RIESZ-TRANSFORM-BASED DEMODULATION OF NARROWBAND SPECTROGRAMS OF VOICED SPEECH

Haricharan Aragonda and Chandra Sekhar Seelamantula

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

Emails: haricharan@ieee.org, chandra.sekhar@ieee.org.

ABSTRACT

Narrowband spectrograms of voiced speech can be modeled as an outcome of two-dimensional (2-D) modulation process. In this paper, we develop a demodulation algorithm to estimate the 2-D amplitude modulation (AM) and carrier of a given spectrogram patch. The demodulation algorithm is based on the Riesz transform, which is a unitary, shift-invariant operator and is obtained as a 2-D extension of the well known 1-D Hilbert transform operator. Existing methods for spectrogram demodulation rely on extension of sinusoidal demodulation method from the communications literature and require precise estimate of the 2-D carrier. On the other hand, the proposed method based on Riesz transform does not require a carrier estimate. The proposed method and the sinusoidal demodulation scheme are tested on real speech data. Experimental results show that the demodulated AM and carrier from Riesz demodulation represent the spectrogram patch more accurately compared with those obtained using the sinusoidal demodulation. The signal-to-reconstruction error ratio was found to be about 2 to 6 dB higher in case of the proposed demodulation approach.

Index Terms— Riesz transform, Spectrogram demodulation.

1. INTRODUCTION

Most speech analysis algorithms work either on the spectral modulations (linear prediction [1], cepstral analysis [2]) or temporal modulations (modulation filtering [3], frequency-domain linear prediction [4]) of speech, independently, and have been quite successful in applications such as speech coding [5, 6] and automatic speech recognition [7, 8]. Some recent results have shown that it is advantageous to work with both spectral and temporal modulations (spectro-temporal modulation) simultaneously [9–15]. These algorithms work in the time-frequency plane and are referred to as two dimensional (2-D) techniques for speech signal analysis. The spectrogram demodulation problem we address in this paper fits into this framework. Local regions of narrowband speech spectrograms can be modeled as an outcome of 2-D modulation process with amplitude modulation (AM) and carrier being related to the vocal tract response and pitch dynamics, respectively [10, 16–20]. The aim in demodulation is to estimate the AM and carrier given a spectrogram patch. Figure 1 illustrates the demodulation and modulation of spectrograms. Spectrogram demodulation has found application in problems such as speaker separation [17] and formant estimation [18]. Before proceeding with other applications, we need

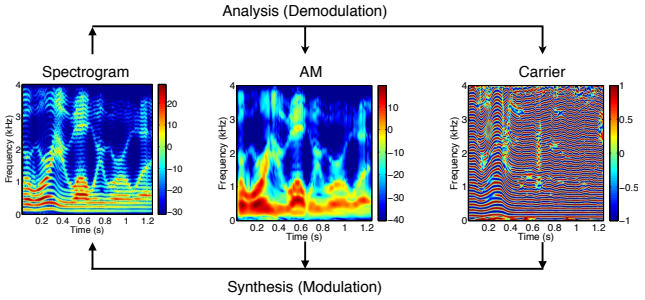

Fig. 1. Illustration of the 2-D demodulation and modulation process.

to develop accurate methods of spectrogram demodulation. Current methods of spectrogram demodulation require an estimate of the underlying carrier [10, 17, 19] to estimate the AM. In this paper, we propose an incoherent approach (not requiring carrier estimates) based on the Riesz transform, which is an extension of the Hilbert transform to 2-D, to address the problem of spectrogram demodulation. Riesz transform was recently introduced in the optics community, where it is also known as the spiral-phase quadrature transform [21]. Riesz-transform-based methods have found applications in problems such as fingerprint analysis [22] and demodulation of digital holograms [23]. In this paper, we employ the Riesz transform for demodulating narrowband speech spectrograms.

The following notations are used in this paper: $S(\mathbf{m})$ is used to denote a spectrogram, where $\mathbf{m} = (\ell, m)$ with ℓ and m denoting the frame and frequency indices, respectively. A patch of spectrogram is denoted by $S_W(\mathbf{m})$ and is obtained by multiplying $S(\mathbf{m})$ with a 2-D window $W(\mathbf{m})$. Fourier transform of $S_W(\mathbf{m})$ is denoted by $\hat{S}_W(\mathbf{\Omega})$, $\mathbf{\Omega} = (\Omega_\ell, \Omega_m)$, with Ω_ℓ and Ω_m denoting the spatial frequency variables along ℓ (time axis) and m (frequency axis), respectively.

