Airaksinen, Manu; Juvela, Lauri; Bollepalli, Bajibabu; Yamagishi, Junichi; Alku, Paavo

A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis

Published in:
IEEE/ACM Transactions on Audio Speech and Language Processing

DOI:
10.1109/TASLP.2018.2835720

Published: 01/09/2018

Document Version
Peer reviewed version

Please cite the original version:
Airaksinen, M., Juvela, L., Bollepalli, B., Yamagishi, J., & Alku, P. (2018). A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis. IEEE/ACM Transactions on Audio Speech and Language Processing, 26(9), 1658-1670. https://doi.org/10.1109/TASLP.2018.2835720

A comparison between STRAIGHT, glottal, and sinusoidal vocoding in statistical parametric speech synthesis

Manu Airaksinen, Lauri Juvela, Bajibabu Bollepalli, Junichi Yamagishi, *Senior Member, IEEE* and Paavo Alku, *Senior Member, IEEE*

***Abstract*—A vocoder is used to express a speech waveform with a controllable parametric representation that can be converted back into a speech waveform. Vocoders representing their main categories (mixed excitation, glottal, sinusoidal vocoders) were compared in this study with formal and crowd-sourced listening tests. Vocoder quality was measured within the context of analysis-synthesis as well as text-to-speech (TTS) synthesis in a modern statistical parametric speech synthesis framework. Furthermore, the TTS experiments were divided into synthesis with vocoder-specific features and synthesis with a shared envelope model, where the waveform generation method of the vocoders is mainly responsible for the quality differences. Finally, all of the tests included four distinct voices as a way to investigate the effect of different speakers on the synthesized speech quality.**

The obtained results suggest that the choice of the voice has a profound impact on the overall quality of the vocoder-generated speech, and the best vocoder for each voice can vary case by case. The single best-rated TTS system was obtained with the glottal vocoder GlottDNN using a male voice with low expressiveness. However, the results indicate that the sinusoidal vocoder PML (pulse model in log-domain) has the best overall performance across the performed tests. Finally, when controlling for the spectral models of the vocoders, the observed differences are similar to the baseline results. This indicates that the waveform generation method of a vocoder is essential for quality improvements.

***Index Terms*—Speech synthesis, vocoder, statistical parametric speech synthesis.**

I. Introduction

A VOCODER is used to express a speech waveform with a parametric representation that can be converted back into a speech waveform. Furthermore, the parametric representation enables the statistical modeling of speech and it also makes it possible to manipulate speech, for example, to enhance its intelligibility [1]. These properties make vocoders flexible tools that can be applied in several areas of speech technology such as statistical parametric speech synthesis [2], voice transformation and modification [3], musical applications [4], and even low bit-rate speech coding [5]. However, in order to be used with generic parameterization techniques, vocoders discard a part of speech information, for instance the phase of the excitation signal [1]. Thus signal reconstruction by vocoders is lossy, not only with respect to quantization, which is a fundamental difference from modern high-quality speech codecs [8].

In statistical parametric speech synthesis (SPSS), a vocoder is one of the traditional backbones of the framework that enables the statistical modeling of context-dependent speech using parameters (see Figure 4). The majority of vocoders used in SPSS are based on some form of the source-filter model of speech production [6], which assumes that speech is a convolution between an excitation signal and a filter. This separation is effective in allowing modifications to produce a vast space of perceptually different speech sounds. This versatility of vocoders has been key for their adaption within SPSS as vocoders can be used to reliably and efficiently transform generated speech parameters into stable, continuous waveforms of different characteristics. In addition to SPSS, vocoders have been utilized in voice modification tasks due to the versatility of the parametric representation [9].

The main drawback of vocoding has been—and still continues to be—the artifacts generated by over-simplified source-filter modeling, caused by poor separation of speech into excitation and filter [10], and especially by too simple modeling of the excitation: A straightforward voiced excitation model, consisting only of an impulse train controlled by the fundamental frequency (f_0), results in a "robotic," "buzzy" voice that is perceived as highly unnatural by human listeners [2]. This problem has been addressed in several vocoder-oriented SPSS studies aiming at synthetic speech of better segmental voice quality.

The vocoders developed can be categorized roughly into three groups: 1) mixed excitation with a spectral envelope (e.g., STRAIGHT [10], [11], WORLD [12], DSM [13]), 2) glottal vocoders (e.g., GlottHMM [14], GSS [15]), and 3) sinusoidal vocoders (e.g., HMPD [16], HNM [17], Vocaine [18]). The mixed excitation and glottal vocoders operate under the assumption of the source-filter model of speech production (see Section II-B), but the interpretation of the main spectral properties between the source and filter are different: In mixed excitation, the source is assumed to be a spectrally flat, "impulse plus noise" signal, with all of the spectral envelope information contained in the filter. Glottal

M. Airaksinen, L. Juvela, B. Bollepalli, and P. Alku are with the Department of signal processing and acoustics, Aalto University, Finland. e-mail: manu.airaksinen@aalto.fi, paavo.alku@aalto.fi.

J. Yamagishi is with the National Institute of Informatics, Japan.

Manuscript received November 2, 2017, revised March 29, 2018

[1] In this study the term *vocoder* refers to the abstraction of the *channel vocoder* [6], and not, for instance, the *phase vocoder* [7] that synthesizes speech from encoded phase derivatives with the cost of general modifiability.

vocoders in turn use a source signal that mimics the true acoustical excitation of voiced speech, the glottal volume velocity waveform (or its derivative), that is generated by the fluctuating vocal folds. Since the vocal folds are physiological organs whose fluctuation mode is controlled by the talker, the spectral envelope of the excitation in glottal vocoders is not flat but varies depending on, for example, the speaking style and phonetic stress (e.g., the position of the underlying utterance within the word and sentence). In sinusoidal vocoders, the harmonic structure is modeled by individual sinusoidal components, with a complementary stochastic noise envelope. In SPSS applications, however, sinusoidal vocoders have to resort to source-filter based parameterization after the analysis [19]. This is because regression tasks in acoustic models require constant-length target vectors, whereas the number of sinusoids in the acoustic signal is not constant but depends on f_0. Thus, sinusoidal vocoders commonly convert the sinusoidal amplitudes into a spectral envelope representation from which the sinusoid amplitudes (with minimum phase response) are sampled according to f_0 during synthesis. As an alternative to vocoding, SPSS research has recently ventured into direct waveform modeling frameworks (e.g., WaveNet [20], Tacotron [21]) that aim to circumvent the problematic signal processing-based waveform synthesis of vocoders by treating speech synthesis more as an end-to-end (text to waveform) machine learning problem. The segmental speech quality achieved by these systems has been shown to surpass vocoder-based SPSS, but with many orders of magnitude higher computational costs and little controllability over the generated voice.

Vocoders from the three main categories described above (mixed excitation, glottal, sinusoidal vocoders) are compared in the current study with two separate listening test arrangements (formal and crowd-sourced). Even though vocoder comparisons have been published in several articles (e.g., [19]), the current research is motivated as follows. The main overall objective of this study is to evaluate vocoder performance by *simultaneously* comparing representative vocoders from each of the three main categories using the latest SPSS engines based on deep learning. To the best of our knowledge, this kind of comparison has not been conducted before. In addition, vocoder performance is known to vary from voice to voice (e.g., based on gender, speaking style, the audio quality of the recordings) [22]. However, there are no previous investigations that have studied how much the known vocoders belonging to the three main categories are affected by the characteristics of the selected voice. Therefore, as a supplement to the above overall objective, the current text-to-speech (TTS) study aims to find out how synthesis quality, achieved with vocoders belonging to the three main categories, depends on the voice selected for TTS.

Vocoder selection of this study is limited to one per vocoder family, with the exception of glottal vocoding where two techniques, GlottHMM and GlottDNN, are included. The GlottHMM vocoder [14] is regarded as a state-of-the-art glottal vocoder because it was already published some years ago, after which it has been successfully used in several SPSS studies [22]. GlottDNN is a new glottal vocoder based on the cumulative evolution of studies over recent years [23]–[26], including, for example, new spectral modeling and excitation generation techniques. Given this background, the current article is structured so that the first comprehensive description of the GlottDNN vocoder is presented first in Section II. The other vocoders selected for the study—namely STRAIGHT, pulse model in log-domain (PML) [27], and GlottHMM—are known techniques that have been described in previous articles and therefore they will be only briefly presented in Section III. Sections IV, V, and VI contain the conducted experiments, results, and discussion respectively.

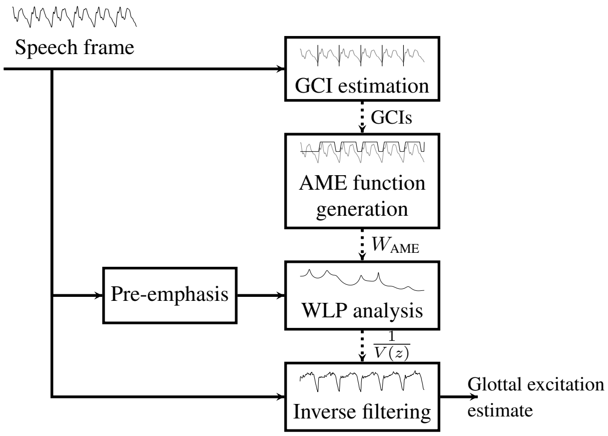


Fig. 1. A block diagram of the QCP glottal inverse filtering method.

II. The GlottDNN vocoder

A. General

As discussed in Section I, GlottDNN is the most recent glottal vocoder and it has been developed in a cumulative evolution from the GlottHMM vocoder in a series of studies conducted over the past five years [23]–[26]. Compared to GlottHMM, GlottDNN utilizes the latest glottal inverse filtering (GIF) methodology to estimate the glottal source from the speech signal. The GIF methodology used in GlottDNN is quasi-closed phase (QCP) analysis [28], which enables computing the high-quality physiologically motivated source-filter separation of speech into the glottal excitation and vocal tract transfer function. Moreover, GlottDNN synthesis is performed by using a deep learning-based generation of the vocoder's excitation waveform.

