A COMPARISON OF AUDITORY FEATURES FOR ROBUST SPEECH RECOGNITION

Finnian Kelly, Naomi Harte

Department of Electronic and Electrical Engineering,
Sigmedia Group,
Trinity College Dublin, Ireland
{kellyfp, nharte}@tcd.ie

ABSTRACT

This paper presents a detailed comparison of the performance of two auditory based feature extraction algorithms for 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 comparison. Although front-ends have been compared in previous papers, this work focuses on two of the most promising algorithms for noise robustness. The performance of all features is reported on the TIMIT database using a HMM system. It is found that the PNCC features outperform MFCC in clean conditions and are 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 typical speech recognition system consists of two main elements, a front end processor and a recognition engine. The front end processing is referred to as *feature extraction*. The task of feature extraction 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 [1]. Humans have an impressive ability to recognise speech even in the most adverse environmental conditions. Thus an approach to achieving robust 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 with the recognition engine provided by the HMM Tookit (HTK).

2. ZCPA FEATURES

ZCPA features were first proposed by Kim [1] as a adaptation of the Ensemble Interval Histogram model [2]. The motivation 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.

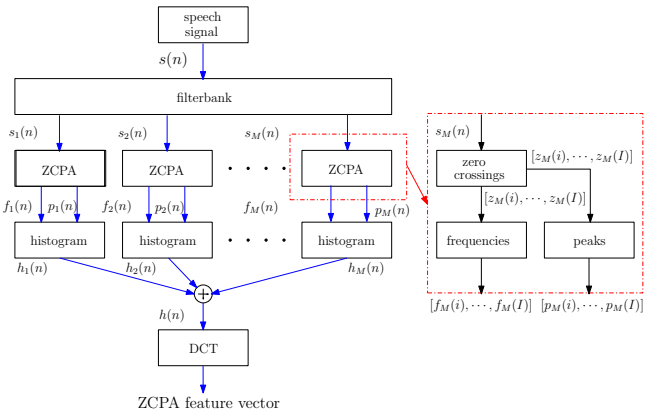

Figure 1: ZCPA extraction scheme

2.1 Auditory Filters

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.

2.1.1 Cochlear Filterbank

A carefully designed cochlear filterbank replicating the basilar membrane response was presented by Seneff in [3], 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 [4] was used in this evaluation. Its frequency response is shown in Figure 2.

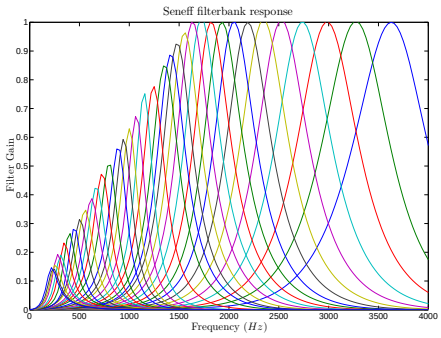

Figure 2: Seneff's Cochlear Filterbank Frequency Response

2.1.2 ERB Filterbank

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. The bandwidth and spacing of adjacent channels is equal on the ERB scale.

The frequency response of the filterbank is shown in Figure 3.

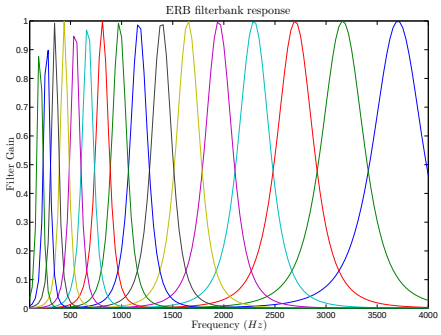

Figure 3: ERB Filterbank Frequency Response

2.1.3 FIR Filterbank

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 [1]. Subsequent implementations [5, 6] 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. Filter bandwidths were equal to 2 bark - reported to be an optimal spacing by Gajic [5]. The frequency response of this filterbank is shown in Figure 4

