Neural “Spike rate spectrum” as a noise robust, speaker invariant feature for automatic speech recognition

Kiran G.V., Krishna A.G., Sreenivas T.V.

Speech and Audio Group
Department of Electrical Communication Engineering
Indian Institute of Science
Bangalore, India
{kirangv,krishna}@protocol.ece.iisc.ernet.in, tvsree@ece.iisc.ernet.in

Abstract

A new feature set for ASR called Rate-Spectrum (RS) is proposed. RS is a spectral representation obtained using a computational auditory model. The feature is noise-robust and considerably speaker invariant. RS matches the smoothed log spectrum both in shape and dynamic range variation. DCT is used to reduce dimensionality. Comparison of the proposed features with MFCC is done using an Isolated word recognition experiment on the TIDigits database, for clean and noisy speech cases. For speakers seen during training, RS and RS-DCT outperform MFCC in noisy case while matching that of MFCC in the clean case. For unseen speakers, RS does better than MFCC in the clean case, RS-DCT outperforms MFCC in the noisy case. We have thus shown that the proposed feature for ASR is noise robust and speaker invariant.

1. Introduction

Traditionally all automatic speech recognition systems, be it simple isolated digit recognition or a large vocabulary continuous speech recognition system, are divided into front-end and back-end components. While all the statistical modeling and classification is handled by the back-end the feature extraction is handled by the front-end system. Although there are many issues wherein the back-end design influences the front-end implementation, feature extraction has remained essentially an independent front-end function. Although parallels to the human auditory system in terms of perception, physiology and the issues of noise robustness have provided new direction to the search for new features, there isn’t as yet an optimal or perfect feature set.

The human auditory system is undoubtedly the ultimate speech recognition system given its ability to generalise, adapt and provide noise immunity in all kinds of environment. Given this, it is but natural that we seek to mimic the functionality of the auditory system to as close a level as possible. Although not much is still known about the different aspects of formation of a percept in the auditory system, a lot of work has been carried out on the peripheral auditory system which includes the action of the basilar membrane, the signal transduction and the subsequent firing of the auditory neurons due to the bending of the inner hair cells and the transmission of the neural spikes into the higher levels.

Auditory models are essentially signal processing or mathematical models of the physiological processes in the auditory system. These are developed based on either physiological measurements or hypotheses based perception experiments. The speech recognition performance of the human auditory system is enough motivation for using these systems for feature extraction in speech recognition systems. MFCC is a good example of such an approach leading to its acceptance as the best feature set at present. Although most of the auditory models proposed in the literature have been used for extensive modeling of the neural firing, only recently many models have been used as front ends for speech recognition systems including its performance in noise and other environments[1] [2]. Also less extensive yet reasonably accurate auditory models based again on the firing patterns in the auditory nerve have also been used for speech recognition[3].

In this paper we report the use of a spectral representation derived from a similar auditory model. We show that not only is the representation accurate but when used as a feature it outperforms MFCC in a noisy environment and also for the case of unseen speakers. Section 2 describes the auditory model in detail and explains the method of obtaining the rate-spectrum; section 3 discusses the effectiveness of the rate spectrum in modeling the signal. Section 4 describes the new feature in an Isolated word recognition system along with comparisons to an MFCC based front end.

2. Auditory model

A block diagram of the auditory model is shown in Fig.1. The auditory model is composed of a first stage filterbank which consists of spectrally overlapping second order resonant filters followed by a halfwave rectifier and a second stage that consists of a sequence of level crossing detectors followed by an average firing rate computing block. This average firing rate as a function of the subband channel number gives the rate-spectrum. With an overlapping frame by frame analysis we get a sequence of these rate-spectra which are then used as a feature for the isolated word recognition system.

