Interspeech 2018
2-6 September 2018, Hyderabad

Multicomponent 2-D AM-FM Modeling of Speech Spectrograms

Jitendra Kumar Dhiman, Neeraj Sharma, and Chandra Sekhar Seelamantula

Department of Electrical Engineering, Indian Institute of Science, Bangalore - 560 012, India

jkdiith@gmail.com, neerajww@gmail.com, chandra.sekhar@ieee.org

Abstract

In contrast to 1-D short-time analysis of speech, 2-D modeling of spectrograms provides a characterization of speech attributes directly in the joint time-frequency plane. Building on existing 2-D models to analyze a spectrogram patch, we propose a multicomponent 2-D AM-FM representation for spectrogram decomposition. The components of the proposed representation comprise a DC, a fundamental frequency carrier and its harmonics, and a spectrotemporal envelope, all in 2-D. The number of harmonics required is patch-dependent. The estimation of the AM and FM is done using the Riesz transform, and the component weights are estimated using a least-squares approach. The proposed representation provides an improvement over existing state-of-the-art approaches, for both male and female speakers. This is quantified using reconstruction SNR and *perceptual evaluation of speech quality* (PESQ) metric. Further, we perform an overlap-add on the DC component, pooling all the patches and obtain a time-frequency (t-f) aperiodicity map for the speech signal. We verify its effectiveness in improving speech synthesis quality by using it in an existing state-of-the-art vocoder.

Index Terms: multicomponent 2-D AM-FM modeling, aperiodicity parameter, Riesz transform.

1. Introduction

Speech signals feature a temporally evolving spectral content, evident in spectrogram visualizations [1]. The widely used approaches in speech processing assume quasi-stationarity and extract spectral features on successive short-time segments, usually 10 or 25 ms long [2]. These features and their correlations over a temporal context, usually 1 s, are found to capture speaker and phoneme attributes and serve as front-end features in applications such as speech activity detection, speaker identification [3], and speech recognition [4]. However, over the past decade, analysis of the temporally evolving spectral content using joint time-frequency (t-f) analysis of the spectrogram has gained interest [5–8]. Mathematically, this provides a means to tackle the 1-D spectral nonstationarity in speech using 2-D stationary spectral analysis. Interestingly, auditory neuroscience findings suggest that the neurons in the auditory cortex are tuned to distinct spectrotemporal patterns in speech spectrograms [9–11]. Arguably, 2-D modeling of the spectrogram can provide a useful framework for analysis of spectrotemporal features associated with perceived speech attributes.

Consider the narrowband speech spectrogram shown in Figure 1. In this paper, by default, the term spectrogram refers to the narrowband flavor. From psychoacoustics [12], it is well established that the gliding spectrotemporal striations are associated with the time-varying pitch in natural speech [1]. The relative strengths of these striations, seen as color contrast, encode the uttered phonemes. The key idea in spectrotemporal analysis is to model these patterns with a 2-D signal model. This

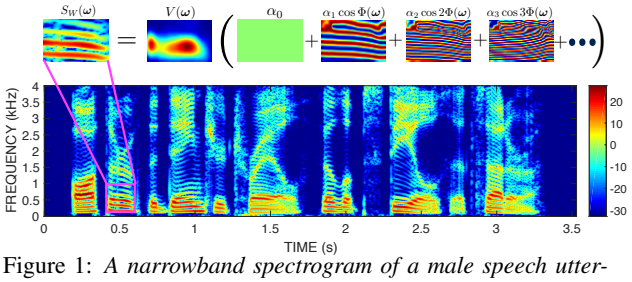

Figure 1: *A narrowband spectrogram of a male speech utterance. The highlighted patch is modeled using the proposed model. The computed parameters are* $\alpha_0 = 0.83, \alpha_1 = 1, \alpha_2 = 0.26$*, and* $\alpha_3 = 0.01$*. The estimation procedure will be described in Sec. 2.2.*