The paper is organized as follows: In Section 2, we discuss the signal model and formulate the 2-D demodulation problem. We next present the Riesz transform and develop a Riesz-transform-based spectrogram demodulator in Section 3. The demodulation algorithm is tested on real speech data and the results are compared with that of sinusoidal demodulation in Section 4. We conclude with Section 5, where the results are summarized and the relative merits and demerits of the proposed algorithm compared with other demodulation algorithms are discussed.

2. PROBLEM FORMULATION

We adopt the spectrogram patch model similar to that used by Wang and Quatieri in [17] but with additional flexibility. Our model allows

This work is sponsored in part by the Department of Information Technology, Government of India project entitled, "Speech-based access," and the Indian Space Research Organization—Indian Institute of Science Space Technology cell project, ISTC/EEE/CSS/293.

the spatial frequency and the orientation of the 2-D carrier to be a function of $\mathbf{m}$. This generalization of the carrier allows us to model pitch dynamics more accurately. That is, $S_W(\boldsymbol{\omega})$ can be expressed as

$$\begin{aligned} S_W(\mathbf{m}) &= V(\mathbf{m})\left(D + \cos\Phi(\mathbf{m})\right), \\ &= \underbrace{V(\mathbf{m})D}_{S_{W,l}(\boldsymbol{\omega})} + \underbrace{V(\mathbf{m})\cos\Phi(\mathbf{m})}_{S_{W,b}(\boldsymbol{\omega})}, \end{aligned} \tag{1}$$

where $\Phi(\mathbf{m}) = \omega(\mathbf{m})\left[\ell\cos\theta(\mathbf{m}) + m\sin\theta(\mathbf{m})\right]$. $\omega(\mathbf{m})$ and $\theta(\mathbf{m})$ denote the spatial frequency and orientation at the point $\mathbf{m} = (\ell, m)$. We address the problem of estimating the AM, $V(\mathbf{m})$, and the carrier, $\cos\Phi(\mathbf{m})$, given $S_W(\boldsymbol{\omega})$. $S_{W,l}(\boldsymbol{\omega})$ and $S_{W,b}(\boldsymbol{\omega})$ in (1), are the lowpass and the bandpass components of $S_W(\mathbf{m})$, respectively.

For the pitch harmonics to be modeled as a 2-D cosine, we have empirically observed that the size of the 1-D window used should be between 3 to 6 times of the pitch period. In order to satisfy the requirement, we have used 20 ms and 30 ms windows for female and male speakers, respectively, for computing spectrograms.

3. RIESZ-TRANSFORM-BASED DEMODULATION OF SPEECH SPECTROGRAMS

The Riesz transform is a 2-D extension of Hilbert transform [24], and is associated with frequency response $\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega})$:

$$\hat{h}_{\mathcal{R}}(\boldsymbol{\Omega}) = \frac{-\mathrm{j}\Omega_\ell + \Omega_m}{\sqrt{\Omega_\ell^2 + \Omega_m^2}}. \tag{2}$$

From (2), we see that the Riesz transform is a unitary operator, that is, it has an all pass behavior. The phase response associated with Riesz transform is shown in Figure 2. Given a 2-D signal of the form $a(\mathbf{m})\cos\Phi(\mathbf{m})$, its Riesz transform is given by [25]

$$\mathcal{R}\left\{a(\mathbf{m})\cos\Phi(\mathbf{m})\right\} = e^{\mathrm{j}\beta(\mathbf{m})}a(\mathbf{m})\sin\Phi(\mathbf{m}), \tag{3}$$

