Auditory Features Revisited for Robust Speech Recognition

Finnian Kelly, Naomi Harte
Dept. of Electronic and Electrical Engineering
Sigmedia Group, Trinity College Dublin, Ireland
{kellyfp, nharte}@tcd.ie

Abstract

Auditory based front-ends for speech recognition have been compared before, but this paper focuses on two of the most promising algorithms for noise robustness in automatic speech recognition (ASR). The feature sets are Zero-Crossings with Peak Amplitudes (ZCPA) and the recently introduced Power-Law Nonlinearity and Power-Bias Subtraction (PNCC). Standard Mel-Frequency Cepstral Coefficients (MFCC) are also tested for reference. The performance of all features is reported on the TIMIT database using a HMM-based recogniser. It is found that the PNCC features outperform MFCC in clean conditions and are most robust to noise. ZCPA performance is shown to vary widely with filterbank configuration and frame length. The ZCPA performance is poor in clean conditions but is the least affected by white noise. PNCC is shown to be the most promising new feature set for robust ASR in recent years.

1 Introduction

The task of feature extraction in speech recognition is to obtain a compact representation of a speech signal that compresses the useful information into a small number of measures or coefficients. The information held by the coefficients must be sufficient to allow different elements of speech to be distinguished from one another. Typically this information is about the distribution of energy in the different frequency bands of the signal and how these vary with time. Conventional features such as Mel Frequency Cepstral Coefficients (MFCC) perform this task effectively in ideal operating conditions. However, it is well established that their performance degrades severely when there is a mismatch between the training and testing conditions, typically due to background noise [7]. Humans have an impressive ability to recognise speech even in the most adverse environmental conditions. Thus an approach to achieving robustness in ASR is to use an understanding of human speech processing in feature extraction. Such features, which can be based on physiological or perceptual aspects of human speech processing are referred to as auditory features.

In this paper, the performance of an established auditory feature type, Zero-Crossings with Peak Amplitudes (ZCPA), with several filterbank configurations, along with a recently developed one, Power-Law Nonlinearity and Power-Bias Subtraction (PNCC), are evaluated in clean and noisy conditions. The performance of standard MFCC features are included for comparison. Results are reported on the TIMIT database.

2 ZCPA Features

ZCPA features were first proposed by Kim [7] as a adaptation of the Ensemble Interval Histogram model [4]. The idea is to model the neural firing patterns of the human cochlea. In the proposed model, the speech signal is filtered with a set of auditory filters, then the output of each filter is passed through a zero-crossing detector. The distance between adjacent upward going zero-crossings is used to give a frequency estimate. The resulting frequencies are collected in a histogram, with the weight of each histogram entry being given by a non-linear compression of the peak amplitude between the zero-crossings. The histograms across all filter channels are then summed to produce the feature vector. A schematic for the algorithm is shown in Figure 1.

The auditory filterbank aims to simulate the frequency selectivity behaviour of the human cochlea. It comprises of multiple channels with bandwidth and spacing determined by some non-linear scale. In this paper three different auditory filters were evaluated. A carefully designed cochlear filterbank replicating the basilar membrane response was presented by Seneff in [8], as part of her synchrony / mean-rate model of
speech processing. The evaluated filterbank had 36 channels with a bandwidth of approximately 0.5 Bark and centre frequencies from 130 to 3400kHz. An implementation of Seneff’s model in Matlab included as part of an ‘Auditory Toolbox’ for Matlab by Slaney [9] was used in this evaluation. The frequency response is shown in Figure 2.

An Equivalent Rectangular Bandwidth (ERB) filterbank can be viewed as a cochlear filterbank providing a more simple modelling of the basilar membrane response than Seneff’s filterbank. Slaney’s Auditory Toolbox includes an ERB filterbank implementation which was used in this evaluation. The filterbank evaluated had 16 filters with centre frequencies ranging from 200Hz to 3400Hz (see Figure 3). The bandwidth and spacing of adjacent channels is equal on the ERB scale.

Despite the fact that cochlear filters are designed to replicate a physiological response to speech, it has been reported that a filterbank of FIR filters can exceed the performance of the cochlear filters in a ZCPA implementation (Kim [7]). Subsequent implementations [2, 1] use FIR filterbanks exclusively. A filterbank was designed with 16 Hamming FIR filters of order 61. The centre frequencies of the filters were spread evenly on the Bark scale from 200Hz to 3400Hz (see Figure 4). Filter bandwidths were equal to 2 bark - reported to be an optimal spacing by Gajic [2].

3 PNCC Features

This a recent feature extraction algorithm introduced in [6]. It can be seen as a variant on MFCC feature extraction with different stages of the conventional algorithm replaced with auditory motivated elements. A schematic of its operation is shown in fig 5. Firstly, the triangular filterbank used by MFCC is replaced with a gammatone filterbank. The novel aspects to the algorithm are the use of a Power Function Nonlinear-

The nonlinearity of the human auditory system has been well established, and the use of a nonlinear function in feature extraction methods is common. MFCCs pass the filter outputs through a log nonlinearity. PNCC adopts a power function which aims to better model peripheral nonlinearities than a log function. Taking a closer look at accurate auditory models [10], the graph relating decibels to auditory nerve firing rate is S-shaped. For decibels below a certain threshold, the firing rate is almost constant. Above this the increase in decibels with firing rate is almost linear, until it reaches a saturation point. If a log nonlinearity is adopted then there is no lower threshold. Thus small changes at a low power can result in large changes at the output of the log function. With power function however - when the input level is close to zero, so too is the output level. This is what is observed in the human auditory system. The target function then will be close to zero up to a threshold, and then increase linearly. Because the dynamic behaviour of the output does not depend critically on the input amplitude, this ideal piecewise-linear curve is approximated with an a MMSE-based best fit power function. The power nonlinearity is described by the equation

$$y = x^{\alpha_0}$$

The best-fit value of the exponent was found to be $\alpha_0 = 0.1$ by [6].

