Zero-Crossings with Adaptation for Automatic Speech Recognition

Serajul Haque, Roberto Togneri, Anthony Zaknich

School of Electrical, Electronic and Computer Engineering
University of Western Australia, Australia
{serajul,roberto,tonko}@ee.uwa.edu.au

Abstract

An auditory model based on zero-crossings with peak amplitudes (ZCPA) was used as a front-end for automatic speech recognition (ASR) with the perceptual property of adaptation as determined by psychoacoustic observations. The model performance was evaluated on the isolated digits (TIDIGITS) database using continuous density HMM recognizer in additive noise. Experimental results indicate that the ASR performance of the ZCPA may be improved with adaptation over the static baseline performance in white Gaussian and factory noise. The perceptual front-end was also evaluated with dynamic (delta and delta-delta) features added to the adaptation. It was observed that adaptation with dynamic features performed better in factory, babble and car noise over a wide range of SNR values. The recognition performances were compared with the baseline MFCC. The performance of the dynamic ZCPA with adaptation was better than the dynamic MFCC in white Gaussian noise.

1. Introduction

The ability of the human auditory system to perceive speech under widely varying and adverse conditions has motivated researchers to include auditory-based feature extraction methods for speech processing and automatic speech recognition. For ASR, however, it is essential that the important speech-like features such as formant transitions, stops and fricatives, and other phonemic information are preserved in speech signals. It should also be robust under unexpected acoustic and channel conditions, speaker variability, coarticulation and adverse noise conditions. Popular parametrization for ASR such as MFCC and PLP employ auditory features like variable bandwidth filter banks and magnitude compression to simulate compressive nonlinearity. Computational auditory models simulate the transformation of the mechanical vibrations of the basilar membrane into neural representations through a series of nonlinear transformations, such as response saturation and rapid and short-term adaptation. The “instantaneous” discharge rate of auditory-nerve fibres is often significantly highest during the initial 15 ms of acoustic stimulation. It decreases thereafter, until it reaches a steady-state level approximately 50 ms after the signal onset. This decrease in response rate, referred to as “adaptation”, has been determined by psychoacoustic experiments as observed responses to pure tones (Smith & Zwislocki, 1975, Westerman & Smith, 1984)

Perceptual features including adaptation have been employed in ASR by several researchers with encouraging results. Holmberg, Gelbart & Hemmert (2006) incorporated a simplified model of synaptic adaptation into MFCC features as a competitive strategy to RASTA and the cepstral mean subtractuion (CMS). It showed improved ASR performance compared to baseline MFCC, MFCC processed by RASTA, and MFCC processed by CMS. Strope & Alwan (1997) described a dynamic model with logarithmic adaptation stage based on forward masking data. It showed improvement in robustness to background noise when used as a front-end for DTW and hidden Markov model (HMM) based recognition. Ghulam, Fukuda, Horikawa & Nitta (2005) implemented pitch synchronous processing of the ZCPA and demonstrated that combining forward and backward masking with the ZCPA may improve recognition. Adaptation has also been employed in auditory models duplicating the functionalities of the inner ear and the cochlea. Lyon & Mead (1988) proposed a computational cochlear model based on a simple cascade/parallel filterbank network. It was followed by compression based on half-wave rectification and automatic gain control to include the effects of adaptation and masking. The generalized synchrony detector (GSD) model proposed by Seneff (1988) simulates several nonlinear transductions in the cochlea such as saturation, adaptation and forward masking. But the ASR performance of these models are not well documented.

