of harmonics N is chosen to be maximum, extending to the Nyquist frequency, since the subsequent glottis filter will naturally attenuate the high frequency harmonics.

In a perfect pulse train, each harmonic is a multiple of the fundamental frequency f_0 with zero phase, and all harmonics have the same amplitude. In order to add variability to the pulse train, random variations are added to frequency, phase, and amplitudes of harmonics. The whisper excitation does not contain harmonics. Hence, frequency and phase variations cannot be estimated from it and must be modelled using rules. However, harmonic amplitudes can be derived from the whisper excitation amplitude on frequency bands. The next two sections detail firstly the addition of phase and frequency variability, and secondly the addition of amplitude variability.

2) Phase and frequency variability: Random phase and frequency variation in the pulse train are based on rules:

$$\begin{cases} \phi_i &= \pi \quad [1 + A_\phi \mathcal{N}_i(0,\sigma)] \\ f_i(t) &= i f_0(t) \quad \left[1 + A_f \mathcal{N}_i(0,\sigma) u(i\overline{f_0} - F_{\text{vuv}})\right] \end{cases} \tag{5}$$

where $\overline{f_0}$ is the mean f_0 value over the total duration of the signal and u is the step function defined as:

$$\begin{cases} u(i\overline{f_0} - F_{\text{vuv}}) &= 1 \quad \text{if} \quad i\overline{f_0} \geq F_{\text{vuv}} \\ u(i\overline{f_0} - F_{\text{vuv}}) &= 0 \quad \text{if} \quad i\overline{f_0} < F_{\text{vuv}} \end{cases} \tag{6}$$

The first factors of each product of eqn. 5 are the expected values for a perfect pulse train. In particular, the speech glottal flow derivative presents negative valleys of maximum magnitude. To comply with this model, the phases of each harmonic are set to π so the pulses are negative. The second factors of the products are the added random variations. Raitio *et al.* used a uniform noise distribution to generate random phase signals that led to completely inharmonic sounds [50]. To keep the excitation periodic, the noise variations are centred by the use of zero-mean Gaussian noises $\mathcal{N}_i(0,\sigma)$ for each harmonic. We choose $\sigma = 0.3$ so that 99.9% of the noise values fall into the interval [-1,1]. The phase and frequency noises are weighted by the coefficients A_ϕ and A_f, respectively. Moreover, it has been shown that only the addition of random variations to high-frequency harmonics is perceived as the expected roughness, while inharmonicity in low frequencies results in the perception of individual partials [54]. Therefore, only harmonics above a frequency threshold are altered with the help of the step function u (eqn. 6). In a similar fashion to the Harmonic-plus-Noise Model [55], we denote this frequency threshold the voiced/unvoiced frequency F_{vuv}.

The three values A_f, A_ϕ and F_{vuv} along with the fundamental frequency f_0 are system calibration parameters. Increasing values of A_f, A_ϕ and a smaller value of F_{vuv} will reduce the harmonicity of the excitation and lead to rougher speech. A default of $A_\phi = 0.5$ is chosen for experiments. This allows the phase to approximately vary within a range of $\pi \pm \pi/2$, the limit above which the harmonics would be in phase opposition. A value of $A_f = 1\%$ is chosen for frequency variation, above the threshold of perception of inharmonicity [54]. Finally, a typical value of $F_{\text{vuv}} = 4\,\text{kHz}$, as reported in [55], is chosen in this study.

3) Amplitude variability: Due to the approximately logarithmic frequency resolution of human hearing over much of its range [54], we did not add amplitude variability to individual harmonics but to harmonic groups belonging to the same Mel-frequency band. Moreover, we aim to provide variation of amplitude over time, by using the sub-band temporal envelopes (STE) of the whisper excitation, calculated as follows. First, we are only interested in the relative amplitude variation between different frequency bands for each instant. Therefore, the overall amplitude of the whisper excitation for each frame is removed by normalising the signal by its RMS value (see top-left of panel "Speech excitation synthesis" in Fig. 2). Then, both whisper excitation e_w and pulse train excitation $e_{\sqcup\!\sqcup}$ are passed through a Mel filter bank. We use 24 triangular filters with centre frequencies $\{F_{\text{mel}i}\}_{i=1:24}$ equally spaced from 0 to $F_s/2$ on the Mel scale with 50% bandwidth overlap. Finally, the time-dependent amplitude envelopes of each of the 24 narrowband whisper excitation signals (i.e., the STE) are derived from the RMS of a sliding time window of size $F_s/\{F_{\text{mel}i}\}_{i=1:24}$ on each signal, as shown on the "Speech excitation synthesis" panel in Figure 2.

The extracted STE show high variability for each frequency band. Nevertheless, as stated above, introducing high variability on the low-frequency harmonics of the pulse train severely degrades signal harmonicity. As a consequence, the high variations of STE are filtered with different gains depending on the frequency band by means of high-shelf filters[1] [56] shown in the left part of Fig. 4. We empirically define the filter transition at 5 Hz with the gain G_{BF} of frequencies below this set to one, i.e., the low variations of the amplitude envelopes remain unchanged for all bands. The gains $\{G_{\text{HF}i}\}_{i=1:24}$ of amplitude variation frequencies above 5 Hz increase with frequency band. In other words, high-frequency amplitude variations are removed on low-frequency bands, and left unchanged on the highest frequency bands. The gains $\{G_{\text{HF}i}\}_{i=1:24}$ are derived from the frequency response of a first order Butterworth high-pass filter with cut-off frequency F_{vuv}, as shown in the right part of Fig. 4, and are defined as the amplitude of the frequency response of the filter evaluated at the centre frequencies of each band $\{F_{\text{mel}i}\}_{i=1:24}$.

Finally, each narrow-band pulse train excitation signal is multiplied by the corresponding filtered STE. The modulated narrowband pulse train excitations are then summed to give the final synthesised speech excitation.