where $\mathcal{R}$ denotes the Riesz operator, and $\beta(\mathbf{m})$ indicates local orientation angle of $S_W(\mathbf{m})$ at $\mathbf{m}$. $\beta(\mathbf{m})$ gives the angle of the vector in the direction of minimum change in a 2-D function. The concept of orientation is explained with the help of Figure 3, where in we show a synthetic cosine oriented at $\frac{\pi}{4}$ radians to the horizontal axis, and a spectrogram corresponding to real speech signal. In both cases arrows indicate the local orientation, which is defined as the direction along which the local variation is minimum. While in the case of a synthetic cosine, the orientation is constant throughout, the orientation is function of $\mathbf{m}$ in the case of a real spectrogram. From the figure, we see that the local orientation is related to pitch dynamics. Multiplying both sides of (3) with $e^{-\mathrm{j}\beta(\mathbf{m})}$, we get that

$$e^{-\mathrm{j}\beta(\mathbf{m})}\mathcal{R}\{a(\mathbf{m})\cos\Phi(\mathbf{m})\} = a(\mathbf{m})\sin\Phi(\mathbf{m}). \tag{4}$$

The operator on the left hand side of (4), called the Vortex operator [21], is denoted by $\mathcal{V}\{\cdot\} = e^{-\mathrm{j}\beta(\mathbf{m})}\mathcal{R}\{\cdot\}$. Vortex operator exhibits quadrature property similar to that of the 1-D Hilbert transform. We use the quadrature property of the vortex operator to carry out spectrogram demodulation.

Figure 4 shows the block diagram of the Riesz-transform-based demodulator of $S_W(\mathbf{m})$. The spectrogram patch $S_W(\mathbf{m})$ is passed through a bandpass filter to retain only $S_{W,b}(\mathbf{m})$ component of $S_W(\mathbf{m})$. Since $S_{W,b}(\mathbf{m})$ is of the form $V(\mathbf{m})\cos\Phi(\mathbf{m})$, the output of Vortex operator can be written as $V(\mathbf{m})\sin\Phi(\mathbf{m})$.

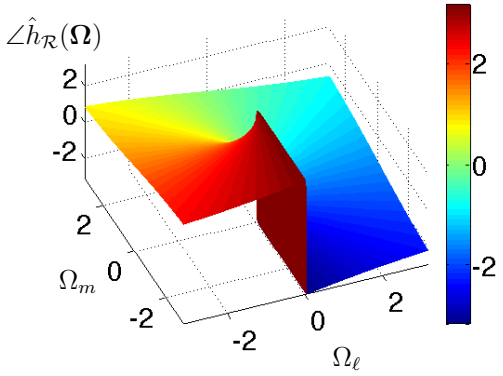

Fig. 2. Phase response associated with the Riesz transform

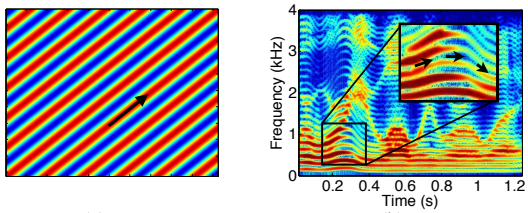

Fig. 3. (Color in electronic version) Illustration of the concept of orientation in spectrograms. (a) shows a 2-D cosine with arrow indicating its orientation; (b) illustrates the concept of orientation with respect to a real spectrogram patch. We observe that the local orientation changes with $\mathbf{m}$ and is related to the pitch dynamics.

The outputs are then combined to form a 2-D complex signal, $S_{W,c} = S_{W,b} + \mathrm{j}\mathcal{V}\{S_{W,b}\} = V(\mathbf{m})e^{\mathrm{j}\Phi(\mathbf{m})}$, from which the AM and carrier are extracted. $\beta(\mathbf{m})$ is computed using 2-D principal component analysis, which is equivalent to the structure tensor method [23] in image processing

Let $\tilde{V}(\mathbf{m})$ and $\tilde{\Phi}(\mathbf{m})$ denote the estimated AM and phase of $S_W(\mathbf{m})$, and let $\tilde{S}_W(\mathbf{m})$ denote the estimate of $S_W(\mathbf{m})$ obtained from $\tilde{V}(\mathbf{m})$ and $\tilde{\Phi}(\mathbf{m})$. Then,

$$\tilde{S}_W(\mathbf{m}) = \tilde{V}(\mathbf{m})[\tilde{D} + \cos\tilde{\Phi}(\mathbf{m})], \tag{5}$$

where $\tilde{D} = \arg\min_D ||S_W(\mathbf{m}) - \tilde{S}_W(\mathbf{m})||_2^2$ [17]. Let $\tilde{S}_W^{i,j}(\mathbf{m})$ denote (i,j)th reconstructed spectrogram patch, $S(\mathbf{m})$ is reconstructed from $S_W^{i,j}(\boldsymbol{\omega})$ corresponding to different values of i and j using overlap-add in the least-squares sense (OLA-LSE) [17].

