FREQUENCY-WARPED TIME-WEIGHTED LINEAR PREDICTION FOR GLOTTAL VOCODING

Manu Airaksinen[1]*, Bajibabu Bollepalli*[1]*, Jouni Pohjalainen*[2]*, Paavo Alku*[1]

[1]Aalto University, Finland
[2]University of Passau, Germany

ABSTRACT

Auto-regressive modeling is a prevalent source-filter separation method of speech. Conventional linear prediction (LP) and its derivatives such as weighted linear prediction (WeLP) produce parametric spectral models within a linear frequency scale, whereas frequency-warped linear prediction (WaLP) can be used to take into account the frequency sensitivity of the human auditory system. From the perspective of glottal vocoding, the principles behind WeLP have been found to be beneficial for an accurate separation of the glottal source signal and the vocal tract transfer function, but this approach can not utilize the auditory benefits of frequency warping. On the other hand, the WaLP approach suffers from less accurate source-filter separation properties. In this study, a generalized frequency-warped time-weighted linear prediction (WWLP) analysis is proposed. Experiments with WWLP are performed within the context of glottal vocoding. The subjective listening test results show that WWLP-based spectral envelope modeling is able to increase quality over previously developed methods in some of the test cases.

Index Terms— Linear prediction, vocoder, glottal inverse filtering, speech synthesis

1. INTRODUCTION

The source-filter model of speech production [1] is a prevalent method of decomposing speech into a representation that enables efficient compression and modeling of the signal in applications such as speech coding [2] and vocoding in statistical parametric speech synthesis (SPSS) [3]. The source-filter model assumes that the speech signal is produced by a convolution between an *excitation signal* and a filter representing the *spectral envelope* of speech. In speech coding, linear prediction (LP) [4] -derived codecs utilizing residual codebooks have been used to obtain high quality speech with low bit-rates [2]. In the field of SPSS-related vocoding, the code-excited linear prediction approach cannot be used as such because of the sparse representation of the residual that is unsuitable for statistical modeling. Commonly in vocoding approaches based on mixed excitation, such as the widely used STRAIGHT vocoder [5, 6], the spectral envelope of the signal is modeled as mel-cepstral (MCEP) coefficients that take into account the frequency sensitivity of the human auditory system [7]. In these vocoders, the voiced excitation signal is modeled as a spectrally flat waveform which is generally parameterized with fundamental frequency (f_0) and some type of a band aperiodicity measure [6].

LP-based vocoding approaches of SPSS have been particularly used in glottal vocoders [3]. In these vocoders, the source-filter model is adopted by separating the (voiced) speech signal into the glottal volume velocity signal, the vocal tract transfer function, and the lip radiation effect. The separation is computed with glottal inverse filtering (GIF) algorithms [8] that almost exclusively use auto-regressive (AR) techniques on a linear frequency scale in modeling of the vocal tract, thereby making the AR coefficients appropriate for the vocoder implementation. With wide-band speech (i.e. sampling frequency $F_s > 16kHz$), a trade-off must be made between the accuracy of the GIF-estimated vocal tract envelope and the auditory accuracy of the AR model: to obtain AR models capable of modeling the most important lowest 3-4 formants accurately, GIF algorithms need to use relatively large AR filter orders (e.g. $p = 80$ for $F_s = 48$ kHz) thereby also allocating extensive resources to model the higher, auditorily less important frequencies. To a certain extent this can be solved with frequency-warped linear prediction (WaLP) [9]. Using WaLP in GIF, however, deteriorates the accuracy of the source-filter decomposition because auditory-based frequency warping has a tendency to focus on harmonic peaks of the excitation in the most important lower frequency bands [10] as illustrated in Figure 1. An alternative approach was studied in [11], where band-wise processing of the speech signal in linear frequency scale was explored in glottal vocoding, thereby achieving improved quality over a previous WaLP-based system.

On the other hand, time-weighted linear prediction (WeLP) [12] has been previously proposed to specifically de-emphasize the contribution of the harmonic peaks in AR envelope modeling. The use of WeLP-based techniques has been proven to be beneficial in glottal inverse filtering and formant estimation [13, 14]. In this study, a fusion of frequency-warped LP (WaLP) and time-weighted LP (WeLP) is proposed to obtain a new AR modeling technique called generalized frequency-warped time-weighted linear prediction (WWLP). With the proposed fusion, we aim to combine psychoacoustical benefits of WaLP to the accurate formant extraction properties of WeLP thereby resulting in new vocal tract models to be used in glottal vocoders. Section 2.1 presents the formulation of WWLP and its weight function, Section 3 briefly documents the use of WWLP in the GlottDNN vocoder [11], Section 4 presents the subjective listening test experiments and results obtained with the WWLP-based vocoder, and Section 5 provides discussion and conclusions of the WWLP approach.