B. GIF with QCP analysis

Voiced speech is produced by the air-flow streaming from the lungs, generating oscillations at the vocal folds. This airflow is further modulated by the resonances of the vocal tract and finally radiated at the lips to produce the speech pressure signal. In GIF, this process is commonly modeled with the linear source-filter model of speech production [6]:

$$S(z) = G(z)V(z)L(z), \tag{1}$$

where $S(z)$ is the speech signal, $G(z)$ is the glottal excitation, $V(z)$ is the vocal tract transfer function, and $L(z)$ is a first-order differentiator modeling the lip radiation effect. Based on this, most GIF algorithms focus on the task of estimating the transfer function $V(z)$ and the first time-derivative of glottal flow is regarded as the effective driving excitation $E(z)$ of voiced speech:

$$E(z) = G(z)L(z) = \frac{S(z)}{V(z)}, \tag{2}$$

which is also the glottal excitation model assumed in the GlottDNN vocoder.

It is known that $V(z)$ can be represented with great accuracy for most speech sounds as an all-pole filter (i.e., as an autoregressive [AR] process), with the exception of nasal sounds [6]. Furthermore, the glottal excitation serves mainly as a maximum-phase component in the production of the speech signal [29] and this excitation has its largest impact on short intervals in the vicinity of glottal closure instants (GCIs). Based on this information, the recently proposed GIF method, QCP analysis [28], aims to attenuate the effect of the glottal excitation in the estimation of $V(z)$ by using temporally weighted linear prediction (WLP) [30] with a specific attenuated main excitation (AME) weighting function [31]. The goal of AME weighting is to de-emphasize the prominent effect of the prediction error in the vicinity of GCIs (see Section II-C) so that the filter coefficients optimized are more prone to model $V(z)$ and not the effects of the periodic excitation signal that cause harmonic bias to estimated formants. It should be emphasized that the AME weighting is done on the squared prediction error signal in the optimization of the underlying AR model and should not be mixed up with the traditional short-time windowing (e.g., Hamming windowing) used for reducing truncation effects.

A block diagram describing the estimation of the glottal flow with QCP is shown in Figure 1. For each frame, the GCIs are estimated and the corresponding AME function is generated. Next, WLP analysis is performed for the pre-emphasized frame. The pre-emphasis is used to flatten the spectral tilt of the glottal excitation [6]. Finally, the original frame is inverse filtered with $\frac{1}{V(z)}$ to obtain the estimate for the glottal flow derivative $e(n)$.

C. Speech analysis and parametrization

A block diagram of the analysis stage of the GlottDNN vocoder is shown on the left-hand side of Figure 2. The speech signal is analyzed in frames of length t_f at intervals of t_s. The fundamental frequency f_0 (in Hz) and windowed energy (in dB) of the frame are extracted and added to the feature vector. Next, the vocal tract model $V(z)$ is obtained with *frequency-warped time-weighted linear prediction* (WWLP) that uses GCIs to construct the AME weighting function (see Section II-B) for the filter optimization. The estimate of $V(z)$ is converted into line spectral frequencies (LSFs) [32] and added to the feature vector. Finally, the frame is inverse filtered with $\frac{1}{V(z)}$ to obtain the glottal excitation estimate $e(n)$. From the glottal excitation estimate, two sets of parameters are computed: First, the spectral tilt, which is modeled with low-order linear prediction (LP) analysis (parameterized as LSFs), and second, the harmonic-to-noise ratio (in dB) of the glottal excitation that is compressed using equivalent rectangular bandwidths (ERB) [33].

For unvoiced frames, the analysis pipeline is identical, except in vocal tract modeling where WWLP with a constant weighting function is used without pre-emphasis in computing the vocal tract transfer function.

Vocal Tract Modeling with WWLP: The AME weighting function was originally formulated to be used in WLP to compute all-pole spectral envelopes that are less prone to show formants biased by the harmonic peaks of the excitation [31]. However, WLP is formulated in the linear frequency domain that is not auditorily justified, particularly for full-band speech (i.e., $f_s \approx 48$ kHz). This is because the most important acoustical contents of speech are present at lower frequencies (e.g., 1 kHz to 3 kHz) [6], so it is desirable to have better modeling accuracy in those frequencies. In warped LP [34], the AR model can be optimized using a warped frequency scale that approximates the human auditory system. Unfortunately, warped LP also results in larger biasing of formants by harmonics [35]. WWLP [26], however, is a recently proposed method that can be regarded as a fusion between weighted LP and warped LP: WWLP takes advantage of the AME weighting function to reduce the biasing of formants by harmonics in the filter optimization in order to obtain frequency-warped AR estimates for the vocal tract transfer function. The formulation of WWLP, originally presented in [26], is as follows.

In WWLP, the AR model predicts the sample s_n at time-index n as a linear combination of p previous *warped* samples:

$$s_n = \sum_{k=1}^{p} a_k y_{k,n} + G e_n, \tag{3}$$

where $\{a_k\}$ denote the *prediction coefficients*, G is the filter gain, e_n is the excitation, and $y_{k,n}$ is the output of s_n convolved k times through a general function $D(z)$ that models the warped delay line. In WWLP, $D(z)$ is of the form:

$$D(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{4}$$

where λ is the warping coefficient. Based on this, $y_{k,n}$ can be expressed as:

$$y_{k,n} = \begin{cases} s_n & , \quad k = 0 \\ \underbrace{d_n * d_n * \cdots * d_n}_{k\text{-fold convolution}} * s_n & , \quad 1 \leq k \leq p, \end{cases} \tag{5}$$

where d_n is the impulse response of $D(z)$.

To obtain the optimal coefficients $\{a_k\}$ of the model in Eq. 3, an optimization criterion must be selected. The approach taken in WWLP is to use a *time-weighted* squared sum as the optimization criterion:

$$E_{\text{WWLP}} = \sum_n W_n e_n^2 = \sum_n W_n (s_n - \sum_{k=1}^{p} a_k y_{k,n})^2, \tag{6}$$

where for voiced speech W_n is the AME weighting function discussed in Section II-B. For unvoiced speech $W_n = 1.0$, which reduces the analysis into conventional warped LP. With

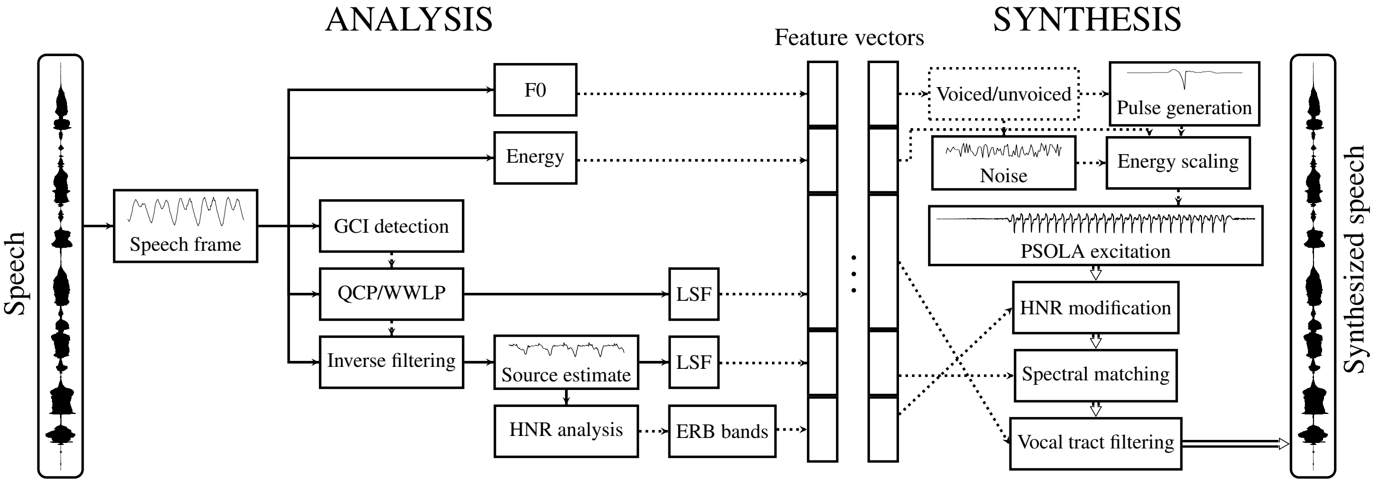


Fig. 2. A block diagram of the GlottDNN vocoder.

the error criterion set, the analytic solution for the optimal coefficients can be obtained as [26]:

$$\mathbf{a}_{\text{opt}} = \left(\sum_n W_n \mathbf{y}_n \mathbf{y}_n^\intercal \right)^{-1} \left(\sum_n W_n s_n \mathbf{y}_n \right), \qquad (7)$$

where $\mathbf{a} = [a_1, a_2, \ldots, a_p]^\intercal$ and $\mathbf{y}_n = [y_{1,n}, y_{2,n}, \ldots, y_{p,n}]^\intercal$. When summing from $n = 0$ to $n = N - 1 + p$, where N is the frame length in samples, the matrix to be inverted in Eq. 7 can be interpreted as the *frequency-warped* and *time-weighted* autocorrelation matrix of the analyzed signal.

Harmonic-to-Noise Ratio Estimation and Modeling: The degree of voicing, i.e., the level difference between the harmonic components produced by the periodic vibrations of the vocal folds and the aperiodic noise of the glottal source, is represented as the harmonic-to-noise ratio (HNR). HNR is computed as the relative difference between the DFT upper and lower envelopes of the windowed glottal source estimate. The upper envelope is estimated by dynamic peak picking of harmonics from a high-resolution FFT magnitude spectrum and the lower envelope is estimated by averaging samples half-way between the harmonic peaks. Finally, HNR is converted into dB, and averaged across ERB bands of a selected parameter order [33].