In this paper we investigated the effects of adaptation in a zero-crossing auditory model used as an ASR front-end and evaluated the isolated word recognition performance. The zero-crossing analysis (ZCA) of speech waveforms has several advantages over autocorrelation, power spectrum and linear prediction methods. In these methods data extraction by sampling a time waveform depends on the maximum frequency content in the time signal, whereas ZCA requires a number of extracted samples determined by the average rate of zero-crossing intervals. The ZCA is amenable to simple transformations instead of complex transformations between time and frequency domains. An auditory model which utilizes the zero-crossing principle as an ASR front-end is the Ensemble Interval Histogram (EIH) (Ghitza, 1994). An enhancement to the EIH is the ZCPA auditory model (Kim, Lee & Kil, 1999) which replaces multiple levels of EIH with a single zero level for frequency estimation. In addition, it utilizes a peak amplitude detector to extract the intensity information. The ZCPA features were shown to perform better than LPCC, MFCC, PLP, SBCOR and EIH features on

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

a small-vocabulary isolated word recognition task in presence of additive noise (Kim et al. 1999). The disadvantage of the model is the increased processing time since the processing are done in the temporal domain.

The base ZCPA auditory model does not utilize the perceptual property of adaptation in the interval histogram construction. We applied rapid and short term adaptation to the base ZCPA (ZCPA_ADP) and evaluated the ASR performance using a continuous density HMM recognizer. It was observed that with the adaptation strategy, the ASR performance may be improved. To reduce computational time, we implemented a simplified model with fewer number of filters with some optimization of the parameters and feature extraction algorithm. The performance is further enhanced by the integration of dynamic features (delta and delta-delta).

The paper is organized as follows. In section 2, the ZCPA model is described with the adaptation features. The ASR performance of isolated digits in clean and four types of additive noise (white Gaussian, factory, babble and car) are presented. In section 3, the performance enhancement obtained with dynamic features added to the adaptation is described. In sections 4 and 5, we compare the performance of the perceptual ZCPA with the static and dynamic MFCC features and analyse the computation cost of the ZCPA adaptation. Finally, section 6 presents an overall discussion and the conclusions drawn from the experiments.

2. The ZCPA with adaptation for ASR

According to the temporal representation of auditory processing, the auditory nerve fibres tend to fire in synchrony with the stimulus periods corresponding to the formant frequencies and their harmonics. This synchronous firing, also known as the phase-lock phenomena, contains useful frequency information. In the ZCPA, a synchronous neural firing is simulated as the up-ward going zero-crossing event of the signal. The inverse of the time interval between adjacent zero-crossings is collected in a frequency histogram.

Figure 1 shows the block diagram of the ZCPA. If we denote $y_i(n)$ as the output of the ith filter and $w(n)$ as a window of finite length, then the value of $y_i(n)$ at the frame index m is given by

$$y_i(n;m) = y_i(n)\, w_i(m-n) \qquad (1)$$

The output of the ZCPA for time t as a function of frequency f is obtained by summing the frequency bin weights over all the channels within the time interval and is given by

$$y(t,f) = \sum_{i=1}^{C} \sum_{k=1}^{Zi-1} \log(P_{ik}+1)\delta_{bjk} \qquad 1<b<N \qquad (2)$$

where δ_{bj} is the Kronecker delta, C the number of bandpass cochlear filters, N is the number of frequency bins, z_i are the number of upward-going zero-crossings in a windowed frame $y_i(n;m)$ and P_{ik} is the peak amplitude between the k-th and the (k+1)-th zero-crossing within the frame. Corresponding to this zero-crossing interval, the frequency histogram is increased at jk frequency bin by a value which is proportional to the logarithm of the peak within the interval.