$$\tilde{S}(\mathbf{m}) = \frac{\sum_{i,j}\tilde{S}_W^{i,j}(\mathbf{m})W(Fj - m, Ti - n)}{\sum_{i,j}W^2(Fj - m, Ti - n)}, \tag{6}$$

where T and F denote the step size of the 2-D window along the time and frequency axes, respectively. $\tilde{S}(\mathbf{m})$ is combined with the phase of the original STFT, which is then inverted using OLA-LSE criterion [26] to get an estimate of the speech signal, $\tilde{s}(n)$. The inversion formula is given by,

$$\tilde{s}(n) = \frac{\sum_l \tilde{s}_w(n,l)w(Tl - n)}{\sum_l w^2(Tl - n)}, \tag{7}$$

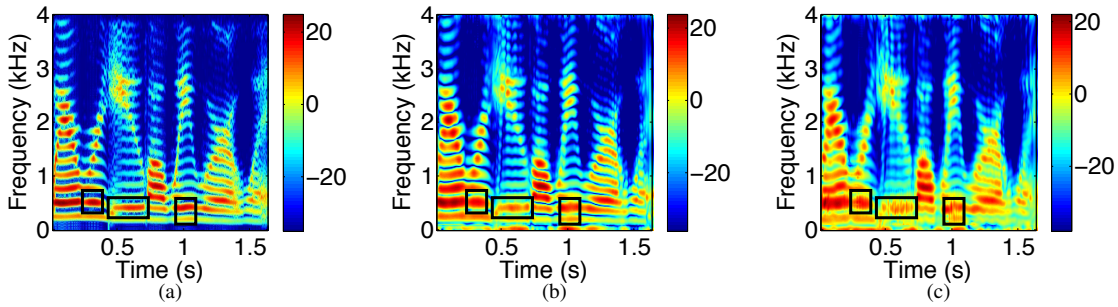

Fig. 6. (Color in electronic version) Spectrogram corresponding to the speech of fS1: (a) Original spectrogram, computed using 20 ms Hamming window and 512-point DFT, (b) Spectrogram reconstructed from the AM and carrier obtained using Riesz-transform-based demodulator, (c) Same as (b) but using sinusoidal demodulation. Rectangles indicate some regions where Riesz-transform-based demodulator gives accurate estimates compared with sinusoidal method.

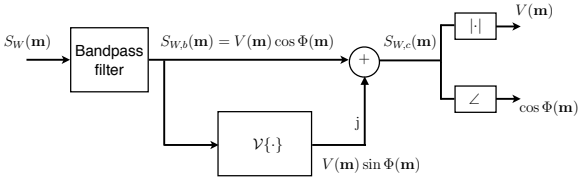

Fig. 4. Block diagram illustrating the Riesz-transform-based demodulation. $|\cdot|$ and $\angle(\cdot)$ denote the modulus and angle, respectively.

where $\tilde{s}_w(n,\ell)$ is the inverse Fourier transform of the ℓth frame of estimated STFT.

4. RESULTS

The proposed demodulation algorithm is tested on real speech data taken from the TIMIT database [27]. Speech files corresponding to all-voiced sentences "S1: Where were you while we were away" and "S2: He will allow a rare lie" were chosen. Male and female speakers are distinguished by a prefixes 'm' and 'f' before the sentence label. 'mS1,' 'fS1,' 'mS2,' and 'fS2' correspond to speakers with speaker ID 'DAC2,' 'GJD0,' 'GJF0,' and 'JMG0', respectively, in the TIMIT database. Before performing demodulation, silence regions were removed manually and the speech was downsampled to 8 kHz. Speech signal is normalized to have peak time-domain magnitude of one. Preprocessing steps were carried out using Praat [28]. Since the model assumes that the 1-D window duration is 3 to 6 times the pitch period, a 30 ms Hamming window is used for male speakers and a 20 ms Hamming window is used for female speakers. Spectrogram is computed using 512-point discrete Fourier transform (DFT) and is demodulated in patches of size 600 Hz in frequency and 100 ms in time. A 2-D Hamming window is used to select a spectrogram patch.