2. FREQUENCY-WARPED TIME-WEIGHTED LINEAR PREDICTION (WWLP)

2.1. WWLP optimization

In AR modeling, sample s_n at time-index n is modeled as a linear combination of p previous samples:

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + G u_n \quad (1)$$

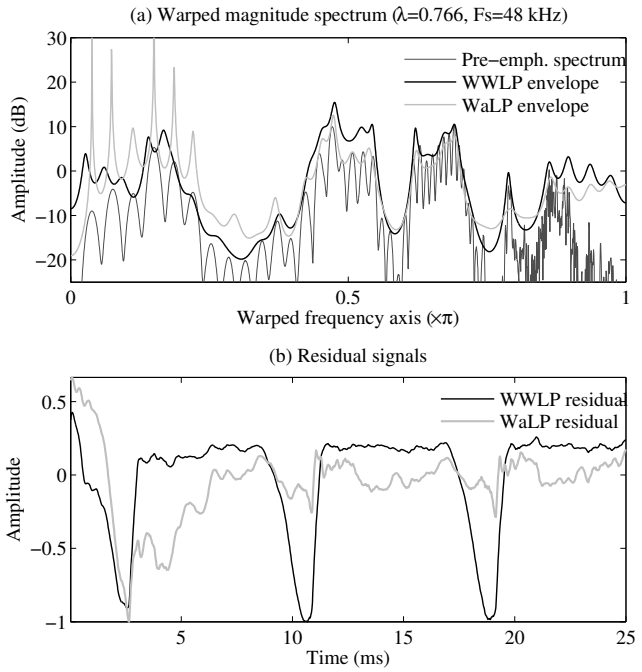


Fig. 1. Example of WWLP- and WaLP -derived (a) spectral envelopes and the corresponding (b) residual signals.

where a_k are denoted as the *prediction coefficients*, G is the filter gain, and u_n is the excitation signal of the AR process [1].

In WaLP, [9] the AR model predicts each sample based on the p previous *warped* samples:

$$s_n = \sum_{k=1}^{p} a_k y_{k,n} + G u_n \tag{2}$$

where $y_{k,n}$ is the output of some general function $D_k(z)$ that models the warped delay line. In WaLP, a cascade of first-order all-pass elements is used to model the delay line:

$$D_k(z) = \prod_{i=1}^{k} \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}}, \tag{3}$$

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

$$y_{k,n} = \begin{cases} s_n & , \quad k = 0 \\ \sum_{m=0}^{\infty} d_{k,m} s_{n-m} & , \quad 1 \le k \le p \end{cases} \tag{4}$$

where $d_{k,m}$ is the mth sample of the impulse response of $D_k(z)$.

To obtain the optimal coefficients a_k of the model in Eq. 2, an optimization criterion must be selected. In conventional LP and WaLP, the squared sum of the prediction error e_n (i.e. the residual) is minimized:

$$E_{\text{WaLP}} = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} (s_n - \sum_{k=1}^{p} a_k y_{k,n})^2, \tag{5}$$

where N denotes the frame length. However, the approach taken with WWLP is to use a *weighted* squared sum as the optimization criterion, akin to that used in Weighted Linear Prediction (WeLP) [12, 15]:

$$E_{\text{WWLP}} = \sum_{n=0}^{N-1} W_n e_n^2 = \sum_{n=0}^{N-1} W_n (s_n - \sum_{k=1}^{p} a_k y_{k,n})^2 \tag{6}$$

where W_n is a temporal positive-valued weight function that enables emphasizing or de-emphasizing residual energy function samples (e_n^2) when optimizing the AR coefficients a_k. With the error criterion set, an analytic solution for the optimal coefficients can be obtained by setting the differentials of Eq. 6 with respect to the prediction coefficients a_k to zero. This leads to a set of normal equations:

$$\sum_n W_n \left(\sum_{k=1}^{p} a_k y_{k,n} \right) y_{j,n} = \sum_n W_n s_n y_{j,n}, 1 \le j \le p \tag{7}$$