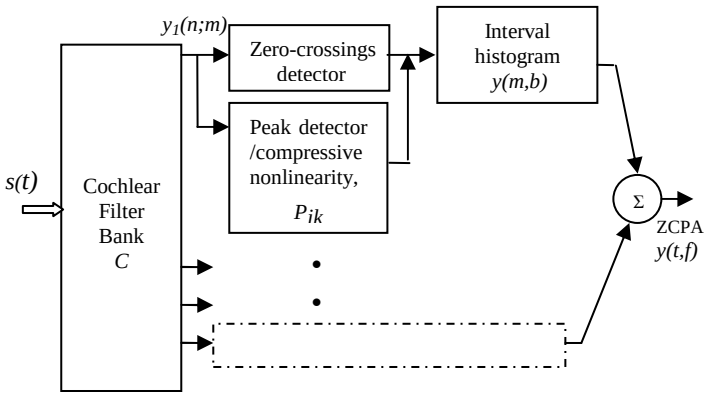


Figure 1: Schematic diagram of the ZCPA

2.1. The adaptation model

Speech transitions or dynamics of response to non-steady-state signals are important cues for ASR. In response to tone bursts, single auditory-nerve fibres exhibit an increased firing rate in the initial 15 ms. This decays monotonically in time, reaching a steady-level within about 50 ms. The decay consists of an initial rapid phase with a time constant of 3 ms (rapid adaptation). It is followed by a slower exponential decay with a time constant of about 40 ms (short-term adaptation) (Smith & Zwislocki, 1975).

Figure 2 shows the rapid and the short-term adaptation response with a time constant of 40 ms to a 2 kHz tone of 150 ms duration followed by a tone of 625 ms duration.

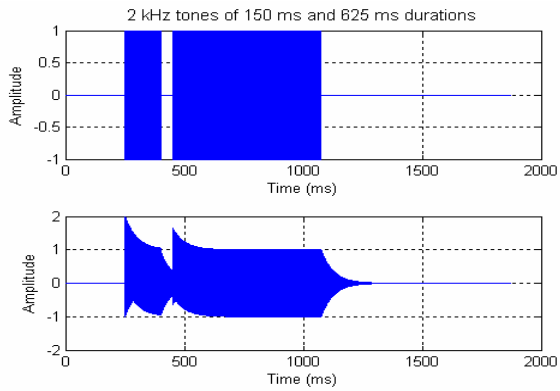


Figure 2: Rapid and short-term adaptation to a 2 kHz tone bursts of 150 and 625 ms durations. The short-term adaptation time constant is 40 ms.

Adaptation accentuates signal onsets by following a high initial firing rate. A rapid adaptation component appears to primarily enhance voicing. Short-term adaptation was found to improve the immunity of the system to noise (Abdelatty, Spiegel & Mueller, 2002). By enhancing the temporal changes it helps to attenuate stationary noise. This effect may be

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

implemented by high-pass filtering, which eliminates any stationary source of distortion. This principle is used in RASTA processing and shown to improve the robustness of the system (Hermansky & Morgan, 1994).

Holmberg et al. (2006) have proposed a method for introducing synaptic adaptation into the MFCC feature extraction. A first-order infinite impulse response (IIR) highpass filter was used to represent the decaying exponential effects of the rapid and the short-term adaptation. Our approach was similar to the one used by Holmberg. However, it differed from this in two aspects. Firstly, it operated in the time domain rather than in the spectral domain. Secondly, rapid adaptation enhances the temporal fine structure of speech signals, but this fine structure is removed by the MFCC feature extraction. Therefore rapid adaptation was not included in the Holmberg filter. Since in the ZCPA the temporal structures are preserved, rapid adaptation may be preserved in the ZCPA model. We defined the first order high pass IIR filter function as

$$H(z) = \frac{5\tau f_r (1 - z^{-1})}{(5\tau f_r + 0.05) + (5\tau f_r - 0.05) z^{-1}} \qquad (3)$$

where τ is the time constant in seconds and f_r is the frame rate equal to 200 Hz.

Forward masking can be viewed as a consequence of auditory adaptation (Dau & Püschel, 1996). Ghulam et al. (2005) combined forward and backward masking with the pitch synchronous ZCPA, which was half-wave rectified and centre clipped. Our approach differed from this method in that we did not use half-wave rectification since this introduces substantial higher order harmonics of the formant frequencies (Seneff, 1988). Moreover, Kim et al. (1999) has shown that higher level values than the zero level result in higher sensitivity in the estimated intervals and frequencies for the ZCPA. Moreover, we have related the process of adaptation with a time constant as observed by psychoacoustic experiments.