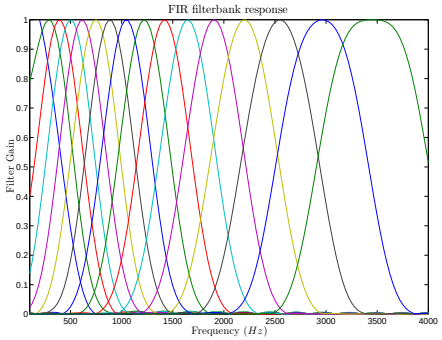

Figure 4: FIR filterbank

2.2 Zero-Crossing and Peak Amplitude Detector

An upward going zero-crossing is assumed to signal a neural firing event on the basilar membrane. The intervals between these events and the peak value across this interval is used to build a frequency histogram at the next stage. The output of the mth filter channel, $s_m(n)$, is passed into a zero-crossings detector. Each of I (upward-going) zero crossings $z_m(i)$ is passed to a peak detector and a frequency estimator. The peak $p_m(i)$ is calculated by:

$$p_m(i) = \max_{z_m(i) \le n < z_m(i+1)} \{s_m(n)\}$$

2.3 Feature Histogram

In the final stage, the frequency $f_m(i)$ between adjacent zero crossings is calculated by:

$$f_m(i) = \frac{1}{z_m(i+1) - z_m(i)}$$

The entry for the jth bin of histogram $h_m(n)$, where I_m is the number of zero crossings for the m_{th} filter output, is then given as:

$$\sum_{i=1}^{I_m - 1} \psi_j \{f_m(i)\}$$

$$where \;\; \psi_j \{f_m(i)\} = \begin{cases} log(p_m(i)+1) & \text{if } f_m(i) \in bin\, j \\ 0 & \text{otherwise} \end{cases}$$

The final histogram $h(n)$ is given as the sum of the corresponding entries in all sub-band histograms:

$$h(n) = \sum_{m=1}^{M} h_m(n)$$

The frequency computed from intervals between zero-crossings can be seen as corresponding to the point of excitation on the membrane. This is the dominant frequency in

that sub-band, and the peak value over its interval is an indication of its power. Thus the resulting histogram represents the dominant energies in the signal, which hold important phonetic information. The histogram was allocated 60 bins, equally spaced on the Bark scale from 0 Hz to 4kHz. A DCT of the final histogram is computed to decorrelate the features.

3. PNCC FEATURES

This a recent feature extraction algorithm introduced in [7]. It can be seen as a variant on MFCC feature extraction with different stages of the conventional algorithm replaced with auditory motivated elements. Firstly the triangular filterbank used by MFCC is replaced with a gammatone filterbank. The novel aspects of the algorithm are the use of a Power Function Nonlinearity (replacing MFCC's log nonlinearity) and the use of Medium-Duration Power Bias Subtraction to suppress the effects of background excitation. A schematic is given in Figure 5.

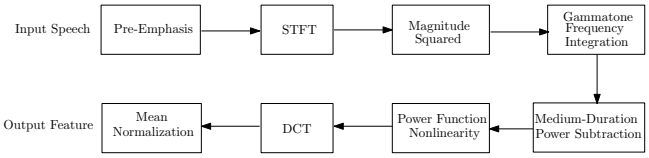

Figure 5: PNCC extraction scheme

3.1 Power Function Nonlinearity

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 [8], 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

$$y = x^{a_0}$$

The best-fit value of the exponent was found to be $a_0 = 0.1$ by [7].

3.2 Medium-Duration Power Bias Removal

This stage of the algorithm subtracts a 'bias' from the speech segment that is assumed to represent an unknown level of background excitation. The adjusted power $\tilde{P}(m,n)$ of the mth channel and nth frame is given by

$$\tilde{P}(m,n) = \left(\frac{1}{2m_r+1} \sum_{m'=max(m-m_r,1)}^{min(m+m_r,M)} w(m',n) \right) P(m,n)$$

Where $P(m,n)$ is the original power of the frame, $w(m',n)$ 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. Full details of the algorithm are given in [7]. For smoothing purposes, $w(m',n)$ is averaged across a range of channels specified by m_r. The value of m_r used was 5. The total number of Gammatone channels, M, was 40.

4. PERFORMANCE EVALUATION

4.1 Experimental Conditions

The test corpus used to evaluate the performance of the feature types was the widely used TIMIT database [9]. The HMM Toolkit (HTK) was used for creating 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 feature 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 testing 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.

4.2 Experimental Results

Recognition performance is measured by comparing the output transcriptions to reference transcriptions. The measure of performance used is phoneme-level accuracy, as defined by HTK in [10] and given by (1).

$$\%accuracy = \frac{L-D-S-I}{L} \times 100 \qquad (1)$$

where L is the total number of labels (phonemes) in the reference transcriptions, D is the number of *deletions*, S is the number of *substitutions* and I is the number of *insertions*.