In matrix notation, this set of equations is equivalent to

$$\left(\sum_n W_n \mathbf{y}_n \mathbf{y}_n^\intercal \right) \mathbf{a} = \sum_n W_n s_n \mathbf{y}_n \tag{8}$$

where $\mathbf{a}^{p\times 1} = [a_1, a_2, \ldots, a_p]^\intercal$, and $\mathbf{y}_n^{p\times 1} = [y_{1,n}, y_{2,n}, \ldots, y_{p,n}]^\intercal$. Using the following matrix definition

$$\mathbf{R}^{p\times p} = \sum_n W_n \mathbf{y}_n \mathbf{y}_n^\intercal, \tag{9}$$

the optimal solution can be expressed as

$$\mathbf{a}_{\text{opt}} = \mathbf{R}^{-1} \left(\sum_n W_n s_n \mathbf{y}_n \right). \tag{10}$$

When summing from n=0 to n=N-1+p, matrix $\mathbf{R}$ in Eq. 9 can be interpreted as the *frequency-warped* and *time-weighted* autocorrelation matrix of the analyzed signal. Contrary to the matrix obtained in conventional autocorrelation LP or WaLP, $\mathbf{R}$ does not have a Toeplitz structure, so the Levinson-Durbin algorithm [1] can not be used to compute the optimal WWLP coefficients. Instead, the solution must be computed by other methods such as the Cholesky decomposition [1].

2.2. Warping coefficient

The warping coefficient λ is used to determine the amount of all-pass phase warping that takes place in a single unit delay step. For warping values $-1 < \lambda < 1$, the system can be deemed stable [16]. For $\lambda = 0$, the formula of the all-pass element presented in Eq. 3 reduces to a unit delay (z^{-1}), and no frequency warping occurs. For $\lambda > 0$, the frequency resolution is increased for lower bands, and for $\lambda < 0$ the resolution is increased for higher bands.

The frequency warping model of Eq. 3 can be used to satisfactorily represent the frequency sensitivity of the human auditory system with a close matching to the Bark scale [17]. The best matching λ value for the Bark scale for a given sampling frequency F_s is given by [17, 10]

$$\lambda_{F_s} \approx 1.067 \left(\frac{2}{\pi} \arctan(0.06583 f_s/1000) \right)^{1/2} - 0.1916 \tag{11}$$

For example, $\lambda_{48kHz} \approx 0.766$, and $\lambda_{16kHz} \approx 0.576$. While Eq. 11 defines λ simply as a function of F_s, it should be noted, however, that the corresponding value of λ might not be optimal considering

the chosen end-user performance metric, which in many cases is the perceived speech quality. Studies utilizing WaLP have indeed reported that using a value of λ that is lower than that determined in Eq. 11 results in the best perceptual results [18].

2.3. Weight function

The selection of the weight function W_n for the computation of the time-weighted autocorrelation matrix $\mathbf{R}$ is of particular interest especially for voiced speech signals, where speech is produced by the coupling of the glottal excitation signal with the resonances of the vocal tract. In many applications (e.g. formant tracking [14], glottal inverse filtering [13]) it is desirable to decouple the glottal source signal and the vocal tract transfer function. WeLP, with a properly selected W_n, has been shown to provide an effective method to compute such a decoupling and the results show improved accuracy both in formant tracking [14, 19] and glottal inverse filtering [13].

Originally, the short-term energy (STE) function was proposed for WeLP [12], and it has also been used in [20]:

$$W_{n,\mathrm{STE}} = \sum_{k=1}^{M} s_{n-k}^2 \tag{12}$$

where M is the length of the energy window. The STE function can be computed with ease straight from the input signal. With the typical choice of $M = p$, STE generally assumes large values after the glottal closure instants (during the glottal closed phase), and small values during the glottal open phase of the speech signal [12].

Recently, a more precisely crafted attenuated main excitation (AME) weight function was proposed for WeLP in [14] and [13] to obtain more accurate estimates of the vocal tract transfer function. In the quasi-closed phase analysis (QCP) glottal inverse filtering method, the AME weight function is formulated so that the samples within the vicinity of the GCIs of the analysis frame are given (near) zero weighting, and the samples corresponding roughly to the glottal closed phase are given a constant weight. QCP has been shown to provide state-of-the-art results when tested with speech of good recording quality [13, 21]. The effect of the AME weight function to the obtained WWLP spectral envelope is illustrated in Figure 1(a), where it can be seen that even though the lower frequency area of the spectrum is expanded, the envelope avoids excessively modeling the harmonic peaks within the low bands, and the obtained residual signal in Figure 1(b) looks like a glottal flow derivative waveform.

If W_n is constant, the analysis generalizes into conventional LP (for $\lambda = 0$) or WaLP ($\lambda \neq 0$) analysis, which is suitable for unvoiced speech. Table 1 summarizes the various forms of LP analysis that can be achieved by adjusting W_n and λ.