Medium-Duration Power Bias Removal subtracts a ‘bias’ from the speech segment that is assumed to represent an unknown level of background excitation. The adjusted power $\hat{P}(i, j)$ of the $i$th channel and $j$th frame is given by

$$\hat{P}(i, j) = \left(1 + \frac{\min(i+N,C)}{2N+1} \sum_{i'=\max(i-N,1)}^{\min(i+N,C)} w(i', j) \right) P(i, j)$$

Where $P(i, j)$ is the original power of the frame, $w(i', j)$ is the power normalization gain given by the
ratio of the normalized power to the average power of a frame. The normalized power is found from the power bias, which can be defined as the smallest power which makes the arithmetic mean to geometric mean ratio of the segment the same as that of clean speech. For smoothing purposes, $w(i', j)$ is averaged across a range of channels specified by $N$. $C$ is the total number of channels. Full details are given in [6].

4 Performance Evaluation

The test corpus used to evaluate the performance of the feature types was the widely used TIMIT database [3]. The HMM Toolkit (HTK) was used to create the recognition engine. The TIMIT database was divided into training and testing subsets as recommended in the documentation. The size of the ZCPA and PNCC frames was 50ms while a 25ms frame length was used for MFCC. The step size in both cases was 10ms. Each vector contained 36 coefficients: 12 static; 12 first-order dynamic; and 12 second-order dynamic coefficients. Cepstral Mean Normalisation was applied in each case. This resulted in 5 features sets - MFCC, PNCC and ZCPA with three filterbanks denoted Seneff, ERB and FIR. Context-independent Monophone HMMs with 8 Gaussian Mixtures were trained using features extracted from the training set. The test utterances were passed to the trained HMMs and a phoneme level transcription was generated. The procedure was carried out for clean speech (the original TIMIT recordings) and for noisy speech (White Gaussian noise added to the recordings at SNRs of 10dB and 0dB). No noise was added to the training material so that a mismatch between training and testing conditions would be simulated.

4.1 Experimental Results

The measure of performance used is phoneme-level accuracy, as defined by HTK in [11]. Figure 6 shows the performance of each feature set for clean speech and SNRs of 10dB and 0dB where FIR, ERB and Sen-eff denote ZCPA features with an FIR, ERB and Seneff cochlear filterbank respectively.

It is clear that MFCC performance degrades significantly in noise, with accuracy decreasing by 45% in 10dB white noise and by 74% in 0dB white noise. PNCC is the most effective of the front-ends, with the highest accuracy in clean conditions and in both levels of noise. Accuracy drops by 47% from clean to 0dB white noise. All ZCPA front-ends perform similarly, with Seneff’s cochlear filterbank being marginally the most accurate. They all perform poorly in clean speech, but show impressive robustness, with the FIR configuration dropping only 18% from clean to 0dB white noise.

To provide more detailed results, a phonetic breakdown of the performance was created. The 39 phonemes were divided into 5 groups, as proposed in [5], with an additional group for silences (sil). The 5 phoneme classifications are: vowels/semi-vowels (v/sv), nasals/flaps (n/f), strong fricatives (sf), weak fricatives (wf) and stops (st). The breakdown in performance is presented in Figures 7,8 and 9. It is notable that weak fricatives are recognised with consistently low accuracy ($\approx 40\%$) by all features in all conditions.
5 Conclusions

Auditory features are clearly more robust than conventional MFCC features in the presence of white noise. PNCC features are significantly more effective than any of the other features tested, with a high recognition accuracy in clean speech and relative robustness in noise. This supports Kim & Stern’s work [6] which showed PNCC to outperform MFCC and Perceptual Linear Prediction (PLP) feature sets on the DARPA Resource Management (RM1) database.

ZCPA perform poorly in clean speech but show a high level of robustness. Contradictory to previous suggestions, the FIR implementation did not provide superior performance to the cochlear filterbanks. The best ZCPA result was observed by the most complex filterbank - Seneff’s. The ZCPA model has many variable parameters. The scope of this paper covers a limited number of optimisations (using Gajic’s recommended values [2] as a starting point). However there are many further potential adaptations possible which may bring about increased performance - frame length, delta window, filterbank parameters, histogram parameters, number of coefficients etc. It must be questioned whether carrying out such exhaustive optimisations are justified given the superior performance and the less complex nature of the PNCC algorithm.

These tests were run with context-independent monophone HMMs. Further study should explore the relative improvements achieved by adopting context-dependent triphone models. Viewing the phonetic breakdown, the weak fricatives are responsible for a large part of the performance degradation. Further study should identify methods to capture these phonemes more reliably. Considering computational complexity, auditory features are more costly to extract in general than conventional features. PNCC is far more computationally efficient than ZCPA however. Overall, PNCC are a more promising development than ZCPA in achieving robust ASR performance.

References