IV. High quality glottal flow synthesis

Glottal flow synthesis comprises two parts. First, as explained in section II-A2, the whisper glottis filter tends to show higher glottal formant frequency and bandwidth and spectral tilt cut-off frequency than in speech. Consequently, a speech glottis filter is synthesised to replace the whisper glottis filter. Second, a voiced/unvoiced alternation is introduced according to the sequence of vowels and consonants (unlike the input which is continuously unvoiced). Previous studies showed few spectral differences between spoken and whispered unvoiced consonants [47], [48]. Hence, the whisper glottal flow derivative is used for unvoiced consonants while voiced synthetic speech serves for the vowels and voiced consonants. The process of mixing between the two is described at the end of this section.

[1]A high shelf filter equally attenuates or boosts all frequencies above the cut-off frequency and leaves unaffected the frequencies that are below.

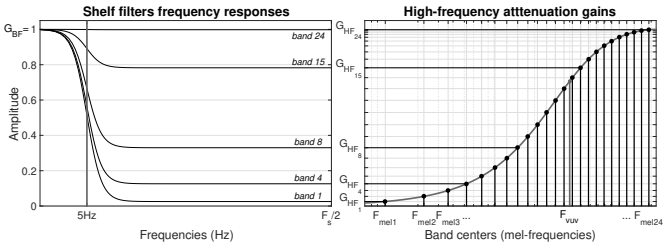

Fig. 4. Left: example of high-shelf filters for frequency bands (1,4,8,15,24). The low-frequency gain G_{BF} equals 1 in all cases. The high-frequency attenuation gains $\{G_{\text{HF}i}\}_{i=1:24}$ are calculated for each band. Right: calculation of the high-frequency gains (black dots) evaluated from the frequency response of a first order high pass filter with cut-off frequency F_{vuv} (grey curve) at the band centre frequencies. Both abscissas use a log-frequency scale.

A. Glottis filter synthesis

1) Linear model of source production: Most existing glottis models are defined in the time-domain [51], [57] and are therefore incompatible with our excitation plus filter decomposition of speech and whisper paradigm. Feugère *et al.* introduced the linear model (LM) of source production [58] that models the glottis as a linear filter following eq. 1 and which we adopt here. The model is driven by three parameters: vocal effort and tension levels $E \in [0,1]$ and $T \in [0,1]$, respectively, and the fundamental frequency f_0. T drives the timbre differences between a relaxed (T=0) and a tensed voice (T=1), by increasing the frequency F_g and bandwidth B_g of the glottal formant. E drives the timbre differences between a soft voice (E=0) and a loud voice (E=1) by adding high frequencies in the spectrum in increasing the frequencies of the glottal formant F_g and the spectral tilt F_{st}. Relations between E, T, F_g, B_g, and F_{st} can be found in [58].

2) Glottis amplitude: In their LM implementation, Feugère *et al.* had direct user input control on vocal effort parameter E. In this paper, E is derived from the whisper GFD through an intensity parameter I_g. To assign a specific vocal effort range to the signal w.r.t. the intensity range requires a mapping between intensity and vocal effort:

$$E(t) = (E_{max} - E_{min}) I_g(t) / \max(I_g) + E_{min} \quad (7)$$

From this mapping, E ranges from E_{min} to E_{max} when I_g evolves from 0 to its maximum. For instance, setting $[E_{min}, E_{max}] = [0.5, 1]$ synthesises a loud voice, while setting $[E_{min}, E_{max}] = [0, 0.5]$ synthesises a softer voice from the same whisper intensity input. To summarise, E_{min}, E_{max}, T and f_0 are inputs of the system, while I_g is extracted from the whisper, as shown on the middle panel of Fig. 2.

In particular, synthesised glottis amplitude I_g is computed from the RMS value of the whisper GFD on each analysis frame, and is processed before driving the glottis filter because

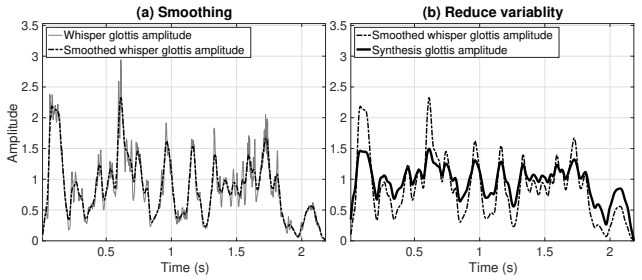

Fig. 5. Glottis amplitude synthesis examples to illustrate (a) smoothing of the whisper glottis amplitude (grey line) into the smoothed whisper glottis amplitude (black dotted line). (b) a reduction in variability of the smoothed whisper glottis amplitude (black dotted line) into the synthesised glottis amplitude (black line). All intensities are normalised by their RMS values.

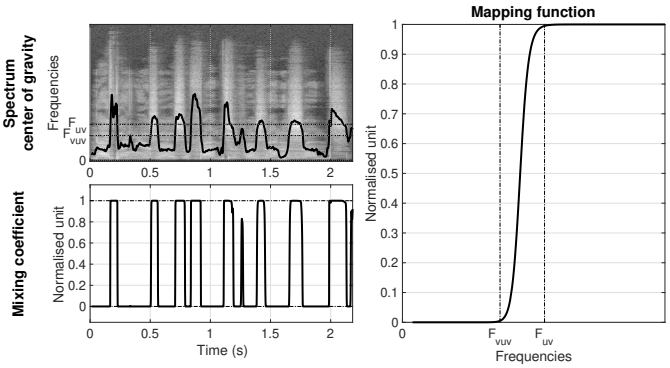

Fig. 6. Mixing coefficient computed from whisper spectrum centre of gravity for the sentence *"Nothing is as offensive as innocence"*. Top-left: whisper spectrum centre of gravity plotted over the spectrogram. Right: mapping function. Dotted lines indicate F_{vuv} and F_{uv}. Bottom-left: mixing coefficient.

the whisper amplitude has high-frequency variations that need to be smoothed. Also, the amplitude of different whispered vowels in an utterance presents high variability, requiring some homogenisation of the amplitude envelope.

