A SPECTRO-TEMPORAL TECHNIQUE FOR ESTIMATING APERIODICITY AND VOICED/UNVOICED DECISION BOUNDARIES OF SPEECH SIGNALS

Jitendra Kumar Dhiman and Chandra Sekhar Seelamantula

Department of Electrical Engineering, Indian Institute of Science, Bangalore 560012, India
Emails: jkdiith@gmail.com, chandra.sekhar@ieee.org

ABSTRACT

In contrast to a 1-D short-time analysis of speech, 2-D approaches aim at characterizing the speech signal attributes jointly in time and frequency. In this paper, we focus on the quasi-periodicity of a voiced spectro-temporal patch and quantify it by proposing an *aperiodicity measure* defined using the underlying frequency modulations in the patch. We further propose a time-frequency *aperiodicity map* obtained by overlapping and adding the aperiodicity measures across patches. The proposed aperiodicity map is utilized to obtain band-wise aperiodicity parameters, which are essential for high-quality speech synthesis. The aperiodicity in unvoiced patches is addressed by identifying them using the *coherence* of the patch. In addition, the proposed technique also provides voiced/unvoiced decisions boundaries of a speech signal. The effectiveness of the proposed band-wise aperiodicity parameters and voiced/unvoiced decisions is verified by incorporating them in an existing state-of-the-art vocoder for speech synthesis. Subjective listening tests show that the quality of the reconstructed speech is on par with that of the state-of-the-art WORLD vocoder in terms of mean opinion score, indicating that spectrotemporal approaches are highly promising for speech analysis and synthesis applications.

Index Terms— aperiodicity in 2-D, band-wise aperiodicy parameters, carrier spectrogram, coherence map.

1. INTRODUCTION

Joint spectrotemporal analysis (or 2-D analysis) of speech offers a means to analyze the time-frequency (t-f) variations in the signal. Findings in auditory neuroscience [1] suggest that certain neurons in the primary auditory cortex are tuned to specialized spectrotemporal patterns of the speech spectrogram [2]. Motivated by these studies, several models have been proposed for 2-D analysis of speech signals. The key idea is to represent a spectrotemporal patch with a 2-D signal model. Ezzat et al. [3,4] proposed a 2-D amplitude/frequency modulation (or AM-FM) model for a patch using a 2-D Gabor filterbank analysis approach. They showed that the 2-D AM encodes the phonetic attributes and the 2-D carrier encodes the speaker attributes. A seminal contribution for 2-D speech modeling was made by Wang and Quatieri [5], who modeled a spectrogram patch using a 2-D Fourier series with amplitude-modulated (AM) 2-D cosine carrier signals. The demodulation of AM and 2-D carriers was done by using the 2-D Fourier transform magnitude spectrum, which is also referred to as the *grating compression transform* (GCT) [6, 7]. The stationary 2-D cosine carrier signal used in [5] was generalized to the nonstationary case by Aragonda and Seelamantula [8] who employed a demodulation technique based on the complex Riesz transform [8–10]. It was shown that the AM component corresponds to the vocal tract response and underlying the 2-D carrier component characterizes the excitation signal attributes such as the time-varying pitch, the quasi-harmonic (voiced sounds) and inharmonic (unvoiced sounds) spectro-temporal patterns. The 2-D AM-FM model [8] was shown to have a superior modeling accuracy over the 2-D Fourier series model [5]. In a recent work, we employed the model for pitch estimation [11] and periodic-aperiodic decomposition of the speech signal [12] using the 2-D carrier component. Arguably, 2-D analysis of speech signals can reveal several interesting properties. In this paper, by default, the term spectrogram refers to a narrowband spectrogram.

The voiced excitation is quasi-periodic and contributes to pitch harmonics in the speech spectrogram. Thus, the underlying carrier in the spectrogram exhibits harmonically related partials. However, the voiced excitation is only quasi-periodic. The effects of jitter and shimmer cause aperiodicity in the voiced excitation. The jitter represents perturbations in periodicity among laryngeal cycles, while shimmer reflects variations among epochal amplitudes across laryngeal cycles [13]. Together, they cause aperiodicity in the excitation signal and disrupt the harmonic structure of the pitch partials. On the other hand, unvoiced sounds don't exhibit any harmonic structure due to the absence of pitch and are noise-like with a high degree of aperiodicity.

