A Probabilistic Analysis of the Ratio Spectrum

Mark D. Skowronski  
Computational Neuro-Engineering Lab  
University of Florida  
Gainesville, FL 32611  
markskow@cnel.ufl.edu

John G. Harris  
Computational Neuro-Engineering Lab  
University of Florida  
Gainesville, FL 32611  
harris@cnel.ufl.edu

Abstract

Recently, Lim and Harris have developed a novel spectral representation called the ratio spectrum [1]. The ratio spectrum is formed by taking the ratio of the power of a low-pass filtered signal to the power of the original unfiltered signal for all filter cutoff frequencies. Features extracted from the ratio spectrum have proven to be promising for phoneme recognition [2, 3] and speech compression [4, 5]. In this work we have shown that the ratio spectrum feature sets are competitive with standard feature sets used for automatic speech recognition. To understand the underlying properties of the ratio spectrum that contribute to the generation of a useful feature set, we derived the analytic expressions that describe the behavior of the ratio spectrum for a white noise model.

1. Introduction

The constant-Q bandpass filter bank technique is commonly used to extract features from a signal for a variety of applications, notably signal compression and speech recognition. The center frequencies of the bandpass filters may be linearly, logarithmically or otherwise spaced but these fixed sampling schemes necessarily waste samples in areas of the spectrum with little signal energy. More complex adaptive bandpass filter schemes of the LMS-type gradient descent variety find difficulty in complex adaptive control and local minima in their attempts to move samples in regions of high signal energy. The ratio spectrum method provides an efficient method for extracting features which tend to occur in regions of high signal energy. Lim and Harris introduced the ratio spectrum in continuous and discrete form [1]. In this paper we consider the FFT-based discrete form.

Consider $x(n)$, a segment of a discrete-time stochastic signal. The discrete-time ratio spectrum of $X(k)$ is

$$\mathcal{R}(m) = \frac{\text{Output Power}}{\text{Input Power}} = \frac{\sum_{k=0}^{N-1} |H_m(k)X(k)|^2}{\sum_{k=0}^{N-1} |X(k)|^2} \quad (1)$$

where $X(k)$ is the DFT of $x(n)$, $H_m(k)$ is a low-pass filter with cutoff frequency at $m$, $N$ is the length of $X$, and $\mathcal{R}(m)$ is the ratio spectrum of $X$, where $m$ ranges from 0 to $N-1$. The ratio spectrum in Equation 1 has many desirable properties:

- Monotonic, non-decreasing function that is bounded between 0 and 1
- Makes no a priori assumptions about $x(n)$
- Simple concept, easy to realize (amenable to powerful FFT algorithms)

The continuous-time version of the ratio spectrum shares similar characteristics but also includes a log-frequency blurring which can be tuned to approximate the Equivalent Rectangular Bandwidth of human hearing [2].
We can generalize the ratio spectrum using an $L_p$ norm as

$$\mathcal{R}_p(m) = \frac{\sum_{k=0}^{N-1} |H_m(k)X(k)|^p}{\sum_{k=0}^{N-1} |X(k)|^p}$$  \hspace{1cm} (2)

In this way, $p = 2$ reduces to Equation 1 and $p = 1$ gives a ratio spectrum using the magnitude spectrum rather than the power spectrum. Note that the desirable properties listed above still hold for $\mathcal{R}_p(m)$ for arbitrary $p$. The ratio spectrum is related to the concept of spectral distribution [6, 7]. Here, the integral of the power spectrum is mainly used to avoid the treatment of impulse functions in the spectral description of a signal. The ratio spectrum is motivated by computational advantages over standard techniques.

Figure 1 illustrates the ratio spectrum concept of feature extraction. Uniform sampling along the frequency axis extracts equally-spaced samples from the spectrum in Figure 1(b); it is clearly seen in Figure 1(a) that the features extracted uniformly along the ratio spectrum axis tend to occur in regions of high signal energy, while samples taken uniformly along the frequency axis may occur in regions of little signal energy. The continuous-time ratio spectrum has been realized in analog hardware [1, 2, 3], while the discrete-time version has been successfully applied to speech recognition [2], speech/audio coding [4] and implemented in real-time DSP hardware [5]. This paper demonstrates the use of ratio spectrum feature sets with an investigation of phoneme recognition. To understand the source of the ratio spectrum’s power as a feature extractor, we described quantitatively the statistical properties of the ratio spectrum on all white Gaussian noise (AWGN).

2. Feature Extraction From Phonemes

Phoneme recognition is useful for word recognition in modern automatic speech recognition systems, and it offers a quantitative means of comparing a ratio spectrum-based feature set with sets from other feature extraction techniques. For this investigation, we chose the popular TIMIT database for recognition of American English phonemes. This database was designed for independent automatic speech recognition and includes hand labeling of phoneme end points. Two popular feature sets used today in automatic speech recognition are mel frequency cepstral coefficients (mfcc) and perceptual linear prediction (plp) coefficients, and they are included for comparison in this phoneme recognition investigation. For this test, we made two simplifications: discrete-time phoneme samples were not divided into overlapping windows—each phoneme was represented by one window taken from the middle of the phoneme sample; training was performed over a subset of the TIMIT database, selecting 620 test and 1835 train phonemes over a set of 8 vowels. The promising results that following warrant a more in-depth experiment.

