Synthesis and Analysis-By-Synthesis of Modulated Diplophonic Glottal Area Waveforms

Philipp Aichinger, *Member, IEEE*, and Franz Pernkopf, *Senior Member, IEEE*

***Abstract*—Diplophonia is a type of disordered voice in which two simultaneous pitches are perceived. Most commonly in diplophonic voices, the vocal folds are divided into two parts that vibrate at different frequencies. The glottal area is the projected area of the space between the vocal folds. The glottal area in time is referred to as the glottal area waveform (GAW). The GAW is modeled for diplophonic voice by superimposing two partial GAWs (pGAWs) that are trains of single-peak pulses with different pulse frequencies, i.e., fundamental frequencies (f_os). In current kinematic models of diplophonic vocal fold vibration, the pGAWs are assumed to be quasiperiodic. This assumption is mitigated here by modulating pulse-to-pulse cycle length and amplitude. Both random and deterministic modulations are considered. Deterministic modulations depend on the difference of the pGAWs' instantaneous phases. Model GAWs are fitted to input GAWs using an analysis-by-synthesis approach which we refer to as 'modulated pulse trains decomposition' (MPD). MPD is shown to be applicable to diplophonic as well as to nondiplophonic types of dysphonia, which include multi-pulse patterns, random timing behaviours, and chaos. It is mostly robust against modulations but degraded by large random modulations. MPD is compared to a deep autoencoder neural network, and the WaveGlow neural network. In terms of time-domain fitting errors, MPD outperforms the other two approaches unless random modulations are large. MPD outperforms the best of the other two approaches by up to approximately 5 dB. For large random modulations, the deep autoencoder network achieves the smallest fitting errors. In terms of magnitude spectrum fitting errors, WaveGlow is superior except for natural input GAWs containing only nondiplophonic types of dysphonia. Also pulse timing errors are shown to be advantageous for MPD.**

***Index Terms*—Diplophonia, glottal area waveforms, kinematic model, modulation, pathological voice, vocal fold vibration.**

I. Introduction

Disorders of the human voice associated with degraded voice quality and disability to talk normally may result in reduced job opportunities, loss of quality of life, and even social isolation. Hence, it is necessary to clinically treat disordered voices, which relies on accurate and reliable diagnostic analysis of voice function. The current diagnostic evaluation of voice function includes imaging of the larynx and the vocal folds, signal processing-based analysis of audio recordings, and auditory evaluation of speech and voice sounds. Fig. 1 illustrates the examination of the human vocal folds by transoral endoscopy and simultaneous recording of the audio signal.

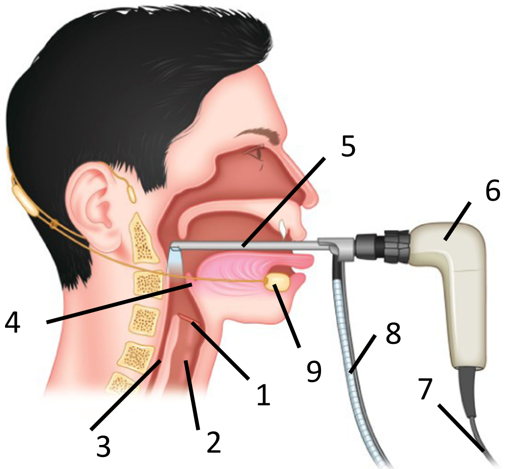


Fig. 1. Midsagittal cross-section view of endoscopic laryngeal high-speed video examination. (1) vocal folds, (2) trachea, (3) esophagus, (4) epiglottis, (5) endoscope, (6) camera handle, including photo sensor, (7) data cable, (8) light conductor, and (9) headworn microphone.

In normal vocal fold vibration, the transglottal air flow is minimal during the vocal folds' constricted/closed phase. In this phase, the subglottal pressure builds up because the lung pushes the air upstream. The pressure exerts opening forces on the vocal folds, which cause the folds to move laterally. During the open phase the subglottal pressure drops, and the transglottal flow and supraglottal pressure increase. The Bernoulli effect and elastic forces cause the vocal folds to close again, which results in a decrease of airflow and supraglottal pressure, and an increase of subglottal pressure. The process of opening and closing repeats cyclically in terms of a self-sustained oscillation. Cyclic increase and decrease of the supraglottal pressure is the 'voice source', which results in a sound wave travelling upwards the vocal tract. This wave is superimposed with the downward travelling wave reflected by the vocal tract [1], [2].

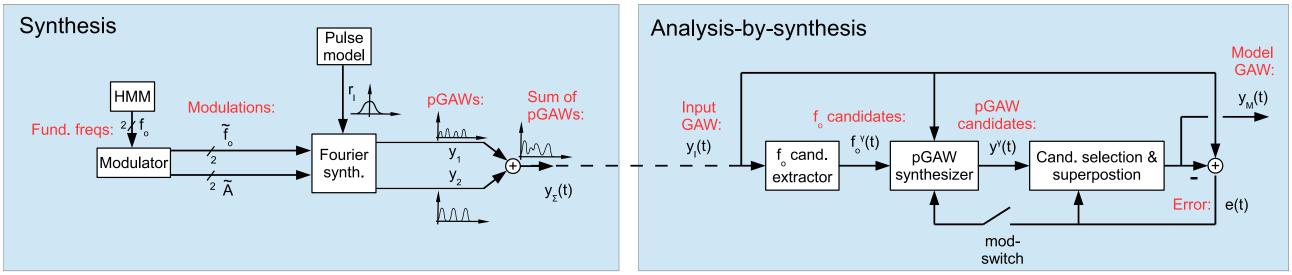
Divergences from normal voice quality are often caused by vocal fold vibration patterns, of which models and analyses are

Manuscript received August 28, 2020; revised November 23, 2020 and January 14, 2021; accepted January 17, 2021. Date of publication January 21, 2021; date of current version February 19, 2021. This work was supported by the Austrian Science Fund (FWF): KLI 722-B30. The computational results presented were achieved in part using the Vienna Scientific Cluster (VSC). The handling of WaveGlow was supported by Markus Huber of sonible GmbH (Graz, Austria). The Associate Editor coordinating the review of this manuscript and approving it for publication was Prof. Juan Ignacio Godino-Llorente. *(Corresponding author: Philipp Aichinger.)*

Philipp Aichinger is with the Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, the Medical University of Vienna, 1090 Vienna, Austria (e-mail: philipp.aichinger@meduniwien.ac.at).

Franz Pernkopf is with the Signal Processing and Speech Communication Laboratory (SPSC), the Graz University of Technology, 8010 Graz, Austria (e-mail: pernkopf@tugraz.at).

Digital Object Identifier 10.1109/TASLP.2021.3053387



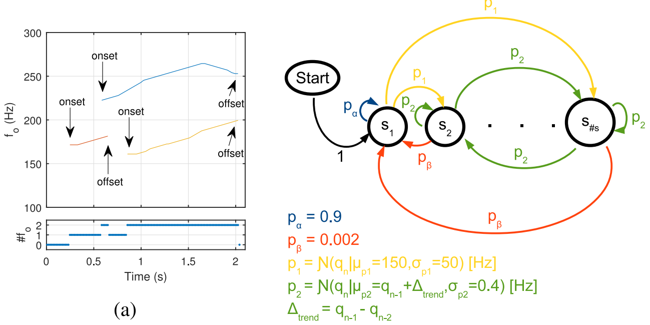
(a)

(b)

Fig. 2. (a) Overview of the synthesis method. Modulated pGAWs y_1 and y_2 are obtained using a Fourier synthesizer. Amplitude and frequency modulation is applied. Examples of deterministic and random modulations are shown in Figs. 5 and 6 respectively. The sum of the pGAWs is $y_\Sigma(t)$, i.e., the output of the synthesizer, which is input to the analysis-by-synthesis. (b) Overview of the analysis-by-synthesis method, referred to as 'modulated pulse trains decomposition' (MPD). Fundamental frequency candidates f_o^γ with candidate index γ are extracted from the input GAW $y_I(t)$. pGAW candidates $y^\gamma(t)$ are synthesized for each f_o^γ. A detailed block diagram of the pGAW synthesizer is shown in Fig. 7. Candidates are selected via superposition of pGAWs and minimizing the time-domain fitting error $e(t)$ which is fed back to the candidate selection process. The error is also fed back to the pGAW synthesizer if modulation analysis is switched on (mod-switch). The output of the analysis-by-synthesis is referred to as model GAW.

not fully mature. This results in a lack of understanding of the causal relations between abnormal vocal fold vibration and auditory consequences. For example, diplophonia is a type of voice in which two simultaneous pitches are auditorily perceived. It may be a sign of a voice disorder and is most commonly caused by the spatial splitting of the vocal folds into two parts, each of which vibrates at a different fundamental frequency (f_o). Little attention has been paid on the properties of amplitude and frequency modulations of individual diplophonic glottal oscillators, but they may have relevant effects on auditory perception. In particular, modulations may influence auditory detection of diplophonia, as they influence perceived pitch strength [3]. The knowledge of modulation properties may thus aid clinical assessment of diplophonic voice as well as the investigation of its pathophysiology and contribute to the planning of treatment for individual patients.

In this paper, we propose a kinematic model-based approach to synthesis and analysis-by-synthesis of GAWs. The kinematic model does not use classical physiological parameters typically used in dynamical modeling such as subglottal pressure, vocal folds' masses, stiffnesses and damping parameters, because in vivo validation data are not available for these. Analysis-by-synthesis is tested using as input (i) synthetic GAWs obtained with a stand-alone synthesizer, and (ii) natural GAWs obtained with a laryngeal high-speed camera. The analysis-by-synthesis is evaluated by comparing its output, i.e., a model GAW, with the input GAW. A modulating and a non-modulating version of the analysis-by-synthesis are compared with two data-driven approaches, i.e., a deep autoencoder and WaveGlow [25]. The synthesis of GAWs and their analysis-by-synthesis are overviewed in Fig. 2(a) and (b) respectively. Input GAWs are synthesized as follows. First, up to two f_os are obtained by a hidden Markov model (HMM). The f_os are allowed to switch on and off independently of each other and up to two f_os may be active simultaneously. An example of a phonation's f_o-tracks is shown in Fig. 3(a). A frequency and amplitude modulated partial GAW (pGAW), i.e., a train of single-peak pulses, is obtained for each f_o using a Fourier synthesizer with a pulse model. The input GAW is a superposition of the pGAWs. The synthesis is described in detail in Section III-A.