In this paper, we focus on estimating aperiodicity of the voiced patches of the spectrogram by modeling them using frequency modulated 2-D sinusoids. We rely on structural differences due to frequency modulations of voiced sounds and propose an *aperiodicity measure* to quantify the degree of aperiodicity in a voiced spectrotemporal patch. On the other hand, unvoiced spectrotemporal patches are highly aperiodic and we address this issue by identifying them using the notion of coherence, which we proposed in [12]. For unvoiced patches, the aperidocity measure is set to a fixed value (Sec 3.2) indicating a high degree of aperiodcity. In addition, the technique also provides the voiced/unvoiced decision boundaries of the speech signal. The importance of computed parameters and voiced/unvoiced decisions is validated by incorporating them in an existing state-of-the-art WORLD vocoder for a copy synthesis experiment (i.e., without any statistical modeling of the parameters).

This paper is organized as follows. In Section 2, we define aperiodicity for a 2-D sinsusoid and propose the measure to quantify it. Following this, in Section 3, we compute the aperiodicity of voiced spectrotemporal patches for natural speech spectrogram and explain the method to separate voiced/unvoiced spectrotemporal patches us-

ing coherence. In Section 4, we present the experimental condition and performance evaluation of proposed parameters for speech synthesis. We conclude in Section 5, highlighting the key contributions and future directions.

2. APERIODICITY IN 2-D

Let $\boldsymbol{\omega} = (t, \omega) \in \mathbb{R}^2$, where t and ω denote the continuous time and frequency variables, respectively. Consider the frequency modulated 2-D cosine $C(\boldsymbol{\omega}) : \mathbb{R}^2 \rightarrow \mathbb{R}$ given by

$$C(\boldsymbol{\omega}) = \cos\big(\Omega(\boldsymbol{\omega})(t\cos\beta_0 + \omega\sin\beta_0)\big), \tag{1}$$

where β_0 and $\Omega(\boldsymbol{\omega}) = \Omega_0 + k_f\Delta\Omega(\boldsymbol{\omega})$ represent the local orientation and spatial frequency of the cosine, respectively. The strength of the frequency modulations is given by the modulation index k_f. Let $\boldsymbol{\Omega} = (\Omega_t, \Omega_\omega) \in \mathbb{R}^2$ denote the dual frequency variable of $\boldsymbol{\omega}$. A schematic of a 2-D cosine with no frequency modulation ($k_f = 0$) and its Fourier transform is shown in Fig. 1.

We consider a 2-D sinusoid to be aperiodic if $k_f\Delta\Omega(\boldsymbol{\omega}) \neq 0$ and $\|\nabla_{\boldsymbol{\omega}}\Delta\Omega(\boldsymbol{\omega})\| \ll \Omega_0$ and use frequency modulations of a 2-D sinusoid to quantify the degree of aperiodicity.

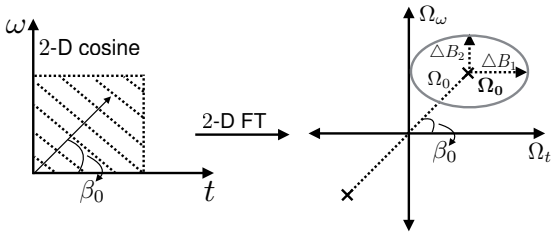

Fig. 1. Schematic of a 2-D sinusoid (left) and its Fourier transform (right). The spectrotemporal frequency and orientation of sinusoid are denoted by Ω_0 and β_0, respectively. A mask around the peak located at $\boldsymbol{\Omega}_0 = (\Omega_0\cos\beta_0, \Omega_0\sin\beta_0)$ is illustrated by an ellipse.

