INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Time-Frequency Coherence for Periodic-Aperiodic Decomposition of Speech Signals

Karthika Vijayan[1,2]*, Jitendra Kumar Dhiman*[2]*, and Chandra Sekhar Seelamantula*[2]

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2]Department of Electrical Engineering, Indian Institute of Science, Bangalore, India

vijayan.karthika@nus.edu.sg, jkdiith@gmail.com, chandra.sekhar@ieee.org

Abstract

Decomposing speech signals into periodic and aperiodic components is an important task, finding applications in speech synthesis, coding, denoising, etc. In this paper, we construct a time-frequency coherence function to analyze spectro-temporal signatures of speech signals for distinguishing between deterministic and stochastic components of speech. The narrowband speech spectrogram is segmented into patches, which are represented as 2-D cosine carriers modulated in amplitude and frequency. Separation of carrier and amplitude/frequency modulations is achieved by 2-D demodulation using Riesz transform, which is the 2-D extension of Hilbert transform. The demodulated AM component reflects contributions of the vocal tract to spectrogram. The frequency modulated carrier (FM-carrier) signal exhibits properties of the excitation. The time-frequency coherence is defined with respect to FM-carrier and a coherence map is constructed, in which highly coherent regions represent nearly periodic and deterministic components of speech, whereas the incoherent regions correspond to unstructured components. The coherence map shows a clear distinction between deterministic and stochastic components in speech characterized by jitter, shimmer, lip radiation, type of excitation, etc. Binary masks prepared from the time-frequency coherence function are used for periodic-aperiodic decomposition of speech. Experimental results are presented to validate the efficiency of the proposed method.

Index Terms: Spectro-temporal patterns, Riesz transform, Hilbert transform, 2-D demodulation, Carrier spectrogram, Coherence function.

1. Introduction

Speech signal is a composite signal containing information related to linguistic content, speaker identity, speaker characteristics such as gender, age and height, emotional and health state of the speaker, etc. The periodicity characteristics of speech signals play a vital role in conveying these information and providing naturalness to human speech. The requirement to characterize the periodicity properties arises in speech synthesis, speech coding, voice analysis and musical acoustics [1]. Analyzing spectro-temporal characteristics is extremely important in inferring valuable cues regarding periodicity of speech signals.

Typically, one-dimensional (1-D) processing is employed for studying speech in frequency and time domains using short-time Fourier transform (STFT) analysis and amplitude modulation-frequency modulation (AM-FM) analysis, respectively [2, 3]. In STFT analysis, one relies on the short-time stationarity of the speech signal and segments it into short time-frames to study local spectral characteristics [2]. On the other hand, AM-FM analysis represents speech as a collection of resonances modulated in their amplitudes and frequencies, and decomposes it into AM and FM parts to study temporal characteristics [3,4]. These methodologies study either time or frequency domain characteristics of speech. However, studying the time and frequency domain properties together will yield deeper insights into the spectro-temporal characteristics of speech.

Analyzing spectro-temporal receptive fields (STRF) using speech spectrograms is important for understanding auditory cortical processing [5] and the STRFs were successfully used for denoising, detection, recognition and perceptual studies of speech signals [6–8]. The conventional AM-FM analysis was extended to multiband demodulation analysis (MDA) of speech signals to study spectro-temporal patterns for formant tracking [9]. Spectro-temporal patterns were studied by considering STFT coefficients from adjacent frames of speech as a multi-dimensional time series, and modeling using a generalized autoregressive conditional heteroskedasticity (GARCH) process [10]. The group-delay function was employed to enhance spectral resolution to finely capture spectro-temporal patterns in speech signals [11]. Feature extraction was performed by merging 1-D spectral and temporal features to a joint spectro-temporal representation for speech, speaker and emotion recognitions [12–14]. All these methods study the spectro-temporal patterns in speech by using 1-D processing.