Smoothing uses a mean filter of size W_{win} applied forward and backward across frames to avoid distorting the phase of the signal. I_{g_w} is the whisper glottis intensity while $< I_{g_w} >$ is its smoothed version. Homogenisation follows the observation that normal speech intensity standard deviation (STD) should be about 3 dB [59]. So the STD of the full whisper intensity contour I_{sw} is computed on the intensity in decibels:

$$\text{STD}_{I_{sw}} = \text{STD}\left[20\log_{10}(I_{sw})\right] \tag{8}$$

Then the standard deviation of the smoothed whisper glottis intensity contour is set to 3 dB by:

$$I_g(t) = \left[< I_{g_w}(t) >\right]^{\frac{3}{\text{STD}_{I_{sw}}}} \tag{9}$$

In practice, the frame-based real time paradigm used in our system does not allow computation of standard deviation over a full sentence. Therefore, we adopt a simple calibration process to compute whisper intensity STD over several sentences uttered by the user before initiating WSC.

Fig. 5 illustrates glottis amplitude envelope synthesis. Smoothing of the whisper glottis amplitude (left panel) allows low-frequency variations to be maintained, at the phone rate. The variability reduction (right panel) keeps relative variations across time so that the global prosody is preserved, while making the amplitude contour smoother.

B. Mixing synthesised and whisper glottis

To avoid having a constant harmonic source across the signal, it is necessary to alternate between voiced and unvoiced sources depending on the phonemic pattern. We thus derive a time-dependent mixing coefficient that allows us to define the weight of the synthesised voiced and whisper unvoiced sources accordingly, to form the final glottis signal.

1) Mixing coefficient: Information must be obtained from whisper analysis to indicate transitions from what must be voiced phones and unvoiced phones. We have found that a good indicator is the spectrum centre of gravity defined by:

$$F_{\text{CoG}} = \frac{\int_0^{F_s/2} f|S(f)|^2 df}{\int_0^{F_s/2} |S(f)|^2 df} \tag{10}$$

where $S(f)$ is the spectrum of the full whisper on a time frame, f is the frequency, and F_s is the sampling frequency. The top-left of Fig. 6 shows an example of whisper spectrum centre of gravity. One can see that the latter follows the alternation between broadband sounds (consonants) and low-frequency energy regions (vowels). In particular, the most prominent formants produced during voiced speech tend to lie below 4 kHz [10]. Therefore, we expect that whisper vocalic sounds have a centre of gravity below this threshold. Conversely, unvoiced consonants fill the high-frequency spectrum, and we expect their centre of gravity to be much higher than this threshold. For coherence with the previous parts, we again use the voiced/unvoiced frequency F_{vuv} to distinguish between vowels and consonants and want the mixing coefficient to be 0 below this threshold. Then, the centre of gravity reaches different peak heights depending on the consonant. For instance, the highest peak in Fig. 6 corresponds to a $/f/$ which contains more high-frequency energy than most of the other peaks (e.g., $/s/$). Nevertheless, we expect all peaks to be classified as unvoiced regardless of the consonant. For this sake, we define the unvoiced frequency F_{uv} as the geometric mean of the values of the centre of gravity above F_{vuv}. Both F_{uv} and F_{vuv} are displayed in Fig. 6. We want the mixing coefficient to be 1 above F_{uv}. As for the computation of the whisper intensity standard deviation (section IV-A2), we suggest that F_{uv} is estimated during the same simple calibration process.

Variations of centre of gravity from vocalic sounds to consonants are sometimes slower than what we would expect from a voiced/unvoiced transition, which are usually binary decisions [44]. Therefore, a direct linear mapping from [F_{vuv}, F_{uv}] to [0,1] yields transitions that are not sharp enough in practice. Hence, we propose a non-linear mapping (shown on the right of Fig. 6) based on the sigmoid function:

$$m(t) = 0.5\frac{e^{\lambda\log(F_{\text{CoG}}(t)-\sqrt{F_{\text{vuv}}F_{\text{uv}}})}-1}{e^{\lambda\log(F_{\text{CoG}}(t)-\sqrt{F_{\text{vuv}}F_{\text{uv}}})}+1} + 0.5 \tag{11}$$

The sigmoid is applied on the log-frequency scale to account for the logarithmic perception of frequencies, and is centred on the geometric mean of F_{vuv} and F_{uv}. λ is the slope at mid-range and is computed so that the mixing function fulfils $\tau_r = 99\%$ of this range between F_{vuv} and F_{uv}:

$$\lambda = \frac{2}{\log(F_{\text{uv}}/F_{\text{vuv}})} \log\left\{\frac{1+\tau_r}{1-\tau_r}\right\} \tag{12}$$

2) Mixing the glottis filters: An example mixing coefficient is displayed on the bottom-left of Fig. 6. It shows several non-binary values, implying the simultaneous presence of both voiced and unvoiced signals. In the case of unvoiced phones (coefficient close to 1), the slight presence of a voiced component introduces strong buzziness in the signal because the voiced energy is mainly distributed on low frequencies while the unvoiced component energy is at high frequencies and thus cannot mask the voiced energy. To mitigate this effect, the synthesised LM filter is also mixed with the whisper glottis filter, so that both voiced and unvoiced components share a similar spectral envelope for unvoiced phones, enabling the unvoiced component to mask the voiced component. Mixing occurs in the line spectral pair (LSP) domain [60] due mainly to the high stability of LSPs, making them well suited for mixing filters. LSPs are derived from both pairs of complex conjugate poles of the whisper glottis filter (LSP_w) and the LM filter (LSP_{LM}). These poles model the glottal formant only, as spectral tilt is not included. A geometric mix of LSPs is then performed, according to mixing coefficient m:

$$LSP_s(t) = LSP_{LM}(t)^{(1-m(t))} \times LSP_w(t)^{m(t)} \tag{13}$$

A new pair of complex conjugate poles is derived from the mixed LSP (LSP_s), and associated with the LM spectral tilt real pole to synthesise the final glottis filter, as shown on the bottom of the middle panel of Fig. 2.