The carrier estimate required for sinusoidal demodulation technique, with which we compare the performance of the Riesz-transform-based approach, is estimated from the center frequency of the bandpass component as described in [10]. Butterworth filters of 10th order are used for highpass and lowpass filtering in sinusoidal demodulation. The cutoff frequency of the highpass filter and the bandwidth of the lowpass filter are taken to be half of the estimated spatial frequency. Bandpass filter used in the Riesz-transform-based demodulator is a 10th-order Butterworth filter with its center frequency corresponding to that of the estimated 2-D carrier, and having a bandwidth equal half of the estimated spatial frequency.

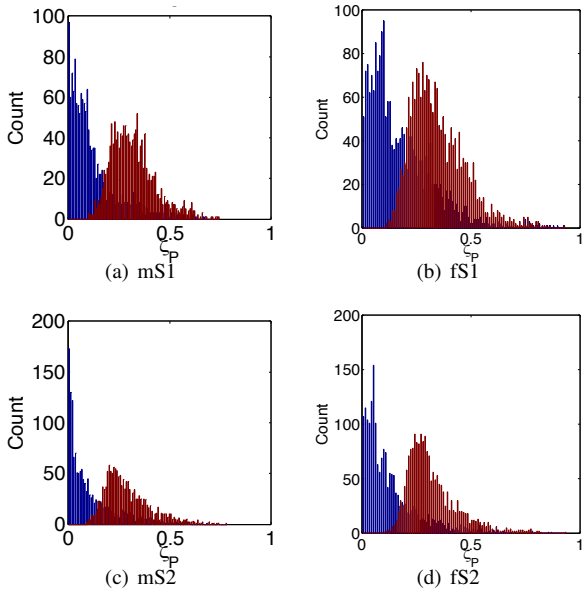

Fig. 5. (Color in electronic version) Histogram of ζ_p corresponding to the different speech files used. Blue and brown colors indicate histograms of Riesz and sinusoidal methods, respectively.

The accuracy of demodulation is measured in terms of how well the estimated AM and carrier can represent the spectrogram patch $S_W(\mathbf{\Omega})$. This is quantified by first estimating the spectrogram patch from the $\tilde{V}(\mathbf{m})$ and $\cos\tilde{\Phi}(\mathbf{m})$. Once we have $\hat{S}_W(\mathbf{m})$, the demodulation performance is then quantified using the metric ζ_p:

$$\zeta_p = \frac{\sum_{\mathbf{m}} |S_W(\mathbf{m}) - \tilde{S}_W(\mathbf{m})|^2}{\sum_{\mathbf{m}} |S_W(\mathbf{m})|^2}. \tag{8}$$

Figure 5 shows the histogram of ζ_p for different speech files for both Riesz-transform-based demodulation (blue) and sinusoidal de-

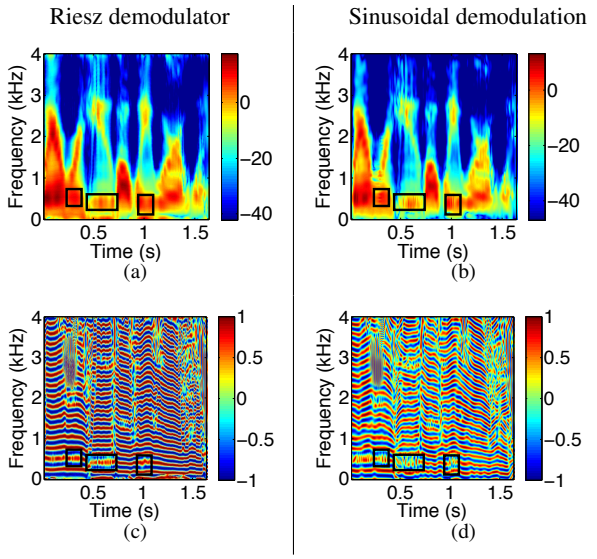

Fig. 7. (Color in electronic version) AM and carrier corresponding to fS1. Top and bottom rows indicate the AM and carrier of 6(a) obtained using Riesz and sinusoidal demodulation techniques. Rectangles are marked at same position as in Figure 6