2.1. Effect of Frequency Modulation on 2-D Spectrum

Let $\Phi_0(\boldsymbol{\omega}) = (t\cos\beta_0 + \omega\sin\beta_0)$ and consider a frequency modulated 2-D cosine:

$$\begin{aligned} C(\boldsymbol{\omega}) &= \cos\big((\Omega_0 + k_f\Delta\Omega(\boldsymbol{\omega}))\Phi_0(\boldsymbol{\omega})\big) \\ &= \cos\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big)\cos\big(k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\big) \\ &\quad - \sin\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big)\sin\big(k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\big). \end{aligned} \tag{2}$$

Under the condition that $\|k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\| \ll 1$, that is, narrowband frequency modulation, (2) may be approximated as follows:

$$C(\boldsymbol{\omega}) \approx \cos\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big) - k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\sin\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big). \tag{3}$$

The first term in (3) is a primary 2-D sinusoid oscillating at frequency Ω_0 and the second term is due to the frequency modulation. The frequency modulation in (3) has the effect of broadening the Fourier magnitude spectrum of the primary sinusoid around Ω_0. For illustration, consider the case where the modulating function is a slowly-varying sinusoid, that is, $\Delta\Omega(\boldsymbol{\omega}) = \cos\big(\Omega_m\Phi_0(\boldsymbol{\omega})\big)$ with $\Omega_m \ll \Omega_0$. Figure. 2 shows the GCTs of a 2-D cosine with a constant frequency versus a frequency-modulated 2-D cosine. One can observe from Fig. 2(d) that the appearance of significant side-band energies implies the spectrum spread around Ω_0 due to the FM. We use the spectral spread property to quantify the degree of aperiodicity.

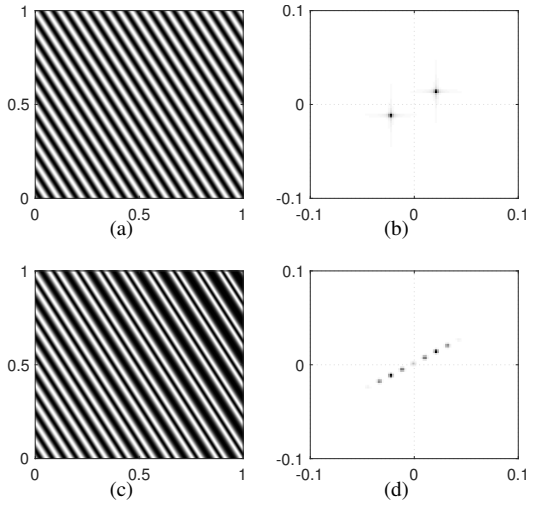

Fig. 2. (a) A 2-D cosine with constant frequency Ω_0, (b) its GCT, (c) a cosine with frequency modulation $k_f\Delta\Omega(\boldsymbol{\omega}) = 0.015\cos(10\pi\Phi_0(\boldsymbol{\omega}))$, and (d) its GCT.

2.2. Aperiodicity Measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$

Consider a windowed 2-D cosine $C_W(\boldsymbol{\omega}) = W(\boldsymbol{\omega})C(\boldsymbol{\omega})$, where $W(\boldsymbol{\omega})$ represents a real 2-D window function. Then, from (3), we have

$$\begin{aligned} C_W(\boldsymbol{\omega}) \approx & W(\boldsymbol{\omega})\cos\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big) \\ & - \underbrace{k_f\Delta\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})W(\boldsymbol{\omega})}_{F(\boldsymbol{\omega})}\sin\big(\Omega_0\Phi_0(\boldsymbol{\omega})\big). \end{aligned} \tag{4}$$

Denoting $\boldsymbol{\Omega}_0 = (\Omega_0\cos\beta_0, \Omega_0\sin\beta_0) \in \mathbb{R}^2$ and taking 2-D Fourier transform on both sides in (4), we have

$$\begin{aligned} \hat{C}_W(\boldsymbol{\Omega}) \approx & \hat{W}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) + \hat{W}(\boldsymbol{\Omega} + \boldsymbol{\Omega}_0) \\ & + \mathrm{j}(\hat{F}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_0) - \hat{F}(\boldsymbol{\Omega} + \boldsymbol{\Omega}_0)). \end{aligned} \tag{5}$$

