INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

On the Suitability of the Riesz Spectro-Temporal Envelope for WaveNet Based Speech Synthesis

Jitendra Kumar Dhiman[1], *Nagaraj Adiga*[2], *Chandra Sekhar Seelamantula*[1]

[1]Department of Electrical Engineering, Indian Institute of Science, Bangalore-12, India
[2]Department of Computer Science, University of Crete, Greece

jkdiith@gmail.com, nagaraj@csd.uoc.gr, chandra.sekhar@ieee.org

Abstract

We address the problem of estimating the time-varying spectral envelope of a speech signal using a spectro-temporal demodulation technique. Unlike the conventional spectrogram, we consider a pitch-adaptive spectrogram and model a spectro-temporal patch using an amplitude- and frequency-modulated two-dimensional (2-D) cosine signal. We employ a demodulation technique based on the Riesz transform that we proposed recently to estimate the amplitude and frequency modulations. The amplitude modulation (AM) corresponds to the vocal-tract filter magnitude response (or envelope) and the frequency modulation (FM) corresponds to the excitation. We consider the AM and demonstrate its effectiveness by incorporating it as an acoustic feature for local conditioning in the statistical WaveNet vocoder for the task of speech synthesis. The quality of the synthesized speech obtained with the Riesz envelope is compared with that obtained using the envelope estimated by the WORLD vocoder. Objective measures and subjective listening tests on the CMU-Arctic database show that the quality of synthesis is superior to that obtained using the WORLD envelope. This study thus establishes the Riesz envelope as an efficient alternative to the WORLD envelope.

Index Terms: pitch-adaptive spectrogram, spectrogram demodulation, Riesz transform, WaveNet, speech synthesis.

1. Introduction

A high quality of machine-generated speech directly benefits many speech applications such as text-to-speech synthesis (TTS) system [1, 2], speech translation, car navigation system, smartphone audio interface, screen readers, etc.. An important building block of a TTS system is *a vocoder*, which uses an analysis-by-synthesis approach to synthesize a speech signal. In the literature, there exist two types of vocoders: (1) source-filter-theory-based [3], and (2) a neural vocoder. In the first category, the most widely used vocoders are STRAIGHT [4] and WORLD [5]. The second category consists of neural network based architectures such as sample RNN [6], WaveNet [7], and WaveRNN [8]. Vocoders based on the source-filter model require prior knowledge of the speech production mechanism. Based on this knowledge, the design of these vocoders includes estimation of the instantaneous fundamental frequency (or pitch) of the speaker, the spectral envelope and the aperiodicity [9] parameter, which models the time-frequency noise essential for the synthesized speech to sound close to the natural one [10]. The pitch and aperiodicity parameters are used to model the source excitation signal and the envelope represents the vocal-tract filter's magnitude response.

On the other hand, the neural vocoder architecture uses a wider-receptive field to predict the current samples from the previous sample in a nonlinear auto-regressive fashion. The recently proposed deep learning based WaveNet architecture has shown to be promising as a statistical vocoder [11]. A WaveNet vocoder learns the mapping between acoustic features and the samples of a speech waveform in a supervised setting by solving a nonlinear auto-regression problem. In the literature, many possible combinations of acoustic features for the WaveNet vocoder have been utilized. They represent characteristics of the vocal tract and the sound source excitation. Tamamori *et al.* [11] used STRAIGHT-based vocal-tract filter and source-excitation features in the WaveNet vocoder. Adiga *et al.* [12] computed mel-filterbank coefficients from the short-time-Fourier transform (STFT) magnitude spectrum and used these acoustic features as local conditioning for the WaveNet vocoder. They showed that a better quality of synthesized speech was obtained using acoustic features derived directly from the STFT magnitude spectrum. This shows that the quality of the speech synthesized using WaveNet depends on the choice of the acoustic features.

In this paper, we focus on using acoustic features derived from the spectral envelope obtained by employing a spectrogram demodulation technique [13], which we proposed recently to estimate the 2-D amplitude and frequency modulations of a spectrogram patch. Our technique utilizes the complex Riesz transform [14] for accurate estimation of 2-D amplitude and frequency modulations. In our previous work, we have shown the importance of the frequency modulation (FM) for some of the fundamental speech processing tasks such as pitch estimation [15] and periodic/aperiodic speech decomposition of a speech signal [16]. While the FM component characterizes the source excitation attributes in 2-D, the amplitude modulation (AM) can effectively model the vocal-tract-filter magnitude response.