Table 1. *The LP methods to which WWLP generalizes to based on λ and W_n.*

	$\lambda = 0$	$\lambda \neq 0$
W_n constant	LP	WaLP
W_n non-constant	WeLP	WWLP

3. WWLP-BASED GLOTTAL VOCODING

GlottDNN [11] is a glottal vocoder that utilizes the QCP algorithm [13] to decompose speech into the glottal source signal and the vocal tract transfer function. The AR vocal tract transfer function is parametrized as line spectral frequencies (LSFs) [22]. The glottal source is parametrized with its f_0, energy, spectral tilt (as LSFs), and voiced noise component spectral envelope (LSFs + energy). During the vocoder synthesis, the glottal excitation is generated with a deep neural network (DNN) that takes the vocoder feature vector as its input, and produces a two-pitch period glottal flow derivative pulse as its output [23]. After the initial glottal pulse generation, the noise component is added to the pulse, and the obtained sum signal is filtered with a spectral matching filter to match the target spectral tilt envelope. The final voiced excitation signal is produced with a pitch-synchronous overlap-add (PSOLA [24]) procedure of the generated waveforms. Finally, the excitation signal is filtered with an adaptive all-pole filter based on the vocal tract transfer function LSFs.

In the current study, the WWLP-based vocal tract spectral modeling was implemented to GlottDNN by modifying the conventional QCP algorithm to work with WWLP instead of WeLP as in [13]. A warping coefficient of $\lambda = 0.42$ and a prediction order $p = 60$ were chosen, as they are prevalently used in STRAIGHT-based systems. For the training of the excitation generation DNN, the target glottal flow pulses were computed with the original WeLP-based QCP algorithm to ensure optimal quality of the waveforms. The DNN input vectors were computed with WWLP.

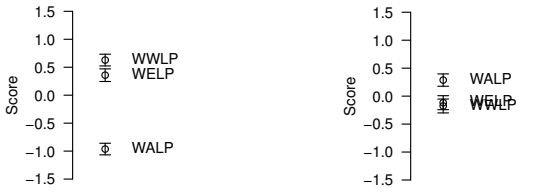
4. EXPERIMENTS

The performance of the WWLP-based GlottDNN vocoder was evaluated in two different subjective listening tests: One for analysis/synthesis quality, and one for text-to-speech (TTS) synthesis quality. For both test types, two different full-band ($F_s = 48$ kHz) voices were used. A male voice, "Nick" [25], and a female voice, "Nancy" [26]. The listening tests were preformed as online tests based on the Beaqlejs application [27].

4.1. Analysis/synthesis listening test

The analysis/synthesis procedure corresponds to parameterizing a speech signal into vocoder parameters, and then re-synthesizing the signal from these parameters. It can be argued that it represents the optimal speech synthesis quality that can be achieved with a given vocoder, as the vocoder parameter trajectories are obtained from natural speech. In the present study, the WWLP-QCP based GlottDNN vocoder was compared to two competing systems in a comparison category rating (CCR) test [28]: One with a straightforward WeLP-QCP based spectral modeling (QCP-based GIF without frequency warping), and one with a WaLP-IAIF based spectral modeling (iterative adaptive inverse filtering [8] based warped GIF without time weighting). The used parameter orders and warping coefficients were kept the same. The motivation for this subjective listening test is to compare the WWLP-based vocal tract envelope estimation method with state-of-the-art LP-based approaches in ideal conditions.

The results of the analysis/synthesis tests are presented in Figure 2. For the "Nancy" voice, the results show that the proposed WWLP method has the best performance, whereas WaLP performs significantly worse. For the "Nick" voice, the differences between the different LP approaches are very small but WaLP has the overall best score. These results indicate that for the chosen p and λ, the more high-pitched female voice suffers from the harmonic peak-modeling problem demonstrated in Figure 1, whereas the male voice is not affected by this problem.



(a) "Nancy" voice. (b) "Nick" voice.

Fig. 2. Subjective listening test results (CCR test) on analysis/synthesis quality for (a) "Nancy", and (b) "Nick"

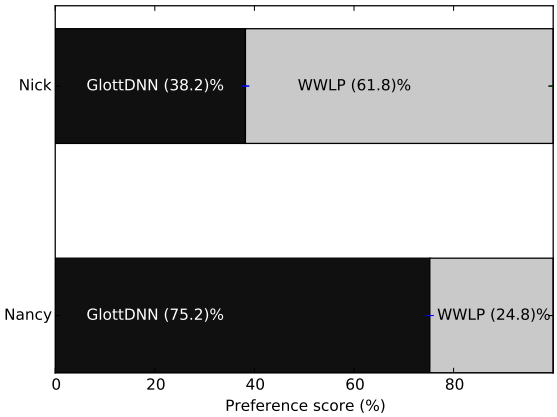


Fig. 3. Subjective listening test results (AB test) on TTS samples.