D. Speech Synthesis

A block diagram of the synthesis stage of the GlottDNN vocoder is presented on the right-hand side of Figure 2. First, the initial excitation signal for the entire utterance is produced by the pitch-synchronous overlap-add (PSOLA) procedure [36]. For voiced speech, the glottal excitation segments are generated as two pitch-period long pulses (based on f_0) that are scaled in energy. The glottal pulse generation block is implemented as a deep neural network (DNN) that is trained with GlottDNN feature vectors as an input and zero-padded two pitch-period glottal flow derivative pulses as an output [23]–[25]. A single pre-computed pulse with interpolation to target f_0 can also be used as the base pulse. For unvoiced speech, a white noise sequence of a pre-determined length is generated for each frame.

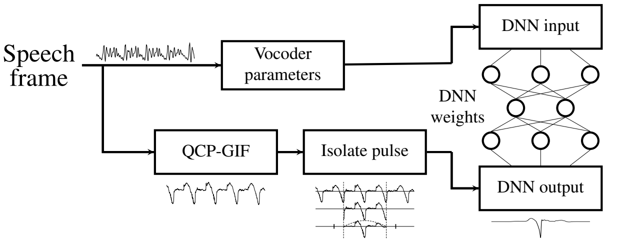


Fig. 3. DNN-based excitation generation framework.

After the initial excitation signal has been generated, it is processed according to the HNR and spectral tilt features. The HNR processing is done by adding noise in the spectral domain using the conventional overlap-add procedure. The spectral tilt is adjusted to the target by modifying the synthesis vocal tract filter according to

$$H_{\text{match}}(z) = \frac{H_{\text{base}}(z)}{H_{\text{target}}(z)}, \qquad (8)$$

where $H_{\text{match}}(z)$ is the matching filter, $H_{\text{base}}(z)$ denotes the LP inverse model of the initial excitation, and $\frac{1}{H_{\text{target}}}(z)$ is the target spectral tilt.

Finally, to obtain the synthesized speech, each frame of the generated excitation signal is filtered in the spectral domain with a time-varying all-pole filter, inverse Fourier transformed and pitch-asynchronously overlap-added. The gain of each filter is adjusted according to the target energy of the current frame so that the energy of the filtered excitation matches the target.

DNN-based Glottal Excitation Generation: A block diagram describing the training of the DNN-based glottal excitation generation is presented in Figure 3. For each speech frame within the training dataset, the GlottDNN vocoder parameters are analyzed and fed to the input vector. The DNN target output generation is more specific: First, the glottal flow derivative, estimated by QCP, is constructed over the entire frame. Second, a two pitch-period segment, delimited by consecutive GCIs, is isolated from the frame. The obtained segment is windowed with a raised cosine window (the square root of a Hann window), and finally zero-padded evenly from both edges to a constant length. The final pulse obtained this way has its middle GCI exactly at the center index of the frame. This aspect is important, as it acts as phase locking for the target pulses, which allows the trained DNN to learn the waveform properties based on the assumption of an identical linear phase component.

It is worth noting that in the context of glottal source generation, a considerably smaller volume of speech data is required for DNN training than, for example, required in the acoustic models used in speech recognition or synthesis. This is for two reasons: First, as two pitch-period long glottal flow waveforms are used as the output of DNN, a large number of glottal pulses can be extracted by using a small volume of speech data (e.g., roughly 50,000 glottal pulses are obtained from 10 minutes of speech data). The second reason is that the glottal pulse to be estimated by the DNN is an elementary (simple) waveform, which is formed at the level of the glottis in the absence of vocal tract resonances.

After the training data has been computed with QCP, the excitation DNN can be trained. Our previous studies show that a simple feed-forward architecture yields satisfactory results [25], [37], but more sophisticated network architectures, such as recurrent [38] and convolutional neural networks, and better training techniques, such as generative adversarial networks [39], have also been explored in order to gain improved results.

III. Other vocoders used in the evaluation and the TTS system

A. STRAIGHT

STRAIGHT [10], [11] is a representative of vocoders that are based on the conventional source-filter model where speech is divided into a spectral envelope filter that is driven by a spectrally flat excitation signal:

$$S(z) = I(z)\hat{V}(z), \tag{9}$$

where $I(z)$ is the spectrally flat excitation signal and $\hat{V}(z)$ is the entire spectral envelope of speech.

The key part of STRAIGHT is its spectral envelope analysis technique that aims to minimize the effect of periodicity interference within and between analysis frames. For each frame, two pitch-adaptive analysis windows, w_p and w_c, are used to produce two complementary representations of the magnitude spectrum, $S_p(\omega, t)$ and $S_c(\omega, t)$:

$$w_p(t) = e^{-\pi(t/t_0)^2} * h(t/t_0), \tag{10}$$

$$w_c(t) = w_p(t)\sin\left(\pi\frac{t}{t_0}\right), \tag{11}$$

where t is the time index, t_0 is the time of the fundamental period, and $h(t)$ is the second order cardinal B-spline function given by:

$$h(t) = \begin{cases} 1 - |t|, & \text{if } |t| < 1, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The magnitude spectra obtained with these windowing functions are combined into the final spectral envelope estimate by:

$$S_U(\omega, t) = \sqrt{S_p^2(\omega, t) + \xi S_c^2(\omega, t)}, \tag{13}$$

where $\xi = 0.13655$ is a blending factor that minimizes the temporal variation of the resulting spectrogram [10].

In principle the estimation of aperiodicity in STRAIGHT is performed similarly to the estimation of HNR in GlottDNN by computing the ratio between the upper and lower spectral envelopes (S_U and S_L respectively). In practice, STRAIGHT uses a table look-up operation from a database of known aperiodicity measurements to compute its aperiodicity spectrum with

$$S_{AP}(\omega) = \frac{\int w_{\text{ERB}}(\lambda;\omega)|S(\lambda)|^2\Gamma\left(\frac{|S_L|^2}{|S_U|^2}\right)d\lambda}{\int w_{\text{ERB}}(\lambda;\omega)|S(\lambda)|^2 d\lambda}, \tag{14}$$

where w_{ERB} is an auditory filter for smoothing the power spectrum at center frequency, ω, $|S(\lambda)|^2$ is the speech power spectrum, and $\Gamma(\ )$ is the table lookup operation.

The inherent STRAIGHT parameters (in addition to f_0) are thus the magnitude spectrogram of the envelope and the spectrogram representing the aperiodicity. For the purposes of SPSS, the STRAIGHT parameters are usually transformed into mel-generalized cepstral coefficients [40] (for the spectral envelope) and into log-average ERB coefficients (for the aperiodicity).

B. PML

PML [27] is a state-of-the art vocoder with its roots in sinusoidal modeling, namely the HMPD vocoder [16]. However, contrary to the traditional sinusoidal vocoding approaches, PML utilizes frequency domain pulse synthesis techniques from parameters that are obtained from sinusoidal modeling-based analysis. The PML analysis begins from the estimation of the f_0 contour (considered continuous and without explicit voicing decisions). Next, the magnitude spectrum is sampled at the harmonic frequencies to obtain a model of the spectral envelope (other spectral envelope extraction methods, such as STRAIGHT, can also be used). Finally, a specific phase distortion deviation (PDD) is computed with the help of harmonic phase distortion (PD) values:

$$\text{PD}_{i,h} = \phi_{i,h+1} - \phi_{i,h} - \phi_{i,1}, \tag{15}$$

where $\phi_{i,h}$ is the phase value at frame i and harmonic h. The PD values are then linearly interpolated to obtain $\text{PD}_i(\omega)$, a continuous spectral representation of phase distortion. $\text{PDD}_i(\omega)$ is then computed as the short-term standard deviation of PD:

$$\text{PDD}_i(\omega) = \sqrt{-2\log\left|\frac{1}{N}\sum_n e^{j(\text{PD}_n(\omega))}\right|}. \tag{16}$$

PDD values show, in a normalized representation, how much phase distortion there is in each frequency bin compared to the estimated fundamental frequency. To simplify this model for speech synthesis applications, the PDD values are quantized into binary values known as the *binary noise mask* $M_i(\omega)$. The quantization is done with a thresholding value of $M_i(\omega) = 1$ if $\text{PDD}_i(\omega) > 0.75$ and otherwise it is zero. The synthesis of the PML vocoder is done straightforwardly in the frequency domain: Based on the (continuous) f_0 contour, a

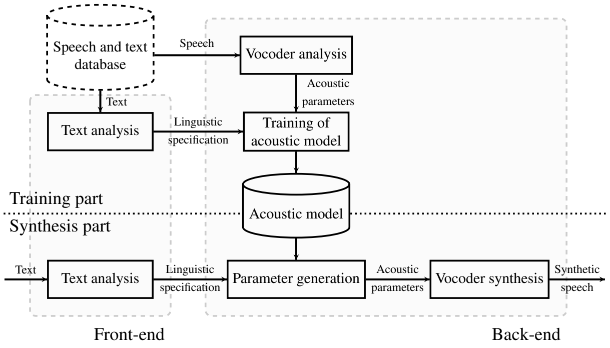


Fig. 4. General block diagram of statistical parametric speech synthesis.

single minimum phase pulse $S_i(\omega)$, of length $\frac{1}{f_0}$, is generated at each pitch mark t_i. The spectrum is set as the minimum phase response of the spectral envelope $V_i(\omega)$, and the phase spectrum values are replaced with random noise at frequency bins where $M_i(\omega) = 1$:

$$S_i(\omega) = e^{-j\omega t_i} \cdot V_i(\omega) \cdot N_i(\omega)^{M_i(\omega)}. \tag{17}$$

This versatile noise model gives the PML vocoder the ability to capture smooth transitions between voiced and unvoiced parts of speech without explicit voicing decisions; as for unvoiced frames (where $M_i(\omega)$ is mostly 1), the phase is mostly random.