The normalized power spectral density is given by $P_C(\boldsymbol{\Omega}) = \frac{|\hat{C}_W(\boldsymbol{\Omega})|}{\|\hat{C}_W(\boldsymbol{\Omega})\|}$. The aperiodicity measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$ of a 2-D sinusoid is defined in terms of $P_C(\boldsymbol{\Omega})$ as

$$\mathcal{A}_{\boldsymbol{\Omega}_0} = 1 - 2\int_{\boldsymbol{\Omega}} P_C(\boldsymbol{\Omega})M_{\boldsymbol{\Omega}_0}(\boldsymbol{\Omega}; \Delta\boldsymbol{B})\mathrm{d}\boldsymbol{\Omega}, \tag{6}$$

where

$$\boldsymbol{M}_{\boldsymbol{\Omega}_0}(\boldsymbol{\Omega}; \Delta\boldsymbol{B}) = \begin{cases} 1, & |\Omega_t - \Omega_{0_1}| < \Delta B_1, |\Omega_\omega - \Omega_{0_2}| < \Delta B_2, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

represents a mask centered around $\boldsymbol{\Omega}_0$ with $\Delta\boldsymbol{B} = (\Delta B_1, \Delta B_2) \in \mathbb{R}^2_+$ and $(\Omega_{0_1}, \Omega_{0_2}) = (\Omega_0\cos\beta_0, \Omega_0\sin\beta_0)$ as displayed in the schematic shown in Fig. 1. ΔB_1 and ΔB_2 denote the main-lobe widths of the window function $W(\boldsymbol{\omega})$ along Ω_t-axis and Ω_ω-axis, respectively. The aperiodicity measure $\mathcal{A}_{\boldsymbol{\Omega}_0}$ takes on continuous values between 0 and 1 with the two extreme values 0 and 1 representing a 2-D sinusoid being perfectly periodic and aperiodic, respectively. A value of $\mathcal{A}_{\boldsymbol{\Omega}_0}$ between 0 and 1 quantifies the degree of aperiodicity (or quasi-periodicity).

3. APERIODICITY OF NARROWBAND SPECTROGRAM PATCHES

Following [8], consider the 2-D AM-FM model for a windowed spectrogram patch $S_W(\boldsymbol{\omega})$:

$$S_W(\boldsymbol{\omega}) = V(\boldsymbol{\omega})\big(\alpha_0 + \cos(\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega}))\big), \tag{8}$$

with $\Phi_0(\boldsymbol{\omega}) = (t\cos\beta_0 + \omega\sin\beta_0)$, $V(\boldsymbol{\omega})$ is a slowly varying t-f envelope, $\alpha_0 \in \mathbb{R}_+$ is a constant that ensures non-negativity of the spectrogram patch, and $\Omega(\boldsymbol{\omega})$ is the FM.

The aperiodicity information due to frequency modulations is present only in the carrier component of (8). Hence, the interference due to AM component $V(\boldsymbol{\omega})$ must be removed prior to aperiodicity estimation. The AM and carrier components are effectively separated by the employing spectrogram demodulation technique proposed in [8]. The AM $V(\boldsymbol{\omega})$ and carrier $\cos\big(\Omega(\boldsymbol{\omega})\Phi_0(\boldsymbol{\omega})\big)$ components obtained by demodulation are shown in Fig. 3(b) and Fig. 3(c), respectively. We focus on the carrier component (or the carrier spectrogram), which captures the dynamics of underlying the 2-D carrier in the t-f plane and is free from AM interference.

3.1. Time-frequency Aperiodicity Map from the Carrier Spectrogram

The carrier spectrogram (cf. Fig. 3(c)) is segmented into smaller overlapping patches such that the 2-D AM-FM model in (8) is not violated. In order to get reliable estimates of aperiodicity, the carrier patches are multiplied by a 2-D Nuttal window function, which has a high side-lobe attenuation (about -100 dB) [14]. The carrier t-f patches are modeled using frequency modulated windowed 2-D cosine (2) and the aperiodicity measure is computed for every carrier patch using (6). Assuming the variations of aperiodicity are negligible within a patch, the aperiodicity value is repeated over the corresponding patch dimension. All such patches are combined using 2-D overlap-add in the least-squares sense [7]. The resulting t-f map is referred to as the *aperiodicity map* and is displayed in Fig 3(d). As shown in Fig. 3(d), the aperiodicity measure is high for unvoiced spectro-temporal regions and relatively low for voiced spectro-temporal regions. This behavior of the aperiodicity map is consistent in accordance with (6). In the next section, we show that the proposed aperiodicity map is useful for obtaining band-wise aperiodicity parameters [15].