It is generally assumed that the human auditory system with longer time constants might be important for speech processing which may give better recognition. Forward masking can last up to 200 ms and the time constant effective in simultaneous masking is also 200 ms. The best time constant for ASR lies between 200 ms and 300 ms (Holmberg, et al. 2006). For our case a time constant of 250 ms was used in all experiments. The corner frequency f_c corresponding to this time constant is given by

$$f_c = \frac{1}{2\pi\tau} \qquad (4)$$

This gave a corner frequency of 0.636 Hz. This is below 1-16 Hz which is the modulation spectrum, considered important for human speech intelligibility (Holmberg et al. 2006). Adaptation has the effect of attenuating the low frequencies in the modulation spectrum. In the ZCPA_ADP, the adaptation filtering was implemented by summing a temporally highpass filtered version of the filter output with the original filter output.

2.2. The ZCPA model with adaptation

Figure 3 shows the adaptation stage added to the ZCPA. Speech frames were pre-emphasized to model the outer and the middle ear functionalities. It was then processed by a bank of 16 finite impulse response (FIR) filters of order 70 uniformly spaced on the equivalent rectangular bandwidth (ERB) scale between 10 Hz and 3.5 kHz. Our choice of FIR filters for frequency processing was motivated by the fact that FIR filters consistently performed better than carefully designed cochlear filters when applied in ASR applications (Kim et al., 1999). The number of filters were kept low to reduce the computational cost, which increases with increased sub-band processing. However, fewer number of filters result in greater frequency overlap of adjacent frequency channels. The use of lower number of filters introduces a histogram bias in the extracted features at the cost of faster processing. In each filter output, inverse of zero-crossing intervals were collected in 26 frequency bins uniformly spaced between 10 Hz and 4 kHz. The interval histogram was weighted by the logarithm of the peak value within the subband.

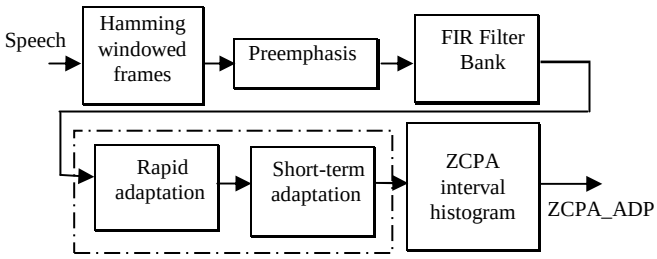


Figure 3: Schematic diagram of the ZCPA with adaptation (ZCPA_ADP).

In temporal processing it is usually required to use large frame sizes especially for lower frequency channels to capture about 10 periods of the signal for the accumulation of temporal information. Although longer time frames give better parameter estimates, the frame size should not be too large to violate the stationarity assumption. Considering these, a frame size of 80 ms was used. One ZCPA frame was obtained every 10 ms. At the filter outputs, the zero-crossing intervals were collected over an analysis window length of $10/f_k$ for lower frequencies and $60/f_k$ for higher frequencies, where f_k was the filter centre frequency (Gajic & Paliwal, 2003). The histogram was normalized to reduce the effects of biasing. Thirteen cepstra without C0 were generated and retained from each speech frame using inverse FFT of log FFT of the extracted features.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006. Copyright, Australian Speech Science & Technology Association Inc.

Accepted after full paper review

Table 1: Comparison of recognition rates (in %) of the base ZCPA, ZCPA with adaptation (ZCPA_ADP) and ZCPA with adaptation and dynamic processing (ZCPA_ADP_DEL) in clean and in four types of additive noise.