The two-dimensional (2-D) analysis of speech signals has been carried out on narrowband (NB) spectrograms to study the time-frequency (t-f) characteristics. A 2-D Gabor filter analysis of small patches in NB spectrograms was employed to analyze harmonicity, formant trajectories and nonstationarity of speech signals [15]. Feature extraction was performed from Gabor filtered NB spectrogram patches for speech recognition and enhancement [16–18]. The idea of grating compression transform (GCT), which is the 2-D Fourier transform applied to small patches of NB spectrogram, was employed in 2-D sinusoidal modeling of speech. The GCT reveals the dominant frequency component in spectrogram patches, which is chosen as the fundamental frequency of harmonically related 2-D sinusoidal components for modeling speech signals [19].

In this paper, we present an extensive analysis of speech signals in 2-D t-f domain with the specific intent of studying deterministic and stochastic parts. Small patches in NB spectrograms of speech are represented as 2-D cosine carriers with AM and FM, which are demodulated using the Riesz transform [20]. The accuracy of Riesz demodulation in estimating the AM and FM components is more than that of Gabor filter analysis and 2-D sinusoidal modeling [20]. In order to study the types of

This work is supported by the Department of Electronics and Information Technology (DeitY) project on "Development of text-to-speech synthesis systems for Indian languages – Phase II." The work of the first author is also supported by NUS Start-up grant with WBS no: R-263-000-C35-133/731.

http://dx.doi.org/10.21437/Interspeech.2017-726

speech sounds, which are mostly characterized by the nature of excitation, the demodulated 2-D FM component is thoroughly analyzed. The FM-carrier structure is observed to be coherent in regions where the speech is nearly deterministic and noncoherent where speech is stochastic. Based on these observations, a t-f coherence map is constructed from FM-carrier spectrogram and is utilized for periodic-aperiodic decomposition (PAPD).

The rest of the paper is organized as follows. In Section 2, we review the 2-D demodulation of spectrogram patches using Riesz transform. The study of FM-carrier to generate the t-f coherence map is explained in Section 3. The application of coherence map to PAPD of speech signals is presented in Section 4. In Section 5, we summarize the key contributions of this paper in formulating a coherence measure for speech analysis.

2. Riesz demodulation of NB spectrogram

Based on the source-filter model of speech production, the speech signal can be represented as a convolution of the impulse response of vocal tract system (VTS) and the excitation source signal. Equivalently, the magnitude spectrum of speech signals comprises pitch harmonics with frequency modulations due to excitation source, superposed with the VTS response modulating the spectral envelope. Hence, the t-f patterns in speech spectrograms are generally the long-term pitch harmonics with formant trajectories superimposed on them. Invoking the short-time stationarity of the speech signal, a small patch of NB spectrogram can be approximated as a 2-D AM-FM model [20] as

$$\begin{aligned} S_W(\boldsymbol{\omega}) &\simeq V(\boldsymbol{\omega})[\alpha_0 + \alpha_1 \cos\Phi(\boldsymbol{\omega})], \\ &= \alpha_0 V(\boldsymbol{\omega}) + \alpha_1 V(\boldsymbol{\omega}) \cos\Phi(\boldsymbol{\omega}), \end{aligned} \tag{1}$$

where the 2-D vector $\boldsymbol{\omega}$ is constituted by indices in time and frequency as $\boldsymbol{\omega} = (t, \omega)$. The functions $S_{W,l}(\boldsymbol{\omega}) = \alpha_0 V(\boldsymbol{\omega})$ and $S_{W,b}(\boldsymbol{\omega}) = \alpha_1 V(\boldsymbol{\omega}) \cos\Phi(\boldsymbol{\omega})$ are the lowpass and bandpass components of $S_W(\boldsymbol{\omega})$, respectively. The functions $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$ are the AM and FM components modulating the 2-D cosine carrier signal, where $V(\boldsymbol{\omega})$ is mostly contributed by the VTS and $\Phi(\boldsymbol{\omega})$ is mostly contributed by the excitation source [20]. We have to perform the 2-D demodulation of $S_W(\boldsymbol{\omega})$ in order to obtain $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$. Since both these components are present in the bandpass component $S_{W,b}(\boldsymbol{\omega})$, we design a bandpass filter to capture this component and demodulate it.