In order to know the suitability of the proposed Riesz spectro-temporal envelope in speech synthesis application, we train the WaveNet vocoder for the task of speech synthesis. We use the proposed envelope as an auxiliary feature for local conditioning in the WaveNet vocoder. For comparison, we train two WaveNet vocoders, one with acoustic features extracted from the proposed envelope and another with acoustic features from the envelope obtained using the state-of-the-art WORLD vocoder [5]. The impact of the proposed acoustic feature is analyzed by quantifying the quality of synthesized speech signals from the two WaveNet vocoders.

STRAIGHT and WORLD vocoders compute vocoder parameters in a pitch-adaptive fashion, which ensures reduced interference among the harmonics in the speech spectrum [17]. Following the same philosophy, we construct a pitch-adaptive spectrogram and model the smaller spectro-temporal patches using 2-D AM-FM cosine carriers. First, we employ Riesz transform-based demodulation technique and obtain an estimate of the AM. Second, we introduce a formant bandwidth correc-

http://dx.doi.org/10.21437/Interspeech.2019-2626

tion mechanism to enhance the suitability of the estimated spectral envelope for speech synthesis.

2. A Pitch-Adaptive Spectrogram

Short-time analysis of a speech signal is a widely used approach to capture its time-varying characteristics where short speech segments are multiplied by a window and then analyzed. Particularly, the STFT is an important tool and has revealed several time-varying speech properties such as formants, pitch, and 2-D time-frequency modulations [18]. The STFT gives a speech spectrogram, which is available in two flavors: *narrowband* and *wideband*. A wider window (20 ms to 25 ms) yields a narrowband spectrogram, whereas the wideband spectrogram is obtained by taking shorter speech segments (4 ms to 6 ms). In both the cases, the analysis window length is set to a fixed duration. However, the analysis window length can be also be varied. Unlike a spectrogram that is computed using a window with a fixed length, a pitch-adaptive spectrogram is obtained by varying the length of analysis window in proportion to the fundamental period of the speaker. The variable-length analysis window slides over the duration of a speech signal with a uniform frame update interval.

A pitch-adaptive STFT of the i^{th} speech frame is written as follows:

$$\hat{s}_w(i,\omega) = \int_{-\infty}^{\infty} s(t)w(t-i\tau)e^{-\mathrm{j}\omega t}\mathrm{d}t, \tag{1}$$

where τ (in seconds) denotes a constant frame-shift and $w(t)$ denotes the analysis window having its time support in the interval $t \in [-\mu T_0, \mu T_0]$ with T_0 denoting the instantaneous fundamental period of the speaker at the instant $i\tau$, and $\mu \in \mathbb{R}^+$. We refer to the square of the absolute of $\hat{s}_w(i,\omega)$ as the *pitch-adaptive spectrogram*. The time support of the analysis window and frame update interval are the design parameters; in this study we empirically choose $\mu = 3$ and $\tau = 1$ ms. For unvoiced speech segments, the window duration is set to 6 ms so that the rapid temporal fluctuations of such sounds can be captured.

3. Demodulation of a Pitch-Adaptive Spectrogram

Following [13], a 2-D AM-FM model for a windowed spectrogram patch $S_W(\boldsymbol{\omega}) : \mathbb{R}^2 \rightarrow \mathbb{R}^+$ is written as follows:

$$S_W(\boldsymbol{\omega}) = V(\boldsymbol{\omega})(\alpha_0 + \cos\Phi(\boldsymbol{\omega})), \tag{2}$$

where $\boldsymbol{\omega} = (t,\omega) \in \mathbb{R}^2$, t and ω denote the time and frequency variables, respectively, and $\alpha_0 \in \mathbb{R}$ is a constant that takes into account the non-negativity of the spectrogram patch. The amplitude modulation and the phase component are denoted by $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$, respectively. The 2-D phase $\Phi(\boldsymbol{\omega})$ of a frequency modulated cosine is modeled as follows:

$$\Phi(\boldsymbol{\omega}) = \Omega(\boldsymbol{\omega})(t\cos\theta(\boldsymbol{\omega}) + \omega\sin\theta(\boldsymbol{\omega})), \tag{3}$$

where $\Omega(\boldsymbol{\omega}) = \Omega_0 + \Delta\Omega(\boldsymbol{\omega})$ denotes the spatial frequency with FM given by $\Delta\Omega(\boldsymbol{\omega})$, and $\theta(\boldsymbol{\omega})$ denotes the local orientation of the 2-D cosine.