3) Mixing the glottal flow derivatives: Finally, the synthesised voiced glottal flow derivative g_v is obtained by filtering the synthesised excitation by the synthesised glottis filter. It is then normalised to the same energy as the whisper glottal flow derivative g_w, and mixed with the latter according to the mixing coefficient m:

$$g_s(t) = \{1 - m(t)\}g_v(t) + m(t)K_w g_w(t) \tag{14}$$

A gain K_w controls the energy balance between voiced and unvoiced components. We empirically define $K_w = 0.1$ to have vowels 20 dB louder than consonants on average [61].

V. Evaluation

We first evaluate the overall contribution of our new speech glottal source in a perceptual experiment, then assess the mixing coefficient and source-filter separation independently.

A. Subjective experiment

The major novelty of this paper is the new source model, including novel excitation plus glottis filter, later called *glottal* source. The first experiment aims to compare this glottal source against previous methods in terms of reconstructed speech naturalness. The choice of the baseline to compare our system against is a critical aspect and is discussed here. In the field of parametric speech generation (e.g., text-to-speech synthesis or statistical voice conversion), the STRAIGHT vocoder is widely used [62]–[64], being a common reference to compare new systems with [65]. However, this is possible only because most recent speech generation systems make use of machine learning techniques that allow them to predict the large number of STRAIGHT parameters, particularly the aperiodic components that describe the degree of aperiodicity included in various frequency bands (typically 25 parameters [66], [67]). By contrast, as summarised in section I-A, many WSC systems are rule-based, and cannot use this aspect of the STRAIGHT vocoder. Moreover, among machine learning-based methods, only a few predict the aperiodic components required by STRAIGHT [11], [12], [16]. The alternatives to STRAIGHT are CELP [25], MELP [21], sinusoidal synthesis [1], and pulse trains [13]–[15], [20], [25]. To our knowledge, pulse trains are the most used method for excitation synthesis in WSC publications. As a result, most authors baseline against either natural signals (e.g., original whispers [11], electrolarynx [1], [25]), or against pulse train generation [13], [15], [16]. Only [12], [16] used a STRAIGHT-based system as a baseline, which is logical because their aim was to improve the STRAIGHT system they compared against.

TABLE III
SYSTEMS USED FOR THE EVALUATION

Source		**Vocal tract**
Baseline	Pulse train + Whisper Glottis	Whisper VT
Glottal	Synthetic Excitation + Synthetic Glottis	Whisper VT
Reference	Speech Excitation + Speech Glottis	Whisper VT
Baseline	Pulse train + Whisper Glottis	Speech VT
Glottal	Synthetic Excitation + Synthetic Glottis	Speech VT
Reference	Speech Excitation + Speech Glottis	Speech VT

It would be possible to replace the prediction of aperiodic components in a rule-based system by empirically defining a constant degree of aperiodicity. However, it has been shown that speech excitation generated by STRAIGHT with multi-band aperiodicity components is similar in terms of quality to a standard pulse train [67]. We hypothesise that the difference between STRAIGHT excitation and a pulse train would be even smaller with an invariant aperiodicity. Therefore, for all these reasons, we follow the majority of rule-based WSC system authors and employ a baseline that uses a pulse train for voiced excitation, combined with the whisper glottis filter. The mixing between synthesised and whisper glottis detailed in section IV-B is applied to both *baseline* and proposed *glottal* sources.

We also compare both sources to a natural source called *reference* (i.e., natural speech excitation and natural speech glottis filter). As the whisper vocal tract is not processed, this severely alters the quality of the reconstructed speech. To isolate the contribution of our new source model, whispers are reconstructed both using the whisper and the natural speech VT. To do this, the latter are aligned to whisper VT using dynamic time warping (DTW) [68] based on manual annotation of phones in speech and whisper with Praat [69].

TABLE IV
SIGNIFICANCE OF THREE FACTORS ON MUSHRA SCORES ASSESSED BY A KRUSKAL-WALLIS RANK SUM TEST USING A χ^2 DISTRIBUTION.

Factor	***df***	χ^2	***p***
Speaker	3	1.51	0.68
System	5	3668	$< 10^{-15}$
Sentence	8	19.0	0.015

The alignment was applied to the LSPs of the normal speech VT. Finally, speech and whisper VTs are used with the three sources (*baseline*, *glottal* and *reference*). Table III summarises the cross-synthesis combinations used to generate the stimuli. Finally, we ensured that all stimuli had identical natural f_0 trajectories so that any artefacts in the synthesised signals are due to the glottal source generation only.

1) Stimuli: The wTIMIT database [70] provides pairs of 450 phonetically balanced sentences uttered with normal speech and whisper by 48 speakers. 9 sentences uttered by 4 native US speakers (2 females, 2 males) were selected from the database for this experiment, sampled at 44.1 kHz. In total, 6 signals were compared for each sentence and subject. Overall, 216 stimuli were generated (4 speakers $\times$ 9 sentences $\times$ 6 systems)[2]. The parameters used for speech reconstruction were $E_{min} = 0$ and $E_{max} = 1$ so that the full range of vocal effort is used, and the tension level was set to $T = 0.4$. The f_0 trajectories of all speech signals were manually extracted with Praat, aligned with the same DTW alignments used for the speech VT, and used as an input parameter. The parameters $\text{STD}_{I_{sw}}$ and F_{uv} were evaluated for each subject on all sentences in the test. A Hanning window of $W_{\text{win}} = 1024$ samples with a shift of $W_{\text{shift}} = 64$ samples was used for the frame-based analysis and synthesis.