C. GlottHMM

GlottHMM [14] is a glottal vocoder that is the precursor to GlottDNN. GlottHMM analysis uses iterative adaptive inverse filtering (IAIF) [41] as an inverse filtering method to separate speech into the glottal source and vocal tract. Similarly to GlottDNN, the spectral tilt of the estimated glottal excitation and the vocal tract filter are parameterized with LSFs, and the HNR of the excitation signal is estimated to model the degree of aperiodicity. The synthesis part of GlottHMM uses a single base pulse to generate the voiced excitation. This base pulse is a hand-picked, high-quality glottal pulse that has been precomputed from a voiced utterance with IAIF. Based on the vocoder parameters, the base pulse is interpolated to target pitch period, matched to the target HNR, and filtered to have the target spectral tilt. After this processing, the pulses are concatenated with white noise that represents the excitation of unvoiced speech sections in order to generate the final excitation signal. The generated excitation is then filtered with a direct form IIR filter based on the vocal tract LSFs.

D. SPSS

In TTS, SPSS refers to a back-end architecture that uses data-driven acoustic modeling as its core technology [2]. The acoustic model is trained as a regression task from front-end provided linguistic specifications (which are language specific, obtained from text input) to a parametric representation of speech, traditionally the vocoder parameters (see Figure 4). Originally, the acoustic model was based on context-dependent HMMs structured by a regression tree [42], but recent approaches have overwhelmingly switched to regression models based on recurrent neural networks, such as long-short terms memory (LSTM) [43] and gated recurrent unit (GRU) [44] networks. During synthesis time, the front-end provided linguistic specifications are fed into the acoustic model that produces the corresponding set of vocoder parameters, and the generated vocoder parameters are then transformed back into the speech waveform by using the vocoder.

IV. Experiments

The experiments carried out in this study were designed to compare the four selected vocoders in terms of their achievable synthesis quality. This design of experiments is many-faceted: First, the performance of a vocoder can vary greatly between voices and/or voice types. For example, in glottal vocoders, the accuracy of the glottal flow estimation with GIF typically decreases for high-pitched speech, thereby degrading the synthesis quality for female voices and for voices with an expressive speaking style. Vocoders based on other paradigms might not suffer from this problem so much. Second, the *potential maximum quality* of vocoded speech, that is the analysis-synthesis quality (synthesis with natural parameter trajectories), might drastically differ from the corresponding *TTS quality* obtained by computing parameter trajectories from acoustic models of the TTS system. However, as the performance of acoustic models (and their post-processing methods [45]) improves, the analysis-synthesis quality also becomes more useful in TTS as a measure of upper bound. Third, many vocoders take advantage of similar analysis procedures (i.e., they have a representation for the f_0, spectral envelope, and aperiodicity), but they typically have greater differences in methods of waveform synthesis. As an example, the PML vocoder is recommended to be used [27] together with the STRAIGHT-estimated spectral envelope for optimal TTS quality. Finally, the test type and its other conditions (e.g., listener population, the quality of audio equipment used in the test) can all affect the outcome of subjective listening tests. For example, naive and expert listeners may focus their attention on different properties in speech quality and naturalness, and the quality of the used listening equipment (e.g., headphones versus laptop speakers) can affect the perceived differences.

The listening tests were designed by taking into account all of the above-mentioned concerns in a balanced way. First, for all of the tests, four different voices (two male and two female) with different characteristics were selected as follows: (1) “Nick” [46], a high-quality British English male voice recording with low expressiveness, (2) “Roger”, a highly expressive British English male voice with audible reverberation present in the recordings, (3) “Jenny”, a high-quality British English female voice recording with moderate expressiveness, and (4) “Nancy” [47], a high-quality American English female voice recording with high expressiveness. Table I shows the number of files employed for training, development, and testing for each voice.

Second, the tests were split into two categories: Tests conducted in a controlled listening environment with relatively experienced listeners and crowd-sourced tests with minimal control of the listeners and listening environment. The controlled tests were conducted as MUSHRA (multiple stimuli

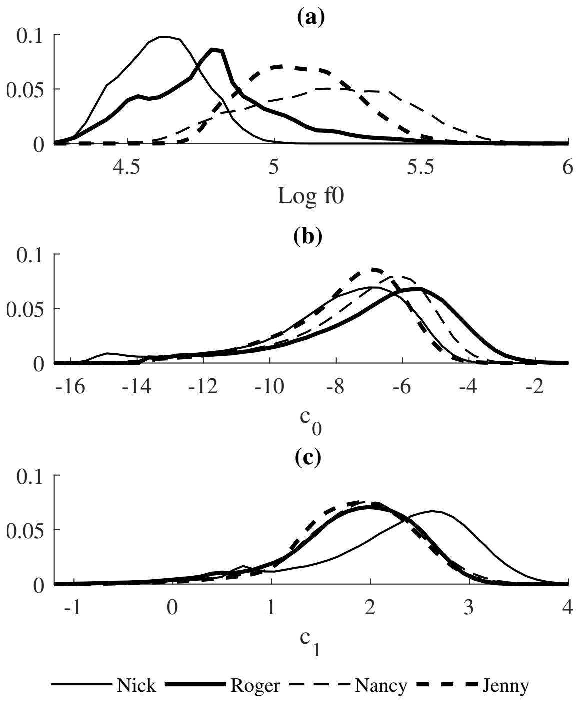


Fig. 5. Normalized parameter distribution histograms computed from log f_0 (a), c_0 (b), and c_1 (c) for the selected voices.

with hidden anchor and reference [48]) tests inside sound proofed listening booths with high-quality sound cards and headphones. The controlled tests (hereafter referred to as *MUSHRA tests*) included the analysis-synthesis quality and the TTS synthesis quality of the evaluated vocoders for all voices. To better facilitate the uncontrolled nature of the crowd-sourcing platform, the crowd-sourced tests utilized the degradation mean opinion score (DMOS) test [49] where a natural recording of an utterance is played as a reference and the subjects rate the quality degradation of the corresponding TTS synthesis sample on a scale from 1 to 5 where a higher score denotes better quality. This test type produces only one data point per test case, which makes single cases easier to evaluate for the listener compared to the MUSHRA test. Furthermore, the choice of the DMOS scale as opposed to the mean opinion score (MOS) or comparative mean opinion score (CMOS) was motivated by the fact that DMOS grounds the conducted tests to the original voice; an ideal vocoder should reproduce the characteristic sound of the original speaker. In the crowd-sourced tests (hereafter referred to as *DMOS tests*), TTS quality was evaluated both using vocoder-specific "out-of-the-box" parameters (i.e., same samples as in the MUSHRA TTS test) and using the special case of shared spectral features based on the STRAIGHT envelope, but with a vocoder-specific synthesis procedure. The use of the shared spectral envelope test is justified because it enables assessing the relevance of different analysis and synthesis procedures on the synthesis quality.

A. Vocoder setup

The speech used in the experiments was processed at a full-band (48 kHz) sampling frequency with a frame length of 25 milliseconds and a frame rate of 5 milliseconds. The vocoder parameter orders were selected to be of equal dimensions in the four vocoders compared. For STRAIGHT and PML, the spectral envelope was parameterized as mel-generalized cepstral coefficients [40] (order $p = 60$) with a warping factor of $\lambda = 0.77$, corresponding to the Bark scale. For GlottDNN and GlottHMM, the vocal tract filter was parameterized with a LSF representation of the order $p = 50$, with a warping factor of $\lambda = 0.54$ corresponding to the mel-scale. Using different warping factors is justified by the fact that in contrast to STRAIGHT and PML where the initial envelope analyzed on a linear scale is warped to the target scale, the glottal vocoders work in the warped frequency domain and are thus more prone to biasing of formants by harmonics that might degrade quality [50]. Additionally, the glottal vocoders used a filter of the order $m = 10$ to represent the spectral tilt of the glottal excitation (parameterized as LSFs).

The aperiodicity coefficients for each vocoder were parameterized with 25 parameters using the log-average value for ERB bands for GlottDNN, GlottHMM, and STRAIGHT. In PML, the mel-compressed average classification values for the binary noise mask values were parameterized.

The f_0 analysis was performed with a vocoder-independent setup using multiple pitch detection algorithms (SWIPE [51], RAPT [52], tempo [10]), with the final f_0 trajectories determined by a median vote of the used methods. All systems shared the same f_0 information in all parts of the experiments (during analysis and synthesis) to eliminate the effect of differing f_0 trajectories on vocoder performance.

Finally, the synthesis output of each vocoder was objectively normalized in terms of loudness according to the ITU-P.56 recommendation [53].

B. Objective analysis of voice characteristics

As discussed in Section IV, the voices of the experiments were selected to exhibit varying speaker characteristics. This is illustrated in Figure 5 using distribution histograms of three general acoustic parameters for all four voices. The analysis was computed from a subset of 2000 utterances for each voice, including only voiced frames. The analyzed features were the log f_0 and the first two cepstral coefficients (c_0 and c_1), which reflect the energy and spectral tilt of speech, respectively. The log f_0 distributions (Figure 5 (a)) clearly illustrate that "Roger" and "Nancy", the two voices described as highly expressive, have wider distributions compared to "Nick" and "Jenny". This difference is also reflected in the c_0 distributions (Figure 5 (b)), which are slightly more skewed towards right (i.e. high energy) for the more expressive voices 'Roger" and "Nancy'. Finally, by looking at the distributions for c_1 (Figure 5 (c)), it can be clearly seen that "Nick" shows a large occurrence of high c_1 values. This means that compared to the other voices, "Nick" has more energy in low frequencies. The differences in spectral tilt are mainly due to the phonation mode of the glottal excitation. Therefore, the c_1 histogram of "Nick" indicates that this speaker has generally a softer, less pressed mode of phonation.