Figure 6 shows the performance of each feature set for clean speech and SNRs of 10dB and 0dB.

It is clear that MFCC performance degrades significantly in noise, with its 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. Its accuracy drops

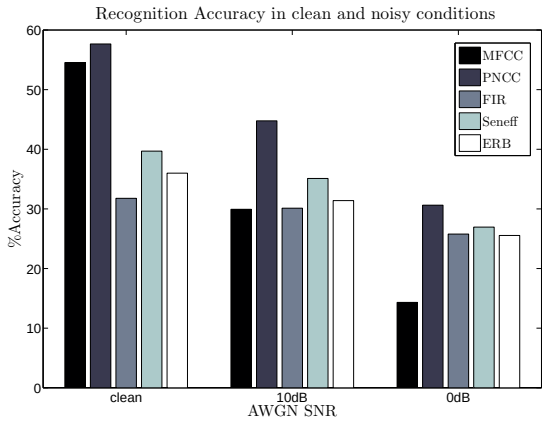

Figure 6: Recognition Accuracy of different features in different environments, where *FIR*, *ERB* and *Seneff* are ZCPA features with FIR, ZCPA and Seneff cochlear filterbank configurations respectively

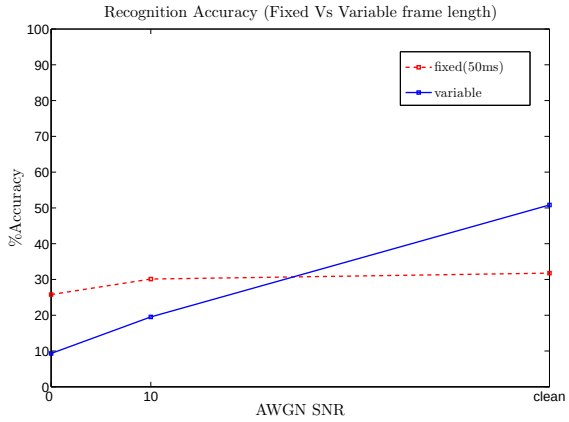

Figure 7: Accuracy of Fixed vs Variable frame lengths

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.

As a fixed frame length was used for the ZCPA implementation, each histogram entry was divided by its frequency (prior to the nonlinear compression) to prevent biasing towards higher frequencies. However, having a fixed frame length for each filter means that some accuracy may be lost - particularly at lower frequencies where there are fewer zero-crossings in the 50ms frame window. To investigate how a variable frame length would affect performance, this was implemented and tested with the FIR filterbank. The frame length L_j is given by (2).

$$L_j = \frac{40}{\sqrt{F_{cj}}} \tag{2}$$

where F_{cj} is the centre frequency of the j_{th} filter. The square root of this is taken to compress the longer frame lengths while keeping the shorter frame lengths from becoming unreasonably short. With the scaling factor of 40 and the given FIR centre frequencies, this resulted in frame lengths of between 43 and 177ms. The results of this adaptation are compared with the original FIR implementation in Figure 7. The variable frame length provides significantly higher accuracy in clean conditions but is less robust to noise as the fixed frame length.

To provide more detailed results, a phonetic breakdown of the performance was created. The 39 phonemes were divided into 5 groups, as proposed in [11], 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 classification is shown in Table 1.

The breakdown in performance is presented in Figures 8,9 and 10. It is interesting that weak fricatives are recognised with consistently low accuracy ($<\approx 40\%$) by all features in all conditions

grouping	phonemes
v/sv	iy ih eh ey ae aa aw ay ah oy ow uh uw er l r w y
n/f	m n ng dx
sf	jh ch z s sh
wf	hh v f dh th
st	b p d t g k
sil	sil

Table 1: Grouping of phonemes

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 [7] 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. Adopting a variable vs fixed frame length gave widely different results - both with advantageous trends - clearly this behaviour must be explored further. The ZCPA model has many variable parameters. The scope of this paper covers a limited number of optimisations (Using Gajic's suggestions [5] 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 (the current standard in HMMs for ASR).