We use an existing pitch extraction algorithm (*Harvest* [20]) to obtain the instantaneous pitch and compute the pitch-adaptive spectrogram using a Hamming window. The pitch-adaptive spectrogram is divided into smaller time-frequency (t-f) patches of dimensions 600 Hz × 100 ms. The patch dimension along the frequency axis covers at least 3 harmonics for both male and female speakers so that the patch model given in (3) remains valid.

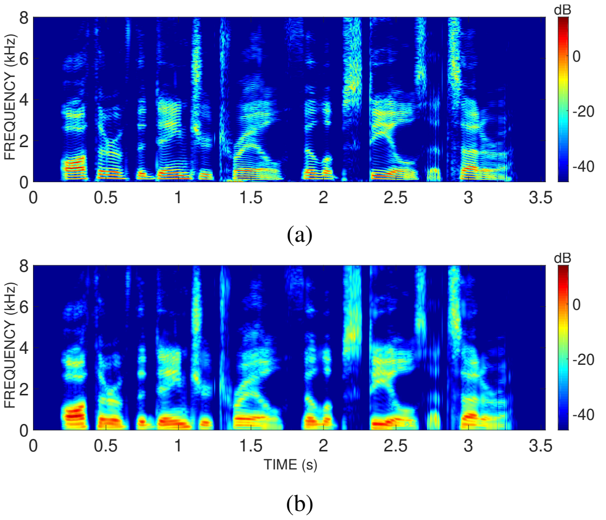

Figure 1: *(Color online) (a) A pitch-adaptive spectrogram; and (b) the time-frequency envelope corresponding to a continuous speech utterance, "Author of the danger trail, Philip Steels, etc.," spoken by a male speaker, taken from the CMU-Arctic database [19].*

3.1. Spectrogram Patch Demodulation

We focus on estimating the slowly varying time-frequency (t-f) envelope $V(\boldsymbol{\omega})$ for a given patch, which is achieved by employing a highly accurate 2-D demodulation technique that we introduced in [13]. The idea behind 2-D demodulation is to compute the quadrature component of a 2-D AM-FM cosine signal. In 1-D, this is done by using the Hilbert transform, and, in 2-D, the quadrature of a 2-D sinusoid can be obtained by using the complex Riesz transform [14]. A spectrogram patch is first subjected to a 2-D bandpass filter (Sec. 3.1.1), which retains only the dominant AM-FM cosine component corresponding to the instantaneous fundamental frequency of the speaker. The Riesz transform followed by orientation compensation [21] on the output of the bandpass filter gives the quadrature of a 2-D AM-FM cosine signal: $V(\boldsymbol{\omega})\sin\theta(\boldsymbol{\omega})$. By construction, the quadrature signal and the output of the 2-D bandpass filter are combined in complex number format, which yields the equivalent 2-D analytic signal representation [22]:

$$\begin{aligned} S_{W,a}(\boldsymbol{\omega}) &= V(\boldsymbol{\omega})\cos\theta(\boldsymbol{\omega}) + \mathrm{j}V(\boldsymbol{\omega})\sin\theta(\boldsymbol{\omega}), \\ &= V(\boldsymbol{\omega})e^{\mathrm{j}\theta(\boldsymbol{\omega})}. \end{aligned} \tag{4}$$

The estimates of the envelope and the 2-D cosine phase are obtained by applying absolute and angle operator on $S_{W,a}(\boldsymbol{\omega})$, respectively. We pool all the 2-D AM envelopes corresponding to all the patches of the spectrogram and perform a 2-D overlap-add in the least-squares-sense (OLA-LSE) [23] on the AM components, which gives the full t-f envelope. Figure 1 displays a pitch adaptive spectrogram along with a full t-f envelope. From the figure, one can observe that the t-f envelope effectively captures the temporal variations of the formants. However, the bandwidth of the envelope around formant frequencies