2) Protocol: A MUSHRA-based test [71], [72] was used to assess the preferences of listeners for the different reconstruction methods. The test started with a training session where subjects could listen freely to any of the 216 stimuli to get an overview of the range of naturalness of the reconstructed speech. The training session was limited to 5 min per subject and they were asked to listen to as many stimuli as they could during that time. During the evaluation, subjects were presented with 36 screens (one for each sentence and speaker). On each screen, the subjects could listen to the 6 reconstructed speech types and were asked to rate their naturalness on a continuous scale from 0 to 100. A natural speech reference was also given, and subjects were free to listen to the reference and the 6 stimuli as much as they wanted. As one of the 6 stimuli was actually natural speech (*reference* source + speech vocal tract), the subjects were asked to rate one of the 6 signals as being 100 on the naturalness scale, but were not told which one. Screens, and the stimuli on each screen, were presented in random order. Sounds were played to subjects using a Sennheiser HD205 headset in a quiet room. Rests were allowed after any screen and the full evaluation took around 45 mins per subject. 21 listeners (5 female, 16 male, average age 31, with self-reported normal hearing) participated. 9 subjects had some experience of audio processing but none of speech reconstruction. They were compensated with a £10 gift card.

[2]Examples of the 6 stimuli generated for one male and one female speaker are given as supplementary material available at http://ieeexplore.ieee.org

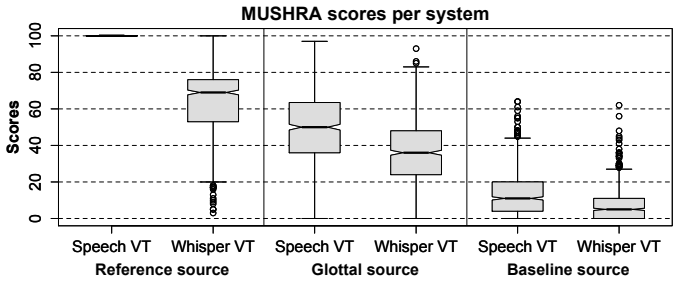

Fig. 7. MUSHRA scores comparing the 6 methods across all listeners, speakers and sentences. Each box represents the interquartile range of the distributions and the thick horizontal lines are the medians. The whiskers include the full distribution omitting outliers represented by the circles.

3) Results: A Kruskal-Wallis rank sum test was used to assess the significance of the factors *Speaker* (4 levels), *System* (6 levels), and *Sentence* (9 levels). Table IV displays the analysis results. We first observe that the *Speaker* factor does not significantly explain the difference between the listeners' scores ($\chi^2 = 1.51$, $p = 0.68$), indicating that over all sentences, all WSC systems apply the same way to all speakers. The number of speakers was small but balanced, suggesting that reconstruction works with both male and female voices.

The main difference between scores is explained by the *System* factor ($\chi^2 = 3668$, $p < 10^{-15}$). Figure 7 depicts the overall listeners' ratings for each system, across all speakers and sentences. A Wilcoxon rank-sum test assessed the difference between each pair of distributions relative to the synthesis method, and all were judged significantly different ($p < 10^{-15}$). First, we can notice that all listeners recognised the natural speech example for all sentences and speakers, without exception. The perceptive difference between the reference speech and the other signals was therefore clear. Then, we note an effect of the source, where the *reference* source was given the best scores (medians of 100 and 69 for the speech and whisper VT, respectively); the *baseline* source was rated with the lowest scores (medians of 11 and 5 for the speech and whisper VT, respectively); the *glottal* source was rated in-between with medians of 50 and 36 for speech and whisper VT, respectively. We infer from these results that while our new *glottal* source does not match the quality of the natural source, it significantly outperforms the commonly used pulse train excitation. We also observe an effect of the VT. Looking at the reference source, the application of whisper VT on a natural speech source significantly degrades speech naturalness. This justifies the need for VT adaptation in WSC. Nevertheless, the improvement gained by using the speech VT (difference between *baseline* source distributions) is much smaller than that gained by applying our *glottal* source (difference between *baseline* and *glottal* sources with whisper VT distributions). This corroborates our belief that source

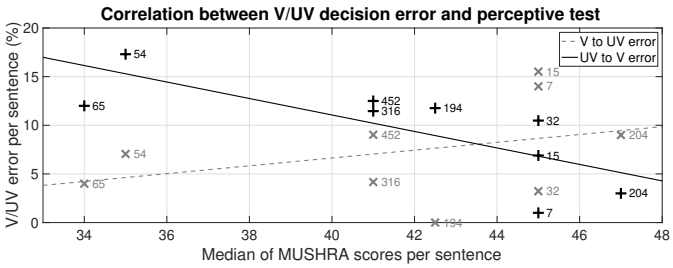

Fig. 8. UV-to-V (black +) and V-to-UV (grey ×) error rates depending on MUSHRA scores per sentence (numbers). Respective correlations are plotted with plain black and dotted grey lines.

reconstruction is more beneficial than VT adaptation.

Finally, the *Sentence* factor has a small but significant effect on the listeners' score ($\chi^2 = 19.0$, $p = 0.015$). Median scores of distributions for each sentence range from 34 to 46. The best results were obtained for the sentences having the fewest unvoiced consonants, suggesting that the output speech quality may be sensitive to voiced/unvoiced (V/UV) decisions.

B. Evaluation of the voiced/unvoiced decisions

The V/UV decisions are fully dependent on the mixing coefficient, and were evaluated for all the stimuli from the subjective experiment. Since the manually annotated V/UV ground truth is binary for each phone, we needed to convert the continuous mixing coefficient values to binary. To do this we identified segments of consecutive 1 or 0 mixing coefficient values on each phone. If the longest segment was 1, the phone was classified as unvoiced, else it was classified as voiced.