The GCT reveals the dominant frequency present in spectrogram patches [19], which represent the carrier frequency. Thus, GCT is employed to identify the center frequency of the bandpass filter and the bandpass component $S_{W,b}(\boldsymbol{\omega})$ is filtered out from the spectrogram patch. For 2-D AM-FM demodulation, we employ an extension of the 1-D Hilbert transform demodulation [21]. The Riesz transform is the 2-D isotropic extension of the Hilbert transform and is capable of providing the quadrature component of a 2-D sinusoid [22]. The complex Riesz transform response is given by [20]:

$$\hat{h}(\boldsymbol{\Omega}) = \hat{h}_t(\boldsymbol{\Omega}) + \mathrm{j}\hat{h}_\omega(\boldsymbol{\Omega}) = \frac{-\mathrm{j}\Omega_t + \Omega_\omega}{||\boldsymbol{\Omega}||} \tag{2}$$

where $\boldsymbol{\Omega} = (\Omega_t, \Omega_\omega)$ represents the frequencies along time and frequency axes of the spectrogram, and $\hat{h}_t(\boldsymbol{\Omega})$ and $\hat{h}_\omega(\boldsymbol{\Omega})$ are the impulse responses of filters associated with Riesz transform along time and frequency axes, respectively. The Riesz transform for the bandpass component $V(\boldsymbol{\omega}) \cos\Phi(\boldsymbol{\omega})$ is its quadrature component $V(\boldsymbol{\omega}) \sin\Phi(\boldsymbol{\omega})$, using which we construct a 2-D complex signal $Z(\boldsymbol{\omega}) = V(\boldsymbol{\omega}) e^{\mathrm{j}\Phi(\boldsymbol{\omega})}$. Now the 2-D AM and FM components are deduced as the magnitude and phase of the complex signal [21], given by

$$\begin{aligned} \hat{V}(\boldsymbol{\omega}) &= \sqrt{\mathscr{R}\{Z(\boldsymbol{\omega})\}^2 + \mathscr{I}\{Z(\boldsymbol{\omega})\}^2}, \text{and} \\ \hat{\Phi}(\boldsymbol{\omega}) &= \tan^{-1}\left\{\frac{\mathscr{I}\{Z(\boldsymbol{\omega})\}}{\mathscr{R}\{Z(\boldsymbol{\omega})\}}\right\} \end{aligned} \tag{3}$$

respectively, where $\mathscr{R}(\cdot)$ and $\mathscr{I}(\cdot)$ denote the real and imaginary parts, respectively, of the complex argument. Thus, a patch-wise 2-D demodulation of NB spectrogram of speech signals is realized.