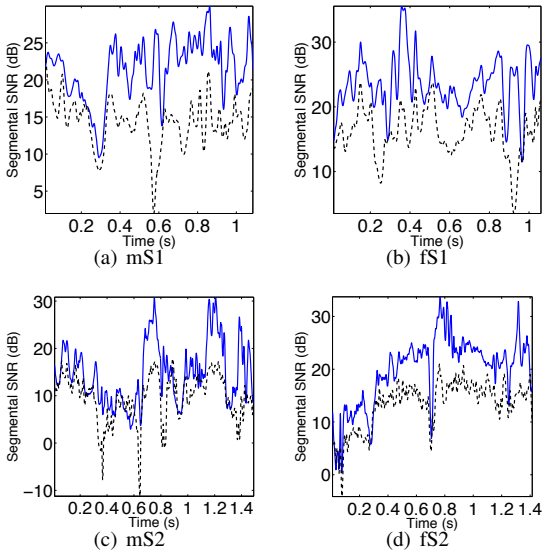

Fig. 8. (Color in electronic version) Segmental SNRs of speech files reconstructed from the demodulated AM and carrier. Thick blue line and black dashed lines correspond to Riesz based method and sinusoidal methods, respectively.

modulation (brown). The histograms of ζ_p corresponding to Riesz-transform-based demodulation are centered at lower values of ζ_p compared with that of sinusoidal demodulation indicating accurate demodulation by the proposed method.

Figure 6 shows the original spectrogram, $S(\mathbf{m})$ corresponding to 'fS1,' and its estimates, $\tilde{S}(\mathbf{m})$ obtained using Riesz and sinusoidal demodulation algorithms[1]. The Riesz-transform-based demodulator gives accurate estimate of spectrogram compared with that the sinusoidal region, rectangular boxes enclose some regions where the performance of the two algorithms differ significantly. The deterioration in the performance of the sinusoidal method is largely due to errors in carrier estimation. In Figure 7 we show the AM and carrier (reconstrcuted from AM and carrier of patches using equation 6) corresponding to the spectrogram in Figure 6. Comparing the AM and carrier obtained using the two methods, we see that Riesz-transform-based demodulation gives relatively smooth estimates of AM and carrier and preserves time-frequency continuity. This is particularly evident in the regions enclosed by the rectangles.

An estimate of speech signal $\hat{s}(n)$ is obtained using (7). Figure 8 shows the segmental SNRs of reconstructed speech. Speech estimated from AM and carrier obtained using Riesz-transform-based demodulation has higher segmental SNR compared with that obtained using sinusoidal method in most frames. Regions where the segmental SNRs drop sharply correspond to inharmonic regions in the spectrograms. Global SNR and average segmental SNR for different speech files are given in Table 1. Riesz-transform-based method performs more accurately compared with the sinusoidal demodulation in terms of both the global SNR as well as average segmental SNR.

[1]Results on other files are available at sites.google.com/site/rdemod

Filename	Global SNR		Average segmental SNR	
	Sinusoidal	Riesz	Sinusoidal	Riesz
mS1	14.60	**19.35**	14.33	**21.54**
fS1	11.08	**13.38**	9.61	**14.74**
mS2	16.62	**22.68**	16.16	**23.52**
fS2	11.22	**14.99**	13.34	**19.81**

Table 1. Comparison of global and average segmental SNRs of speech reconstructed from sinusoidal and Riesz-transform-based demodulation techniques.

5. CONCLUSIONS

We have developed a demodulation algorithm for narrowband speech spectrograms using the Riesz transform, which is a 2-D extension of the Hilbert transform. We have modeled the spectrogram patch as a 2-D AM-FM signal, this model is similar to that used by Wang and Quatieri in [17], but allows for the 2-D carrier to be frequency modulated to model the pitch dynamics accurately. In contrast to some 2-D demodulation algorithms such as Max-Gabor demodulation [19], which uses scattered data interpolation to estimate the AM, and sinusoidal demodulation [17] which uses sinusoidal demodulation, the proposed demodulation algorithm does not require 2-D carrier estimates, making AM estimation independent of carrier estimation errors. Experimental results have shown that Riesz method gives more accurate estimate of AM and carrier compared with the sinusoidal method. As part of future work, we would like to extend the signal model and the demodulation algorithm to handle arbitrary 1-D window sizes along the lines of [10]. We would also like to examine the effect of improved accuracy in AM and carrier estimates provided by the Riesz-transform-based demodulator on applications such as speaker separation and formant estimation.