(a)

(b)

Fig. 3. (a) Example of a vowel's f_o-tracks generated using a hidden Markov model (HMM). Three f_o-tracks are generated. The number of simultaneous f_os ($\#f_o$) is either 0 (unvoiced), 1 (monophonic), or 2 (diplophonic). (b) The state diagram of the HMM used to obtain the f_o-tracks shown in (a). The state of inactivity is s_1, i.e., no f_o, or unvoiced. The other states s_2 to $s_{\#_s}$ map to $f_o = 100$ Hz to $f_o = 350$ Hz. The HMM is used in the synthesis of input GAWs, which is overviewed in Fig. 2(a).

The analysis-by-synthesis is referred to as 'modulated pulse trains decomposition' (MPD). In analysis-by-synthesis, input GAWs are fitted using a modified version of the synthesizer combined with means for extracting synthesis parameters from input GAWs. In particular, a few f_o candidates and corresponding pGAWs candidates are obtained by spectral analysis/synthesis. The pGAW candidates are superimposed and heuristically selected by minimizing the time-domain fitting error, i.e., the samplewise difference between the input GAW and the model GAW. Modulation analysis is included in a separate step that refines the model GAW, where the time instances and magnitudes of the pGAW pulses' maxima are optimized systematically pulse-by-pulse. The optima of both the maximas' time instances and magnitudes are also found by minimizing the time-domain fitting error. The analysis-by-synthesis approach is described in detail in Section IV. For convenience, we refer to time instances and magnitudes of the pGAW pulses' maxima as 'times' and 'heights' respectively throughout the manuscript.

The remainder of the article is structured as follows. In Section II, related prior work is summarized. The corpora are described in Section III, which includes the description of the synthesis. The analysis-by-synthesis is described in Section IV. The benchmarks, i.e., a deep autoencoder neural network and the WaveGlow neural network, are introduced in Section V. In Section VI, the model performance is reported with regard to groups of modulation size and voice quality type. Additionally, some qualitative comparisons are shown. A discussion and a conclusion are provided in Section VII.

II. Related Work on Vocal Fold Modeling and Speech Synthesis

Dynamical models have been used to study voice production throughout the past decades, but the GAW is kinematic only. Particular emphasis was given in the past to the dynamical 'two-mass-model' (2MM), which comprises two coupled mass-spring-damper oscillators per vocal fold. Much effort has been put into understanding the relation between (aero-)dynamical parameters and the resulting kinematics. A particular task was to model diplophonic voices in silico, by using two coupled oscillators with different natural frequencies. In the following, dynamical models capable of producing diplophonic waveforms are reviewed.[1]

Ishizaka and Isshiki proposed a dynamical model that used two coupled mass-spring-damper systems per vocal fold, i.e., the so-called two-mass model (2MM) [4]. The model was capable of making the vocal folds lax or tense individually. Thus, vocal fold thickness and stiffness were controlled, which enabled the modeling of changes in vocal fold stretch and tension. Perceived diplophonia was reported.

Steinecke and Herzel used a similar model to simulate superior and recurrent laryngeal nerve paralysis individually, which was associated with decreased activity of the cricothyroid muscle and the vocalis muscle respectively [5]. The cricothyroid stretches the vocal fold, resulting in scaling up the springs and scaling down the masses. The vocalis muscle is the body of the vocal folds and contracts them. The lower masses and springs were thought to represent the folds' bodies. An $i:j$ labeling was used, where $i, j \in \mathbb{N}$ are the numbers of distinct maxima of the left and the right fold within a period, reflecting (i) frequency ratios if $i \neq j$ and (ii) entrainment/phase locking if $i = j$. There, 'period' refers to the periodicity of the whole system, as opposed to periods or cycles of individual oscillators.

Later investigations using dynamical models include the following. Mergell *et al.* used the 2MM with a time-variant ratio of the vocal folds' natural frequencies [6]. In particular, an exponentially decaying function was manually fitted to the ratio, resulting in alternation of commensurate and incommensurate vocal fold frequencies. Döllinger *et al.* proposed an automatic fitting procedure that minimizes a spectral distance measure between reference and model trajectories [7]. Tokuda and Herzel trained three-layer feedforward neural networks to output waveforms that are similar to the outputs of the 2MM [8]. Lucero *et al.* used one mass-spring-damper system per vocal fold [9]. Good agreement with oscillatory phenomena observed in water-filled latex tube vocal fold replicas was reported. In contrast to modeling left-to-right differences, anterior-posterior differences were also considered in higher dimensional models than the 2MM [10]–[12].

Another type of mechanical models are kinematic models. Kinematic models and analyses relevant to diplophonia are reviewed as follows. The phase-delayed overlapping sinusoid (PDOS) model uses one sinus for the upper margin and one for the lower margin of each vocal fold at a particular sagittal position [2]. The upper sinus is delayed with respect to the lower sinus. Using more general cyclic shapes of trajectories of vocal fold margins, it was fitted in a least squares sense to observed data [13]. Additionally, the PDOS model was used to synthesize GAWs of diplophonic phonations in the absence of modulations [14]. The kinematic model of the mucosal wave by Kumar *et al.* can also be understood as an extension to the PDOS model, since it uses a multitude of surface points instead of only two margins, and includes vertical components of vibration [15]. The Fourier transform was applied to pixel-intensity time series of laryngeal high-speed videos to obtain false color images of vibration frequency [16], [17]. These images enabled visual evaluation of glottal vibration frequencies. In particular, the existence of simultaneous f_os was enabled without making any assumptions about their spatial arrangement, i.e., left-to-right, anterior-to-posterior, or mixed. Empirical orthogonal function analysis of vocal fold deflections was also proposed, which was based on a vibration mode model of the vocal folds [18], [19]. The model uses vibration modes similar to the modes observed in the vibration of strings and membranes. In the analysis, the trajectories are projected onto the orthogonal eigenvectors of the trajectories' covariance matrix, which decomposes the trajectories into spatial modes, i.e., so-called 'topos,' and temporal driving functions, i.e., so-called 'chronos'. The use of a mixture of two pulse trains was proposed for modeling diplophonic GAWs, where sawtooth-shaped and sinusoidal amplitude modulators were used [20]. A few observed GAWs were fitted by manually adjusting the model parameters, and perceptual experimentation was carried out. The distinction between deterministic and random noise was proposed for single-f_o GAWs [21] and impulse sequences [22].

Much improvement of the naturalness of synthesized speech has recently been accomplished by using large neural networks. For example, 'WaveNets' take as input raw audio, and are trained to predict in each time step the next sample from the previous samples, i.e., by means of autoregressive signal modeling [23]. A WaveNet uses a deep feedforward dilated convolutional neural network that enables modeling long-term and short-term interdependencies of the samples. WaveNets were used to create raw audio including speech and music, but also to model the voice source input to a vocal tract filter [24]. In this work, we compare MPD to a deep autoencoder, as well as WaveGlow [25], which extends WaveNet.

An f_o-tracker for up to two simultaneous f_os observed in diplophonic voices was proposed, trained, and tested in [26]. A clinical corpus of 29 vowels recorded during laryngeal high-speed video endoscopy was used. The f_o tracker obtained f_os from input audio signals by (i) spectral analysis and multiple

[1] It should be noted that due to terminological divergences, their output was not always termed 'diplophonic' in past publications, even if two simultaneous f_os were observed.

execution of the Viterbi algorithm, (ii) Fourier synthesis of candidate waveforms, and (iii) heuristic candidate selection using majority voting. In particular, f_os were extracted by fitting audio waveforms in a least squares sense. Median error rates between 6.52% and 11.11% were observed using reference f_os obtained by marking times of vocal folds' maximal lateral deflections in videokymograms.

We used an earlier version of MPD for the purpose of detecting extra pulses from 125 synthetic GAWs [27] and for modeling modulation of pulse times and heights in diplophonic pGAWs and audio waveforms [28], where it was tested using 29 clinically observed vowels.

III. Corpora

We use two synthetic corpora and a natural corpus of input GAWs in this work.

The first synthetic corpus is used for testing the MPD and for training and testing the deep autoencoder. 864 input GAWs are generated using the synthesizer described in Section III-A. Each input GAW is 2.048 s long. The corpus is organized in three groups of deterministic modulation sizes M_d, each of which is further divided into three groups of random modulation sizes M_r. Thus, the 3 × 3 modulation size groups contain 96 input GAWs each. M_d and M_r are uniformly distributed and lie between 0 and 5% for small modulation, 5 and 10% for medium size modulation, and 10 and 20% for large modulation. The second synthetic corpus has the same properties, but is 20 times larger. It is used for training and testing WaveGlow (see Section V-B).

The natural corpus is sampled from a database of laryngeal high-speed videos [29]. A total of 72 natural input GAWs from 35 different patients are used. 29 of these input GAWs contain diplophonation, whereas the remaining 43 do not contain diplophonation but a wide variety of other dysphonic voice qualities. The corresponding videos are 2.048 seconds long. Fig. 1 shows an illustration of the video recording. The endoscope was inserted through the mouth of a patient into the pharynx, while the tongue was softly held by a clinician. The endoscope illuminated the larynx. The video was previewed during examination in real time on a computer monitor, and the endoscope was positioned to allow direct sight onto the vocal folds. The patient was instructed to phonate an /i/ such that the epiglottis moved out of sight. An HRES ENDOCAM 5562 (Richard Wolf GmbH, Knittlingen, Germany) was used with a video frame rate of 4 kHz and a spatial resolution of 256x256 pixels. RGB color videos were recorded. Input GAWs were extracted from the videos using a software [30] for graphical segmentation that facilitates seeded region growing described in [31], [32]. The modulation sizes M_d and M_r are unknown for the natural corpus.

A. Synthesis

Input GAWs of sustained diplophonations are synthesized as outlined in the overview shown in Fig. 2(a). First, f_o-tracks are generated using an HMM. Second, deterministic and random modulators of instantaneous frequencies and amplitudes are obtained. Third, prototype pGAW pulses are obtained using a single-peak GAW pulse model with five parameters [33]. Finally, modulated pGAWs are obtained using a Fourier synthesizer and added together. A detailed explanation is given in the sequel.

1) HMM-Based Genesis of f_o-Tracks: An example of a vowel's f_o-tracks generated using the synthesis HMM is shown in Fig. 3(a). The HMM is applied to obtain simultaneous f_o-tracks as follows. Sequences $Q = \{q_1, q_2, \ldots, q_{\#_f}\}$, where $\#_f$ is the number of frames, are drawn from random processes reflected by conditional distributions encoded in the transition probability matrix $A_{i,j}$. One sequence corresponds to one f_o-track. The length of the sequences is 2.048 s, and the frame duration is 16 ms. The f_os of one vowel are obtained by repeated sampling of state sequences from the HMM. In particular, the sampling of sequences Q is repeated for each vowel until two sequences are active simultaneously in at least 10 % of the vowel's duration.