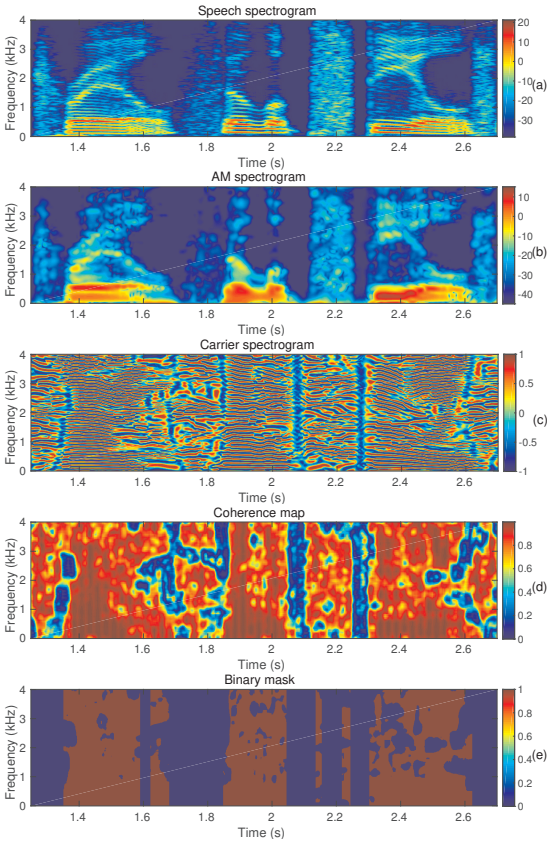

Figure 1: *2-D AM-FM demodulation using the Riesz transform. (a) NB spectrogram, (b) AM spectrogram, (c) Carrier spectrogram, (d) Coherence map, and (e) Binary mask for periodic-aperiodic decomposition.*

Figure 1 illustrates the effectiveness of the 2-D Riesz demodulation of NB spectrograms. The AM spectrogram and carrier spectrogram constructed from the AM components and FM-carriers ($\cos\hat{\Phi}(\boldsymbol{\omega})$) of overlapping patches of speech spectrogram, are shown in Figure 1(b) and Figure 1(c), respectively. The AM spectrogram demonstrates the presence of formant tracks indicating the VTS response, and the carrier spectrogram exhibits pitch harmonics due to the excitation source. Thus, the NB spectrogram of speech signal is demodulated and the contributions due to VTS and excitation are decoupled.

3. Time-frequency coherence

The carrier spectrogram is further investigated to study different components of speech sounds, which are mainly characterized by the nature of excitation. A 2-D cosine signal is repre-

sented by its amplitude, frequency, and orientation. Likewise, the FM-carrier is represented by its unit amplitude, FM component $\Phi(\boldsymbol{\omega})$ and local orientation $\beta(\boldsymbol{\omega})$. The direction in which FM carriers of spectrogram patches are oriented can be better represented by the structure tensor matrix, which indicates the dominant direction of gradient of carrier within a small neighborhood of points. The structure tensor matrix is given by [20] $J(\boldsymbol{\omega}) =$

$$\begin{bmatrix} \psi(\boldsymbol{\omega}) * S^2_{W,t}(\boldsymbol{\omega}) & \psi(\boldsymbol{\omega}) * S_{W,\omega}(\boldsymbol{\omega}) S_{W,t}(\boldsymbol{\omega}) \\ \psi(\boldsymbol{\omega}) * S_{W,\omega}(\boldsymbol{\omega}) S_{W,t}(\boldsymbol{\omega}) & \psi(\boldsymbol{\omega}) * S^2_{W,\omega}(\boldsymbol{\omega}) \end{bmatrix}. \tag{4}$$

The $S_{W,t}(\boldsymbol{\omega})$ and $S_{W,\omega}(\boldsymbol{\omega})$ are the Riesz transform components of $S_{W,b}(\boldsymbol{\omega})$ along the time and frequency axes, respectively. The localizing function $\psi(\boldsymbol{\omega})$ represents a small neighborhood of the t-f point under consideration [20]. The structure tensor $J(\boldsymbol{\omega})$ indicates the definitive structure of spectrogram. Based on the structure tensor, we define the coherence of a t-f patch of the carrier spectrogram as [22]

$$\chi(\boldsymbol{\omega}) = \left[\frac{\lambda_{\max}(\boldsymbol{\omega}) - \lambda_{\min}(\boldsymbol{\omega})}{\lambda_{\max}(\boldsymbol{\omega}) + \lambda_{\min}(\boldsymbol{\omega})} \right]^2, \tag{5}$$

where $\lambda_{\max}(\boldsymbol{\omega})$ and $\lambda_{\min}(\boldsymbol{\omega})$ are the maximum and minimum eigenvalues, respectively, of $J(\boldsymbol{\omega})$. The concept of coherence was utilized to define inter- and intra- spectral and temporal relations in the analysis of wide-sense stationary random processes [23–27]. Here, we present the 2-D t-f coherence using Riesz demodulation of speech spectrograms.