4.2. Text-to-speech listening test

The subjective TTS quality of the proposed WWLP-QCP vocoder was tested in an AB listening test against the previously proposed GlottDNN vocoder (QMF-QCP) with linear band-wise processing. In [11], QMF-QCP was reported to achieve state-of-the-art TTS quality for the "Nick" voice when compared against the STRAIGHT vocoder and the GlottHMM vocoder [3] while using a DNN-based TTS system [29]. In the present study, the TTS system uses the long short-term memory (LSTM) architecture presented in [30]. The "Nick" voices were trained with 2,400 sentences, and the "Nancy" voices were trained with 11,800 sentences of training data.

The results of the AB listening tests are presented in Figure 3. For the "Nick" voice, WWLP improves the quality over the GlottDNN system, but for "Nancy", the GlottDNN system is preferred. Unexpectedly, this result differs from that obtained in the analysis/synthesis tests presented in Figure 2, where the WWLP system was generally preferred for the "Nancy" voice.

5. DISCUSSION

A novel spectral modeling method was presented based on combining frequency-warped linear prediction and weighted linear prediction. The new method, denoted as frequency-warped time-weighted linear prediction (WWLP), generates auto-regressive envelope models that utilize auditory frequency-warping and avoids the excessive modeling of harmonic peaks in the speech spectrum. These properties are ideally suited particularly for speech synthesis-oriented applications that rely on high-quality source-filter separation, such as glottal vocoding.

WWLP was used as a vocal tract modeling technique in the GlottDNN vocoder [11] and the system was evaluated in analysis/synthesis and TTS experiments using one male voice ("Nick") and one female voice ("Nancy"). The results show that WWLP-based spectral modeling gives better synthesis quality compared to more conventional vocoding approaches which do not take advantage of the combination of frequency-warping and time-weighting. The improved quality, however, was not achieved consistently for all the cases studied. In particular, the proposed new method did not improve the TTS quality of the "Nancy" voice when compared to a recently proposed baseline vocoder. We argue that this result, which differs from that obtained in the analysis/synthesis evaluation, might have been caused by using post-filtering settings that were not properly tested before being used in the WWLP-based vocoder in the present study. We believe that by modifying the post-filter to better fit the WWLP-based vocoder, the proposed fusion approach is expected to improve synthesis of female speech also in TTS.

Future plans with the WWLP-based GlottDNN vocoder include a full release version with the features presented in [11]. In addition, we will study the performance of the new vocoder in adaptation to varying speaking styles.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the Academy of Finland (project no. 256961, 284671).

7. REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*, Prentice-Hall, 1978.

[2] M.R. Schroeder and B.S. Atal, "Code-excited linear prediction (celp): high-quality speech at very low bit rates," in *Proc. ICASSP*, 1985, vol. 10, pp. 937–940.

[3] 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.

[4] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.

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

[6] 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.

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

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

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

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

[11] 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*, 2016.

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

[13] 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.

[14] 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.

[15] J. Pohjalainen, H. Kallasjoki, K.J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 2009.

[16] A. Härmä, "Linear predictive coding with modified filter structures," *IEEE Trans. Speech and Audio Proc.*, vol. 9, no. 8, pp. 769–777, November 2001.

[17] 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.

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

[19] D. Gowda, M. Airaksinen, and P. Alku, "Quasi closed phase analysis of speech signals using time varying weighted linear prediction for accurate formant tracking," in *Proc. ICASSP*. IEEE, 2016.

[20] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[21] M. Airaksinen, T. Bäckström, and P. Alku, "Glottal inverse filtering based on quadratic programming," in *Proc. Interspeech*, 2015.

[22] 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.

[23] 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*, 2016.

[24] F. Charpentier and M. Stella, "Diphone synthesis using an overlap-add technique for speech waveforms concatenation," in *Proc. ICASSP*, 1986.

[25] M. Cooke, C. Mayo, and C. Valentini-Botinhao, "Hurricane natural speech corpus," 2013, LISTA Consortium.

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

[27] S. Kraft and U. Zlzer, "BeaqleJS: HTML5 and JavaScript basedFramework for the Subjective Evaluation of Audio Quality," in *Linux Audio Conference*, 2014.

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

[29] Z. Wu, C. Valentini-Botinhao, O. Watts, and S. King, "Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis," in *Proc. ICASSP*. IEEE, 2015, pp. 4460–4464.

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