A detailed description of mfccs and plp coefficients may be found in [8]. For ratio spectrum features, three extraction techniques were included: 1) Use standard LP analysis to find an envelope function for the phoneme power spectrum, then compute the ratio spectrum from this envelope (RS LP-15); 2) Construct the ratio spectrum using a 30th-order low-pass Butterworth filter (RS LPF-15); 3) Construct the ratio spectrum using an ideal low-pass filter (no smoothing of pitch-related signal information) (RS ideal-15). The number following each feature set name denotes the number of features in the set. The following table summarizes the phoneme recognition results from this investigation.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>% Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC-13</td>
<td>30.2</td>
</tr>
<tr>
<td>RS LP-15</td>
<td>30.7</td>
</tr>
<tr>
<td>RS LPF-15</td>
<td>30.8</td>
</tr>
<tr>
<td>PLP-13</td>
<td>33.6</td>
</tr>
<tr>
<td>RS ideal-15</td>
<td>39.1</td>
</tr>
</tbody>
</table>

The ratio spectrum feature sets are frequencies, originally from the linear frequency scale (by nature of the FFT algorithm). The sets were easily mel scaled after extraction and before classification. A nearest neighbor algorithm using Euclidean distance was used to classify the phonemes. It is evident from the table that two of the feature sets based on the ratio spectrum perform as well as feature sets based on mfccs or plp coefficients. Both of these sets come from smoothed power spectra with suppressed pitch information—the LP transfer function is naturally smooth, and the Butterworth low-pass filter used in RS LPF-15 blurs the pitch-related characteristics of the speech power spectrum. Both the mfcc and plp algorithms integrate the power spectrum over critical bands which also reduces the effects of pitch on feature set values. Recognition error is higher when the power spectrum is not smoothed prior to feature extraction, as seen in RS ideal-15. Note that RS LPF-15 makes no assumptions about the signal, aside from pre-emphasis.

3. The Ratio Spectrum And AWGN–A Statistical Approach

The competitive results for ratio spectrum feature sets are due to the adaptive nature of the ratio spectrum—signal energy distributed over a spectrum changes over time, features extracted using the ratio spectrum move with the signal energy. Without adapting, the
Figure 1: (a) Typical $|X(k)|^2$ function with feature sets from uniform sampling of the ratio axis and frequency axis. (b) $\Re(k)$ from (a) showing acquisition of each feature set.

frequencies present in the feature set would reside in regions of less energy. To quantify this energy gap, we consider the average signal amplitude of AWGN at extracted frequencies before and after applying the ratio spectrum.

AWGN is a random signal with Gaussian probability density whose samples are independent and uncorrelated (strictly white noise). These properties make it a viable signal for statistical study and manipulation using the ratio spectrum. Consider an AWGN input sequence $x(n)$ with zero mean and variance $\sigma^2$. Its DFT $X(k)$ of length $N$ has a real and imaginary part

$$\Re \{X(k)\} = \sum_{n=0}^{N-1} x(n) \cos \left( \frac{2\pi}{N} kn \right) = a(k)$$

$$\Im \{X(k)\} = -\sum_{n=0}^{N-1} x(n) \sin \left( \frac{2\pi}{N} kn \right) = b(k)$$

Since $a(k)$ and $b(k)$ are linear combinations of Gaussian i.i.d. random variables, $a(k)$ and $b(k)$ have normal densities $p(a)$ and $p(b)$ with the following properties:

$$p(a) = G(0, N\sigma^2), \quad p(b) = G(0, N\sigma^2/2)$$

where $G(\eta, \sigma^2)$ is a Gaussian density of mean $\eta$ and variance $\sigma^2$. The dependence of $p(a)$ and $p(b)$ on $k$ arises from the harmonic characteristic of $X(k)$: $b(k)$ vanishes for $k = 0, N/2$. To simplify the analysis, we assume that these densities $p(a)$ and $p(b)$ for $k = 0, N/2$ are identical to those for $k \neq 0, N/2$. These probability densities involve only two points out of an $N$-length DFT and the numerical results following show this to be a valid approximation. By applying standard statistical techniques (for an excellent reference on this topic, see [9]), the relevant density functions are derived and listed in the following table:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Probability density</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x(n)$</td>
<td>$p(x) = G(0, \sigma^2)$</td>
</tr>
<tr>
<td>$a(k) = \Re {X(k)}$</td>
<td>$p(a) = G(0, N\sigma^2/2)$</td>
</tr>
<tr>
<td>$b(k) = \Im {X(k)}$</td>
<td>$p(b) = G(0, N\sigma^2/2)$</td>
</tr>
<tr>
<td>$y(k) = a^2(k)$</td>
<td>$p(y) = \frac{1}{\sqrt{\pi N\sigma^2}} e^{-y/N\sigma^2} U(y)$</td>
</tr>
<tr>
<td>$z(k) = b^2(k)$</td>
<td>$p(z) = \frac{1}{\sqrt{\pi N\sigma^2}} e^{-z/N\sigma^2} U(z)$</td>
</tr>
<tr>
<td>$S(k) = y(k) + z(k)$</td>
<td>$p(S) = \frac{1}{\sqrt{\pi N\sigma^2}} e^{-S/N\sigma^2} U(S)$</td>
</tr>
<tr>
<td>$c(k) = \sqrt{S(k)}$</td>
<td>$p(c) = \frac{1}{2\sqrt{\pi N\sigma^2}} e^{-c^2/2N\sigma^2} U(c)$</td>
</tr>
</tbody>
</table>