is pursued by analyzing the spectrogram in a patch-wise fashion. Wang and Quatieri [6] modeled each patch as an element-wise product of a 2-D spectral envelope and a 2-D sinusoidal-series carrier. Ezzat et al. [5] proposed a 2-D AM-FM model for each patch using a 2-D Gabor filter-bank analysis approach. It was shown that the 2-D AM encodes the phonetic attributes and the 2-D carrier encodes the speaker attributes. Building on these models, Aragonda and Seelamantula [8] proposed a 2-D AM-FM model, which generalized the 2-D stationary FM assumption in [6] to spatially varying FM. This requires the design of an accurate 2-D AM and FM estimation technique, and the complex Riesz transform [13] based demodulation approach was used for achieving this goal. The generalized model in [8] gives a 2 to 4 dB benefit in reconstruction SNR in comparison with [6], and has also been used for pitch estimation [14] and periodic/aperiodic separation of speech [15].

In this paper, we further analyze the 2-D AM-FM model in [8] and make two contributions. First, we generalize the model to contain multiple 2-D AM-FM sinusoids. An illustration of the proposed modeling is shown in Figure 1. The motivation lies in the observation that a 2-D patch containing periodic striations will be modeled more accurately by using a weighted sum of harmonically related 2-D sinusoidal carriers, analogous to 1-D Fourier series modeling of periodic signals. Also, spectrogram patches that do not predominantly exhibit any spectrotemporal structure can be modeled more accurately by including more components in the model—this generalization improves the modeling accuracy. Second, we analyze the patch-wise DC components (α_0 in Figure 1) obtained from the proposed model. Existing findings [5, 6, 8] have shown the role of the 2-D envelope and the 2-D carrier in characterizing the vocal tract and the vocal-fold vibrations, respectively. Our findings suggest that along with the AM and carrier components, the DC component is also informative in analyzing the speech attributes. An overlap-add on the DC components in patch-wise fashion provides a 2-D visualization of the aperiodicity structure in the joint t-f plane. We evaluate the effectiveness of

10.21437/Interspeech.2018-1937

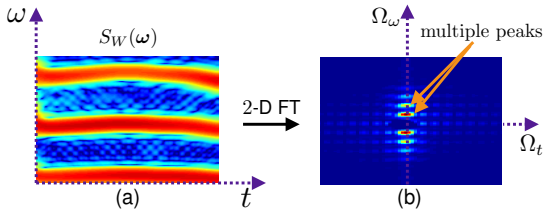

Figure 2: *(a) A spectrogram patch, and (b) its 2-D Fourier transform magnitude spectrum. For better visualization of the peaks, the DC component energy has been removed by applying a high pass filter on the magnitude spectrum.*

the obtained t-f aperiodicity map by providing it as input to the STRAIGHT vocoder [16].

This paper is organized as follows. In Section 2, we present the proposed signal model and the technique for estimation of its components. Following this, in Section 3, we apply the signal model to natural speech utterances, and make performance comparisons with the model proposed in [8]. In Section 4, we present analysis of the DC component of the model and derive a t-f aperiodicity map, which is then evaluated for speech synthesis. We conclude in Section 5, highlighting the key contributions and directions for future work.

2. The Proposed Signal Model

A monocomponent 2-D AM-FM model for a windowed spectrogram patch $S_W(\boldsymbol{\omega})$ is given by [8],

$$S_W(\boldsymbol{\omega}) \approx V(\boldsymbol{\omega})\Big(\alpha_0 + \cos\Phi(\boldsymbol{\omega})\Big), \tag{1}$$

where $\boldsymbol{\omega} = (t, \omega)$ with t and ω denoting continuous time and frequency variables, respectively. The frequency modulation (FM) and amplitude modulation (AM) of a 2-D cosine with spatial frequency $\Omega_0(\boldsymbol{\omega})$ and local orientation $\beta(\boldsymbol{\omega})$ are represented by $\Phi(\boldsymbol{\omega}) = \Omega_0(\boldsymbol{\omega})(t\cos\beta(\boldsymbol{\omega}) + \omega\sin\beta(\boldsymbol{\omega}))$ and $V(\boldsymbol{\omega})$, respectively. As an example, the 2-D Fourier transform of a patch $S_W(\boldsymbol{\omega})$ drawn from the voiced region in a spectrogram is shown in Figure 2. The 2-D cosine in (1) models the first peak in the 2-D Fourier transform. However, as can be seen, there are multiple peaks, and including them will allow for a more accurate modeling of $S_W(\boldsymbol{\omega})$. Towards this, we generalize the monocomponent model to a multicomponent one as follows:

$$\begin{aligned} S_W(\boldsymbol{\omega}) &\approx V(\boldsymbol{\omega})\Big(\alpha_0 + \sum_{k=1}^{K}\alpha_k\cos k\Phi(\boldsymbol{\omega})\Big) \\ &= \underbrace{\alpha_0 V(\boldsymbol{\omega})}_{\text{low-pass component}} + \underbrace{\alpha_1 V(\boldsymbol{\omega})\cos\Phi(\boldsymbol{\omega})}_{\text{fundamental band-pass component}} \\ &\quad + \underbrace{\alpha_2 V(\boldsymbol{\omega})\cos 2\Phi(\boldsymbol{\omega}) + \ldots}_{\text{higher-order band-pass components}}, \end{aligned} \tag{2}$$

where K is the model order, $\alpha_0 \in \mathbb{R}$ is referred to as the DC component and $\{\alpha_k\}_{k=1}^{K} \in \mathbb{R}$ act as weights on the carrier and its harmonics. In line with the model in (1), we assume $\alpha_1 = 1$ and let $\boldsymbol{\theta} = [\alpha_0\ 1\ \alpha_2\ \ldots\alpha_K]^{\mathrm{T}}$. We refer to $V(\boldsymbol{\omega})$ and $\cos\Phi(\boldsymbol{\omega})$ as the AM and FM components, respectively. Given a windowed spectrogram patch $S_W(\boldsymbol{\omega})$, the goal is to estimate $V(\boldsymbol{\omega})$, $\Phi(\boldsymbol{\omega})$, and $\boldsymbol{\theta}$.

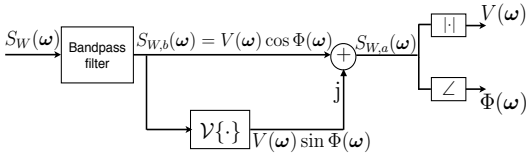

Figure 3: *A block diagram showing the Riesz transform approach for estimating the AM and FM components.*

2.1. Estimation of AM and FM Components

The AM and FM components for a spectrogram patch are estimated using the approach illustrated in Figure 3. The technique is a 2-D extension of the 1-D Hilbert transform [17] and was proposed in [8]. The spectrogram patch is first subjected to a 2-D band-pass filter designed to pick the fundamental band-pass component $V(\boldsymbol{\omega})\cos\Phi(\boldsymbol{\omega})$ denoted by $S_{W,b}(\boldsymbol{\omega})$. The vortex operator $\mathcal{V}$ [13] is applied to $S_{W,b}(\boldsymbol{\omega})$ to obtain its quadrature component [18,19]. Following this, a 2-D complex analytic signal $S_{W,a}(\boldsymbol{\omega})$ is constructed using the band-pass component and its quadrature counterpart. The modulus and angle operations on the 2-D analytic signal yield the estimates of the AM $V(\boldsymbol{\omega})$ and FM $\Phi(\boldsymbol{\omega})$.

2.2. Estimation of Parameter $\boldsymbol{\theta}$

After obtaining estimates of $V(\boldsymbol{\omega})$ and $\Phi(\boldsymbol{\omega})$ for a patch, the parameter set $\boldsymbol{\theta}$ is estimated by using the least-squares approach. Let $\boldsymbol{m} = (l, k)$ denote the discretization of $\boldsymbol{\omega} = (t, \omega)$. For a given spectrogram patch $S_W(\boldsymbol{m})$, we vectorize it into a column vector denoted by $\boldsymbol{s}$. Similarly the column vectors corresponding to $V(\boldsymbol{m})$ and $\Phi(\boldsymbol{m})$ are denoted by $\boldsymbol{v}$ and $\boldsymbol{\phi}$, respectively. We consider the problem

$$\arg\min_{\boldsymbol{\theta}} \left\| \boldsymbol{s} - \boldsymbol{v} \odot \left(\alpha_0 + \sum_{j=1}^{K}\alpha_j\cos j\boldsymbol{\phi}\right)\right\|^2, \tag{3}$$