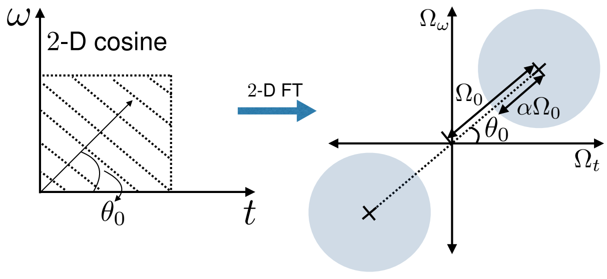

Figure 2: *A schematic showing the placement and choice of the bandwidth of the* 2*-D bandpass filter for Riesz-transform-based demodulation. The peaks in the Fourier transfom domain corresponding to a* 2*-D sinusoid are indicated by a* ×*.*

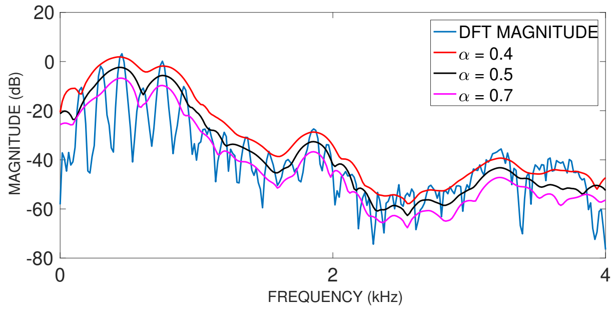

Figure 3: *(Color online) Illustration of different forms of envelopes obtained by varying the bandwidth of the* 2*-D bandpass filter. The envelopes are manually separated by introducing a bias only to aid visualization.*

depends on the circular bandwidth of the 2-D bandpass filter used for the demodulation.

3.1.1. The 2*-D Bandpass Filter for Demodulation*

The bandwidth of the 2-D bandpass filter affects the degree of smoothness and the formant bandwidths in the t-f envelope. We use a 10^{th}-order 2-D circular Butterworth filter, which is designed in the 2-D Fourier transform domain. The center frequency of the filter is located at $\mathbf{\Omega}_0$, which is the dominant peak in the 2-D Fourier transform domain of the 2-D sinusoid. A schematic illustrating the placement of the 2-D bandpass filter for a 2-D sinusoid is shown in Fig. 2. The schematic also illustrates the choice of the filter bandwidth, which is controlled by a scalar $\alpha \in (0, \frac{1}{\sqrt{2}})$. In order to show a few examples of the demodulated envelope, we consider the cases with $\alpha = 0.3, 0.5,$ and 0.7. An all voiced speech sentence, "Where were you while we were away?" spoken by a male speaker is demodulated for varying α values and the result for a voiced speech frame is shown in Fig. 3. From the figure, one can observe that a smaller value of α yields an envelope with wider formant bandwidths compared to the case of a higher value of α where the formant bandwidths are narrower. A correctly estimated formant bandwidth is desirable for high-quality speech synthesis [24]. Hence, we propose a method for formant bandwidth correction of the demodulated envelope.

3.2. Formant Bandwidth Correction Using Weighted Central Difference

Let t_i denote the time index of the i^{th} speech frame and $V(t_i, \omega)$ denote the demodulated envelope. Prior to bandwidth correction, small fluctuations in $V(t_i, \omega)$ are removed by applying a moving-average filter. The smoothed envelope is given by

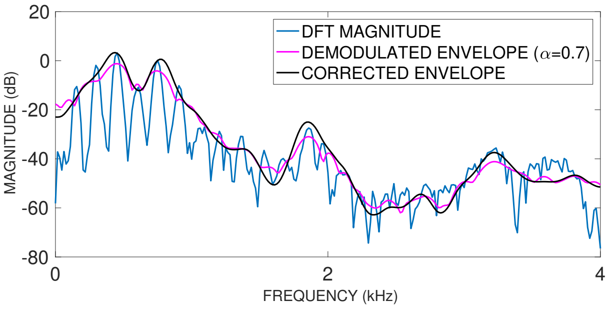

Figure 4: *(Color online) Illustration of the demodulated envelope and the envelope after smoothing plus bandwidth correction for a voiced speech segment.*