where $U$ is the unit step function. Using the density functions describing the AWGN spectrum, the expected values of the power spectrum and magnitude spectrum are

$$E\{S(k)\} = E\{|X(k)|^2\} = N\sigma^2 \equiv \eta_S$$

$$E\{c(k)\} = E\{|X(k)|\} = \sqrt{\frac{\pi}{4} N\sigma^2} \equiv \eta_c$$

We next investigate the ratio spectrum of AWGN. Assume $H_m(k)$ in Equation 1 is an ideal low-pass filter. Consider the plot of $\Re(k)$ in Figure 2. For a point $t$ chosen on the ratio axis, the probability of selecting a sample $S(k)$ is proportional to the amplitude of $S(k)$ as well as $p(S)$. A large $S(k)$ is less likely to be selected.
Figure 2: The discrete ratio spectrum of $X(k)$. The probability of a point $t$ projecting onto a step height $S(k)$ is proportional to the step height $|S(k)|$ as well as to the density of that step height $p(S)$.

if it is less likely to occur in the distribution of $S(k)$. The same holds true when the ratio spectrum is based on magnitude. From Figure 2, we conclude that the probability densities of these samples are

$$p_S(S) = A_S S p(S) \quad \text{where} \quad A_S = \frac{1}{\eta_S}$$

$$p_S(c) = A_c c p(c) \quad \text{where} \quad A_c = \frac{1}{\eta_c} \quad (5)$$

This trial of selecting samples is repeated many times, and we maintain that the mean of the selected samples is

$$E\{R_2(k)\} = \int_{-\infty}^{\infty} S p_S(S) dS = \frac{1}{\eta_S} \int_{-\infty}^{\infty} S^2 p(S) dS$$

$$E\{R_1(k)\} = \int_{-\infty}^{\infty} c p_S(c) dc = \frac{1}{\eta_c} \int_{-\infty}^{\infty} c^2 p(c) dc \quad (6)$$

For AWGN, we know $p(c)$ and $p(S)$, and we calculate the expected values for $R_2(k)$ and $R_1(k)$ as

$$E\{R_2(k)\} = 2N \sigma^2 = 2 \eta_S$$

$$E\{R_1(k)\} = \sqrt{\frac{4}{\pi} N \sigma^2} = \frac{4}{\pi} \eta_c \quad (7)$$

To verify these theoretical results, a Matlab simulation was performed to generate AWGN and find the means of the magnitude and power spectra before and after applying the ratio spectrum. An AWGN signal of zero mean, unity variance, and length 1024 samples was generated $10^9$ times, and each magnitude and power spectrum was randomly sampled 100 times before and after applying the ratio spectrum. The following table summarizes the results:

<table>
<thead>
<tr>
<th></th>
<th>Theory</th>
<th>Simulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean($\eta_1(k)$)</td>
<td>36.108</td>
<td>36.103</td>
</tr>
<tr>
<td>mean(S)</td>
<td>1024</td>
<td>1024.7</td>
</tr>
<tr>
<td>mean($\eta_2(k)$)</td>
<td>2048</td>
<td>2045.8</td>
</tr>
</tbody>
</table>

The simulation confirms the theoretical expressions and shows quantitatively that the averages of the extracted features using the ratio spectrum as compared to uniform sampling of the frequency axis are 27% larger for the magnitude spectrum and 100% larger for the power spectrum.

4. Discussion

We have quantitatively shown that the ratio spectrum tends to extract features from regions of high energy. Thus, the ratio spectrum adapts elegantly to temporal changes in the spectrum of a signal. We used the ratio spectrum to extract features for phoneme recognition, and we quantified the ratio spectrum’s power as a feature extractor by analyzing the ratio spectrum of AWGN. With phoneme recognition, error rate reduces when pitch-related information is suppressed in the speech spectrum. This is evident by observing the decreasing trend in error rate for increasing smoothness of the original magnitude spectrum. The problem of smoothing the original magnitude spectrum has not been fully investigated in this study. The promising phoneme recognition results suggest further development.

References