TABLE I
DATA SPLIT (IN NUMBER OF UTTERANCES) FOR EACH VOICE AND TOTAL DURATION OF THE CORPUS.

Voice	Training	Development	Testing	Total Duration
Nick	2400	70	72	1hr 47mins
Roger	4358	150	150	7hrs
Jenny	4063	150	150	7hrs 51mins
Nancy	11682	200	200	16hrs 44mins

C. The TTS system setup

The Merlin speech synthesis toolkit [54] (with a few modifications to ensure vocoder compatibility) was used to build the TTS voices. To get the phone durations, force alignment based on hidden Markov models (HMM) was done at state-level. Five-state HMM duration models were trained using mel-frequency cepstral coefficients (MFCCs) for each speaker. Mono-phoneme labels were converted to full-context phoneme labels using the Festival toolkit [55]. To map the ASCII text onto phonemic sound units, combilex [56] lexicon was used. The full contextual labels were mapped onto binary and real values at frame level using an HTS-style question file [57]. The dimension of the input labels was 335, of which the last nine values contain information about the duration of the phoneme, such as the position of the current frame in the current phoneme. Min–max normalization was applied on input features that scaled the features into the range of [0.01 to 0.99]. The output features, which depend upon the vocoder parameters, were scaled using the mean-variance normalization technique. The F0 was linearly interpolated before modeling, and a binary feature was used to record the voiced/unvoiced information for each frame.

For acoustic modeling, models based on LSTM [43] were employed to map the linguistic features to acoustic vocoder features (including static frame-level vocoder features and the corresponding Δ and $\Delta\Delta$ features). The architecture of the neural network consists of four hidden layers of the sizes 256, 128, 512, and 512 hidden units. The first three hidden layers were feed-forward layers with a tanh activation function and the last hidden layer was an LSTM layer. The final output layer had a linear activation function. The mean square error between predicted and actual acoustic parameters was used as a cost function. The stochastic gradient decent (SGD) optimization algorithm was used to learn the parameters. The learning rate was set to a constant of value 0.002 for the first 10 epochs and afterwards it was decreased by half for each epoch. The initial momentum value was set to 0.3 and later increased to 0.9 after the first 10 epochs. The mini-batch size was set to 256 and the models were trained for 25 epochs. To increase the generalization accuracy, an early stopping criterion was employed.

Within synthesis time, oracle durations based on the forced alignment of the test set utterances were utilized to ensure minimal prosodic quality degradations and to enable the use of the original utterances as listening test references. This was also done to emphasize the effect of the vocoder differences in the listening tests. The parameters generated by the acoustic model were finalized with the maximum likelihood parameter generation (MLPG) algorithm [58], and finally, straightforward post-filtering [59] was applied to the spectral features to increase formant dynamics: For PML and STRAIGHT, the values of the cepstral coefficients (outside the first two) were multiplied by a constant of 1.4. For GlottDNN and GlottHMM, the spectral valleys of the synthesized vocal tract magnitude envelopes were multiplied by a constant of 0.3 [60]. Both of the utilized post-filtering methods are highly similar in their function (they boost the dynamics of the spectral envelope peaks with a constant factor), and they can be considered as well-known baseline methods. More advanced post-filtering techniques have been developed (e.g., [45], [61]), but we chose the baseline methods to keep the focus of the study on the vocoder differences, which should not be affected by the choice of the post-filtering method.

D. The MUSHRA test setup

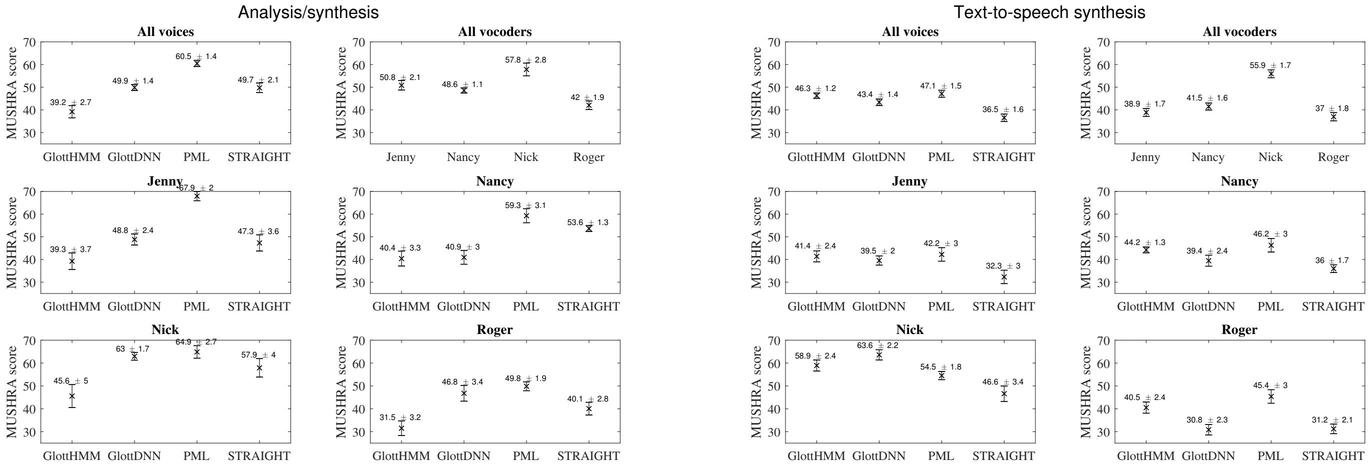
Both of the MUSHRA tests included 40 test utterances (10 per voice) with samples from four vocoders, alongside hidden references (original speech utterances) and anchors (vocoded speech with over-simplified impulse excitation). Sixteen proficient English speakers (international university students, out of whom 10 were native English speakers) were recruited for the tests where the task was to rate (on a scale from 0 to 100) the *overall quality* of the samples compared to the real speech reference. The subjects performed the tests inside sound proofed listening booths using Sennheiser HD 650 headphones and a MOTU UltraLite Mk3 audio interface.

The difference between the two MUSHRA tests was straightforward: In the analysis-synthesis test, the original vocoder parameter trajectories were used to synthesize the test utterances and in the TTS test the synthesized parameter trajectories (with oracle durations) were used.

E. The DMOS test setup

The crowd-sourced DMOS tests were carried out on the CrowdFlower platform [62]. For each test, a total of 40 test utterances were selected (10 for each voice), alongside reference samples to test subject attentiveness. To participate in the test, the subject had to pass a pre-screening test consisting of reference pairs and anchor samples (the same as in the MUSHRA test) with at least 70% accuracy (i.e., on a discrete scale from 1 to 5, the quality of a reference sample had to be rated as 4 or 5, and an anchor sample had to be rated as 1 or 2). Approximately 200 participants (219 and 194 for the first and second test, respectively) attended each test, averaging approximately 30 evaluations per listener (or about 150 evaluations per test utterance).

1) The TTS synthesis quality test with vocoder-specific parameters: The test samples for the "out-of-the-box" vocoder quality test were generated with the same system as those in the MUSHRA test on TTS quality. This overlap acts as a control on the test setup facet of our tests: Any differences seen between this test and the MUSHRA TTS test are caused by either the test type or audio conditions at the listeners' site.



(a) The results of the MUSHRA analysis-synthesis test.

(b) The results of the MUSHRA TTS synthesis test.

Fig. 6. MUSHRA test results (means and their 95% confidence intervals) obtained with repeated measures ANOVA model ($N = 16$). For all presented panels the rANOVA $p \ll 10^{-3}$.

TABLE II
STUDENT'S t-TEST RESULTS FOR THE MUSHRA ANALYSIS-SYNTHESIS TEST WITH t-VALUES REPORTED FOR DEGREES OF FREEDOM = 15. STATISTICALLY SIGNIFICANT RESULTS WITH BONFERRONI CORRECTION ($p < 0.0083$) ARE SHOWN IN A BOLD FONT.

All voices	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = -10.2, p \approx 10^{-8}}$	$\mathbf{t = -14.4, p \approx 10^{-10}}$	$\mathbf{t = -11.1, p \approx 10^{-8}}$
GlottDNN	—	$\mathbf{t = -11.6, p \approx 10^{-8}}$	$t = 0.13, p = 0.90$
PML	—	—	$\mathbf{t = 10.0, p \approx 10^{-8}}$
Jenny	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = -5.7, p \approx 10^{-5}}$	$\mathbf{t = -15.3, p \approx 10^{-10}}$	$\mathbf{t = -5.6, p \approx 10^{-5}}$
GlottDNN	—	$\mathbf{t = -16.4, p \approx 10^{-10}}$	$t = 0.89, p = 0.39$
PML	—	—	$\mathbf{t = 12.0, p \approx 10^{-9}}$
Nancy	GlottDNN	PML	STRAIGHT
GlottHMM	$t = -0.52, p = 0.61$	$\mathbf{t = -7.0, p \approx 10^{-6}}$	$\mathbf{t = -8.9, p \approx 10^{-7}}$
GlottDNN	—	$\mathbf{t = -7.4, p \approx 10^{-6}}$	$\mathbf{t = 10.1, p \approx 10^{-8}}$
PML	—	—	$\mathbf{t = 3.3, p = 0.005}$
Nick	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = -8.7, p \approx 10^{-7}}$	$\mathbf{t = -9.0, p \approx 10^{-7}}$	$\mathbf{t = 8.6, p \approx 10^{-7}}$
GlottDNN	—	$t = -1.8, p = 0.09$	$t = 2.8, p = 0.01$
PML	—	—	$\mathbf{t = 4.7, p \approx 10^{-4}}$
Roger	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = -9.8, p \approx 10^{-7}}$	$\mathbf{t = -14.7, p \approx 10^{-10}}$	$\mathbf{t = -4.6, p \approx 10^{-4}}$
GlottDNN	—	$t = -2.0, p = 0.07$	$\mathbf{t = 3.2, p = 0.006}$
PML	—	—	$\mathbf{t = 8.1, p \approx 10^{-6}}$
All vocoders	Nancy	Nick	Roger
Jenny	$t = 2.6, p = 0.025$	$\mathbf{t = -6.2, p \approx 10^{-5}}$	$\mathbf{t = 6.1, p \approx 10^{-5}}$
Nancy	—	$\mathbf{t = -9.5, p \approx 10^{-7}}$	$\mathbf{t = 6.0, p \approx 10^{-5}}$
Nick	—	—	$\mathbf{t = 9.5, p \approx 10^{-7}}$

TABLE III
STUDENT'S t-TEST RESULTS FOR THE MUSHRA TTS TEST WITH t-VALUES REPORTED FOR DEGREES OF FREEDOM = 15. STATISTICALLY SIGNIFICANT RESULTS WITH BONFERRONI CORRECTION ($p < 0.0083$) ARE SHOWN IN A BOLD FONT.