6. REFERENCES

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

[2] A. Oppenheim and R. Schafer, "Homomorphic analysis of speech," *IEEE Trans. Audio Electroacoust.*, vol. 16, no. 2, pp. 221–226, Jun. 1968.

[3] Q. Li and L. Atlas, "Coherent modulation filtering for speech," *in Proc. IEEE Int. Conf. on Acoust. Speech and Signal Process.*, pp. 4481–4484, Apr. 2008.

[4] S. Ganapathy, *Signal Analysis Using Autoregressive Models of Amplitude Modulation*, Ph.D. thesis, Johns Hopkins, Jan. 2012.

[5] B. Atal and J. Remde, "A new model of LPC excitation for producing natural-sounding speech at low bit rates," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, vol. 7, pp. 614–617, May 1982.

[6] M.S. Vinton and L.E. Atlas, "Scalable and progressive audio codec," vol. 5, pp. 3277–3280, 2001.

[7] H. Hermansky, "Perceptual linear prediction (PLP) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, pp. 1738–1752, 1990.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[9] Y. Shiga and S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis," *in Proc. Eurospeech*, pp. 1737–1740, 2003.

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

[11] P. Smaragdis, "Discovering auditory objects through nonnegativity constraints," *in Proc. Stat. and Perc. Audio Process. (SAPA)*, Oct. 2004.

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

[13] B. T. Meyer and B. Kollmeier, "Optimization and evaluation of gabor feature sets for ASR," *in Proc. Interspeech*, pp. 906–909, 2008.

[14] G. J. Mysore, *A non-negative framework for joint modeling of spectral structure and temporal dynamics in sound mixtures*, Ph.D. thesis, Stanford University, 2010.

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

[16] T. F. Quatieri, "2-D processing of speech with application to pitch estimation," *in Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Sep. 2001.

[17] T. T. Wang and T. F. Quatieri, "Towards co-channel speaker separation by 2-D demodulation of spectrograms," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, pp. 65–68, Oct. 2009.

[18] T. T. Wang and T. F. Quatieri, "High-pitch formant estimation by exploiting temporal change of pitch," *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 1, pp. 171–186, Jan. 2010.

[19] T. Ezzat, J. Bouvrie, and T. Poggio, "AM-FM demodulation of spectrograms using localized 2D Max-Gabor analysis," *in Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, vol. 4, pp. 1061–1064, Apr. 2007.

[20] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-temporal analysis of speech using 2-D Gabor filters," *in Proc. Interspeech*, pp. 506–509, 2007.

[21] K. G. Larkin, D. J. Bone, and M. A. Oldfield, "Natural demodulation of two-dimensional fringe patterns. I. general background of the spiral phase quadrature transform," *J. Opt. Soc. Amer. (A)*, vol. 18, no. 8, pp. 1862–1870, 2001.

[22] K. G. Larkin and P. A. Fletcher, "A coherent framework for fingerprint analysis: Are fingerprints holograms?," *Opt. Exp.*, vol. 15, pp. 8667–8677, 2007.

[23] 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. Amer. (A)*, vol. 29, no. 10, pp. 2118–2129, Oct. 2012.

[24] E. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, 1971.

[25] H. Aragonda and C. S. Seelamantula, "Quadrature approximation properties of the spiral-phase quadrature transform," *Proc. IEEE Int. Conf. Acoust. Speech and Signal Process.*, pp. 1389 –1392, May 2011.

[26] T. F. Quatieri, *Discrete-Time Speech Signal Processing: Principles and Practice*, Prentice Hall, Upper Saddle River, NJ, 2001.

[27] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: Specifications and status," *in Proc. DARPA Workshop Speech Recogn.*, pp. 93–99, 1986.

[28] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.44)," .