When the underlying FM-carrier has a consistent structure or a preferred orientation, the value of coherence is high and vice versa. The coherence $\chi(\boldsymbol{\omega})$ computed from small patches of the carrier spectrogram are overlap-added to form the coherence map, as shown in Figure 1(d). In this paper, we propose to utilize the t-f coherence map to distinguish between speech sounds primarily characterized by the excitation.

The voiced excitation is generated by quasi-periodic vibrations of vocal folds, and it holds considerable energy. The unvoiced excitation resembles random noise, generated by constricted air flow through open vocal folds and supraglottal acoustic noise sources [2]. The speech sounds produced by these two types of excitation can be distinguished by the coherence map. The unvoiced excitation (noise-like) fails to induce any definitive pattern in speech spectrogram. Hence, the resultant carrier spectrogram exhibits an inconsistent t-f pattern, as can be seen from Figure 1(c) between 1.7 s to 1.8 s and 2.1 s to 2.3 s. Consequently, the coherence map possess relatively low values. However, it is observed that certain high frequency regions of unvoiced speech illustrate nearly consistent carrier patterns, which are different from that of pitch harmonics.

The voiced excitation is quasi-periodic and contributes to pitch harmonics in the speech spectrogram. Thus, the carrier spectrogram exhibits harmonically related pitch partials, resulting in a consistent structure. In other words, the carrier structure corresponding to voiced speech is highly coherent. Consequently, the coherence map exhibits high values (see t-f region 1.4 s to 1.6 s and 1.9 s to 2.1 s in Figure 1(d)). However, the voiced excitation is only quasi-periodic. The effects of jitter and shimmer cause aperiodicity in voiced excitation. The jitter represents perturbations in periodicity among laryngeal cycles, while shimmer reflects variations among epochal amplitudes across laryngeal cycles [2]. Together, they cause aperiodicity in excitation signal and disrupts the harmonic structure of pitch partials. This disrupts the coherent structure in carrier spectrogram, which are effectively captured by the coherence map (see t-f region 2.4 s to 2.6 s and 1.5 kHz and 3.5 kHz in Figure 1(d)).

Thus, the stochastic components of speech are constituted by unvoiced sounds and aperiodic portions of voiced sounds. The deterministic components of speech are constituted by periodic portions of voiced sounds contributing to the pitch harmonics. The coherence map captures information distinguishing deterministic and stochastic components of speech signal.

4. Periodic-aperiodic decomposition

The problem of PAPD of speech signals is important to speech analysis, synthesis, speech coding, voice analysis and study of musical acoustics [1, 28]. As coherence map developed in this work exhibits information regarding aperiodicity in speech sounds, we construct a binary mask based on coherence for PAPD. Adaptive thresholding of the coherence map is performed for creating the binary mask.

The subband coherence map within 0 Hz and 500 Hz is considered to separate voiced and unvoiced sounds. The lower subband of coherence map is chosen for this study, as the carrier spectrogram portrays nearly persistent structures in certain higher bands of unvoiced speech leading to wrong voicing decisions. The mean value of subband coherence across all short-time frames is chosen as the threshold. All frames with their mean coherence value greater than this threshold are pronounced as voiced frames.

For characterizing aperiodicity within voiced speech, the entire frequency range of coherence map is considered. The histogram of coherence map is generated to determine the number of t-f points having different ranges of coherence values. The adaptive threshold for PAPD is chosen as the coherence value corresponding to the histogram bin, which has as many number of t-f points as a scale factor of that in the last histogram bin. All t-f points having a coherence value higher than the threshold are chosen as parts of the periodic component. The final binary mask for PAPD is prepared by combining the voicing decisions and periodicity decisions within voiced frames of speech. The binary mask has the value of 1 for the t-f regions corresponding to periodic component and a value of 0 for the regions corresponding to aperiodic component, as shown in Figure 1(e).