Fig. 3(b) shows the state diagram of the synthesis HMM and the transition probability distributions. The state space $S = \{s_1, s_2, \ldots, s_{\#_s}\}$ spans $\#_s$ discrete f_o states. State s_1 is the state of inactivity, i.e., no f_o, and the remaining states are active states. Active states linearly map the f_os from 100 Hz to 350 Hz in steps of 0.5 Hz, resulting in the number of states $\#_s = 502$. The random variable $q_n \in S$ corresponds to the f_o at frame n. Voice onsets and transitions between voiced states, i.e., states between s_2 and $s_{\#_s}$, are modeled using Gaussian probability distributions p_1 and p_2, respectively. The distribution p_2 preserves the instantaneous trend of f_o, i.e., Δ_{trend}. In particular, the mean of p_2 is increased by Δ_{trend}. The probabilities of staying inactive and of voice offset are p_α and p_β, respectively.

The currently sampled sequence Q is discarded if one of the following applies: (i) short f_o-tracks and exclusively inactive tracks, i.e., sequences with fewer than two elements unequal s_1 are discarded, (ii) to avoid clipping f_os, sequences that include s_2 or $s_{\#_s}$ are discarded, (iii) to maintain the number of simultaneous f_os less or equal to two, the currently sampled sequence is discarded if more than two sequences include active states in the same frame, and (iv) to avoid small f_o differences between simultaneous f_o-tracks, the currently sampled sequence is discarded if the smallest relative frequency difference between simultaneous f_o-tracks is below 5% at any time.

In summary, we obtain up to two simultaneous f_o-tracks using an HMM. In contrast to the HMM used previously [26], the HMM maintains the f_o trend, which enables the generation of realistic f_o-tracks.

2) Modulation: This section introduces and motivates the deterministic and random modulation of frequency and amplitude. Fig. 4 shows a kymogram of a representative diplophonic vocal fold vibration from [29]. The top and bottom of the image show the left and right vocal fold respectively, which vibrate at different frequencies. The dark area in between is the glottal gap, i.e., the space between the vocal folds. Lateral maximal deflections of the vocal folds are marked by green crosses. Medial maximal deflections of the right and left vocal folds are marked by blue and red crosses, respectively. The vocal folds vibrate at different frequencies. In particular, four cycles are observed in the left vocal fold, while five cycles are observed in the right vocal fold. Black vertical lines are

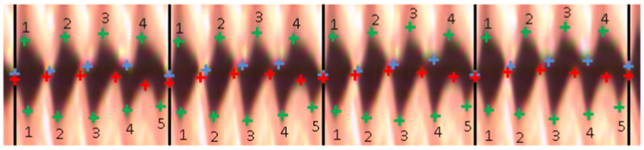
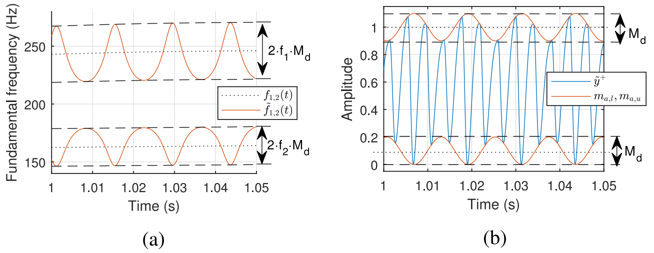


Fig. 4. Kymogram of a representative diplophonic phonation annotated with regard to cyclicity and modulation over time. The crosses mark lateral and medial maximal deflections of the vocal folds.



(a) (b)

Fig. 5. (a) Deterministic frequency modulation for a modulation size $M_d = 0.2$. Shown are the unmodulated frequencies $f_1(t)$and $f_2(t)$, the modulated frequencies $\tilde{f}_1(t)$ and $\tilde{f}_2(t)$, and the boundaries of modulation (dashed lines). (b) Deterministic amplitude modulation of a pGAW for a modulation size $M_d = 0.2$. Shown are a modulated pGAW pulse train $\tilde{y}$ (blue curve) and the lower and upper boundaries of modulation $m_{a,l}$ and $m_{a,u}$ (red curves).

boundaries of metacycles, i.e., cycles of the whole dynamic system, as opposed to cycles of the individual oscillators. The left and right vocal fold cycles are enumerated from 1 to 4 and 1 to 5, respectively, starting with 1 for the first cycle after the boundary of the metacycle. Deterministic modulation of magnitudes and times of deflection maxima are observed at the frequency of the metacycle. Deterministic modulations are superimposed with random perturbations thereof.

Modulation of the magnitudes and times of maximal lateral and medial deflections are observed. In particular, for both vocal folds, the green crosses '3' marking maximal lateral deflections are located most laterally among all green crosses, i.e., closest to the top or bottom. For the right vocal fold, the green crosses '5' marking the maximal medial deflections are located most medially among all green crosses, i.e., closest to the vertical center of the image. For the left vocal fold, either 1 or 4 is most medial. Regarding maximal medial deflections of the left vocal fold, the blue crosses '1' are most medial, and the blue crosses '4' are most lateral. Regarding maximal medial deflections of the right vocal fold, the most medial red crosses are '1' or '5,' and the most lateral red crosses are '3'. In summary, deterministic modulations of times and magnitudes of maximal deflections are observed at the frequency of the metacycle, and they are superimposed with random perturbations. These deterministic and random modulations were not considered in our previous kinematic model of diplophonic vocal fold vibrations [26].

a) Deterministic modulation. The frequency and amplitude are modulated deterministically using modulators $m_f(t)$ and $m_a(t)$, respectively. The modulators are the sine and cosine of the glottal oscillators' instantaneous phase difference, respectively. M_d is the size of deterministic modulation.

The modulated instantaneous phases of the pGAWs are obtained iteratively by Algorithm 1. The frequency modulator $m_f(t)$ is obtained as the sine of the pGAWs' instantaneous phase difference, which creates the need of a recursive algorithm. The modulated pGAWs' instantaneous frequencies $\tilde{f}_{1,2}(t)$ are obtained using M_d, i.e., here the relative maximum increase or decrease of instantaneous frequency. The modulated pGAWs' instantaneous phases $\tilde{\Theta}_{1,2}(t)$ are obtained by cumulating the modulated instantaneous frequencies of the pGAWs. The instantaneous phases of the pGAWs are iteratively updated until convergence or until 100 iterations have been executed. Convergence is determined by using root mean squared differences c between phases of the current and the previous iteration. In particular, the iteration stops if the current c is smaller than 0.1% of its cumulated sum. If c does not converge within 100 repetitions, the procedure restarts with a new M_d. Examples of modulated instantaneous frequencies of pGAWs are shown in Fig. 5(a).

Algorithm 1: Deterministic Frequency and Phase Modulation.

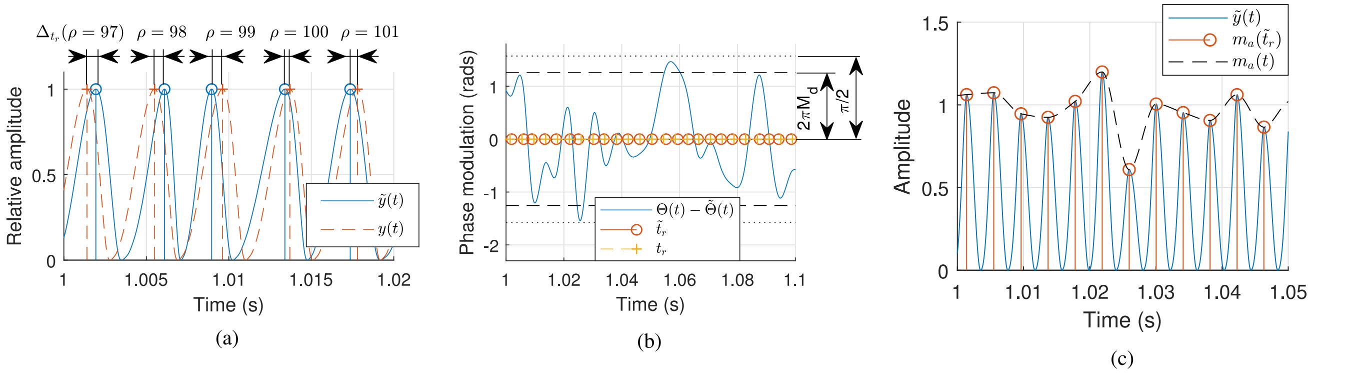
```
$\tilde{\Theta}_{1,2}(t) \leftarrow \Theta_{1,2}(t)$
**repeat**
  $M_d \leftarrow \mathcal{U}(b_l, b_u)$
  **repeat**
    $i \leftarrow 1$
    $m_f(t) \leftarrow \sin[\tilde{\Theta}_1(t) - \tilde{\Theta}_2(t)]$
    $\tilde{f}_1(t) \leftarrow f_1(t) \cdot [1 - M_d \cdot m_f(t)]$
    $\tilde{f}_2(t) \leftarrow f_2(t) \cdot [1 + M_d \cdot m_f(t)]$
    $\tilde{\Theta}^{old}_{1,2}(t) \leftarrow \tilde{\Theta}_{1,2}(t)$
    $\tilde{\Theta}_{1,2}(t) \leftarrow \sum_{t'=0}^{t} \tilde{f}_1(t')$
    $c(i) \leftarrow rms[\tilde{\Theta}^{old}_1(t) - \tilde{\Theta}_1(t)] +$
      $rms[\tilde{\Theta}^{old}_2(t) - \tilde{\Theta}_2(t)]$
    $i \leftarrow i + 1$
  **until** $c(i) < 0.1\% \cdot \sum_{i'=1}^{i} c(i')$ **or** $i = 100$
**until** $i < 100$
**return** $\tilde{\Theta}_{1,2}(t)$
```

For deterministic modulation of amplitudes illustrated in Fig. 5(b), the modulator $m_a(t)$ is obtained as the cosine of the pGAWs' instantaneous phase difference. A pGAW $y(t)$ provided by the Fourier synthesizer is modulated as $\tilde{y}(t) = y(t) \cdot \{1 - M_d \cdot [0.5 - m_a(t)]\} + (M_d/2) \cdot [1 - m_a(t)]$.