3.2. The Band-wise Aperiodicity Parameters

Band-wise aperiodicity parameters are essential for high quality speech synthesis using vocoders [15, 16], which are based on the source-filter theory of speech production. In these vocoders, typically, the Nyquist frequency band is divided into smaller sub-bands and the aperiodicity is defined for each sub-band. For voiced speech sounds, the band-wise aperiodicity value is the same as the value in the proposed aperiodicity map at the center of the corresponding sub-band. In order to get a t-f map of band-wise aperiodicity parameters, the parameters are linearly interpolated on the log scale by appending an aperiodicity value of -60 dB at zero frequency [15] and then converted back to the linear scale. The resulting t-f map is shown in Fig. 4 where the aperiodicity values are set to 1 in the unvoiced t-f regions identified using the coherence map as explained in the subsequent section.

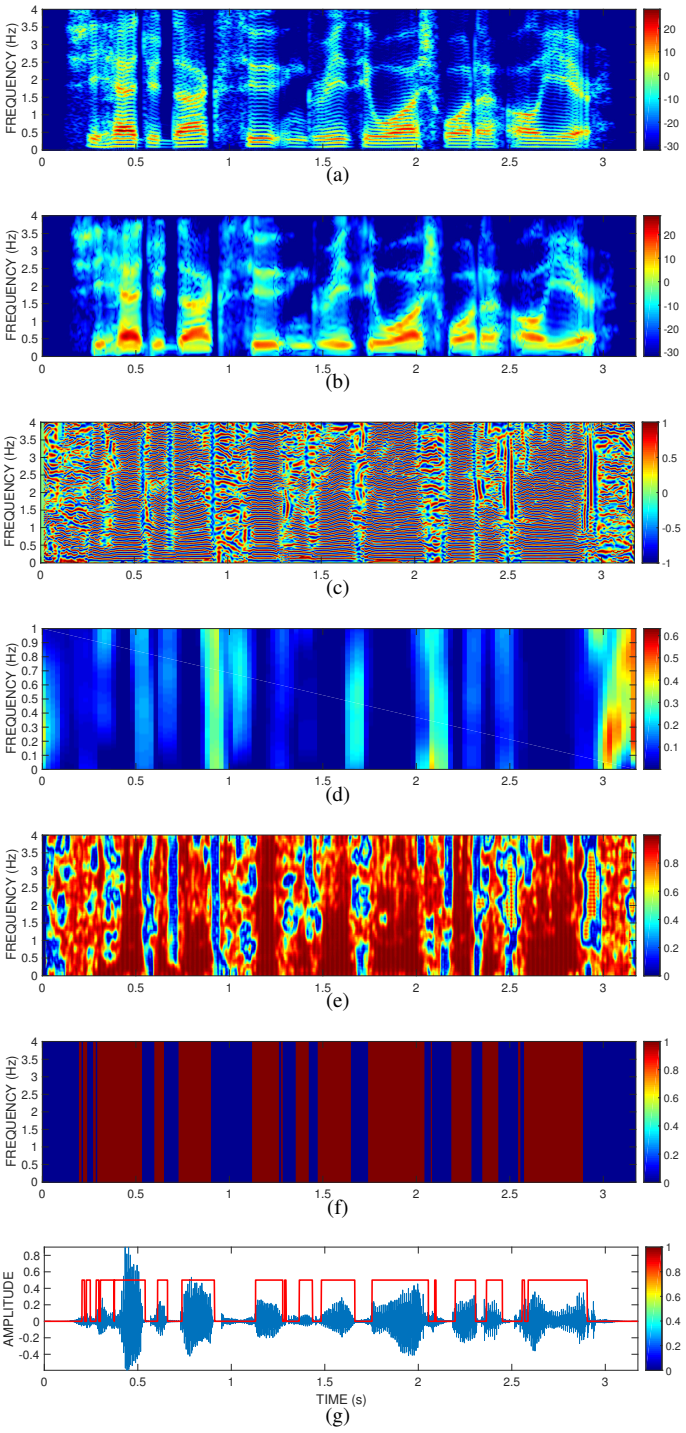

Fig. 3. (a) Spectrogram, (b) amplitude modulation (AM), (c) carrier spectrogram, (d) t-f aperiodicity map, (e) coherence map, (f) binary mask derived from coherence map, and (g) voiced/unvoiced (VUV) speech segments.