2.1. Filterbank

The filtering action of the basilar membrane which is highly frequency tuned is simulated by the first stage filterbank. The human auditory system has about 3000 auditory nerves innervating the basilar membrane and each of these auditory nerves, by virtue of their position along the BM, are tuned to a different frequency and also have different bandwidths. Their frequency responses are highly overlapping amongst adjacent filters and also these filters are more closely spaced near the low frequency regions while being more farther spaced in the high frequency
regions. For a given number of filters and sampling frequency of 16KHz, the \( f_c \) or the center frequencies(CF) for the filters and the effective rectangular bandwidth (ERB) are calculated as [7]\[4],

\[
f_c(n) = -228.8 + e^{-n(0.02)} 8228.8 \quad (1)
\]
\[
ERB(n) = 24.7 \left( \frac{f_c(n)}{1000} + 1 \right) \quad (2)
\]

for \( n = 1 \ldots N \) (number of filters). Given the CF and the bandwidth of the filters, each filter is implemented as a simple second order resonant filter whose frequency response is given by

\[
H(z) = \frac{b_0}{(1 - r e^{j\omega_0} z^{-1})(1 - r e^{-j\omega_0} z^{-1})} \quad (3)
\]

where \( r \) is the magnitude of the pole and \( \omega_0 \) is its angular frequency. The parameters \( \omega_0 \) and \( r \) can be designed for a given resonant frequency \( f_c = \omega_0 \) and bandwidth \( \Delta \omega \) as

\[
\Delta \omega \approx 2(1 - r) \quad (4)
\]
\[
\omega_0 = \cos^{-1} \left[ \frac{\cos(\omega_0) - 2r}{1 + r^2} \right] \quad (5)
\]

for values of \( r \) close to 1. As emphasis is more on a new feature set for speech recognition rather than a detailed computational auditory model, we have resorted to the use of simple second order resonant filters instead of the physiologically more accurate gammatone or the more recent gammachirp filters[6].

2.2. Rate Spectrum

The rate-spectrum is determined based on the firing rate information which is obtained through a levelcrossing analysis. In the auditory nerve fibers whose center frequencies are below 1kHz it is seen that the nerve firings exhibit phase locking to the stimulus. Since at low frequencies, the bandwidth of the auditory filters are small, the bandpass filtered output can be treated as a sinusoid (carrier) with an amplitude modulation.

The neural firing is nearly synchronous with the positive cycle of the sinusoid and the rate reaches a maximum during the maximum of the sinusoid and then the firing rate decreases. The same behaviour is seen on repetitive cycles and this behaviour is called “phase locking”. Interestingly, beyond 1kHz the physical properties of the auditory nerve fibers do not show such a phaselocking.

Most of the auditory model based representations reported have relied explicitly on the timing information (periodicity of the neural spikes) in the firing pattern obtained in each channel. Such a processing strategy is a consequence of the assumption that the auditory system is extracting periodicity and other time related information from the firing in each channel. But, if we take into account that no “phase lock” is seen in the auditory nerve fibers above 1kHz and still the information present in these frequencies are being used by the auditory system we have to frpmf more on the rate-place theory.

2.2.1. The rate-place theory

The rate-place theory proposes that the frequency information is by virtue of the position of the fiber along the BM and it is only the rate of firing in each channel that conveys the magnitude information (strength of the frequency component). In this context, it is interesting to note that a similar and maybe more meaningful spectral representation, to that obtained using the timing information, can be obtained by using the rate information; i.e. the accumulated number of neural spikes per unit time. Since the firing rate information is related to the energy of the sub-band signal, a meaningful representation can be expected by aggregating the rate information from different fibres.

Since each auditory nerve synapses with the inner hair cell in a particular part of the BM, the position of each AN is frequency coded. That is, by virtue of its position of synapse along the BM, the AN has in its firing, information corresponding to a particular frequency and a certain bandwidth about it. Drawing an analogy with the DFT of a signal, the position of an AN can be thought of as a frequency point (the abscissa) in the DFT. Here, the strength of the frequency point is obtained through the rate information, in otherwords the average neural firing rate for that fiber is the strength of the contribution at that frequency (CF of the nerve fiber).

2.2.2. Computing the rate spectrum

The average firing rate information is obtained through a series of levelcrossings detectors. A given number of level detectors are distributed uniformly along the hearing range of 0 to 120dB (on the log scale). Assuming a generic AM model for the subband signal \( x(t) \) we have

\[
x(t) = A(t) \sin(2\pi f_c t) \quad (6)
\]

where \( A(t) \) carries the amplitude modulation information. Everytime the subband filtered waveform in a channel crosses a particular level \( l \), the time instant is marked and a firing is generated as \( \delta(t_k) \) as shown in Fig. 2.

\[
x(t_k) = A(t_k) \sin(2\pi f_c t_k) = \ell \quad \text{for } \ell = 1 \ldots L \quad (7)
\]