In summary, we propose modulating amplitude and frequency based on the cosine and sine of the difference of the pGAWs' instantaneous phases. This couples the individual glottal oscillators.

b) Random modulation. In addition to deterministic modulations, pGAW pulse times and heights are randomly modulated pulse-to-pulse independently. Time shifts and scaling factors are drawn separately for each pGAW pulse from truncated Gaussian distributions. Random modulations are motivated, e.g., in [34], [35], and described as follows.

Regarding random modulation of frequency, an example of which is shown in Fig. 6(a), pGAW pulse times t_r are obtained from pGAWs' instantaneous phases $\Theta(t)$. pGAW pulse



(a)

(b)

(c)

Fig. 6. (a) Random frequency modulation. Modulated and unmodulated pGAWs (curves) and pulse times (stems) are shown. pGAW pulse times are indicated by dashed and solid stems for the unmodulated and modulated pGAWs, respectively. (b) Random phase modulation. Shown are the phase modulation $\Theta(t) - \tilde{\Theta}(t)$, i.e., the phase difference between the unmodulated pGAW and its modulated version, the intervals $\pm 2\pi M_d$ and $\pm\pi/2$, and pGAW pulse times $\tilde{t}_r$ and t_r of the modulated and unmodulated pGAW, respectively. (c) Random amplitude modulation. Shown are the modulated pGAW $\tilde{y}(t)$, and the modulator $m_a(t)$ obtained by interpolating the supporting points $m_a(\tilde{t}_r)$.

times t_r are times at which $\Theta(t)$ is an odd integer multiple of π, i.e., $\Theta(t_r) = \pi, 3\pi, 5\pi, \ldots$. Modulated pGAW pulse times $\tilde{t}_r$ are obtained by adding to t_r a jitter vector, i.e., $\tilde{t}_r(\rho) = t_r(\rho) + \Delta_{t_r}(\rho)$, where ρ is the pulse index. The elements of a jitter vector are random numbers drawn from a truncated zero-mean Gaussian distribution with a standard deviation of M_r, which is the jitter size given as a fraction of the cycle length T_0. The distribution is truncated at $\pm T_0/4$. A randomly modulated pGAW's instantaneous phase $\tilde{\Theta}(t)$ is obtained at the pGAW pulse times $\tilde{t}_r$ as $\tilde{\Theta}(t = \tilde{t}_r) = \pi, 3\pi, 5\pi, \ldots$ and by spline interpolation in between. Spline interpolation results in a differentiable instantaneous frequency, in contrast to piecewise-constant or piecewise-linear interpolation proposed in [36]. An example of phase modulation resulting from random frequency modulation is shown in Fig. 6(b).

Regarding the random modulation of amplitude illustrated in Fig. 6(c), a modulated pGAW $\tilde{y}(t) = m_a(t) \cdot y(t)$, where $m_a(t)$ is a modulator, and $y(t)$ is a carrier pGAW. The modulator's supporting points at pGAW pulse times t_r are drawn from a Gaussian distribution with a mean of 1 and a standard deviation of M_r. The distribution is truncated to the interval [0.1, 1.9]. The modulator is interpolated using shape-preserving piecewise cubic splines between pGAW pulse times.

In summary, we propose to randomly modulate the timing of pGAW pulses (resulting in modulation of frequency and phase), as well as pGAW pulse heights on a pulse-to-pulse time scale. Timing of pGAW pulses is modulated by small random time shifts, and amplitude is modulated by scaling pGAW pulse heights by random numbers close to 1.

3) Fourier Synthesis of pGAWs: Prototype pGAW pulses r_l are obtained using the single-peak GAW pulse model by Chen *et al.* [33]. Here, the samples of r_l are shifted circularly such that the maximum of r_l is centered, i.e., $\text{argmax}(r_l) = T_0/2$. This aligns maxima of the pGAWs with instantaneous phases that are odd integer multiples of π.

pGAWs are obtained using a Fourier synthesizer. In particular, prototype pGAW pulses r_l and instantaneous frequencies are provided to the synthesizer. The pGAW pulses are Fourier transformed to obtain their Fourier coefficients $R_k \in \mathbb{C}$, which are fed into the Fourier synthesizer together with the pGAW's instantaneous phase $\Theta(t)$. A pGAW is obtained as

$$y(t) = \Re(R_0) + \sum_{k=1}^{10} \{\Re(R_k) \cdot \cos[k \cdot \Theta(t)] + \Im(R_k) \cdot \sin[k \cdot \Theta(t)]\} \quad (1)$$

using the real and imaginary parts of R_k, i.e., $\Re(R_k)$, and $\Im(R_k)$, where k is the discrete frequency given in integer multiples of f_o. The use of Fourier coefficients R_k models a pGAW pulse shape as a sum of sinusoids, the frequencies of which are integer multiples of f_o. To enable smooth modulation of pGAW pulse shapes, the coefficients R_k are upsampled before synthesis from 62.5 Hz to 50 kHz, using shape-preserving piecewise cubic spline interpolation.

4) Superposition of pGAWs: To obtain the input GAWs, the pGAWs are superposed in the time-domain as $y_\Sigma(t) = a_\Sigma \cdot [a_1 \cdot y_1(t) + a_2 \cdot y_2(t)]$, where a_1 and a_2 are random scaling factors drawn from a truncated Gaussian distribution. The distribution has a mean of 0.3, a standard deviation of 0.15 and is truncated to the interval [0.1, 0.6]. The scaling factor $a_\Sigma = \sqrt{1/(a_1 + a_2)}$ when two f_os are active simultaneously, and $a_\Sigma = 1$ elsewhere. Smooth switching is applied.

The additional scaling factor a_Σ enables nonlinear superposition of the pGAWs, which has the following physiological interpretation. When only one f_o is active, all parts of the vocal folds vibrate in synchrony, and the whole glottal area is assigned to only one f_o. When two f_os are active, a part of the vocal folds vibrate at a different frequency than the rest of the vocal folds, e.g., the left and the right vocal folds vibrate at different frequencies and share the glottal area. Qualitative comparisons have shown that pGAWs scaled in the proposed nonlinear way enable better fitting of natural GAWs than linearly superposed pGAWs. In particular, a decrease in the overall amplitude is needed if two pGAWs with large amplitudes are superposed, whereas an increase in the overall amplitude is needed if two gGAWs with small amplitudes are superposed.

In summary, we propose using scaling factors that enable the control of relative amplitudes of pGAWs and nonlinear superposition of the pGAWs.

IV. Analysis-By-Synthesis

An analysis-by-synthesis method denoted as 'modulated pulse trains decomposition' (MPD) is applied for fitting the input GAWs. Fig. 2(b) shows an overview of the analysis-by-synthesis method. First, f_os and pGAWs are estimated. The fitting involves (i) extraction of f_o candidates by using spectral peak picking and an HMM, (ii) genesis of pGAW candidates for each of the f_o candidates via Fourier synthesis, and (iii) selection of f_os and pGAWs via superposition. f_o candidates are extracted using the HMM combined with repetitive execution of the Viterbi algorithm. Ten sequences Q^{γ} are obtained by running Viterbi ten times, and f_o candidates f_o^{γ} are obtained by mapping Q^{γ} back to the frequency-domain, where γ is the candidate index. Instead of selecting f_os and pGAWs via majority voting-based preselection combined with brute force [26], a candidate-wise switching method [27] is adapted here to allow up to two simultaneous f_os. Second, the model is refined by enabling modulations of the pGAW pulse times and prominences by closing the mod-switch shown in Fig. 2(b).

In summary, an analysis-by-synthesis method specialized on GAWs of diplophonic voices is proposed. The features to which the analysis-by-synthesis maps are (i) up to two simultaneous f_o-tracks, (ii) pGAW pulse shapes that vary on a frame-by-frame time scale, and (iii) fluctuations of pGAW pulse times and heights on a pulse-by-pulse time scale. Details of the analysis-by-synthesis are disclosed hereafter.

A. Genesis of pGAW Candidates (pGAW Synthesizer)

For each f_o candidate, a candidate pGAW $y^{\gamma}(t)$ is obtained by Fourier synthesis. Estimated pGAW pulse shapes and instantaneous frequencies are provided to the Fourier synthesizer. Fig. 7 shows an overview of the pGAW synthesizer used in the analysis-by-synthesis. The pGAW pulse shapes are estimated and provided to the Fourier synthesizer that is driven by the instantaneous frequency.

First, the input GAW $y_I(t)$ is scaled to peak-to-peak amplitudes of 1e-3, and its moving average is subtracted.

Second, instantaneous phases of the quasiunit-pulse train $u(t)$ and the input GAW $y_I(t)$ are linearized as described in Appendix A.

Third, the pGAW pulse shapes r_l are obtained by normalized cross-correlation of a quasiunit-pulse train $u(t)$ with the input GAW $y_I(t)$, i.e., $r_l = (1/\Sigma u(t)) \cdot \Sigma_n[u(t) \cdot y_I(t-l)]$. The unit-pulse train $u(t)$ is obtained as $u(t) = \Sigma_{\rho}\delta(t - \rho \cdot T_0 - \Delta_{\Phi})$, where δ is a unit-pulse, i.e., $\delta(t=0) = 1$ and $\delta(t \neq 0) = 0$, ρ is the pulse index, the fundamental period $T_0 = \lfloor f_s/f_o \rceil$ (samples), and the phase shift $\Delta_{\Phi} = \text{argmax}(r_l)$. The pulse shape r_l and the shift Δ_{Φ} are iteratively updated until $\text{argmax}(r_l) = 0$, or 10 times at most. As a result of shifting by Δ_{Φ}, the quasiunit-pulses align with local maxima of the pGAW. The pGAW pulse shapes r_l are obtained using Hann-windowed unit-pulse trains and input GAWs that are 32 ms long and overlap by 16 ms.

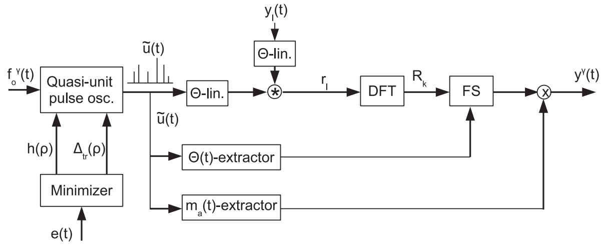


Fig. 7. pGAW synthesizer used in the analysis-by-synthesis (see also Fig. 2(b)). To extract the time-domain pGAW pulse shape r_l, a phase-linearized version of the quasiunit-pulse train $\tilde{u}(t)$ with frequency $f_o^{\gamma}(t)$ is cross-correlated with a phase-linearized version of the input GAW $y_I(t)$. The amplitude modulation vector $h(\rho)$ and the pGAW pulse times modulation vector $\Delta_{t_r}(\rho)$ are estimated by minimizing the time-domain fitting error $e(t)$. The instantaneous phase $\Theta(t)$ and the amplitude modulator $m_a(t)$ are extracted from $\tilde{u}(t)$. The pGAW pulse shapes' Fourier coefficients R_k and the phase $\Theta(t)$ are provided to a Fourier synthesizer (FS). The output of the FS is multiplied by $m_a(t)$, which yields the candidate pGAW $y^{\gamma}(t)$.