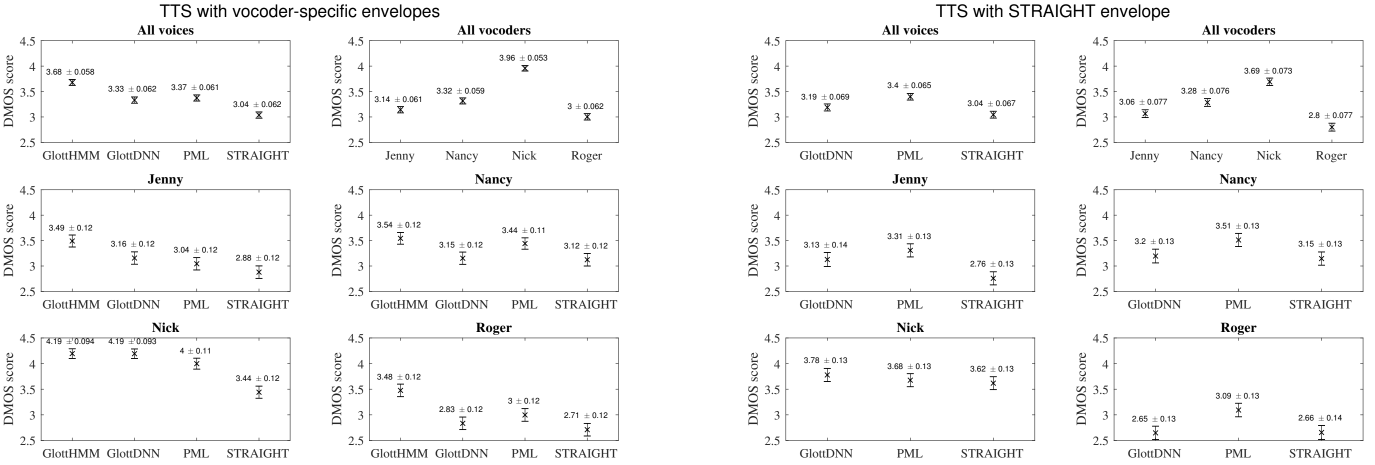
All voices	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = 3.3, p = 0.005}$	$t = -0.94, p = 0.36$	$\mathbf{t = 12.9, p \approx 10^{-9}}$
GlottDNN	—	$\mathbf{t = -6.3, p \approx 10^{-5}}$	$\mathbf{t = 7.7, p \approx 10^{-6}}$
PML	—	—	$\mathbf{t = 14.5, p \approx 10^{-10}}$
Jenny	GlottDNN	PML	STRAIGHT
GlottHMM	$t = 1.3, p = 0.21$	$t = -0.48, p = 0.64$	$\mathbf{t = 5.3, p \approx 10^{-4}}$
GlottDNN	—	$t = -2.3, p = 0.04$	$\mathbf{t = 4.8, p \approx 10^{-4}}$
PML	—	—	$\mathbf{t = 8.4, p \approx 10^{-7}}$
Nancy	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = 4.2, p \approx 0.001}$	$t = -1.5, p = 0.15$	$\mathbf{t = 10.5, p \approx 10^{-8}}$
GlottDNN	—	$\mathbf{t = -4.9, p \approx 10^{-4}}$	$\mathbf{t = 3.9, p = 0.001}$
PML	—	—	$\mathbf{t = 8.7, p \approx 10^{-7}}$
Nick	GlottDNN	PML	STRAIGHT
GlottHMM	$t = -2.8, p = 0.012$	$\mathbf{t = 3.6, p = 0.003}$	$\mathbf{t = 8.1, p \approx 10^{-7}}$
GlottDNN	—	$\mathbf{t = 8.4, p \approx 10^{-7}}$	$\mathbf{t = 10.7, p \approx 10^{-8}}$
PML	—	—	$\mathbf{t = 6.6, p \approx 10^{-5}}$
Roger	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{t = 9.1, p \approx 10^{-7}}$	$\mathbf{t = -3.7, p = 0.002}$	$\mathbf{t = 10.8, p \approx 10^{-8}}$
GlottDNN	—	$\mathbf{t = -10.3, p \approx 10^{-8}}$	$t = -0.3, p = 0.78$
PML	—	—	$\mathbf{t = 9.5, p \approx 10^{-7}}$
All vocoders	Nancy	Nick	Roger
Jenny	$t = -2.5, p = 0.025$	$\mathbf{t = -18.2, p \approx 10^{-11}}$	$t = 1.7, p = 0.10$
Nancy	—	$\mathbf{t = -14.6, p \approx 10^{-10}}$	$\mathbf{t = 4.7, p \approx 10^{-4}}$
Nick	—	—	$\mathbf{t = 16.0, p \approx 10^{-10}}$

2) *The TTS synthesis quality test with a STRAIGHT envelope:* For the test on TTS synthesis quality with the STRAIGHT envelope, the GlottDNN, PML, and STRAIGHT vocoders were used with the post-filtered STRAIGHT envelopes. GlottHMM was omitted from this test, because it does not support synthesis with a generic spectral envelope model. For GlottDNN, the synthesis filtering with the STRAIGHT spectral model was implemented as follows: First, the glottal excitation signal was generated as described in Section II-D. Next, the excitation was filtered in the frequency domain with a fixed frame rate so that the envelope of the magnitude spectrum matches the STRAIGHT spectrum (no additional post-filtering was performed). The phase is set according to the phase of the excitation and the minimum phase response of the STRAIGHT envelope, compensated by the spectral tilt of the excitation. The resulting spectra are inverse transformed into the time domain and overlap-added in order to produce the speech signal. The synthesis models of PML and STRAIGHT directly support the STRAIGHT envelope representation, so no modifications were needed in these two vocoders.

V. RESULTS

A. *The MUSHRA test results*

The results for the MUSHRA tests are presented in Figure 6. The results were computed as marginal means and their 95% confidence intervals based on the repeated measures analysis of variance (rANOVA) model. In the first row, the overall results for all speakers for the compared vocoders are



(a) The results of the DMOS test with vocoder-specific envelopes.

(b) The results of the DMOS test with STRAIGHT envelopes.

Fig. 7. The results of the crowd-sourced DMOS listening tests ($N \approx 150$). Means and confidence intervals were obtained by fitting them to a normal distribution.

TABLE IV
THE MANN-WHITNEY U-TEST RESULTS FOR THE DMOS TEST ON VOCODER QUALITY WITH VOCODER-SPECIFIC ENVELOPES. STATISTICALLY SIGNIFICANT RESULTS WITH BONFERRONI CORRECTION ($p < 0.0083$) ARE SHOWN IN A BOLD FONT.

All voices	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{z = 10.8, p \approx 10^{-27}}$	$\mathbf{z = 6.8, p \approx 10^{-11}}$	$\mathbf{z = 17.4, p \approx 10^{-68}}$
GlottDNN	—	$\mathbf{z = -4.3, p \approx 10^{-5}}$	$\mathbf{z = 6.6, p \approx 10^{-28}}$
PML	—	—	$\mathbf{z = 11.0, p \approx 10^{-28}}$
Jenny	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{z = 4.0, p \approx 10^{-4}}$	$\mathbf{z = 5.2, p \approx 10^{-7}}$	$\mathbf{z = 9.3, p \approx 10^{-20}}$
GlottDNN	—	$z = 1.4, p \approx 0.17$	$\mathbf{z = 5.9, p \approx 10^{-9}}$
PML	—	—	$\mathbf{z = 4.4, p \approx 10^{-5}}$
Nancy	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{z = 6.7, p \approx 10^{-11}}$	$z = -0.1, p \approx 0.90$	$\mathbf{z = 6.3, p \approx 10^{-10}}$
GlottDNN	—	$\mathbf{z = -7.0, p \approx 10^{-12}}$	$z = -0.4, p \approx 0.68$
PML	—	—	$\mathbf{z = 6.5, p \approx 10^{-11}}$
Nick	GlottDNN	PML	STRAIGHT
GlottHMM	$z = 1.1, p \approx 0.28$	$\mathbf{z = 2.9, p \approx 0.003}$	$\mathbf{z = 10.2, p \approx 10^{-24}}$
GlottDNN	—	$z = 1.8, p \approx 0.07$	$\mathbf{z = 9.2, p \approx 10^{-20}}$
PML	—	—	$\mathbf{z = 7.6, p \approx 10^{-14}}$
Roger	GlottDNN	PML	STRAIGHT
GlottHMM	$\mathbf{z = 10.6, p \approx 10^{-26}}$	$\mathbf{z = 6.3, p \approx 10^{-10}}$	$\mathbf{z = 10.5, p \approx 10^{-25}}$
GlottDNN	—	$\mathbf{z = -4.7, p \approx 10^{-6}}$	$z = -0.3, p \approx 0.79$
PML	—	—	$\mathbf{z = 4.5, p \approx 10^{-5}}$
All vocoders	Nancy	Nick	Roger
Jenny	$\mathbf{z = -5.0, p \approx 10^{-6}}$	$\mathbf{z = -24.8, p \approx 10^{-135}}$	$\mathbf{z = 4.2, p \approx 10^{-5}}$
Nancy	—	$\mathbf{z = -20.5, p \approx 10^{-93}}$	$\mathbf{z = 8.7, p \approx 10^{-18}}$
Nick	—	—	$\mathbf{z = 26.9, p \approx 10^{-159}}$

TABLE V
THE MANN-WHITNEY U-TEST RESULTS FOR THE DMOS TEST ON VOCODER QUALITY WITH THE STRAIGHT ENVELOPE MODEL. STATISTICALLY SIGNIFICANT RESULTS WITH BONFERRONI CORRECTION ($p < 0.0167$) ARE SHOWN IN A BOLD FONT.