Over all stimuli, the V-to-UV error rate was 7.6% and UV-to-V error rate was 10.1%, leading to a total V/UV error rate of 17.7%. Among the reviewed rule-based WSC methods, only one implements a voicing decision based on a phoneme classifier, but was not evaluated objectively [25]. Conversely, most machine learning based methods include a voicing classifier, with best reported V/UV error rates about 9% using GMMs [13], [17] and 6.8% using neural networks [17]. Those classifiers were all trained from parallel databases of speech and whisper of about 150 to 200 utterances. Our untrained voicing prediction, by contrast, is obtained from a single function of the spectrum centre of gravity. Therefore, while its performance has not yet reached that of machine learning methods, the results are highly encouraging given its efficiency of implementation.

To assess the effect of voicing prediction on perceived quality, we computed Pearson's correlations between the medians of the MUSHRA scores per sentence from the subjective experiment, the V-to-UV error rate per sentence ($r = 0.35$, $p = 0.34$) and UV-to-V error rate per sentence ($r = -0.76$, $p = 0.018$), respectively, shown in Fig. 8. No significant correlation is observed between V-to-UV error rate and MUSHRA scores, indicating that when the system fails to add voicing, it has little effect on listeners' scores. By contrast, a significant negative correlation exists between UV-to-V error rate and MUSHRA scores (the higher the error, the lower the scores). This suggests that speech quality is sensitive to the presence of wrong-voiced phones, causing buzziness in the output signal. This indicates that future work should focus primarily on reducing UV-to-V error rate by including more whisper information in the mixing coefficient computation.

C. Evaluation of the source-filter decomposition

The authors previously evaluated GFM-IAIF on static phones [42], but the current paper is its first use with continuous speech and whispers. Thus, we need to demonstrate that GFM-IAIF correctly decomposes the glottis and VT elements for those signals – specifically that the glottal parameters account for the main difference between V and UV versions of the same phone, while VT parameters remain broadly similar. To assess this, we computed VT spectral tilt for whispered and corresponding voiced phones on a frame-by-frame basis. Across all test speakers and utterances, this showed a mean difference of just 0.40 dB/decade, and a standard deviation of 5.58 dB/decade, providing a clear indication that both VTs derive from the same model, and that the whisper/speech difference is largely encoded in the glottal component.

VI. Discussion and Conclusion

This paper has presented a new rule-based WSC method that focuses on the generation of a high quality phonated glottal source. The system employs GFM-IAIF to achieve consistent source-filter decomposition to accurately separate VT components from whisper glottis signals. A high-quality spectral glottis model is used to shape an excitation signal, while whisper variability is incorporated into the speech excitation, via a small set of empirically-defined parameters that drive reconstruction naturalness in a way inspired by the high-quality Cantor Digitalis singing synthesiser [58]. We note that a full range of values for these parameters is likely to provide satisfying sound quality, but allow different degrees of voice quality (e.g., roughness). Eventually, the design of a mapping between whisper features and all voice quality parameters (i.e., the excitation roughness, and the glottis filter tension and effort parameters) would help to increase the reconstructed speech naturalness further. Overall, the entire WSC system was evaluated using a MUSHRA-based test, yielding results that show a statistically significant improvement over the baseline reconstruction method.

To conclude, this paper presented a system for whisper-to-speech conversion that perceptively improves on state-of-the-art methods that use pulse train glottal sources. The use of a voice-oriented source-filter model allows synthesis of a high-quality glottis signal, independent of the VT. Whisper variability is introduced in the harmonic generation, allied with a linear glottis model. There has been much recent research on VT adaptation and absolute pitch frequency prediction for WSC, but none to our knowledge on glottal source reconstruction. Since those components are independent in a source-filter model of speech, the glottal source proposed in this paper has the potential to enhance the performance of a wide variety of different VT adaptation + f_0 prediction methods.

ACKNOWLEDGMENT

This work is supported by the French National Research Agency in the framework of the "Investissements d'avenir" programs ANR-15-IDEX-02 and ANR-11-LABX-0025-01.

REFERENCES

[1] I. V. McLoughlin, H. R. Sharifzadeh, S. L. Tan, J. Li, and Y. Song, "Reconstruction of phonated speech from whispers using formant-derived plausible pitch modulation," *ACM Trans. on Accessible Computing*, vol. 6, no. 4, pp. 12:1–12:21, 2015.

[2] M. A. Morris, S. K. Meier, J. M. Griffin, M. E. Branda, and S. M. Phelan, "Prevalence and etiologies of adult communication disabilities in the United States: Results from the 2012 national health interview survey," *Disability and Health Journal*, vol. 9, no. 1, pp. 140–144, 2016.

[3] P. Carding, A. Welch, S. Owen, and F. Stafford, "Surgical voice restoration," *The Lancet*, vol. 357, no. 9267, pp. 1463–1464, 2001.

[4] H. Takahashi, M. Nakao, Y. Kikuchi, and K. Kaga, "Alaryngeal speech aid using an intra-oral electrolarynx and a miniature fingertip switch." *Auris Nasus Larynx*, vol. 32, no. 2, pp. 157–162, 2005.

[5] H. Liu and M. L. Ng, "Electrolarynx in voice rehabilitation," *Auris Nasus Larynx*, vol. 34, no. 3, pp. 327–332, 2007.

[6] R. Kaye, C. G. Tang, and C. F. Sinclair, "The electrolarynx: voice restoration after total laryngectomy," *Medical Devices (Auckland, N.Z.)*, vol. 10, pp. 133–140, 2017.

[7] G. Fant, *Acoustic Theory of Speech Production*. Mouton, 1970.

[8] B. Doval, C. d'Alessandro, and N. Henrich, "The spectrum of glottal flow models," *Acta Acustica united with Acustica*, vol. 92, no. 6, pp. 1026–1046, 2006.