Fourth, the pGAWs' instantaneous phases $\Theta(t)$ at pGAW pulse times t_r are odd integer multiples of π, i.e., $\Theta(t_r) = \pi, 3\pi, 5\pi, \ldots$ and are obtained by spline interpolation in between. The corresponding pGAW pulse times of overlapping frames are weighted averaged to obtain $t_r = [w_1(t_{r,1}) \cdot t_{r,1} + w_2(t_{r,2}) \cdot t_{r,2}]/\{2 \cdot [w_1(t_{r,1}) + w_2(t_{r,2})]\}$, where w_1 and w_2 are Hann-windows of overlapping frames, and $t_{r,1}$ and $t_{r,2}$ are times of corresponding pGAW pulses of overlapping frames.

Fifth, a candidate pGAW $y^{\gamma}(t)$ is obtained for each f_o candidate by Fourier synthesis using (1). The pGAW pulse shapes r_l are Fourier transformed as $r_l \overset{\text{DFT}}{\leftrightarrow} R_k$, with the discrete frequency $k = 1, 2, \ldots, 10 \cdot f_o$, and are upsampled before synthesis from 62.5 Hz to 50 kHz using shape-preserving piecewise cubic spline interpolation.

In summary, candidates of pGAWs are obtained. They are compared to the input GAW as described below. In particular, up to two candidate pGAWs are added together and compared to the input GAW. pGAW candidates are selected, which belong to the model GAW that is optimal in a least squares sense.

B. Candidate Selection

A switching method is used for candidate selection. The length of the binary candidate selection vector $K = \{\kappa_{\gamma}\}_{\gamma=1}^{\gamma=\Gamma} \in \{0, 1\}$ equals the number of candidates Γ. The optimal candidate selection vector K_{opt} minimizes objective Δ_K, i.e., $K_{opt} = \text{argmin}(\Delta_K)$. Candidates are selected for which κ_{γ} of K_{opt} is 1. In particular, the model GAW $y_M(t) = \Sigma_{\gamma}[\kappa_{\gamma} \cdot y^{\gamma}(t)]$ is the superposition of all selected candidate pGAWs. The objective Δ_K is the sum of the time-domain fitting error level and a penalty, i.e., $\Delta_K = L_e + \lambda_{pen} \cdot \frac{\#n_{\nu}(K)}{\#_n}$, where $L_e = 20 \cdot \log_{10}[\sqrt{\overline{e_K^2(t)}}/\sqrt{\overline{y^2(t)}}]$ with the fitting error $e_K(t) = y(t) - y_M(t)$, the penalty parameter λ_{pen}, the number of voiced frames $\#n_{\nu}(K)$, and the total number of frames $\#_n$. The penalty parameter $\lambda_{pen} = 1.0108$ dB, which is an optimum found earlier [26], and $\#n_\nu(K)$ counts twice frames with two simultaneous f_os. The penalty's purpose is to regularize the model such that model GAWs with fewer selected candidates are preferred.

The optimal candidate selection vector K_{opt} is found as follows. First, K is initialized as a zero vector. Second, candidate pGAWs $y^\gamma(t)$ are sorted with regard to their energy. Third, the states of the individual elements of K are switched one by one in the order of their energies starting with the candidate with the highest energy. After each switch, it is determined whether (i) at most two f_os are active simultaneously and (ii) the objective Δ_K decreased. If either of the requirements (i) or (ii) are not fulfilled, the most recent switch is reverted. Switching is repeated in a loop until convergence, i.e., until no decrease in Δ_K is observed anymore.

In summary, a linear superposition of previously obtained candidate pGAWs is found, which yields the best resynthesis of the input GAW in a least squares sense. In the following, the selected pGAWs are modulated to further minimize the error.

C. Modulation of pGAW Pulse Times and Heights

The model GAW $y_M(t)$ obtained in the previous section is refined by finding pGAW pulse times and heights that minimize the fitting error, which is fed back to the pGAW synthesizer by closing the mod-switch shown in Fig. 2(b). In particular, the refined model GAW is obtained using for each pGAW a modulated quasiunit-pulse train $\tilde{u}(t) = \Sigma_\rho h(\rho) \cdot \delta(t - \rho \cdot T_0 - \Delta_{t_r}(\rho) - \Delta_\Phi)$ instead of $u(t)$, where $h(\rho)$ are supporting points of the amplitude modulator $\tilde{A}(t)$, and $\Delta t_r(\rho)$ are time shifts of individual quasiunit-pulses. The pGAWs are obtained by modifying (1) such that

$$\tilde{y}(t) = \tilde{A}(t) \cdot \{\Re(\tilde{R}_0) + \sum_{k=1}^{10} [\Re(\tilde{R}_k) \cdot \cos[k \cdot \tilde{\Theta}(t)] + \Im(\tilde{R}_k) \cdot \sin[k \cdot \tilde{\Theta}(t)]]\}, \tag{2}$$

where $\tilde{A}(t)$ is an amplitude modulator, $\tilde{R}$ are the Fourier coefficients of the pGAW pulse shape, and $\tilde{\Theta}(t)$ is the pGAW's instantaneous phase.

The amplitude modulation vector $h(\rho)$ is initialized as 1 for all ρ and time shifts $\Delta_{t_r}(\rho)$ are initialized as 0. The active-set algorithm [37] is used to find $h(\rho)$ and $\Delta_{t_r}(\rho)$ for each Hann-windowed 32 ms frame separately. In frames with one f_o, $h(\rho)$ and $\Delta_{t_r}(\rho)$ are estimated for individual pGAW pulses, whereas in frames with two f_os, $h(\rho)$ and $\Delta_{t_r}(\rho)$ of a pulse of the faster pGAW are estimated together with $h(\rho)$ and $\Delta_{t_r}(\rho)$ of the closest pulse of the slower pGAW. Hence, in frames with one f_o, two parameters are jointly estimated, i.e., the height and time shift of one pGAW pulse, whereas in frames with two f_os, four parameters are jointly estimated, i.e., the heights and time shifts of one pulse of each pGAW. Optimization is conducted for individual pGAW pulses of a frame in a loop that is repeated as long as the framewise improvement of L_e is larger than 0.01 dB.

In summary, the difference between the input GAW and the model GAW is minimized by systematically varying times and heights of pGAW pulses.

V. Benchmarks

A. Deep Autoencoder

A deep autoencoder is used as a benchmark for comparison. The data are sampled at 16 kHz and 32 ms long Hann-windowed frames with an overlap of 16 ms are used. Thus, the input layer size is 512. The subsequent layers have 233, 106, 48, 106, 233, and 512 units. Thus, in the narrowest layer, i.e., the bottleneck, the input signals are encoded into 48 coefficients per frame.[2] Input signals are normalized to a minimum of 0 and a maximum of 1. The layers with 233 units use a logsig activation function, whereas the remaining layers' activations are linear units clipped to 0 and 1. After overlap-and-add, the output GAWs are compared to the input GAWs by means of L_e in dB.

The autoencoder is first trained using the synthetic corpus containing 864 input GAWs. The synthetic data are randomly split into a 70% training set, a 15% validation set, and a 15% test set. The network parameters are obtained by using scaled conjugate gradient backpropagation. The mean squared error with weight decay (L2 norm) and sparsity regularization is used as the loss function. After the network parameters for the synthetic data are obtained, they are used as a starting point for fine tuning with the natural data. The natural data are randomly split for fine tuning into a 50% training set, a 25% validation set, and a 25% test set.

B. Waveglow

WaveGlow is used as an additional benchmark [25]. It is a combination of Glow [38] and WaveNet [23]. WaveGlow learns a multilayer neural network function that maps a zero-mean spherical Gaussian to an output waveform. The function is conditioned on input log magnitude mel-frequency spectrograms. The function is realized via a series of 12 so-called 'steps of flow,' each of which comprises a 1 × 1 invertible convolution and an affine coupling layer. The affine coupling layers contain dilated convolutions in the WaveNet-style, but with three non-causal taps instead of two causal taps. WaveGlow performs as well as WaveNet in terms of human perceptual preference of text-to-speech synthesized samples while being less difficult to train [25].

The model parameters available at [39] are used as the starting point of fine tuning. The synthetic corpus comprising 17,280 GAWs is mixed with the natural corpus, including 72 GAWs (29 diplophonic, 43 other types of dysphonia). For training, 95% of the synthetic corpus is randomly selected, and 50% of the natural corpus is used. The remaining 5% and 50% are used for testing. The Adam optimizer is used for fine tuning WaveGlow. Training is started with a learning rate of 5e-5, which is decreased stepwise down to 1e-7, whenever the loss

[2] This enables a fair comparison to MPD, which uses an average of approximately 44.53 coefficients per frame for our data.

TABLE I
COMPARISON OF THREE METHODS FOR ANALYSIS-BY-SYNTHESIS OF GAWs, WHICH ARE MPD, A DEEP AUTOENCODER, AND WAVEGLOW. MEDIANS OF TIME-DOMAIN FITTING ERROR LEVELS L_e ARE REPORTED WITH REGARD TO INPUT GAW TYPES

	Corpus:	Synthetic									Natural	
	M_d:	Small			Medium			Large			Diplophonic dysphonia	Nondiplophonic dysphonia
	M_r:	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large		
MPD (proposed)	unmod.	-13.47	-8.31	-5.40	-13.63	-8.27	-5.42	-12.15	-8.07	-5.52	-11.89	-12.16
	mod.	**-15.89**	**-10.95**	-7.64	**-15.80**	**-10.76**	-7.07	**-13.96**	**-10.25**	-7.40	**-12.69**	**-13.53**
Deep autoencoder		-11.87	-10.11	**-8.94**	-11.31	-9.92	**-8.81**	-10.27	-9.56	**-8.34**	-8.11	-8.02
WaveGlow	untuned [39]	1.56	1.57	1.62	1.61	1.56	1.68	1.61	1.60	1.67	1.36	1.22
	tuned	2.09	1.99	1.97	2.06	1.98	2.06	2.06	2.13	1.97	0.96	1.06

Sizes of deterministic and random modulations M_d, M_r: Small: $0\% < M_d, M_r < 5\%$, Medium: $5\% < M_d, M_r < 10\%$, Large: $10\% < M_d, M_r < 20\%$. L_e is obtained as $L_e = 20 \cdot \log_{10}[\sqrt{\overline{e^2(t)}}/\sqrt{\overline{y_I^2(t)}}]$ (dB), where $y_I(t)$ is the input GAW and the fitting error $e(t) = y_I(t) - y_M(t)$ is obtained for the model GAW $y_M(t)$. The error levels are obtained for 32 ms Hann-windows with 16 ms overlap. Smaller numbers of L_e reflect better performance.