All voices	PML	STRAIGHT	—
GlottDNN	$\mathbf{z = -6.8, p \approx 10^{-11}}$	$\mathbf{z = 3.3, p \approx 0.001}$	—
PML	—	$\mathbf{z = 10.1, p \approx 10^{-24}}$	—
Jenny	PML	STRAIGHT	—
GlottDNN	$\mathbf{z = -3.6, p \approx 10^{-4}}$	$\mathbf{z = 4.8, p \approx 10^{-6}}$	—
PML	—	$\mathbf{z = 8.3, p \approx 10^{-16}}$	—
Nancy	PML	STRAIGHT	—
GlottDNN	$\mathbf{z = -5.6, p \approx 10^{-8}}$	$z = -0.5, p \approx 0.60$	—
PML	—	$\mathbf{z = 5.1, p \approx 10^{-7}}$	—
Nick	PML	STRAIGHT	—
GlottDNN	$z = 2.3, p = 0.02$	$\mathbf{z = 2.6, p = 0.01}$	—
PML	—	$z = 0.3, p \approx 0.33$	—
Roger	PML	STRAIGHT	—
GlottDNN	$\mathbf{z = -6.8, p \approx 10^{-11}}$	$z = 0.3, p \approx 0.75$	—
PML	—	$\mathbf{z = 7.1, p \approx 10^{-12}}$	—
All vocoders	Nancy	Nick	Roger
Jenny	$\mathbf{z = -5.0, p \approx 10^{-6}}$	$\mathbf{z = -24.8, p \approx 10^{-135}}$	$\mathbf{z = 4.2, p \approx 10^{-5}}$
Nancy	—	$\mathbf{z = -20.5, p \approx 10\ -93}$	$\mathbf{z = 8.7, p \approx 10^{-18}}$
Nick	—	—	$\mathbf{z = 26.9, p \approx 10^{-159}}$

presented (column 1), as well as the overall results for all vocoders for the different speakers (column 2). The bottom two rows present the voice-specific results for the vocoders. Post-hoc tests for statistically significant differences between evaluated pairs, presented in Tables II and III, were performed with Student's t-test with Bonferroni correction (i.e., $p < \frac{0.05}{N_{\text{cases}}}$ is considered statistically significant).

1) The analysis-synthesis quality test: The MUSHRA analysis-synthesis test results are presented in Figure 6a, and the Student's t-test results between all systems are presented in Table II. Looking at the overall results, it can be seen that PML is the best performing vocoder (with a statistically significant margin), followed by GlottDNN and STRAIGHT without statistically significant differences. GlottHMM has the worst performance with a significant difference compared to the other vocoders. By looking at the voice-specific mean scores averaged over all vocoders, it can be observed that the "Nick" voice has the best overall performance by a significant margin, whereas the reverberant "Roger" voice seems to be the least suitable for vocoding.

The voice-specific scores for the vocoders reveal a clear trend for the GlottDNN vocoder: The low-pitched male voices performed relatively well (no significant differences compared to PML), but the more high-pitched female voices dropped in relative performance. For the other vocoders, the voice-specific analysis-synthesis results are consistent between voices.

2) TTS synthesis quality test: The MUSHRA TTS test results are presented in Figure 6b, and the Student's t-test results for all systems are presented in Table III. The overall results indicate PML and GlottHMM to be the best performing vocoders without statistically significant differences. An important thing to note is that when switching from analysis-synthesis to TTS, GlottHMM counter-intuitively *increases* its score within the TTS context, whereas the overall scores drop for the other vocoders. The overall performance of GlottDNN is slightly behind PML and GlottHMM with statistically significant margins, and STRAIGHT is the worst performing vocoder with significant margins. The voice-specific scores averaged across all vocoders indicate that the differences between the voices are smaller than within analysis-synthesis, except for "Nick" whose TTS performance is on a par with the analysis-synthesis MUSHRA score.

Within the voice-specific scores for the vocoders, more specific differences arise than in the analysis-synthesis test. For the "Jenny" voice, the GlottDNN, GlottHMM, and PML do not have statistically significant differences. For "Nancy", PML and GlottHMM have similar results (no significant difference), followed by GlottDNN and STRAIGHT (with significant differences). For "Nick", GlottDNN has the highest score that, however, is not deemed as statistically significant compared to GlottHMM after the Bonferroni correction. For "Roger" the performance of GlottDNN drops greatly compared to the analysis-synthesis performance, with PML having the best quality. We speculate the the drastic relative drop in GlottDNN performance is caused by the reverberation that is present in the recordings. The excitation generation DNN learns to replicate this audio quality, which becomes more apparent in TTS.

B. The DMOS test results

Due to the nature of the crowd-sourcing platform, complete data for all test cases could not be obtained for each listener. Furthermore, since the DMOS test utilizes discrete ranking with sparse categories without anchor points, the linearity of the ranking scale is questionable. Due to this, ANOVA-based statistical methods are not suitable for the analysis of mean opinion score (MOS)-type tests, and the non-parametric Mann-Whitney U-test was used instead. The Mann-Whitney U-test with bias control for listener and utterance variability is recommended in [63] for MOS tests. Since the reasoning behind this test for MOS arise from equivalent concerns as to those in our DMOS test, we applied the recommended procedures as described in [63] for our tests. The means and their confidence intervals are obtained by parameter fitting to a normal distribution based on the raw test data.

1) The TTS synthesis quality test with vocoder-specific parameters: The "out-of-the-box" TTS quality DMOS test results are presented in Figure 7a, and the Mann-Whitney U-test results for all systems are presented in Table IV. GlottHMM gets the best DMOS score with significant margins. The mean results for GlottDNN and PML are highly similar, but a significant difference in favor of PML can be seen from the U-test. The overall voice-specific results are highly similar to the MUSHRA test for TTS quality, as expected, with only the performance of GlottHMM being elevated for all voices. The "Nick" voice receives a DMOS score of 4.19 for the GlottHMM and GlottDNN vocoders, which indicates a peculiarly high quality as the quality is deemed, on average, to be above "good" compared to the corresponding natural speech sample. The U-test does not indicate a statistically significant difference between GlottDNN and PML however.

2) The TTS synthesis quality test with a STRAIGHT envelope: The DMOS test results for TTS with a STRAIGHT envelope are presented in Figure 7b, and the Mann-Whitney U-test results for all systems are presented in Table V. The overall performance of STRAIGHT in both tests is nearly identical, as expected. PML slightly increases its mean score, but the differences are within the error margins, whereas the overall performance of GlottDNN decreases. The overall scores have statistically significant differences between all vocoders. The voice-specific scores averaged over vocoders cannot be directly compared to the results of the first DMOS test due to the absence of GlottHMM within the data. Within the voice-specific scores for the vocoders, the use of the STRAIGHT envelope increases the performance of PML within all of the voices except for "Nick". For GlottDNN, the DMOS score for "Nancy" is slightly increased and for "Nick" it is greatly decreased.

VI. Summary and conclusions

Vocoders from the three main categories (mixed excitation, glottal, sinusoidal vocoders) were compared in the current study with formal and crowd-sourced listening tests. Vocoder quality was measured within the context of analysis-synthesis as well as TTS synthesis. Furthermore, the TTS experiments were divided into synthesis with vocoder-specific features and synthesis with a shared envelope model, where the waveform generation method of the vocoders is mainly responsible for the quality differences. Finally, all of the tests included four distinct voices as a way to investigate the effect of different speakers on the resulting vocoder-generated speech quality.

The results presented in Section V reveal many interesting facets about the role of vocoders in statistical parametric speech synthesis. Most importantly, the choice of the voice has a profound impact on the overall quality of the vocoder-generated voice, and the best vocoder for each voice can vary case by case (e.g., see the performance of GlottDNN for "Nick" compared to the other voices). Furthermore, it can be seen that PML has the best overall performance across all of the performed tests. The overall best performance of "Nick" can be attributed to the voice characteristics presented in Section IV-B, which show that 'Nick" has the lowest average f_0 with the smallest deviation, alongside a larger spectral tilt compared to the other voices. Both of these properties are beneficial for accurate extraction of the spectral envelope. In addition, the steep spectral tilt downgrades the relative importance of high frequencies, where the stochastic components are more prominent, in assessing signal quality. This also explains the good performance of the glottal vocoders with the "Nick" voice.

Second, the analysis-synthesis quality of a vocoder cannot be used as a reliable predictor of TTS synthesis quality with the current acoustic models (see the remarkable absolute and relative performance gains of GlottHMM in TTS compared to analysis-synthesis). Based on this difference for GlottHMM, it can be argued that the shortcomings of the spectral model (from which GlottHMM suffers in analysis-synthesis) are averaged out in current SPSS acoustic models. Furthermore, when comparing the performance differences between GlottHMM and GlottDNN, we speculate that the quality of the DNN-based excitation of GlottDNN is highly voice-specific (for example, the performance of "Roger" decreased considerably because of the picked-up reverberation), as is also concluded in a separate study [64]. The simple excitation generation based on a high-quality glottal pulse utilized by GlottHMM is thus a safer option for more challenging voices.