[9] K. J. Kallail and F. W. Emanuel, "An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects," *Journal of Phonetics*, vol. 12, pp. 175–186, 1984.

[10] H. R. Sharifzadeh, I. V. McLoughlin, and M. J. Russell, "A comprehensive vowel space for whispered speech," *Journal of Voice*, vol. 26, no. 2, pp. 49–56, 2012.

[11] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 20, no. 9, pp. 2505–2517, 2012.

[12] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Improvement to a NAM-captured whisper-to-speech system," *Speech Communication*, vol. 52, no. 4, pp. 314–326, 2010.

[13] J. Li, I. V. McLoughlin, L.-R. Dai, and Z.-h. Ling, "Whisper-to-speech conversion using restricted boltzmann machine arrays," *Electronics Letters*, vol. 50, no. 24, pp. 1781–1782, 2014.

[14] I. V. McLoughlin, J. Li, Y. Song, and H. R. R. Sharifzadeh, "Speech reconstruction using a deep partially supervised neural network," in *Healthcare Technology Letters*, vol. 4, no. 4, 2017, pp. 129–133.

[15] M. Janke, M. Wand, T. Heistermann, and T. Schultz, "Fundamental frequency generation for whisper-to-audible speech conversion," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 4-9 2014, pp. 2579–2583.

[16] G. N. Meenakshi and P. K. Ghosh, "Whispered speech to neutral speech conversion using bidirectional lstms," in *Proceedings of Interspeech*, Hyderabad, India, September 2-6 2018, pp. 491–495.

[17] V.-A. Tran, G. Bailly, H. Lœvenbruck, and T. Toda, "Predicting F0 and voicing from NAM-captured whispered speech," in *Speech Prosody*, Campinas, Brazil, May 6-9 2008, pp. 107–110.

[18] K. Tanaka, H. Kameoka, T. Toda, and S. Nakamura, "Physically constrained statistical F0 prediction for electrolaryngeal speech enhancement," in *Proc. of Interspeech*, Stockholm, Sweden, August 21-24 2017, pp. 1069–1073.

[19] X. Shao and B. P. Milner, "Pitch prediction from MFCC vectors for speech reconstruction," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Montreal, Canada, May 17-21 2004, pp. 97–100.

[20] H. Konno, M. Kudo, H. Imai, and M. Sugimoto, "Whisper to normal speech conversion using pitch estimated from spectrum," *Speech Communication*, vol. 83, pp. 10–20, 2016.

[21] R. W. Morris and M. A. Clements, "Reconstruction of speech from whispers," *Medical Engineering & Physics*, vol. 24, no. 7-8, pp. 515–520, 2002.

[22] H. R. Sharifzadeh, I. V. McLoughlin, and F. Ahmadi, "Regeneration of speech in voice-loss patients," in *Int. Conf. on Biomedical Engineering (ICBME)*, C. T. Lim and J. C. H. Goh, Eds., Singapore, December 3-6 2009, pp. 1065–1068.

[23] ——, "Voiced speech from whispers for post-laryngectomised patients," *IAENG International Journal of Computer Science*, vol. 36, no. 4, pp. 367–377, 2009.

[24] ——, "Spectral enhancement of whispered speech based on probability mass function," in *Advanced International Conference on Telecommunications (AICT)*, Barcelona, Spain, May 9-15 2010, pp. 207–211.

[25] ——, "Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec," *IEEE Trans. on Biomedical Engineering*, vol. 57, no. 10, pp. 2448–2458, 2010.

[26] J. Makhoul, "Linear predicion: A tutorial review," in *Proc. of the IEEE*, vol. 63, no. 4, April 1975, pp. 561–580.

[27] A. V. McCree and T. P. Barnwell, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 4, pp. 242–250, 1995.

[28] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[29] S. Pascual, A. Bonafonte, J. Serrà, and J. A. González López, "Whispered-to-voiced alaryngeal speech conversion with generative adversarial networks," in *Proceedings of IberSPEECH*. Barcelona, Spain: ISCA, November 21-23 2018, pp. 117–121.

[30] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using GMM-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[31] H. Doi, T. Toda, K. Nakamura, H. Saruwatari, and K. Shikano, "Alaryngeal speech enhancement based on one-to-many eigenvoice conversion," *IEEE Trans. Audio, Speech, Lang. Proc.*, vol. 22, no. 1, pp. 172–183, 2014.

[32] G. Fant, A. Kruckenberg, J. Liljencrants, and M. Bavegard, "Voice source parameters in continuous speech. transformation of LF-parameters," in *Int. Conf. on Spoken Language Processing (ICSLP)*, Yokohama, Japan, September 18-22 1994, pp. 1451–1454.

[33] P. Hedelin, "A glottal lpc-vocoder," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, San Diego, CA, USA, March 19-21 1984, pp. 21–24.

[34] D. Vincent, O. Rosec, and T. Chonavel, "A new method for speech synthesis and transformation based on an ARX-LF source-filter decomposition and HNM modeling," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 4, Honolulu, Hawaii, USA, April 15-20 2007, pp. 525–528.

[35] J. P. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195–208, 2013.

[36] G. Degottex, P. Lanchantin, 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, 2013.

[37] D. G. Childers and H. T. Hu, "Speech synthesis by glottal excited linear prediction," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2026–2036, 1994.

[38] Y.-L. Shue, J. Kreiman, and A. Alwan, "A novel codebook search technique for estimating the open quotient," in *Proc. of Interspeech*, Brighton, UK, September 6-10 2009, pp. 2895–2898.

[39] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Prague, Czech Republic, May 22-27 2011, pp. 4564–4567.

[40] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Comparing glottal-flow-excited statistical parametric speech synthesis methods," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, May 26-31 2013, pp. 7830–7834.