TABLE II
COMPARISON OF THREE METHODS FOR ANALYSIS-BY-SYNTHESIS OF GAWs, WHICH ARE MPD, A DEEP AUTOENCODER, AND WAVEGLOW. MEDIANS OF MAGNITUDE SPECTRUM FITTING ERRORS, I.E., LOG-SPECTRAL DISTANCES LSD, ARE REPORTED WITH REGARD TO INPUT GAW TYPES

	Corpus:	Synthetic									Natural	
	M_d:	Small			Medium			Large			Diplophonic dysphonia	Nondiplophonic dysphonia
	M_r:	Small	Medium	Large	Small	Medium	Large	Small	Medium	Large		
MPD (proposed)	unmod.	7.62	9.63	10.33	8.38	10.40	11.03	9.40	10.32	10.97	14.86	10.01
	mod.	6.11	5.67	5.69	5.62	5.68	6.19	5.88	5.97	5.75	10.75	**7.07**
Deep autoencoder		11.74	8.96	7.66	9.59	9.04	8.24	10.50	8.95	8.43	11.76	10.71
WaveGlow	untuned [39]	11.52	9.64	8.14	10.67	9.63	8.18	10.82	9.64	8.25	8.94	9.45
	tuned	**4.82**	**4.83**	**4.71**	**4.88**	**4.85**	**4.79**	**4.91**	**4.76**	**4.72**	**8.54**	9.49

Sizes of deterministic and random modulations M_d, M_r: Small: $0\% < M_d, M_r < 5\%$, Medium: $5\% < M_d, M_r < 10\%$, Large: $10\% < M_d, M_r < 20\%$. The log-spectral distance is obtained as $LSD = \sqrt{\overline{\{10 \cdot [\log_{10}(|Z|) - \log_{10}(|\hat{Z}|)]\}^2}}$ (dB), where $Z \in \mathbb{C}$ and $\hat{Z} \in \mathbb{C}$ is the spectrogram of the input GAW and the model GAW respectively. The spectrograms are obtained for frequencies between 0 and 3 kHz, using a 48 ms Hann window and 40 ms overlap. Smaller numbers of LSD reflect better performance.

does not decrease any more. Approximately 170,000 fine tuning iterations are performed.

VI. RESULTS

Tables I and II report medians of the time-domain fitting error levels L_e (dB) and of the magnitude spectrum fitting errors LSD (dB) with regard to input GAW types and analysis-by-synthesis approaches. For synthetic input GAWs, L_e is reported with regard to the sizes of deterministic and random modulations. For natural input GAWs, L_e is reported separately for diplophonic and nondiplophonic types of dysphonic voice qualities. The approaches are MPD, the deep autoencoder, and WaveGlow. For MPD, the nonmodulating and modulating versions are evaluated. For WaveGlow, the untuned version using model parameters available at [39] and the tuned version are evaluated. Bold numbers reflect columnwise minima of L_e and LSD.

Table I shows that MPD outperforms the other approaches in terms of L_e by up to approximately 4-5 dB, except for input GAWs with large random modulations, for which the autoencoder's L_e is smaller by approximately 1-2 dB. Since MPD has the smallest L_e for natural input GAWs, one may hypothesize that large random modulations are rare in natural GAWs. WaveGlow fails in terms of L_e, reflecting a lack of phase fidelity. However, Table II shows that WaveGlow outperforms our modulating approach by approximately 1 dB, except for the nondiplophonic natural input GAWs, for which MPD has the smallest LSD.

TABLE III
MEDIANS OF pGAW PULSE TIMING ERRORS (SAMPLES) AND MAGNITUDE ERRORS (dB) OF THE OUR MPD APPROACH, WITH REGARD TO MODULATION SIZES OF INPUT GAWs

		Timing errors		Magnitude errors	
M_d	M_r	unmod.	mod.	unmod.	mod.
Small	Small	8.94	**7.11**	1.23	**1.12**
	Medium	18.21	**13.78**	2.34	**2.33**
	Large	28.42	**21.05**	3.43	**3.42**
Medium	Small	10.93	**10.06**	1.61	**1.54**
	Medium	18.68	**14.84**	2.28	**2.23**
	Large	28.56	**20.47**	3.12	**2.84**
Large	Small	11.53	**9.97**	1.84	**1.59**
	Medium	21.22	**13.71**	**1.60**	1.64
	Large	27.84	**17.67**	1.50	**1.17**

Sizes of deterministic and random modulations M_d, M_r: Small: $0\% < M_d, M_r < 5\%$, Medium: $5\% < M_d, M_r < 10\%$, Large: $10\% < M_d, M_r < 20\%$. The timing errors are pGAW pulse timing differences of input pGAWs and model pGAWs. The magnitude errors are pGAW pulse prominence differences of input pGAWs and model pGAWs.

Table III reports pGAW pulse timing and magnitude differences of the input pGAWs and the model pGAWs. The timing and magnitude errors reflect the model's frequency and amplitude modulation fidelity. Timing errors increase with modulation

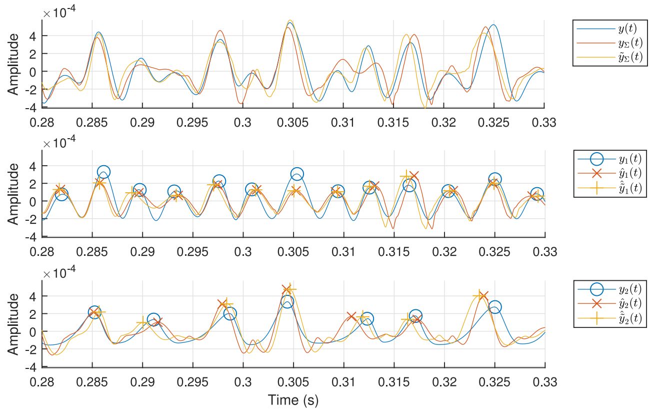
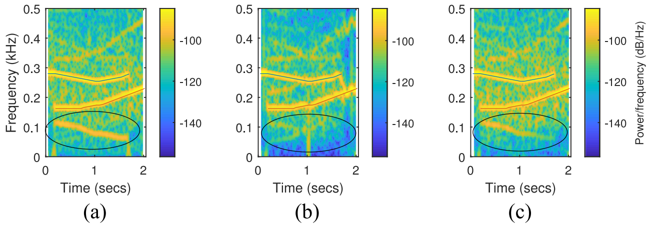


Fig. 8. Top plot: Example of an input GAW with two f_os and its modulated and unmodulated model GAWs. Middle and bottom panel: pGAWs with one f_o each. The pGAWs sum to the GAWs shown in the top panel. Markers indicate the pGAW pulse times. Extra peaks, which arise from crosstalk between the channels, are observed in the model pGAWs (bottom plot).



(a) (b) (c)

Fig. 9. (a) Spectrograms of an example input GAW, (b) its model GAW obtained with nonmodulating version of MPD, and (c) with the modulating version of MPD. Better agreement is observed between the input GAW and the model GAW obtained with the modulating MPD than between the input GAW and the model GAW obtained with the nonmodulating version of MPD. Ellipses mark a combination tone that is weaker or absent in the unmodulated model GAW and better reflected in the modulated model GAW. Additionally, noise speckles are more similar in the modulated model GAW.

size, and the modulating version of MPD performs better in all modulation sizes (bold numbers). Also the magnitude errors are smaller in the modulating version of MPD, with one exception. However, the differences appear to be marginal for some modulation size groups.

Fig. 8 shows an example of GAWs involved in the MPD of an input GAW with two simultaneous f_os. The top plot shows the input GAW (blue curve), its model GAWs obtained with the nonmodulating MPD (red curve), and with the modulating MPD (yellow curve). The model GAW obtained with the modulating MPD looks qualitatively more similar to the input GAW than the model GAW obtained with the nonmodulating MPD. The middle and bottom subplots show the pGAWs of the input GAW (blue curves) and their model pGAWs obtained using the nonmodulating MPD (red curves) and modulating MPD (yellow curve). The model pGAWs of the nonmodulating MPD are not strictly periodic, however, which is due to the frame-to-frame modulations of f_o and the pGAW pulse shapes, as opposed to pulse-to-pulse modulations. Frame-to-frame modulations are inherent to the nonmodulating MPD also, but pulse-to-pulse modulations are only considered in the modulating MPD.

Fig. 9 shows the spectrograms of an input GAW example, its model GAW obtained with the nonmodulating MPD, and its model GAW obtained with the modulating MPD. The spectrogram of the model GAW obtained with the modulating MPD looks more similar to the input GAW's spectrogram than the spectrogram of the model GAW obtained with the nonmodulating MPD. In particular, the encircled narrowband component is better reflected in the spectrogram of the GAW obtained with the modulating MPD. Additionally, noise speckles are reflected better in the spectrogram of the GAW obtained with the modulated MPD.

VII. Discussion and Conclusion

Synthesis and analysis-by-synthesis of modulated diplophonic GAWs are proposed and tested on a corpus comprising 864 synthetic GAWs of sustained vowels and a corpus comprising 72 natural GAWs of 35 individual patients. 29 of the GAWs contain diplophonation, and the remaining 43 do not contain diplophonations but a wide variety of other dysphonic voice quality types. In synthesis, the deterministic amplitude and frequency modulators are simple trigonometric functions of the difference of the pGAWs' instantaneous phases. Random pulse-to-pulse modulations are also considered. In analysis-by-synthesis, pGAW pulse times and heights are varied to minimize the time-domain fitting error in a least squares sense. It is shown that modulation improves the MPD in terms of time-domain fitting error levels and the magnitude spectrum fitting error. MPD is robust against deterministic modulations but is degraded when random modulations are large. MPD is also shown to be generalizable to other types of dysphonia than diplophonia. More added values of MPD are that it is a fully automated parameter estimation/analysis-by-synthesis method. Explicit control of deterministic and random modulations as well as other properties of the pGAWs may be manipulated while the pGAWs do not have to be available separately. The MPD's performance for synthetic and natural input GAWs is comparable, but some differences were observed. In particular, it appears as if large modulation sizes are not as frequent in natural GAWs as it was assumed for generating the synthetic input GAWs.