Only the positive going level crossings are used and \( L \) is the total number of levels. In doing so a series of neural spikes or a
spike train is obtained.

\[ S(t) = \sum_{k=0}^{K-1} \delta(t - t_{k}) \]

(8)

\( K \) is the total number of spikes generated in a fixed interval, say \( T_{w} \). To obtain the average firing rate the total number of neural spikes is integrated or added up across time.

\[ T = \frac{1}{T} \sum_{\ell=0}^{T_{w} - 1} K_{\ell} \]

(9)

\( T \) which is the total number of spikes in a given channel is a function of the channel number and is referred to as \( T(c) \) where \( c \) is the channel index. Since the time interval over which this integration is done is kept the same across all channels \( T(c) \) can in itself be the average firing rate.

2.2.3. Normalisation by the center frequency

Although the firing rate is a function of the energy of the signal in that sub-band, it is also an implicit function of the center frequency (CF) of the channel. This is because the inherently higher periodicities of the high frequency channels will generate higher number of neural firings per given time although this doesn’t necessarily imply that the signal might have crossed more levels. In order to make the average firing rate independent of the CF, it is normalised by the CF of the respective channel.\(^1\)

\[ T_{\text{norm}}(c) = \frac{T(c)}{f_{c}} \quad \text{where} \quad c = 1, \ldots, N \]

(10)

where \( N \) is the number of channels. The rate spectrum is then \( RS = T_{\text{norm}}(c) \) for \( c = 1, \ldots, N \).

3. Signal modeling

Interestingly, the rate-spectrum matches the smoothed log spectrum of the signal in shape and in the dynamic range variations of the spectral peaks as shown in Fig. 3. A synthetic pre-emphasised vowel with closely spaced first and second formant frequencies of 800 Hz and 1100 Hz and a third formant at 2300 Hz was used to test the signal modeling capability of the rate-spectrum. A 10th order LP fit cannot resolve the closely spaced formants while Fig. 3(a) shows the 12th order and 14th order LP spectral fit. Interestingly, the rate-spectrum clearly resolves the two formants as in Fig 3(c) and also, interms of the average firing rate, the dynamic variations in the formant magnitudes match that of the LP spectra. Thus it can be seen that the rate-spectrum is similar to an LP model spectrum of an optimum order. Fig 3(b) and 3(d) show the 10th order LP spectrum and the rate-spectrum of a synthetic pre-emphasised vowel with formant frequencies at 500 Hz, 1700 Hz and 2700 Hz. The fact that the rate-spectrum is a smoothed log spectrum of the signal makes it an attractive feature for speech recognition.

4. Experiments with an IWR system

The auditory model has a 160 channel filterbank and in each channel about 1000 levels are uniformly distributed on the log scale (0 to 120dB). The rate-spectrum is computed by using a overlapping frame scheme on each of the subband outputs. A frame length of 20msec is used to compute the firing rate and the frames shifted by 10msec. Thus a rate-spectrum is computed every 10msec and this forms a feature vector giving a feature vector rate of 100 per second. These feature vectors are then input to a HMM based isolated word recognition (IWR) system.

A HMM based IWR system is used to experiment with the rate-spectrum based feature and to compare it with the standard MFCC features. The database used was the TI-20 alpha digits database\(^5\) which has 8 male speakers and 8 female speakers. Only the words corresponding to the 10 digits are used for implementing and testing the system. Each word is modeled as a 3 state HMM with 5 mixture Gaussian emission densities in each state. To verify the effectiveness of the new feature and to test the claim of speaker invariance, the experiments are done in two modes namely the “seen” speaker mode and the “unseen” speaker mode. The features used in the experiment are 24 dimension MFCC, 160 point rate-spectrum (RS) and 24 point DCT on rate-spectrum (RS-DCT). The HMMs are trained on clean speech feature vectors and tested on either clean speech features or noisy speech features. The experiments are done on both noisy and the clean case. In the noisy case, white gaussian noise at 0dB, 5dB and 10dB is added on the signal and features are obtained from the noisy signal.

4.1. Seen speaker mode

In this mode, the HMMs are trained on 8 male speakers \( M_{1} \) to \( M_{8} \) with 10 repetitions per word and tested on the unseen data from the same speakers \( M_{1} \) to \( M_{8} \) with 16 repetitions per word. The testing is done on both the clean and noisy case and the results are as shown in Table 1.