Viewing the phonetic breakdown, the weak fricatives in particular appear to be responsible for a large part of the per-

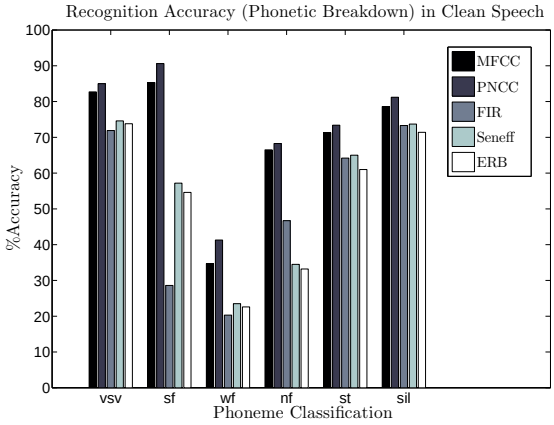

Figure 8: Clean

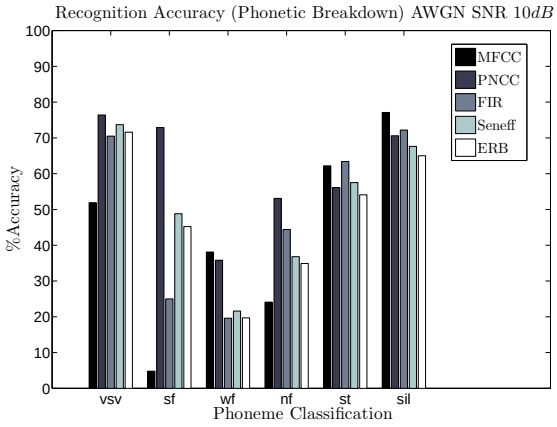

Figure 9: SNR 10dB

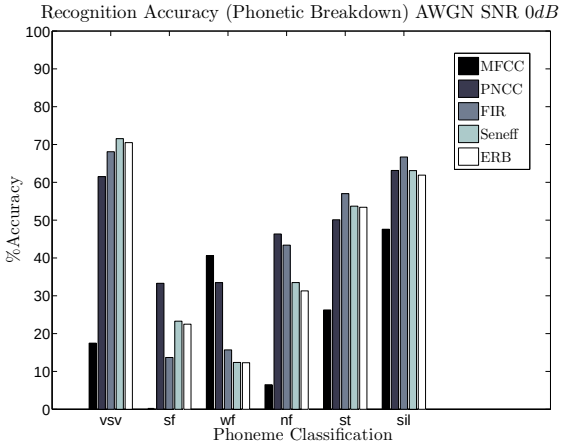

Figure 10: SNR 0dB

formance degradation. Further study should consider 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.

Based on all considerations - PNCC are a more promising development than ZCPA in achieving robust ASR.

REFERENCES

[1] D.-S. Kim, S.-Y. Lee, and R. Kil, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, pp. 55–69, Jan 1999.

[2] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, pp. 115–132, Jan 1994.

[3] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing," *Journal of Phonetics*, vol. 16, pp. 55–76, 1988.

[4] M. Slaney, "Auditory toolbox for matlab, version 2." Interval Research Corporation, 1998.

[5] B. Gajic and K. Paliwal, "Robust speech recognition using features based on zero crossings with peak amplitudes," *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 1, pp. I–64–7 vol.1, April 2003.

[6] O. Cheng, W. Abdulla, and Z. Salcic, "Performance evaluation of front-end algorithms for robust speech recognition," vol. 2, pp. 711–714, 28-31, 2005.

[7] C. Kim and R. M. Stern, "Feature extraction for robust speech recognition using a power-law nonlinearity and power-bias subtraction," *INTERSPEECH 2009*, Sept 2009.

[8] I. C. B. a. L. H. C. X. Zhang, M. G. Heinz, "A phenomenological model for the responses of auditory-nerve fibers: I. nonlinear tuning with compression and suppression," vol. 109, pp. 648–670, Feb 2001.

[9] J. Garofolo et al, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Linguistic Data Consortium, Philadelphia, 1993.

[10] S. Young et al, *The HTK Book (for HTK version 3.2.1)*. Cambridge University Engineering Department, 2002.

[11] A. Halberstadt and J. Glass, "Heterogeneous acoustic measurements for phonetic classification," tech. rep., Spoken Language Systems Group, Massachusetts Institute of Technology.