$$V_s(t_i, \omega) = \frac{1}{\omega_0(t_i)} \int_{-\frac{\omega_0(t_i)}{2}}^{\frac{\omega_0(t_i)}{2}} V(t_i, \omega + \lambda)\mathrm{d}\lambda \qquad (5)$$

where $\omega_0(t_i)$ denotes the instantaneous fundamental frequency. Let $L(t_i, \omega) = \ln V_s(t_i, \omega)$; the weighted central difference of $L(t_i, \omega)$ is given by

$$X(t_i, \omega) = \sum_{k=-1}^{k=1} w_k L(t_i, \omega + k\omega_0(t_i)), \qquad (6)$$

where $\{w_k\}$ represent the weights and satisfy the following relation

$$w_{-1} + w_0 + w_1 = 1. \qquad (7)$$

Without loss of generality, we assume that $w_{-1} = w_1$, and consequently (7) reduces to $w_0 + 2w_1 = 1$. Taking the inverse Fourier transform on both sides of (6) gives

$$\hat{X}(t_i, \eta) = \Big(w_0 + 2w_1 \cos(\omega_0(t_i)\eta)\Big) \hat{L}(t_i, \eta), \qquad (8)$$

where the first term in the product on the right-hand side is the bandwidth correction term and η is dual of ω, which denotes the quefrency variable in the cepstral domain. The corrected envelope is obtained as follows:

$$V_c(t_i, \omega) = e^{\mathcal{F}_\eta\{\hat{X}(t_i,\eta)\}}, \qquad (9)$$

where $\mathcal{F}_\eta$ denotes the forward Fourier transform operator with respect to the variable η. The choice of w_1 determines the degree of bandwidth correction. A negative value of w_1 in (6) ensures formant bandwidth reduction. In this paper, we empirically choose $w_1 = -0.55$ and the value of w_0 is obtained by using (7). Fig. 4 shows the demodulated envelope with $\alpha = 0.7$ and the corrected envelope along with the Fourier magnitude spectrum for a voiced frame of a speech signal. We test the effectiveness of the proposed envelope by incorporating it as an acoustic feature in the WaveNet vocoder.

4. The WaveNet Vocoder System

WaveNet is a statistical vocoder, which learns a nonlinear autoregressive (AR) model to predict the next signal sample from the previous samples and time-varying acoustic features. In this work, we have used the spectral envelope computed from the

Riesz transform as an acoustic feature for local conditioning in the WaveNet vocoder. The statistical vocoder architecture has two main modules: a stack of residual blocks, which acts as a feature extractor, and a post-processing module, which combines the information from the residual blocks to predict the next speech sample. In the r^{th} residual block, the key operation is a gated convolution where a hidden state vector $z^{(r)}$ is computed, and then added to its input $x^{(r-1)}$ to generate the final output $x^{(r)}$:

$$z^{(r)} = \tanh(W_f^{(r)} * x^{(r-1)} + C_f^{(r)}) \odot \sigma(W_g^{(r)} * x^{(r-1)} + C_g^{(r)}), \tag{10}$$

where $C_f^{(r)}$ and $C_g^{(r)}$ are the outputs of block r of the local conditioning network when it is fed with h. The symbol $*$ denotes convolution and the symbol $\odot$ denotes element-wise multiplication. The local condition h is an upsampled version of the acoustic features to get the new time-series with the same resolution as the input raw waveform.

The WaveNet vocoder is trained for 4 different voices taken from the CMU-ARCTIC database [19]. We have used two male speakers (BDL and RMS), and two female speakers (SLT and CLB) for the evaluation. The original sampling frequency of ARCTIC database is 32 kHz, which we downsampled to 16 kHz in order to simplify computation. The database for each of the speakers mentioned above consists of 1132 sentences, out of which 1000 sentences were used for training, 50 for validation, and the remaining for testing. We compute 45-dimensional mel-filterbank features from the Riesz envelope and use them as acoustic features. In order to compare the proposed spectral envelope, we also train another WaveNet vocoder using mel-filterbank features computed from the spectral envelope obtained from the state-of-the-art WORLD vocoder. The local conditioning in WaveNet vocoder is implemented similar to the architecture mentioned in [12]. In both the vocoders, the acoustic features were extracted for a frame rate of 1 ms. While feeding into the WaveNet vocoder, these features were upsampled to the sampling rate of the raw audio sample.