[41] O. Perrotin and I. McLoughlin, "Gfm-voc: A real-time voice quality modification system," in *Proceedings of Interspeech*, Graz, Austria, September 15-19 2019, pp. 3685–3686.

[42] O. Perrotin and I. V. McLoughlin, "A spectral glottal flow model for source-filter separation of speech," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Brighton, UK, May 12-17 2019, pp. 7160–7164.

[43] D. G. Childers, "Vocal quality factors: Analysis, synthesis and perception," *J. Acoust. Soc. Am.*, vol. 90, no. 5, pp. 2394–2410, 1991.

[44] D. O'Shaughnessy, "Linear predictive coding," *IEEE Potentials*, vol. 7, no. 1, pp. 29–32, 1988.

[45] W. F. L. Heeren, "Vocalic correlates of pitch in whispered versus normal speech," *J. Acoust. Soc. Am.*, vol. 138, no. 6, pp. 3800–3810, 2015.

[46] M. F. Schwartz, "Power spectral density measurements of oral and whispered speech," *Journal of Speech, Language, and Hearing Research*, vol. 13, no. 2, pp. 445–446, 1970.
[47] T. Itoh, K. Takeda, and F. Itakura, "Acoustic analysis and recognition of whispered speech," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, FL, USA, May 13-17 2002, pp. 429–432.
[48] T. Ito, K. Takeda, and F. Itakura, "Analysis and recognition of whispered speech," *Speech Communication*, vol. 45, no. 2, pp. 139–152, 2005.
[49] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109–118, 1992.
[50] T. Raitio, L. Juvela, A. Suni, M. Vainio, and P. Alku, "Phase perception of the glottal excitation and its relevance in statistical parametric speech synthesis," *Speech Communication*, vol. 81, pp. 104–119, 2016.
[51] D. H. Klatt and L. C. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, 1990.
[52] T. Drugman and T. Raitio, "Excitation modeling for hmm-based speech synthesis: Breaking down the impact of periodic and aperiodic components," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, Italy, May 4-9 2014, pp. 260–264.
[53] R. Maia, M. Akamine, and M. J. Gales, "Complex cepstrum for statistical parametric speech synthesis," *Speech Communication*, vol. 55, no. 5, pp. 606–618, 2013.
[54] B. C. J. Moore, R. W. Peters, and B. R. Glasberg, "Thresholds for the detection of inharmonicity in complex tones," *J. Acoust. Soc. Am.*, vol. 77, no. 5, pp. 1861–1867, 1985.
[55] Y. Stylianou, "Harmonic plus noise models for speech combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécommunications, Paris, January 1996.
[56] U. Zölzer, *Digital Audio Signal Processing*. Wiley, 1999.
[57] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," Royal Institute of Technologies - Dept. for Speech, Music and Hearing, Quarterly Progress and Status Report 4, 1985.
[58] L. Feugère, C. d'Alessandro, B. Doval, and O. Perrotin, "Cantor Digitalis: Chironomic parametric synthesis of singing," *EURASIP Journal on Audio, Speech, and Music Processing*, 2017.
[59] J. Robbins, H. B. Fisher, E. C. Blom, and M. I. Singer, "A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production," *Journal of Speech and Hearing Disorders*, vol. 49, no. 2, pp. 202–210, 1984.
[60] I. V. McLoughlin, "Line spectral pairs," *Signal Processing*, vol. 88, no. 3, pp. 448–467, 2008.
[61] E. Kennedy, H. Levitt, A. C. Neuman, and M. Weiss, "Consonant–vowel intensity ratios for maximizing consonant recognition by hearing-impaired listeners," *J. Acoust. Soc. Am.*, vol. 103, no. 2, pp. 1098–1114, 1998.
[62] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, no. C, pp. 65–82, 2017.
[63] Y. Ning, S. He, Z. Wu, C. Xing, and L.-J. Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, 2019.
[64] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden markov models," in *Proceedings of the IEEE*, vol. 101, no. 5, 2013, pp. 1234–1252.
[65] S. King, J. Crumlish, A. Martin, and L. Wihlborg, "The blizzard challenge 2018," in *The Blizzard Challenge 2018 workshop*, Hyderabad, India, September 8 2018.
[66] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *ISCA Speech Synthesis Workshop*, Sunnyvale, USA, September 13-15 2016, pp. 202–207.
[67] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *ISCA Speech Synthesis Workshop*, Barcelona, Spain, Aug. 31 - Sep. 2 2013, pp. 135–140.
[68] D. Ellis. (2003) Dynamic time warp (DTW) in Matlab. [Online]. Available: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/
[69] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001. [Online]. Available: http://www.praat.org/
[70] B. P. Lim, "Computational differences between whispered and non-whispered speech," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 2010.
[71] "Method for the subjective assessment of intermediate quality level of audio systems," International Telecommunication Union, Tech. Rep. ITU-R BS.1534-3, October 2015.
[72] E. Vincent. (2005) MUSHRAM - a Matlab interface for MUSHRA listening tests. [Online]. Available: http://c4dm.eecs.qmul.ac.uk/downloads/

Olivier Perrotin received a Ph.D. degree in computer science from Université Paris-Sud, France, in 2015, and a M.Sc. in Signal Processing from Grenoble Institute of Technology, France, in 2012. He has been a researcher at LIMSI, CNRS, France, from 2012 to 2016, at the University of Kent, UK, from 2017 to 2018, and currently at GIPSA-lab, CNRS, France. His main research interests include human-computer interaction, expressive speech and singing synthesis, voice analysis, and voice reconstruction.

Ian McLoughlin (M'94, SM'04) completed his PhD in Electronic and Electrical Engineering at the University of Birmingham, UK in 1997 and was elected a Fellow of the IET in 2012. He is author of two Cambridge University Press textbooks on speech processing, along with many papers and patents related to his research on speech and associated technology.