where $\odot$ denotes element-wise product operation. Taking derivative of the cost function with respect to $\boldsymbol{\theta}$ in (3) and equating it to zero gives us a set of K linear equations,

$$\underbrace{\begin{bmatrix} \|\boldsymbol{v}_0\|^2 & \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_K \\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_0 & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_1 & \cdots & \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{v}_K \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{v}_0 & \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{v}_1 & \cdots & \|\boldsymbol{v}_K\|^2 \end{bmatrix}}_{A_{K\times(K+1)}} \underbrace{\begin{bmatrix} \alpha_0 \\ 1 \\ \alpha_2 \\ \vdots \\ \alpha_K \end{bmatrix}}_{\boldsymbol{\theta}_{(K+1)\times 1}} = \underbrace{\begin{bmatrix} \boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{s} \\ \boldsymbol{v}_2^{\mathrm{T}}\boldsymbol{s} \\ \vdots \\ \boldsymbol{v}_K^{\mathrm{T}}\boldsymbol{s} \end{bmatrix}}_{\boldsymbol{b}_{K\times 1}}, \tag{4}$$

where $\boldsymbol{v}_j = \boldsymbol{v} \odot \cos j\boldsymbol{\phi}$ for $j = 0, 1, 2, \ldots K$. The closed-form least-squares solution is $\boldsymbol{\theta} = A^{\dagger}\boldsymbol{b}$ where $A^{\dagger}$ denotes the pseudo-inverse of A. An illustration of the obtained AM, FM and $\boldsymbol{\theta}$ is shown in Figure 1.

2.3. Choice of the Model Order K

The spectrogram patches corresponding to voiced regions show localized multiple peaks in the 2-D Fourier transform (for instance, see Figure 2). A schematic of the first peak location is shown in Figure 4. The distance of first peak from the origin, that is d_0, is dependent on the fundamental frequency F0 in the patch. This is quantified by

$$d_0 = \frac{N_2}{N_1}\frac{f_s}{\mathrm{F0}\sin\gamma_0}, \tag{5}$$

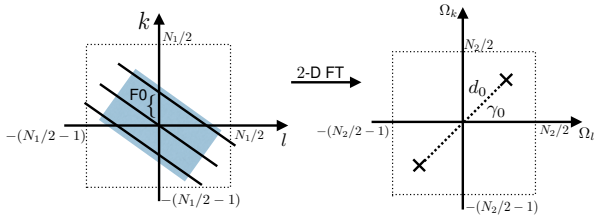

Figure 4: *A schematic for the 2-D Fourier transform of a voiced signal patch.*

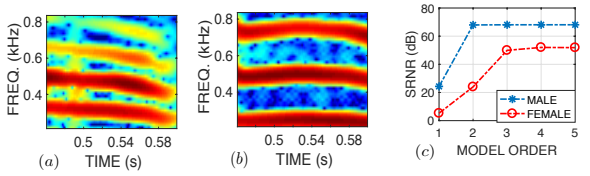

Figure 5: *(a) A spectrogram patch of a male speaker (avg. F0 = 130 Hz), (b) a spectrogram patch of a female speaker (avg. F0 = 190 Hz), and (c) SRNR (dB) vs. model order. The spectrogram is computed using a 40 ms Blackman window and 1 ms frame update interval.*

where f_s, N_1, and N_2 denote the sampling frequency, the number of DFT points used to compute the spectrogram and the 2-D Fourier transform, respectively. The model order K is adapted such that $Kd_0 \sin\gamma_0$ does not imply choosing a frequency bin outside $N_2/2$. Correspondingly,

$$K = \left\lfloor \frac{N_2/2}{d_0 \sin\gamma_0} \right\rfloor = \left\lfloor \frac{\text{F0}\, N_1}{2f_s} \right\rfloor, \tag{6}$$

which shows that for fixed N_1 and f_s, the model order K directly depends on F0, and hence, it is inherently adaptive for each patch according to F0. For instance, female speakers have a higher F0 compared to male speakers and hence K would be correspondingly higher.

The 2-D Fourier transform of an unvoiced patch does not show harmonically separated peaks. Assuming d_0 corresponds to the highest peak location in an unvoiced patch, we use the same selection criterion for the model order as given in (6).