5. Results and Discussion

We evaluated the effectiveness of the proposed envelope for the task of speech synthesis by using two objective measures: STOI (short-time objective intelligibility) [25] and PESQ (perceptual evaluation of speech quality) [26], computed between the given pairs of the test speech signals and the corresponding synthesized speech signals. The STOI lies between 0 and 1 and is a measure for speech intelligibility. The PESQ lies between -0.5 and 4.5 and indicates a perceptual correlate of the quality of the speech signal. A higher value of these scores indicates a better quality of the synthesized speech signal. Table 1 shows the average PESQ scores and STOI scores over the waveforms taken from the CMU-Arctic database for each of the speakers. From the table, one can observe that the objective scores are consistently higher for the proposed method compared with WORLD, which indicates the suitability of the Riesz envelope as an effective acoustic feature for the WaveNet vocoder.

We have also conducted an informal listening test to know the quality of the synthesized speech using mean opinion score (MOS). The subjects rated the sound quality of the speech using a 5-point scale: 5 – excellent, 4 – good, 3 – fair, 2 – poor, and 1 – bad. In the test, 12 subjects in the age group 21 to 32 years participated and the test was conducted in a quiet room with HD 650 Sennheiser headphone. Five test speech utterances were used from each of the speakers: BDL, CLB, RMS, and SLT. From each of the speech files, the Riesz and WORLD envelopes were estimated and 40 speech samples were synthesized using the two vocoders: WaveNet-Riesz and WaveNet-WORLD. A total of 60 speech samples (40 synthesized and 20 original) were evaluated by presenting them to the subjects in a random order.

Table 1: *Objective scores with standard error for the synthesized speech using the envelopes from the proposed and WORLD approaches.*

Method	BDL	CLB	RMS	SLT
	(a) PESQ: Speech quality test			
Proposed	$\mathbf{2.86 \pm 0.25}$	$\mathbf{2.51 \pm 0.52}$	$\mathbf{2.40 \pm 0.41}$	$\mathbf{2.41 \pm 0.24}$
WORLD	2.69±0.41	1.95±0.64	2.22±0.65	2.15±0.23
	(b) STOI: Speech intelligibility test			
Proposed	$\mathbf{0.93 \pm 0.02}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.92 \pm 0.02}$	$\mathbf{0.93 \pm 0.02}$
WORLD	0.93±0.01	0.90±0.03	0.90±0.03	0.92±0.02

Table 2: *MOS with standard error for synthesized speech using the envelopes from the proposed and WORLD approaches.*

Method	BDL	CLB	RMS	SLT
Original	4.67 ± 0.04	4.69 ± 0.04	4.72 ± 0.04	4.76 ± 0.04
Proposed	$\mathbf{3.39 \pm 0.06}$	$\mathbf{4.44 \pm 0.05}$	$\mathbf{3.85 \pm 0.07}$	$\mathbf{4.37 \pm 0.05}$
WORLD	3.07 ± 0.07	3.46 ± 0.07	3.14 ± 0.07	3.47 ± 0.06

The MOS scores indicate the overall quality perceived by the subjects while listening to the synthesized speech samples and the original speech signals. Table 2 shows the average MOS scores obtained for the speech files synthesized using the proposed (Riesz) envelope and the WORLD envelope. The table shows the average MOS scores along with standard error for original speech samples, which were included with the synthesized speech samples during the test. From the table, one can observe that the MOS scores are consistently higher for the proposed Riesz envelope than the WORLD envelope. The superior performance of the proposed features in terms of both objective and subjective measures of the speech quality shows its suitability for incorporation in the WaveNet vocoder for a speech synthesis task . Some synthesized speech samples are available for listening at `https://www.csd.uoc.gr/~nagaraj/riesz.html`.