Deterministic comodulation of the distinct glottal oscillators explains the genesis of combination tones by the source. This comodulation arises from the mechanical and/or aerodynamical coupling of the oscillators. If one vocal fold is split into two oscillators anterior-posteriorly, mechanical coupling via vocal fold tissue is relevant to the genesis of combination tones. If the left and right vocal folds vibrate at different frequencies, aerodynamic coupling via the transglottal airflow is relevant to the genesis of combination tones. In mixed types, both types of coupling are relevant. Deterministic modulators of frequency and amplitude are obtained here as the sine and cosine, respectively, of the oscillator's instantaneous phase difference.

The amplitude modulator is proportional to the cosine of the oscillators phase difference, which is physiologically justified as follows. The larger the phase difference between the oscillators is, the smaller becomes the GAW's variable component. For example, when the oscillators have a 180° phase difference but vibrate equally otherwise, the variable component of the GAW vanishes due to destructive interference. In this case also the variable components of the transglottal airflow as well as the driving forces become small, which essentially causes the self-sustained oscillation to decay. We model this decay using the cosine of the phase difference, because the cosine is maximal and minimal when its argument is 0° and 180°, respectively. The use of a cosine for modeling the amplitude comodulation of coupled oscillators had also been suggested in [40], and a more didactic explanation can be found in [41].

The frequency modulator is proportional to the sine of the oscillators' phase difference, which is physiologically justified as follows. When the oscillators' natural frequencies are approximately equal, their vibration frequencies are also approximately equal and no relevant phase shift nor frequency comodulation is observed. In this case, the frequency modulator is proportional $\sin(0^\circ) = 0$. The oscillators vibrate in synchrony, which is referred to as 'entrainment,' 'phase locking,' or 'limit cycle' [9]. When the oscillators have natural frequencies that are slightly different from each other, they still vibrate at frequencies that are approximately equal and that lie between the natural frequencies, but the phase of one oscillator is shifted with regard to the other. In particular, the oscillator with the lower natural frequency is delayed with regard to the other one, and phase-locked. In this case, the frequency modulator is proportional to, e.g., $\sin(90^\circ) = 1$ for a phase difference of 90°, and added to and subtracted from the frequency of the oscillator with the lower and higher natural frequency, respectively. If the difference of the natural frequencies is large enough such that the phase difference exceeds 180°, the sign of the frequency modulator swaps due to $\sin(\phi) = -\sin(\phi + 180^\circ)$. In this case, the delayed oscillator gets lapped by the other oscillator, and the vibration frequencies differ. Afterwards, the reversal of the modulator's sign for phase differences exceeding 180° promotes the resynchronization of the oscillators. The oscillator with the higher natural frequency is than accelerated and the other one is decelerated until the phase difference approaches $360^\circ \hat{=} 0^\circ$, which may be referred to as resynchronization. This desync/resync process repeats if the difference of natural frequencies remains large enough, which causes the phase difference to once again approach, and exceed, 180°. This is a standard approach to frequency comodulation of coupled oscillators [42].

In addition to deterministic modulation, random modulations are observed. The random modulations arise from random fluctuations of the transglottal airstream due to turbulence and from randomness intrinsic to laryngeal muscle twitches [43]–[45].

The analysis of GAWs and pGAWs instead of individual trajectories has three advantages. First, no sagittal positions of the trajectories have to be preselected. Hence, asymmetries other than the idealized case of left-to-right asymmetry are considered in a straightforward way. Additionally, there is the risk of losing relevant information if only two trajectories are used. Second, GAWs have larger signal-to-noise ratios than trajectories because the relatively coarse spatial resolutions, i.e., pixel resolutions, of standard clinical high-speed cameras result in spatial quantization noise that has higher levels in trajectories than in GAWs. Third, GAWs are more closely related to the radiated sound pressure wave and the consequent auditory perception than trajectories and dynamical properties. Hence, analyzing the GAW smooths the path towards modeling acoustics and perception from clinically observed vocal fold vibration.

We also compared our analysis-by-synthesis approach to neural network-based approaches, i.e., a deep autoencoder, and the recent WaveGlow approach. The deep autoencoder approach was configured to have 48 coefficients at its bottleneck, which makes it comparable to MPD in terms of signal representation sparsity. The available pretrained WaveGlow network uses 80 mel-spectral coefficients. WaveGlow uses magnitude spectrograms as intermediate signal representation, and it is trained by minimizing spectral distances. This results in arbitrary phase differences between the input and model GAWs but competitive values of LSD. Spectral fidelity apparently comes with the reportedly competitive naturalness and intelligibility of synthesized speech samples, and phase fidelity is often thought to be of minor relevance for the purpose of speech synthesis. However, the intraframe temporal fine structure of pathological voices may play an important role in (i) the perception of modulation noise and roughness and (ii) the characterization of vocal fold pathophysiology. An additional advantage of MPD is that it uses intermediate signal representation features that explicitly represent simultaneous f_os, which are characteristic of diplophonia. As a next step, phase-aware (complex valued) neural networks may be used.

The proposed GAW model can potentially be applied for the testing and improvement of clinical voice analysis procedures. In particular, tests using diplophonic GAWs with known properties can help to identify conditions under which general-purpose methods for voice assessment fail. In addition, the availability of a larger number of more diverse and more realistic GAWs may aid in the improvement of such general-purpose methods.

Thus, approaches to addressing problems observed in the coding of hoarse voices [46] may be inspired by MPD. In particular, special attention may be paid to glottal closure instances, which may be extracted and modeled for diplophonic voices quite differently than for voices that are normal or hoarse in a nondiplophonic way.

Limitations of the study include the following. First, auditory experimentation regarding modulation parameters was beyond the scope of this paper, but properties of modulation may be pivotal to the perception of auditory stimuli [47]–[49]. In particular, it would be possible to listen to and compare GAWs, as well as to conduct listening experiments using speech audio predicted from the GAW. We may hypothesize that direct control of modulation properties will in the future enable the creation of more accurate and reliable perceptual models through auditory experimentation. Second, the synthesizer has some hyperparameters that were adjusted by trial and error and qualitative comparisons. To the best of our knowledge, no stronger evidence for suitable hyperparameter values of diplophonic GAWs exists in the literature than for the values used here.

Appendix A
Linearization of Instantaneous Phases for pGAW Pulse Shape Estimation

Due to frequency modulation of the input pGAWs, their pulse shapes are distorted, which needs to be compensated in the analysis-by-synthesis by using in the pulse shape estimation phase-linearized versions of $\tilde{u}(t)$ and $y_I(t)$. In particular, the time-domain pGAW pulse shape is obtained by normalized cross-correlation of the phase-linearized pulse train $\tilde{u}_{lin}(t)$ and the phase-linearized GAW $y_{lin}(t)$, i.e., $\tilde{r}_l = \Sigma_t[\tilde{u}_{lin}(t) \cdot y_{lin}(t-l)]/\Sigma\tilde{u}_{lin}(t)$. The linearized instantaneous phase $\tilde{\Theta}_{lin}(t)$ is obtained by linear regression of $\tilde{\Theta}(\tilde{t}_r)$, i.e., $\tilde{\Theta}_{lin}(t) = b_0^{lin} + b_1^{lin} \cdot t$, where b_0^{lin} and b_1^{lin} are the regression coefficients. The phase-linearized input GAW $y_{lin}(t)$ is obtained by resampling the input GAW $y(\tilde{\Theta}(t))$ at phases $\tilde{\Theta}_{lin}(t)$ using linear interpolation as $y(\tilde{\Theta}) \xRightarrow[\tilde{\Theta}\rightarrow\tilde{\Theta}_{lin}]{} y_{lin}(\tilde{\Theta}_{lin})$, which warps the time scale. Similarly, the quasiunit-pulse train $\tilde{u}_{lin}(t)$ is obtained by resampling $\tilde{u}(\tilde{\Theta}(t))$ at $\tilde{\Theta}_{lin}(t)$ using nearest neighbor interpolation as $\tilde{u}(\tilde{\Theta}) \xRightarrow[\tilde{\Theta}\rightarrow\tilde{\Theta}_{lin}]{} \tilde{u}_{lin}(\Theta_{lin})$.

References

[1] G. Fant, *Acoustic Theory of Speech Production*, The Hague: Mouton & Co., 1960.

[2] I. Titze and F. Alipour, *The Myoelastic Aerodynamic Theory of Phonation*, Iowa City: National Center for Voice and Speech, 2006.

[3] H. Fastl and E. Zwicker, *Psychoacoustics: Facts and Models*, 3rd ed. Berlin: Springer, 2007, pp. 136–141.

[4] K. Ishizaka, N. Isshiki, "Computer simulation of pathological vocal cord vibration," *J. Acoust. Soc.*, vol. 50, no. 5, pp. 1193–1198, 1976.

[5] I. Steinecke and H. Herzel, "Bifurcations in an asymmetric vocal-fold model," *J. Acoust. Soc.*, vol. 97, no. 3, pp. 1874–1884, 1995.

[6] P. Mergell, H. Herzel, and I. R. Titze, "Irregular vocal-fold vibration: High-speed observation and modeling," *J. Acoust. Soc.*, vol. 108, no. 6, pp. 2996–3002, 2000.

[7] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt, "Vibration parameter extraction from endoscopic image series of the vocal folds," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 8, pp. 773–781, Aug. 2002.

[8] I. Tokuda and H. Herzel, "Detecting synchronizations in an asymmetric vocal fold model from time series data," *Chaos-Woodbury*, vol. 15, no. 1, 2005, Art. no. 013702.

[9] J. Lucero, J. Schoentgen, J. Haas, P. Luizard, and X. Pelorson, "Self-entrainment of the right and left vocal fold oscillators," *J. Acoust. Soc. Amer.*, vol. 137, no. 4, pp. 2036–2046, 2015.

[10] D. Wong, M. R. Ito, N. B. Cox, and I. R. Titze, "Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases," *J. Acoust. Soc. Amer.*, vol. 89, no. 1, pp. 383–394, 1991.

[11] R. Schwarz, M. Döllinger, T. Wurzbacher, U. Eysholdt, and J. Lohscheller, "Spatio-temporal quantification of vocal fold vibrations using high-speed videoendoscopy and a biomechanical model," *J. Acoust. Soc. Amer.*, vol. 123, no. 5, pp. 2717–2732, 2008.