4.2. Unseen speaker mode

In this mode, the HMMs are trained on 5 male speakers \( M_{1} \) to \( M_{5} \) and tested on 3 unseen speakers \( M_{6} \) to \( M_{8} \). The words used...
are digits from 0 through 9. The results are shown in Table 2. Also an experiment was conducted across genders with training on male speakers $M_1$ to $M_8$ and testing on unseen female speakers $F_1$ to $F_8$. In this case only the clean data was used and the results are shown in Table 3.

Table 2: Recognition results for the unseen speaker case

<table>
<thead>
<tr>
<th>Feature</th>
<th>Clean(%)</th>
<th>10dB(%)</th>
<th>5dB(%)</th>
<th>0dB(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>96.75</td>
<td>86.25</td>
<td>80.25</td>
<td>68.00</td>
</tr>
<tr>
<td>RS</td>
<td>92.38</td>
<td>88.25</td>
<td>82.25</td>
<td>70.00</td>
</tr>
<tr>
<td>RS-DCT</td>
<td>90.88</td>
<td>84.75</td>
<td>78.75</td>
<td>66.75</td>
</tr>
</tbody>
</table>

Table 3: Results for unseen speaker across gender

<table>
<thead>
<tr>
<th>Feature</th>
<th>Clean (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>38.625</td>
</tr>
<tr>
<td>RS</td>
<td>60.5</td>
</tr>
<tr>
<td>RS-DCT</td>
<td>47.375</td>
</tr>
</tbody>
</table>

5. Discussion

Some very interesting observations can be made from the results of the above experiments. The rate-spectrum (RS) feature and the DCT of the rate-spectrum (RS-DCT) perform nearly as well as MFCC in the clean, seen speaker case. But the RS-DCT feature perform remarkably well in the noisy case and outperform MFCC by a huge margin and in most of the cases the performance is twice as good! In the unseen speaker case both RS and RS-DCT outperform MFCC in the clean case while in the noisy case the RS-DCT does better than MFCC by more than a factor of two! However, the RS performance also degrades similarly as that of MFCC in the noisy case.

While the rate-spectrum feature is seen to do well in the unseen speaker case the reduction in the dimension of the rate spectrum by applying a DCT is seen to be making it robust to noise. Thus, we can conclude from the above experiments that the RS-DCT feature not only improves robustness to noise but also brings with it the speaker invariant properties of the rate-spectrum. Thus the RS-DCT is a truly a “noise robust” and “speaker invariant” feature. The similarity of the LPC log spectrum and the rate-spectrum (obtained log amplitude levels) leads us to interpret RS-DCT as similar to a cepstrum, and hence may be referred to as rate-cepstrum.

5.1. Speaker invariance and noise robustness

A possible explanation for speaker invariance is that the rate-spectrum is similar to the smoothed log spectrum without the periodic excitation that is normally present in the voiced spectrum i.e. the deconvolved filter information only from the source-filter theory. Also, during the computation of the average firing rate only the integrated firing information is used and no timing information is retained. While in the computation of the MFCC the Mel filter bank is applied on the DFT spectrum which has all the periodic information pertaining to the excitation. Since the excitation (the pitch) is one of the main source of information for speaker identification, the rate-spectrum achieves speaker invariance or normalisation by discarding most of the speaker dependent information.

Also the filterbank used in the rate-spectrum computation has more number of filters and smaller bandwidths. This gives it the advantage of lesser white noise because less noise is included in the narrower filters and the subband SNR is more likely to be higher than the Mel filters of the MFCC. This could be a reason for the RS-DCT to outperform in the noisy case.

6. Conclusions

We have shown the use of an auditory model motivated rate-spectrum representation as a feature in a HMM based speech recognition task. We have shown that a cepstrum-like representation, rate-cepstrum outperforms the MFCC by more than a factor of two in the unseen speaker case for noisy speech. We see that rate-cepstrum captures speaker invariance as well as noise robustness. However, the results have to be tested on a larger scale speech recognition task.

The authors would like to thank S. Chandra Sekhar for the motivating discussions on the issue of level crossings and Sup Ldr P. Srinivas for his support and encouragement in the completion of this work.

7. References