2.3.1. Accuracy Versus Model Order

We analyze the significance of the model order by evaluating the patch reconstruction signal-to-reconstruction noise ratio (SRNR) for different values of K. This is expressed as

$$\text{SRNR} = 10\log_{10} \frac{\|S_W(\boldsymbol{\omega})\|^2}{\|S_W(\boldsymbol{\omega}) - \hat{S}_W(\boldsymbol{\omega})\|^2} \quad \text{dB},$$

where $\hat{S}_W(\boldsymbol{\omega})$ denotes the reconstructed patch using the proposed signal model. Figure 5 depicts the evaluation for patches corresponding to a male and a female speaker. It can be seen that increasing the model order from $K = 1$ to $K = 2$ improves the SRNR by about 45 dB for a male speaker. For a female speaker, the SRNR improves by about 42 dB when the model order is increased from $K = 1$ to $K = 3$. Also, we see a saturation in SRNR beyond a certain value of K. Because female speakers have a higher F0, the saturation occurs at a higher value of K (cf. (6)) than for a male speaker.

3. Evaluation on Speech Data

To evaluate the proposed approach for natural continuous speech, we used the Starkey speech database [20]. This database includes a set of 8 male and 8 female American speakers, reading the standard *rainbow passage* [21]. The 16 speakers additionally provide speaking style variability in terms of pitch, speaking rate, and perceived voice quality. From the continuous speech recordings in the dataset, we selected a total of 80 utterances, that is, 5 utterances each from the 16 speakers. Each utterance is about 4 s long. The sound files were downsampled to 8 kHz.

In order to get a narrowband spectrogram for both male and female speakers, we use a 40 ms long Blackman window with 1 ms frame update interval for the computation of short-time Fourier transform (STFT). The spectrogram is segmented into overlapping patches of fixed size (100 ms $\times$ 600 Hz). These patches are subjected to the multicomponent modeling approach introduced in Sec. 2. The reconstructed patches are subjected to a 2-D overlap-add in the least-squares sense (OLA-LSE) [22]. The resulting spectrogram is inverted with the original STFT phase to obtain the reconstructed speech signal. Three objective measures namely global SNR (GSNR), average segmental SNR (SSNR), and PESQ, are used for performance evaluation. These measures are computed between the input signal and the reconstructed speech signal. The GSNR quantifies the reconstruction error in the time domain, that is, GSNR=$20\log_{10}\left(\frac{\|x\|}{\|x-\hat{x}\|}\right)$ dB, where x and $\hat{x}$ are the input and the reconstructed signals, respectively. The average segmental SNR is obtained by averaging the frame-wise SNR over frames of duration 20 ms. The PESQ (perceptual evaluation of speech quality) metric is recommended by the ITU-T P.862 standard as an objective method to test the speech quality. It lies in the range -0.5 to 4.5 with a higher value indicating a quality close to the reference input signal.

Figure 6 shows the objective scores for the files from the dataset obtained using the proposed model (adaptive K) and the previously proposed 2-D AM-FM model [8], which uses $K = 1$. With respect to all three objective measures, we observe that adapting K increases the model accuracy. The gain is significantly more for female speakers, about 3 dB in average GSNR and SSNR. The improvement is justified as follows. Depending on the number of harmonic peaks found within the 2-D Fourier transform of a patch, the model order varies from one patch to another. In order to get an idea of the variation, we analyze the model order required by different patches obtained while analyzing the above dataset. An average normalized count of the model order after pooling all patches is shown in Figure 7. The spectrogram patches corresponding to female speakers feature a higher occurrence of model orders 3, 4 and 5, relative to the patches from male speakers. This is attributed to the higher F0 for female speakers. This also implies that it is advantageous to adapt the model order, as recommended in the proposed approach.