The DMOS tests reveal that when controlling for the spectral models of the vocoders (where the perceived differences arise from the waveform generation procedure), the differences between vocoders are similar to the baseline results. This indicates that the waveform synthesis method of a vocoder is essential for quality improvements, which is also reported in [18]. In PML, the performance gains compared to STRAIGHT are achieved by a more sophisticated noise model, whereas in GlottDNN the main difference is the more physiologically oriented phase information of the excitation (whereas STRAIGHT and PML use minimum phase responses of the overall spectral envelope). This suggests that in future research, the integration of these approaches—the accurate excitation phase generation based on the vocoder features, and a more natural stochastic texture based on the binary noise mask features—could be beneficial.

Acknowledgment

The research leading up to this publication has received funding from the Academy of Finland (project no. 312490, 284671).

References

[1] M. Cooke, S. King, M. Garnier, and V. Aubanel, "The listening talker: A review of human and algorithmic context-induced modifications of speech," *Computer Speech & Language*, vol. 28, no. 2, pp. 543 – 571, 2014.

[2] H. Zen, K. Tokuda, and A. W. Black, "Review: Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[3] Y. Stylianou, "Voice transformation: A survey," in *Proc. ICASSP*. IEEE, 2009, pp. 3585 – 3588.

[4] P. R. Cook, "Toward the perfect audio morph? singing voice synthesis and processing," in *Proc. Workshop on Digital Audio Effects*, 1998.

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

[6] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*, ser. Prentice-Hall signal processing series. Prentice-Hall, 1978.

[7] J. Flanagan and R. Golden, "Phase vocoder," *Bell System Technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[8] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*, ser. Signals and Communication Technology. Springer International Publishing, 2017.

[9] A. Sorin, S. Shechtman, and A. Rendel, "Semi parametric concatenative TTS with instant voice modification capabilities," in *Proc. Interspeech*, 2017, pp. 1373–1377.

[10] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive timefrequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 34, pp. 187 – 207, 1999.

[11] H. Kawahara, J. Estill, and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," in *Proc. MAVEBA*, 2001.

[12] M. Morise, F. Yokomori, and K. Ozawa, "WORLD: A vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 7, pp. 1877–1884, 2016.

[13] T. Drugman and T. Dutoit, "The deterministic plus stochastic model of the residual signal and its applications," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 968–981, 2012.

[14] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 153–165, 2011.

[15] 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, 2014.

[16] G. Degottex and D. Erro, "A uniform phase representation for the harmonic model in speech synthesis applications," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, p. 38, Oct 2014.

[17] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 184–194, 2014.

[18] Y. Agiomyrgiannakis, "Vocaine the vocoder and applications in speech synthesis," in *Proc. ICASSP*. IEEE, 2015.

[19] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *8th ISCA Workshop on Speech Synthesis*, 2013, pp. 155 – 160.

[20] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *CoRR*, vol. abs/1609.03499, 2016.

[21] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. V. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: A fully end-to-end text-to-speech synthesis model," *CoRR*, vol. abs/1703.10135, 2017.

[22] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Synthesis and perception of breathy, normal, and lombard speech in the presence of noise," *Computer Speech & Language*, vol. 28, no. 2, pp. 648 – 664, 2014.

[23] T. Raitio, H. Lu, J. Kane, A. Suni, M. Vainio, S. King, and P. Alku, "Voice source modelling using deep neural networks for statistical parametric speech synthesis," in *Proc. EUSIPCO*, 2014.

[24] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. Interspeech*. ISCA, 2014, pp. 1969–1973.

[25] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. ICASSP*. IEEE, 2016.

[26] M. Airaksinen, B. Bollepalli, J. Pohjalainen, and P. Alku, "Glottal vocoding with frequency-warped time-weighted linear prediction," *IEEE Signal Processing Letters*, vol. 24, no. 4, pp. 446 – 450, 2017.

[27] G. Degottex, P. Lanchantin, and M. Gales, "A pulse model in log-domain for a uniform synthesizer," in *Proc. 9th ISCA Speech Synthesis Workshop*, 2016, pp. 230 – 236.

[28] M. Airaksinen, T. Raitio, B. Story, and P. Alku, "Quasi closed phase glottal inverse filtering analysis with weighted linear prediction," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 3, pp. 596–607, 2014.

[29] C. d'Alessandro, B. Bozkurt, B. Doval, T. Dutoit, N. Henrich, V. N. Tuan, and N. Sturmel, *Phase-Based Methods for Voice Source Analysis*. Springer Berlin Heidelberg, 2007, pp. 1–27.

[30] C. Ma, Y. Kamp, and L. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 2, pp. 69–81, 1993.

[31] P. Alku, J. Pohjalainen, M. Vainio, A.-M. Laukkanen, and B. H. Story, "Formant frequency estimation of high-pitched vowels using weighted linear prediction," *The Journal of the Acoustical Society of America*, vol. 134, no. 2, pp. 1295 – 1313, 2013.

[32] F. Itakura, "Line spectrum representation of linear predictor coefficients of speech signals," *The Journal of the Acoustical Society of America*, vol. 57, no. S1, pp. S35–S35, 1975.

[33] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697–708, 1999.

[34] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[35] A. Härmä and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 579–588, November 2001.

[36] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech communication*, vol. 16, no. 2, pp. 175–205, 1995.

[37] M. Airaksinen, B. Bollepalli, L. Juvela, Z. Wu, S. King, and P. Alku, "GlottDNN — A full-band glottal vocoder for statistical parametric speech synthesis," in *Proc. Interspeech*. ISCA, 2016.

[38] J. Juvela, X. Wang, S. Takaki, M. Airaksinen, J. Yamagishi, and P. Alku, "Using text and acoustic features in predicting glottal excitation waveforms for parametric speech synthesis with recurrent neural networks," in *Proc. Interspeech*. ISCA, 2014.

[39] B. Bollepalli, L. Juvela, and P. Alku, "Generative adversarial network-based glottal waveform model for statistical parametric speech synthesis," in *Proc. Interspeech*. ISCA, 2017, pp. 3394 – 3398.

[40] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. ICSLP*, 1994.

[41] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109 – 118, 1992.

[42] T. Masuko, K. Tokuda, T. Kobayashi, and S. Imai, "Speech synthesis using HMMs with dynamic features," in *Proc. ICASSP*, vol. 1. IEEE, 1996, pp. 389 – 392.

[43] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*. ISCA, 2014.

[44] Z. Wu and S. King, "Investigating gated recurrent networks for speech synthesis," in *Proc. ICASSP*. IEEE, 2016, pp. 5140 – 5144.

[45] T. Kaneko, H. Kameoka, N. Hojo, Y. Ijima, K. Hiramatsu, and K. Kashino, "Generative adversarial network-based postfilter for statistical parametric speech synthesis," in *Proc. ICASSP*. IEEE, 2017, pp. 4910 – 4914.

[46] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus," 2013, LISTA Consortium. [Online]. Available: http://dx.doi.org/10.7488/ds/140

[47] S. King and V. Karaiskos, "The Blizzard challenge 2011," in *Blizzard Challenge 2011 Workshop*, 2011.

[48] ITU, "ITU-R BS.1534 (method for the subjective assessment of intermediate quality levels of coding systems)," 2015.

[49] ——, "Methods for subjective determination of transmission quality," in *International Telecommunication Union, Recommendation ITU-T P.800*, 1996.

[50] T. Raitio, A. Suni, M. Vainio, and P. Alku, "Wideband parametric speech synthesis using warped linear prediction," in *Proc. Interspeech*, 2012.

[51] A. Camacho and J. G. Harris, "A sawtooth waveform inspired pitch estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 124, no. 3, pp. 1638–1652, 2008.

[52] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. Elsevier, 1995.

[53] ITU, "ITU-T P.56 (Objective measurement of active speech level)," 2011.

[54] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *9th ISCA Speech Synthesis Workshop*, 9 2016, pp. 218 – 223.

[55] "Festival." [Online]. Available: http://www.festvox.org/festival/

[56] K. Richmond, R. Clark, and S. Fitt, "On generating combilex pronunciations via morphological analysis," in *Proc. Interspeech*. ISCA, 2010, pp. 1974 – 1977.

[57] K. Tokuda, Y. Nankaku, T. Toda, H. Zen, J. Yamagishi, and K. Oura, "Speech synthesis based on hidden Markov models," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1234 – 1252, 2013.

[58] K. Tokuda, T. Kobayashi, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. ICASSP*. IEEE, 2000, pp. 1315 – 1318.

[59] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Incorporating a Mixed Excitation Model and Postfilter into HMM-based Text-to-speech Synthesis," *Systems and Computers in Japan*, vol. 36, no. 12, pp. 43 – 50, 2005.

[60] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Comparison of formant enhancement methods for HMM-based speech synthesis." in *SSW*, 2010, pp. 334–339.

[61] A. Sorin, S. Shechtman, and V. Pollet, "Uniform speech parameterization for multi-form segment synthesis," in *Proc. Interspeech*. ISCA, 2011, pp. 337–340.

[62] "Crowdflower." [Online]. Available: http://www.crowdflower.com

[63] A. Rosenberg and B. Ramabhadran, "Bias and statistical significance in evaluating speech synthesis with mean opinion scores," in *Proc. Interspeech*. ISCA, 2017, pp. 3976 – 3980.

[64] M. Airaksinen and P. Alku, "Effects of training data variety in generating glottal pulses from acoustic features with DNNs," in *Proc. Interspeech*. ISCA, 2017.