3.3. Voiced/Unvoiced Separation Using Coherence Map

Coherence map [12] is a time-frequency map that captures the relative discrepancy between coherent and incoherent spectro-temporal

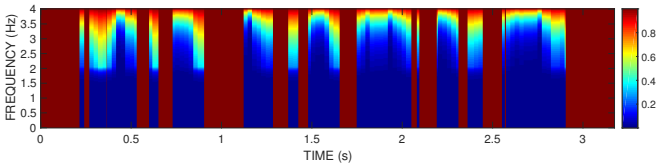

Fig. 4. Time-frequency map of the proposed band-wise aperiodicity parameters for a male speaker.

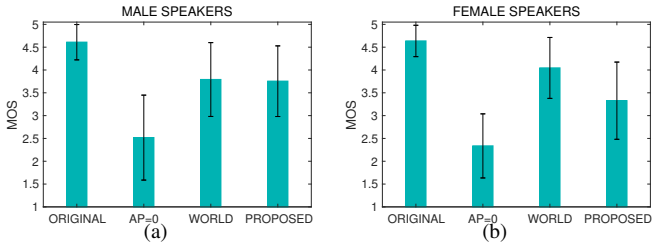

Fig. 5. Mean opinion scores (MOS) (a) for male speakers, and (b) for female speakers. The vertical lines indicate 0.5 standard deviation on either side of the mean.

patches of the carrier. Fig. 3(e) displays coherence map computed from the carrier spectrogram. It exhibits high values for coherent carrier patches and relatively low values for incoherent patches. By means of the coherence map the two goals of separating voiced/unvoiced spectrotemporal patches and segmenting speech signal into voiced/unvoiced parts can be achieved by computing a binary mask as follows. An adaptive thresholding of a subband from the coherence map in the band from 0 to f_1 with $f_1 = 1$ kHz, is attempted. Across the short-time speech frames, the mean value of coherence map $C(t, \omega)$ within the specified subband is computed as follows:

$$\overline{C}(t) = \frac{1}{2\pi f_1} \int_0^{2\pi f_1} C(t, \omega) \mathrm{d}\omega. \tag{9}$$

The threshold value is computed as

$$C_{th} = \frac{1}{T} \int_0^T \overline{C}(t) \mathrm{d}t, \tag{10}$$

where T denotes the signal duration. It is worth noting that the selection of the threshold across the short-time speech frames is inherently adaptive, thus avoiding any manual selection of the threshold. Let t_i denote the frame update interval with i as the frame index. A speech frame is classified as voiced if $\overline{C}(t_i) \geq C_{th}$, otherwise it is classified as an unvoiced frame. Notice that the value of coherence is also high in the silence regions (see the interval $[0, 0.4]$ seconds in Fig. 3(e)). This is due to the fact that the carrier spectrogram exhibits a consistent structure different from the pitch harmonics in the silence regions. To overcome the disparity due to silence regions, voicing decisions from the coherence map are conditioned with a crude short-time energy estimator decisions and a binary mask is created to distinguish between voiced and unvoiced frames in the speech signal which is shown in Fig. 3(f). The binary mask has values either 1 or 0 for voiced and unvoiced regions, respectively and distinguishes between voiced/unvoiced t-f regions of the spectrogram. The band-wise aperiodicity t-f map by the application of binary mask is shown in Fig 4. The speech signal and corresponding voiced and unvoiced segments by the application of the binary mask are shown in Fig. 3(g).