4. DC Component Analysis

The DC component α_0 in the signal model is constant within a patch. Pooling such matrices corresponding to all the patches in the spectrogram and employing OLA-LSE, we obtain a t-f map $A_0(\boldsymbol{\omega})$. We normalize it to lie between 0 and 1 as follows:

$$A_0(\boldsymbol{\omega}) = \frac{A_0(\boldsymbol{\omega}) - \min\{A_0(\boldsymbol{\omega})\}}{\max\{A_0(\boldsymbol{\omega})\} - \min\{A_0(\boldsymbol{\omega})\}}. \tag{7}$$

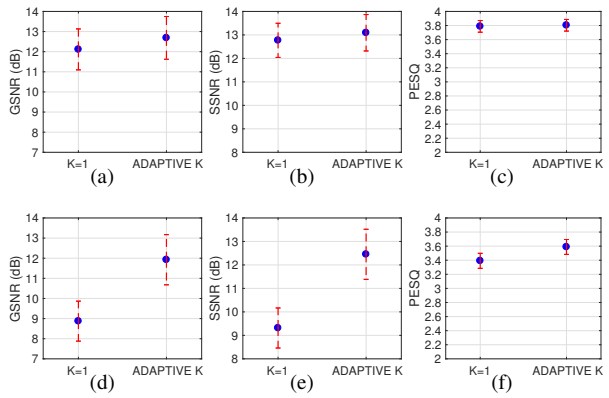

Figure 6: *Evaluation on continuous speech taken from the Starkey speech database. The first row shows the scores for male speakers and the second one for female speakers. The solid circle represents the mean value and the vertical bars indicate* 0.5 *standard deviation on either side of the mean.*

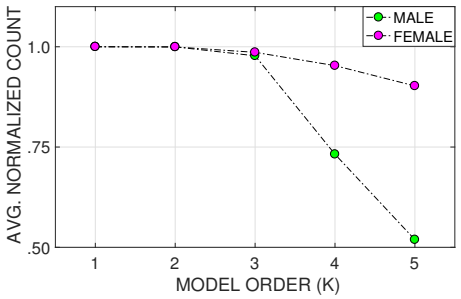

Figure 7: *The model order count is normalized over the number of patches for a given speech file and averaged with respect to the total number of files in the database.*

Figure 8 displays the t-f map of $A_0(\boldsymbol{\omega})$ along with the spectrogram. It can be observed that $A_0(\boldsymbol{\omega})$ captures the residual spectrum left after fitting the harmonics to the spectrogram. Also, it assumes relatively small values for spectrotemporal regions that have prominent harmonic striations and high values otherwise. Owing to this property of $A_0(\boldsymbol{\omega})$, we propose $A_0(\boldsymbol{\omega})$ as a t-f aperiodicity map. We evaluate its effectiveness by deploying it in a speech synthesis task.

4.1. Using $A_0(\boldsymbol{\omega})$ as the Aperiodicity Map in a Vocoder

The band-wise source aperiodicity parameters (AP) have been used for accurate modeling of the noise in the t-f domain required to synthesize natural sounding speech [23, 24]. Particularly, the aperiodicity parameters have been successfully used in vocoder based speech synthesis applications. We use STRAIGHT [16, 25], which is the most widely used vocoder. Given an input speech signal, the STRAIGHT vocoder takes an analysis-by-synthesis approach. It estimates the spectral envelope, fundamental frequency, and AP from the speech signal and uses them to synthesize the speech signal. Focusing on AP, we analyze the following three cases for speech synthesis: (1) discard the aperiodicity parameter (AP=0); (2) use AP estimated by STRAIGHT; and (3) set AP equal to the proposed aperiodicity map $A_0(\boldsymbol{\omega})$.

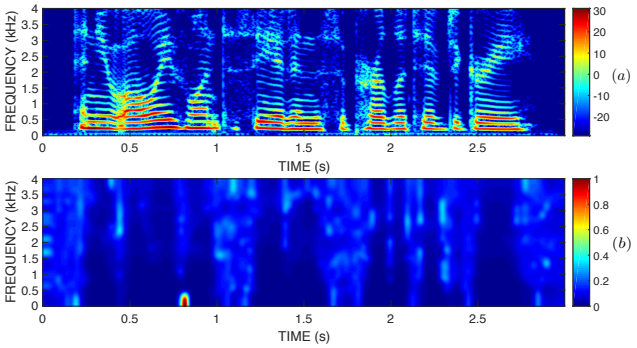

Figure 8: *(a) Spectrogram for the sentence, "And you always wanted to see it in the superlative degree," uttered by a female speaker, and (b) the estimated aperiodicity map* $A_0(\boldsymbol{\omega})$.