6. Conclusions

We proposed a novel technique to estimate the spectral envelope of a speech signal by demodulating a pitch-adaptive spectrogram. The demodulation yielded amplitude and frequency modulations. The estimated amplitude modulation was used to model the spectral envelope corresponding to the vocal-tract-filter magnitude response. We also proposed a novel technique for formant bandwidth correction, which utilized the weighted central difference operator. The acoustic features from the proposed envelope were used as local conditioning in a WaveNet vocoder and compared with another WaveNet vocoder trained in a similar fashion on the envelope features obtained from the state-of-the-art WORLD vocoder. The proposed envelope features showed a better performance in terms of the quality of synthesized speech.

7. References

[1] Y. Wang *et al.*, "Tacotron: A fully end-to-end text-to-speech synthesis model," *arXiv preprint arXiv:1703.10135*, 2017.

[2] T. Dutoit, *An Introduction to Text-to-Speech Synthesis*. Springer Science & Business Media, 1997, vol. 3.

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

[4] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time frequency smoothing and an instantaneous-frequency-based F0 extraction," *Speech Communication*, vol. 27(3-4), pp. 187–207, 1999.

[5] 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. 99, no. 7, pp. 1877–1884, 2016.

[6] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[7] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

[8] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," *arXiv preprint arXiv:1802.08435*, 2018.

[9] 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. Second International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, 2001.

[10] M. Morise, "D4C, a band-aperiodicity estimator for high-quality speech synthesis," *Speech Communication*, vol. 84, pp. 57–65, 2016.

[11] A. Tamamori, T. Hayashi, K. Kobayashi, K. Takeda, and T. Toda, "Speaker-dependent WaveNet vocoder." in *Proc. INTERSPEECH*, 2017, pp. 1118–1122.

[12] N. Adiga, V. Tsiaras, and Y. Stylianou, "On the use of WaveNet as a statistical vocoder," in *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5674–5678.

[13] H. Aragonda and C. S. Seelamantula, "Demodulation of narrowband speech spectrograms using the Riesz transform," *IEEE/ACM Transactions on Audio, Speech, and Language Process.*, vol. 23, no. 11, pp. 1824–1834, Nov. 2015.

[14] C. S. Seelamantula, N. Pavillon, C. Depeursinge, and M. Unser, "Local demodulation of holograms using the Riesz transform with application to microscopy," *Journal of the Optical Society of America*, vol. 29, no. 10, pp. 2118–2129, Oct. 2012.

[15] J. K. Dhiman, N. Adiga, and C. S. Seelamantula, "A spectro-temporal demodulation technique for pitch estimation." in *Proc. INTERSPEECH*, 2017, pp. 2306–2310.

[16] K. Vijayan, J. K. Dhiman, and C. S. Seelamantula, "Time-frequency coherence for periodic-aperiodic decomposition of speech signals," in *Proc. INTERSPEECH*, 2017, pp. 329–333.

[17] H. Kawahara and M. Morise, "Technical foundations of TANDEM-STRAIGHT, a speech analysis, modification and synthesis framework," *Sādhana*, vol. 36, no. 5, pp. 713–727, 2011.

[18] T. T. Wang and T. F. Quatieri, "Two-dimensional speech-signal modeling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 6, pp. 1843–1856, 2012.

[19] J. Kominek and A. W. Black, "The CMU-ARCTIC speech databases," in *Proc. 5th ISCA Speech Synthesis Workshop*, 2004, pp. 223–224.

[20] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. INTERSPEECH*, 2017, pp. 2321–2325.

[21] H. Knutsson, C.-F. Westin, and M. Andersson, "Representing local structure using tensors II," in *Proc. Scandinavian Conference on Image Analysis*, Springer, 2011, pp. 545–556.

[22] M. Felsberg and G. Sommer, "The monogenic signal," *IEEE Transactions on Signal Processing*, vol. 49, no. 12, pp. 3136–3144, 2001.

[23] T. T. Wang and T. F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," in *Proc. IEEE Workshop on Applications of Signal Process to Audio and Acoustics*, Oct. 2009, pp. 65–68.

[24] H. Kawahara, M. Morise, T. Toda, R. Nisimura, and T. Irino, "Beyond bandlimited sampling of speech spectral envelope imposed by the harmonic structure of voiced sounds." in *Proc. INTERSPEECH*, 2013, pp. 34–38.

[25] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[26] ITU-T Recommendation, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.