4. IMPLEMENTATION DETAILS AND VALIDATION

The evaluation is done for 30 speech wavefiles (15 male and 15 female) with sampling frequency 16 kHz taken from standard CMU-ARCTIC database [17]. The narrowband spectrogram is computed at a frame rate of 1 ms with 40 ms Hann window and segmented into patches of size 900 Hz $\times$ 100 ms. The patches are subjected to 2-D demodulation and the carrier spectrogram is obtained by employing 2-D overlap-add in the least-squares sense on carrier patches. Prior to estimating the t-f aperiodicity map, the carrier patches are multiplied by a 2-D Nuttal window, which has a fairly high side-band attenuation (about -100 dB) [14] required for accurate estimation of aperiodicity. Therefore, the values ΔB_1 and ΔB_2 are the main-lobe (cf. Fig. 1) widths of the 1-D Nuttal window along time and frequency axes, respectively. The band-wise aperiodicity parameters are obtained from the aperiodicity map (Sec 3.2) by dividing the Nyquist frequency band (8 kHz) into 3 overlapping sub-bands of widths 2 kHz with an overlap of 1 kHz. The obtained parameters are validated using the WORLD vocoder [18] for speech synthesis. Given an input signal, the WORLD vocoder takes the analysis-by-synthesis approach. It estimates the vocal tract envelope, fundamental frequency, aperiodicity parameters (AP) and voiced/unvoiced (VUV) decisions using which it synthesizes speech. Focusing on AP and VUV, we analyze the following three cases in a copy-synthesis experiment: (1) discard AP i.e. AP = 0, (2) use AP and VUV estimated by WORLD, and (3) set AP and VUV equal to that given by the proposed approach. In each case, speech signals were synthesized and a mean opinion score (MOS) listening test was conducted. 25 listeners participated in the test to rate the overall quality of the synthesized speech compared to the original speech on scale of 1-bad, 2-poor, 3-good, 4-fairly good, and 5-excellent. Fig. 5(a) and Fig. 5(b) show the MOS values for male and female speakers, respectively. The low values of MOS for both male and female speakers for Case 1 (i.e. AP=0) indicate that AP plays an important role for high quality speech synthesis. The figure shows that the MOS values are comparable for male speakers with state-of-the-art vocoder. However, for female speakers, the MOS score is lower than the state-of-the-art by about 0.7. Some synthesized speech signals are available at [19] for listening.

5. CONCLUSIONS

We proposed an aperiodicity measure in 2-D using frequency modulation of a 2-D sinusoid. In addition, we also proposed a method for identifying voiced/unvoiced spectrotemporal patches and segmenting the speech signal into its voiced/unvoiced parts using the coherence of spectrotemporal patches. The proposed aperiodicity map was used to obtain band-wise aperiodicity parameters, which were found useful for vocoder-based speech analysis and synthesis. An evaluation using WORLD vocoder highlighted the effectiveness of new aperiodicity parameters and voiced/unvoiced decisions.

6. REFERENCES

[1] S. A. Shamma, "Speech processing in the auditory system ii: Lateral inhibition and the central processing of speech evoked

activity in the auditory nerve," *Journal of the Acoustical Society of America*, vol. 78, no. 5, pp. 1622–1632, 1985.

[2] S. A. Shamma, J. W. Fleshman, P. R. Wiser, and H. Versnel, "Organization of response areas in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 69, no. 2, pp. 367–383, 1993.

[3] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," in *Proceedings of the Eighth Annual Conference of the International Speech Communication Association*, 2007.

[4] ——, "AM-FM demodulation of spectrograms using localized 2D Max-Gabor analysis," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Process. - ICASSP*, vol. 4, 2007, pp. IV–1061–IV–1064.

[5] 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.

[6] T. F. Quatieri, "2-D processing of speech with application to pitch estimation." in *Proc. Interspeech*, 2002.

[7] 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.

[8] 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.

[9] 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.

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

[11] J. K. Dhiman, N. Adiga, and C. S. Seelamantula, "A spectro-temporal demodulation technique for pitch estimation," in *Proceedings of INTERSPEECH*, 2017.

[12] K. Vijayan, J. K. Dhiman, and C. S. Seelamantula, "Time-frequency coherence for periodic-aperiodic decomposition of speech signals," in *Proceedings of INTERSPEECH*, 2017.

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

[14] A. Nuttall, "Some windows with very good sidelobe behavior," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 1, pp. 84–91, 1981.

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

[16] H. Kawahara, "STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds," *Acoustical Science and Technology*, vol. 27, no. 6, pp. 349–353, 2006.

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

[18] 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.

[19] "Demonstration of speech synthesis using aperiodicity parameters and voiced/unvoiced decisions, [Online]," https://jitendradhiman.github.io/APTEST.html, 2018.