To illustrate the effectiveness of the binary mask prepared from coherence map in realizing PAPD, we perform synthetic experiments. The synthetic speech signal is generated using a formant synthesizer excited with a periodic glottal pulse train following Liljencrants-Fant (LF) model [29], together with pitch-synchronously added noise bursts [1, 30]. The noise burst added is a colored noise in the high frequency range, to simulate the effects of jitter, shimmer and lip radiation, which generally affect the high-frequency spectrum of speech signals [2]. The synthetic speech allows reasonable control over its parameters and also provides ground truth for evaluation, which is not the case with real speech. A synthetic speech signal and its periodic and aperiodic components are shown in Figure 2. For 2-D demodulation of synthetic speech, its NB spectrogram is computed and demodulated using the Riesz transform. The resultant carrier spectrogram and coherence map are shown in Figures 3(b) and 3(c), respectively. The coherence map is adaptively thresholded to construct a binary mask for PAPD, which is given in Figure 3(d).

The periodic and aperiodic components are generated by synthesizing the modified spectrograms, obtained by applying the binary mask and its complement on the original speech spectrogram. The estimated and synthetic periodic and aperiodic

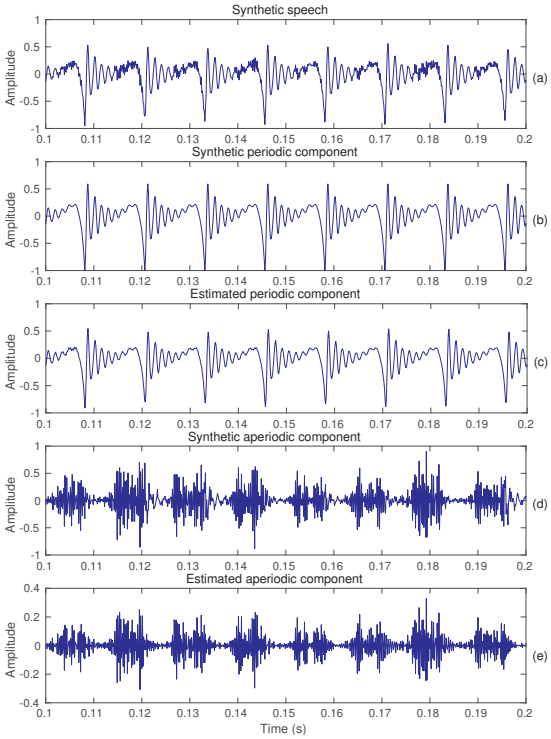

Figure 2: *PAPD of a synthetic signal. (a) Synthetic signal, (b) Synthetic periodic component, (c) Estimated periodic component, (d) Synthetic aperiodic component, and (e) Estimated aperiodic component.*

components closely match with each other (See Figure 2), illustrating the effectiveness of the coherence map in PAPD. The higher range of values in the synthetic speech spectrogram between 2.5 kHz and 3.5 kHz in Figure 3(a) indicate the presence of high frequency colored noise bursts, inducing disruptions in carrier structure as shown in Figure 3(b), which are captured by the coherence map shown in Figure 3(c). The resultant binary mask given in Figure 3(d) portrays the aperiodicity due to noise bursts, indicating the efficacy of Riesz analysis in PAPD.

For quantitative evaluation of the proposed PAPD methodology, we perform experiments by varying parameters of synthetic speech signals of 1 s duration. The fundamental frequency $F0$ and the ratio of duration of noise burst to that of the laryngeal cycle, termed as burst duration ratio (BDR), are varied. Also, different types of noises were added as noise bursts to generate synthetic speech. The performance of the proposed PAPD method is compared with the iterative algorithm reported in [30]. The performance evaluation is done in terms of harmonic-to-noise ratio (HNR), which is the power ratio of periodic to aperiodic components. The HNR values obtained from PAPD algorithms based on Riesz analysis and the iterative technique are reported in Table 1. The HNR values reported are averaged over different types of noise bursts – pink, blue, violet and white noises. For an ideal PAPD, the HNR values of synthetic and estimated components should be the same. It can be observed that the HNR values of the estimated components obtained by Riesz analysis are consistently closer to those of the synthetic components, for almost all evaluation conditions. The proposed PAPD strategy considerably outperforms the iterative algorithm [30], indicating the efficiency of coherence map