[12] T. Wurzbacher, M. Döllinger, R. Schwarz, U. Hoppe, U. Eysholdt, and J. Lohscheller, "Spatiotemporal classification of vocal fold dynamics by a multimass model comprising time-dependent parameters," *J. Acoust. Soc. Amer.*, vol. 123, no. 4, pp. 2324–2334, 2008.

[13] J. Jiang, C. I. B. Chang, J. R. Raviv, S. Gupta, F. M. Banzali, and D. G. Hanson, "Quantitative study of mucosal wave via videokymography in canine larynges," *Laryngoscope.*, vol. 110, no. 9, pp. 1567–1573, 2000.

[14] J. Schoentgen and P. Aichinger, "Glottal area patterns in numerically simulated diplophonia," in *Int. Congr. Phonetic Sci.*, Glasgow, Scottland, 2015, [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0604.pdf

[15] S. P. Kumar and J. Švec, "Kinematic model for simulating mucosal wave phenomena on vocal folds," *Biomed. Signal Proces.*, vol. 49, pp. 328–337, 2019.

[16] M. Kimura, H. Imagawa, T. Nito, K. I. Sakakibara, R. W. Chan, and N. Tayama, "Arytenoid adduction for correcting vocal fold asymmetry: High-speed imaging," *Ann. Otol. Rhinol. Laryngol.*, vol. 119, no. 7, pp. 439–446, 2010.

[17] S. Granqvist and P. Lindestad, "A method of applying fourier analysis to high-speed laryngoscopy," *J. Acoust. Soc. Am.*, vol. 110, no. 6, pp. 3193–3197, 2001.

[18] J. Neubauer, P. Mergell, U. Eysholdt, and H. Herzel, "Spatio-temporal analysis of irregular vocal fold oscillations: Biphonation due to desynchronization of spatial modes," *J. Acoust. Soc. Amer.*, vol. 110, no. 6, pp. 3179–3192, 2001.

[19] D. Berry, H. Herzel, I. R. Titze, and K. Krischer, "Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions," *J. Acoust. Soc. Amer.*, vol. 95, no. 6, pp. 3595–3604, 1994.

[20] K. I. Sakakibara, H. Imagawa, H. Yokonishi, M. Kimura, and N. Tayama, "Physiological observations and synthesis of subharmonic voices," in *Proc. Asia-Pacific Signal and Inf. Process. Assoc. Annu. Summit and Conf.*, Xi'an, China, 2011, pp. 1079–1085.

[21] T. Ikuma, M. Kunduk, and A. McWhorter, "Advanced waveform decomposition for high-speed videoendoscopy analysis," *J. Voice*, vol. 27, no. 3, pp. 369–375, 2013.

[22] N. Malyska and T. F. Quatieri, "Spectral representations of nonmodal phonation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 34–46, Jan. 2008.

[23] A. van den Oord *et al.*, "WaveNet: A generative model for raw audio," 2016, [Online]. Available: http://arxiv.org/abs/1609.03499

[24] L. Juvela, V. Tsiaras, B. Bollepalli, M. Airaksinen, J. Yamagishi, and P. Alku, "Speaker-independent raw waveform model for glottal excitation," in *Proc. Int. Conf. Spoken Lang. Process.*, Hyderabad, India, 2018, pp. 2012–2016.

[25] R. Prenger, R. Valle, and B. Catanzaro, "WaveGlow: A flow-based generative network for speech synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, Brighton, U.K., 2019, pp. 3617–3621.

[26] P. Aichinger, M. Hagmüller, B. Schneider-Stickler, J. Schoentgen, and F. Pernkopf, "Tracking of multiple fundamental frequencies in diplophonic voices," *IEEE/Assoc. Comput. Machinery Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 2, pp. 330–341, Feb. 2018.

[27] P. Aichinger, F. Pernkopf, and J. Schoentgen, "Detection of extra pulses in synthesized glottal area waveforms of dysphonic voices," *Biomed. Signal Proces.*, vol. 50, pp. 158–167, 2019.

[28] P. Aichinger, "Perturbation of times and magnitudes of cycle maxima observed in diplophonic voices," in *Models and Anal. Vocal Emissions for Biomed. Appl.*, Florence, Italy, 2019, pp. 121–124.

[29] P. Aichinger, I. Roesner, M. Leonhard, D. Denk-Linnert, W. Bigenzahn, and B. Schneider-Stickler, "A database of laryngeal high-speed videos with simultaneous high-quality audio recordings of pathological and non-pathological voices," in *Proc. Int. Conf. Lang. Resour. Eval.*, Portorož, Slovenia, 2016, pp. 767–770.

[30] M. Döllinger, "Glottis analysis tools," University Hospital Erlangen, Germany., 2018.

[31] J. Lohscheller, H. Toy, F. Rosanowski, U. Eysholdt, and M. Döllinger, "Clinically evaluated procedure for the reconstruction of vocal fold vibrations from endoscopic digital high-speed videos," *Med. Image Anal.*, vol. 11, no. 4, pp. 400–413, 2007.

[32] J. Lohscheller, U. Eysholdt, H. Toy, and M. Döllinger, "Phonovibrography: Mapping high-speed movies of vocal fold vibrations into 2-D diagrams for visualizing and analyzing the underlying laryngeal dynamics," *IEEE Trans. Med. Imag.*, vol. 27, no. 3, pp. 300–309, Mar. 2008.

[33] G. Chen, Y. Shue, J. Kreiman, and A. Alwan, "Estimating the voice source in noise," in *Proc. Int. Conf. Spoken Lang. Process.*, Portland, OR, USA, 2012, pp. 1600–1603.

[34] J. Schoentgen, "Stochastic models of jitter," *J. Acoust. Soc.*, vol. 109, pp. 1631–1650, 2001.

[35] J. Kreiman and B. R. Gerratt, "Perception of aperiodicity in pathological voice," *J. Acoust. Soc.*, vol. 117, no. 4, pp. 2201–2211, 2005.

[36] N. Malyska, T. F. Quatieri, and R. Dunn, "Sinewave representations of nonmodality," in *Proc. Int. Conf. Spoken Lang. Process.*, Florence, Italy, 2011, pp. 69–72.

[37] *MATLAB Documentation*. Accessed: Apr. 27, 2020. [Online]. Available: https://de.mathworks.com/help/optim/ug/constrained-nonlinear-optimization-algorithms.html#brnox01

[38] D. P. Kingma, and P. Dhariwal, "Glow: Generative flow with invertible 1×1 convolutions," in *Proc. Conf. Neural Inf. Process. Syst.*, Montr´eal, Canada, 2018, pp. 10215–10224.

[39] *WaveGlow Model Parameters: published model - waveglow_256 channels_universal_v5.pt*. Accessed: Aug. 18, 2020. [Online]. Available: https://github.com/NVIDIA/waveglow

[40] A. Pikovsky, M. Rosenblum, J. Kurths, and R. C. Hilborn, "Synchronization: A Universal Concept in Nonlinear Science," Cambridge, U.K.: Cambridge University Press, p. 230, 2001.

[41] A. Balanov, N. Janson, D. Postnov, and O. Sosnovtseva, "Synchronization: From Simple to Complex." Berlin Heidelberg: Springer, p. 440, 2009.

[42] S. H. Strogatz, "Nonlinear dynamics and chaos: With applications to physics, Biology, *Chemistry, and Engineering*," Boulder, CO: Westview Press, p. 276, 2015.

[43] J. Schoentgen, P. Aichinger, "Analysis and synthesis of vocal flutter and vocal jitter," in *Proc. Int. Conf. Spoken Lang. Process.*, Graz, Austria, 2019, pp. 2518–2522.

[44] I. Titze, "A model for neurologic sources of aperiodicity in vocal fold vibration," *J. Speech, Hear. Res*., vol. 34, pp. 460–472, 1991.

[45] I. P. Herman, "*Physics of the Human Body*," Berlin, Heidelberg: Springer, p. 281, 2007.

[46] M. S. Al-Radhi, T. G. Csapó, and G. Németh, "Continuous noise masking based vocoder for statistical parametric speech synthesis," *IEICE Trans. Inf. Syst*., vol. E103.D, no. 5, pp. 1099–1107, 2020.

[47] B. Moore, "Modulation minimizes masking," *Nature.*, vol. 397, pp. 108–109, 1999.

[48] B. Moore, "The roles of temporal envelope and fine structure information in auditory perception," *Acoust. Sci. Technol*., vol. 40, no. 2, pp. 61–83, 2019.

[49] J. Kreiman, M. Garellek, G. Chen, A. Alwan, and B. R. Gerratt, "Perceptual evaluation of voice source models," *J. Acoust. Soc.*, vol. 138, no. 1, pp. 1–10, 2015.

Philipp Aichinger (Member, IEEE) was born in Vienna, Austria, in 1983. He received the M.Sc. degree in electrical engineering and audio engineering from the University of Music and Performing Arts Graz, Graz, Austria, in 2009 and the Ph.D. degree in information and communications engineering from the Graz University of Technology, Graz, Austria, in 2015.

Since 2010, he has been a Research Associate with the Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Medical University of Vienna, Vienna, Austria. He is the first or senior author of 11 journal papers and 19 conference papers. His research interests include voice disorders, acoustics, diagnostic studies, signal processing, psychoacoustics, pattern recognition, artificial intelligence, psychoacoustics, and hearing impairment.

He is a Member of the IEEE Signal Processing Society, Audio Engineering Society (AES), Acoustical Society of America (ASA), Voice Foundation Philadelphia, America Speech-Language-Hearing Association (ASHA), Association for Research in Otolaryngology (ARO), and International Speech Communication Association (ISCA).

Franz Pernkopf (Senior Member, IEEE) received his M.Sc. (Dipl.Ing.) degree in electrical engineering from the Graz University of Technology, Graz, Austria, in summer 1999 and the Ph.D. degree from the University of Leoben, Leoben, Austria, in 2002. In 2002, he was awarded the Erwin Schrödinger Fellowship. From 2004 to 2006, he was a Research Associate with the Department of Electrical Engineering, University of Washington, Seattle, WA, USA. From 2010 to 2019, he was an Associate Professor with the Laboratory of Signal Processing and Speech Communication, Graz University of Technology, Graz, Austria. Since 2019, he has been a Professor of intelligent systems with Signal Processing and Speech Communication Laboratory, Graz University of Technology. His research interests include pattern recognition, machine learning, and computational data analytics with applications in various fields ranging from signal and speech processing to medical data analysis and other data modeling problems from industrial applications. He is particularly interested in probabilistic graphical models for reasoning under uncertainty, discriminative, and hybrid learning paradigms, deep neural networks, and sequence modeling.