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 applications in problems such as speaker separation [17] and formant estimation [18]. Before proceeding with other applications, we need 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: $S(m)$ is used to denote a spectrogram, where $m = (\ell, m)$ with $\ell$ and $m$ denoting the frame and frequency indices, respectively. A patch of spectrogram is denoted by $S_W(m)$ and is obtained by multiplying $S(m)$ with a 2-D window $W(m)$. Fourier transform of $S_W(m)$ is denoted by $\hat{S}_W(\Omega)$, $\Omega = (\Omega_\ell, \Omega_m)$, with $\Omega_\ell$ and $\Omega_m$ denoting the spatial frequency variables along $\ell$ (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
the spatial frequency and the orientation of the 2-D carrier to be a function of \( m \). This generalization of the carrier allows us to model pitch dynamics more accurately. That is, \( S_W(\omega) \) can be expressed as

\[
S_W(m) = \frac{V(m)(D + \cos \Phi(m))}{S_{W,I}(\omega)} + \frac{V(m)D + V(m)\cos \Phi(m)}{S_{W,Q}(\omega)},
\]

where \( \Phi(m) = \omega(m)(\ell \cos \theta(m) + m \sin \theta(m)) \). \( \omega(m) \) and \( \theta(m) \) denote the spatial frequency and orientation at the point \( m = (\ell, m) \). We address the problem of estimating the AM, \( V(m) \), and the carrier, \( \cos \Phi(m) \), given \( S_W(\omega) \). \( S_{W,I}(\omega) \) and \( S_{W,Q}(\omega) \) in (1), are the lowpass and the bandpass components of \( S_W(m) \), 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 \( \hat{h}_R(\Omega) \):

\[
\hat{h}_R(\Omega) = \frac{-j\Omega_x + \Omega_m}{\sqrt{\Omega_x^2 + \Omega_m^2}}.
\]

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 \( e^{j\beta(m)}a(m) \cos \Phi(m) \), its Riesz transform is given by [25]

\[
\mathcal{R}\{a(m)\cos \Phi(m)\} = e^{j\beta(m)}a(m)\sin \Phi(m),
\]

where \( \mathcal{R} \) denotes the Riesz operator, and \( \beta(m) \) indicates local orientation angle of \( S_W(m) \) at \( m \). \( \beta(m) \) 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 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 \( m \) 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 \( e^{-j\beta(m)} \), we get that

\[
e^{-j\beta(m)}\mathcal{R}\{a(m)\cos \Phi(m)\} = a(m)\sin \Phi(m).
\]

The operator on the left hand side of (4), called the Vortex operator [21], is denoted by \( \mathcal{V}\{\cdot\} = e^{-j\beta(m)}\mathcal{R}\{\cdot\} \). 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 \( S_W(m) \). The spectrogram patch \( S_W(m) \) is passed through a bandpass filter to retain only \( S_{W,b}(m) \) component of \( S_W(m) \). Since \( S_{W,b}(m) \) is of the form \( V(m)\cos \Phi(m) \), the output of Vortex operator can be written as \( V(m)\sin \Phi(m) \).

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

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

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

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

\[
\tilde{s}(m) = \frac{\sum_{i,j} S^j_W(m)V(F_j - m, Ti - n)}{\sum_{i,j} W^2(F_j - m, Ti - n)},
\]

where \( T \) and \( F \) denote the step size of the 2-D window along the time and frequency axes, respectively. \( \tilde{s}(m) \) 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,

\[
\tilde{s}(n) = \frac{\sum_{l} \tilde{s}_w(n,l)w(Tl - n)}{\sum_l w^2(Tl - n)},
\]
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.

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

\[
\zeta_p = \frac{\sum_m |S_W(m) - \hat{S}_W(m)|^2}{\sum_m |S_W(m)|^2}.
\]  

Figure 5 shows the histogram of \( \zeta_p \) corresponding to different speech files for both Riesz-transform-based demodulation (blue) and sinusoidal de-
Sinusoidal demodulation is centered at lower values of modulation (brown). The histograms of $\zeta_p$ corresponding to Riesz-transform-based demodulation are centered at lower values of $\zeta_p$ compared with that of sinusoidal demodulation indicating accurate demodulation by the proposed method.

Figure 6 shows the original spectrogram, $S(m)$ corresponding to ‘IS1,’ and its estimates, $\hat{S}(m)$ obtained using Riesz and sinusoidal demodulation algorithms. The Riesz-transform-based demodulator gives accurate estimate of spectrogram compared with that of sinusoidal demodulation indicating accurate demodulation by the proposed method.

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.

---

<table>
<thead>
<tr>
<th>Filename</th>
<th>Global SNR</th>
<th>Average segmental SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>mS1</td>
<td>14.60</td>
<td>19.35</td>
</tr>
<tr>
<td>IS1</td>
<td>11.08</td>
<td>13.38</td>
</tr>
<tr>
<td>mS2</td>
<td>16.62</td>
<td>22.68</td>
</tr>
<tr>
<td>IS2</td>
<td>11.22</td>
<td>14.99</td>
</tr>
</tbody>
</table>

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

---

1 Results on other files are available at sites.google.com/site/rdemod
6. REFERENCES