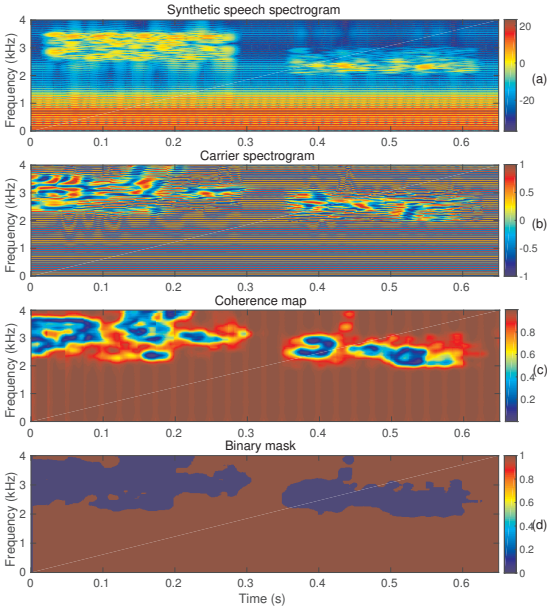

Figure 3: *PAPD of a synthetic signal. (a) Spectrogram, (b) Carrier spectrogram, (c) Coherence map, and (d) Binary mask.*

Table 1: *Evaluation of PAPD of synthetic speech signals.*

Synthetic speech parameters			Estimated HNR (dB)	
$F0$ (Hz)	BDR (%)	HNR (dB)	Riesz	Iter. algo. [30]
120	30	22.57	**21.68**	20.30
120	60	19.58	**18.87**	17.03
120	100	17.36	16.94	**17.21**
200	30	20.00	**19.46**	19.13
200	60	18.41	**17.56**	16.82
200	100	16.71	**15.99**	15.21

in distinguishing between deterministic and stochastic components of speech. Notice that the proposed algorithm efficiently delivers the PAPD of signals having aperiodicities contributed by different types of noise sources and not just the ones that are clearly separated in the t-f plane.

5. Conclusions

In this paper, speech signals were analyzed in 2-D time-frequency domain to segregate periodic and aperiodic components in speech. The NB spectrogram of speech signals was segmented into small patches and demodulated using Riesz transform. The demodulated FM-carrier illustrated properties of the excitation source and the coherence of the carrier structure was computed to build a time-frequency coherence map. Distinct separation between periodic and aperiodic components of speech was exhibited by the coherence map, which was explored to prepare binary masks for periodic-aperiodic decomposition of speech signals. The performance of the proposed strategy was evaluated using synthetic experiments, as ground truth is not available in case of natural speech. The experimental results verified the effectiveness of the coherence map in periodic-aperiodic decomposition, which is advantageous to several applications including, but not limited to, speech event detection, speech synthesis, denoising, and computational auditory scene analysis.

6. References

[1] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana, "Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan 1998.

[2] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1978.

[3] P. Maragos, J. Kaiser, and T. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing*, vol. 41, no. 10, pp. 3024–3051, Oct 1993.

[4] ——, "On amplitude and frequency demodulation using energy operators," *IEEE Transactions on Signal Processing*, vol. 41, no. 4, pp. 1532–1550, Apr 1993.

[5] S. Shamma, "On the role of space and time in auditory processing," *Trends in Cognitive Sciences*, vol. 5, no. 8, pp. 340–348, Nov 2016, doi: 10.1016/S1364-6613(00)01704-6.

[6] N. Mesgarani and S. Shamma, "Denoising in the domain of spectrotemporal modulations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, p. 042357, 2007.

[7] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, May 2006.