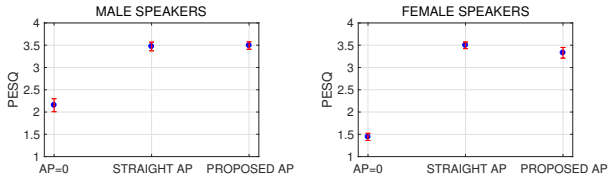

Figure 9: *PESQ scores for synthesized speech signals without aperiodicity parameters (AP=*0*), STRAIGHT AP, and proposed AP. The solid circles represent the average scores.*

We use the Starkey database and synthesize speech signals using the STRAIGHT analysis-by-synthesis approach. We test the quality of synthesized speech signals by computing PESQ scores (cf. Figure 9). Without AP (Case 1), the average PESQ scores (represented by solid circles) are below 2.5 for both male and female speakers – this indicates a poor synthesis quality, far from natural speech. With STRAIGHT AP (Case 2), the PESQ improves significantly and is about 3.5 for both male and female speakers, highlighting the importance of AP in speech synthesis. The PESQ of Case 3 is comparable to that of Case 2. This shows the effectiveness of $A_0(\boldsymbol{\omega})$ in capturing the aperiodicity attribute. Some synthesized speech samples are available at [26] for listening.

5. Conclusions

We proposed a multicomponent 2-D AM-FM model for spectrotemporal analysis of the speech signal. The number of components used is adapted patch-wise. The proposed model builds on an existing state-of-the-art monocomponent 2-D AM-FM model and offers a significant improvement in terms of modeling accuracy for both male and female speakers by making the model-order adaptive. Further, we analyzed the speech attribute captured by the DC component of the signal model and showed that it can be used to obtain a meaningful t-f representation of the aperiodicity. An evaluation carried out using the STRAIGHT vocoder highlighted the effectiveness of the new aperiodicity map.

It would be interesting to quantify the spectrotemporal speech attributes captured by the higher-order model coefficients. It would also be worthwhile analyzing the proposed signal model for noisy and reverberant speech. We hypothesize that the model parameters may aid in quantifying the noise and degree of reverberation in a speech signal.

6. References

[1] H. C. Green, R. K. Potter, and G. A. Kopp, *Visible Speech*. Van Nostrand, 1947.

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

[3] F. K. Soong and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6, pp. 871–879, 1988.

[4] R. V. Shannon, F.-G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, no. 5234, pp. 303–304, 1995.

[5] T. Ezzat, J. Bouvrie, and T. Poggio, "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.

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

[7] ——, "Towards interpretive models for 2-D processing of speech," *IEEE Transactions on Audio, Speech, and Language Process.*, vol. 20, no. 7, pp. 2159–2173, Sept 2012.

[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] D. A. Depireux, J. Z. Simon, J. Klein, David, and S. A. Shamma, "Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex," *Journal of Neurophysiology*, vol. 85, no. 3, pp. 1220–1234, 2001.

[10] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[11] F. E. Theunissen and J. E. Elie, "Neural processing of natural sounds," *Nature Reviews Neuroscience*, vol. 15, pp. 355–366, 2014.

[12] K. N. Stevens, *Acoustic Phonetics*. The MIT Press, 1999, vol. 1st Edition.

[13] 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.

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

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

[16] 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.

[17] T. Quatieri, *Discrete-time Speech Signal Processing: Principles and Practice*, 1st ed. Upper Saddle River, NJ, USA: Prentice Hall Press, 2001.

[18] A. H. Nuttall and E. Bedrosian, "On the quadrature approximation to the Hilbert transform of modulated signals," *Proceedings of the IEEE*, vol. 54, no. 10, pp. 1458–1459, 1966.

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

[20] "Starkey Hearing Technologies. Open access stimuli for the creation of multi-talker maskers, [Online]," http://www.starkeyevidence.com, 2013.

[21] G. Fairbanks, "Voice and articulation drillbook," vol. 2, pp. 124–139, 1960.

[22] 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.

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

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

[25] 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.

[26] "Demonstration of speech synthesis using aperiodicity parameters, [Online]," https://jitendradhiman.github.io/APTEST.html, 2018.