[8] M. Elhilali, T. Chi, and S. A. Shamma, "A spectro-temporal modulation index (STMI) for assessment of speech intelligibility," *Speech Communication*, vol. 41, no. 2–3, pp. 331 – 348, 2003.

[9] A. Potamianos and P. Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," *The Journal of the Acoustical Society of America*, vol. 99, no. 6, pp. 3795–3806, 1996.

[10] I. Cohen, "Modeling speech signals in the time–frequency domain using GARCH models," *Signal Processing*, vol. 84, no. 12, pp. 2453 – 2459, 2004.

[11] Y. Bayya and D. N. Gowda, "Spectro-temporal analysis of speech signals using zero-time windowing and group delay function," *Speech Communication*, vol. 55, no. 6, pp. 782 – 795, 2013.

[12] S. Thomas, S. Ganapathy, and H. Hermansky, "Spectro-temporal features for automatic speech recognition using linear prediction in spectral domain," in *2008 16th European Signal Processing Conference*, Aug 2008, pp. 1–4.

[13] S. Ganapathy, S. Thomas, and H. Hermansky, "Robust spectro-temporal features based on autoregressive models of hilbert envelopes," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4286–4289.

[14] T.-S. Chi, L.-Y. Yeh, and C.-C. Hsu, "Robust emotion recognition by spectro-temporal modulation statistic features," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 1, pp. 47–60, 2012.

[15] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-d gabor filters," in *Interspeech 2007*, Aug 2007, pp. 506–509.

[16] J. Bouvrie, T. Ezzat, and T. Poggio, "Localized spectro-temporal cepstral analysis of speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2008, pp. 4733–4736.

[17] S. Y. Zhao and N. Morgan, "Multi-stream spectro-temporal features for robust speech recognition," in *Interspeech 2008*, Sep 2008, pp. 898–901.

[18] S. Y. Chang, B. T. Meyer, and N. Morgan, "Spectro-temporal features for noise-robust speech recognition using power-law nonlinearity and power-bias subtraction," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7063–7067.

[19] 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, Aug 2012.

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

[21] L. Cohen, *Time-Frequency Analysis*. Prentice Hall PTR Englewood Cliffs, NJ:, 1995.

[22] C. S. Seelamantula, N. Pavillon, C. Depeursinge, and M. Unser, "Local demodulation of holograms using the Riesz transform with application to microscopy," *J. Opt. Soc. Am. A*, vol. 29, no. 10, pp. 2118–2129, Oct 2012.

[23] G. Carter, C. Knapp, and A. Nuttall, "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing," *IEEE Transactions on Audio and Electroacoustics*, vol. 21, no. 4, pp. 337–344, Aug 1973.

[24] J. Benesty, J. Chen, and Y. Huang, "A generalized mvdr spectrum," *IEEE Signal Processing Letters*, vol. 12, no. 12, pp. 827–830, Dec 2005.

[25] A. Jakobsson, S. R. Alty, and J. Benesty, "Estimating and time-updating the 2-d coherence spectrum," *IEEE Transactions on Signal Processing*, vol. 55, no. 5, pp. 2350–2354, May 2007.

[26] C. Zheng, M. Zhou, and X. Li, "On the relationship of non-parametric methods for coherence function estimation," *Signal Processing*, vol. 88, no. 11, pp. 2863 – 2867, 2008.

[27] M. J. Hinich, "A statistical theory of signal coherence," *IEEE Journal of Oceanic Engineering*, vol. 25, no. 2, pp. 256–261, April 2000.

[28] 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 *MAVEBA*, 2001, pp. 59–64.

[29] G. Fant, J. Liljencrants, and Q.-g. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, vol. 4, no. 1985, pp. 1–13, 1985.

[30] B. Yegnanarayana, C. d'Alessandro, and V. Darsinos, "An iterative algorithm for decomposition of speech signals into periodic and aperiodic components," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 1, pp. 1–11, Jan 1998.