Noise SNR (dB)	White			Factory			Babble			Car		
	ZCPA	ZCPA_ ADP	ZCPA_ ADP_DEL	ZCPA	ZCPA_ ADP	ZCPA_ ADP_DEL	ZCPA	ZCPA_ ADP	ZCPA_ ADP_DEL	ZCPA	ZCPA_ ADP	ZCPA_ ADP_DEL
Clean	95.4	95.4	100.00									
40	90.9	95.4	95.4	95.4	90.9	100.0	90.9	90.9	95.4	95.4	95.4	100.0
30	81.8	90.9	90.9	86.3	81.8	95.4	86.3	86.3	90.9	90.9	90.9	100.0
15	77.3	72.7	59.0	81.8	77.3	81.8	72.7	72.7	77.3	86.3	86.3	95.4
10	68.8	59.0	54.5	68.8	72.7	77.3	63.6	59.1	63.6	86.3	81.8	90.9
5	50.0	31.8	22.7	50.0	68.1	72.7	31.8	59.1	22.7	81.8	77.3	81.8

Figure 4 shows the spectrogram of the 35-th frame for the male utterance of the CV (consonant followed by a vowel sound) /ba/. It is observed in 4(b) that high frequency segments are enhanced by the application of adaptation.

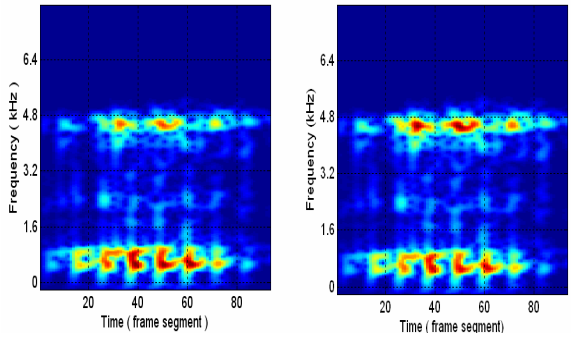
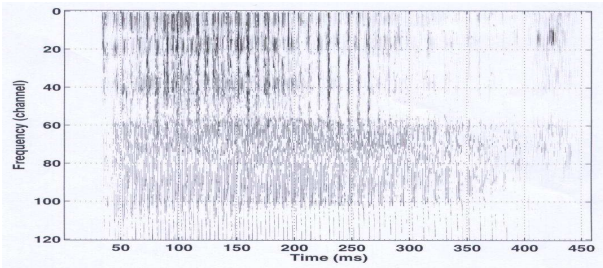


Figure 4: Spectrogram showing the effects of adaptation in the high frequency segments for the 35-th frame of the utterance /ba/ (a) without adaptation, (b) with adaptation (time constant of 250 ms).

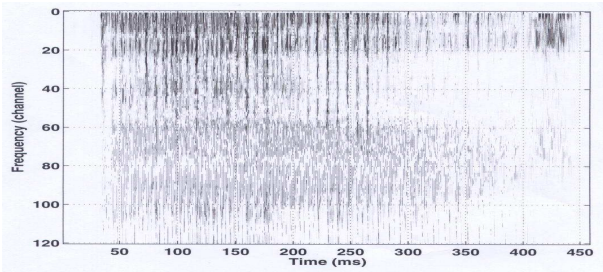
Figure 5 shows a time-frequency (channel) plot of the point process obtained from the simulated firing pattern for the voiced plosive /ba/ in clean condition. The point process was obtained using 120 FIR filters from the upward zero-crossing events of the waveform and was collected over a fixed analysis window of 80 ms with the same frame parameters as for the interval histogram construction. It is seen in 5 (b) that the burst after the closure is emphasized with adaptation in the high frequency regions (lower channels).

2.3. Recognition results

Speaker independent isolated digits from TIDIGITS speech database were used for recognition experiments. There were 55 male speakers in the training set and a separate 55 male speakers in the test set, each with 22 utterances of the digits 1-9, 'oh' and 'zero'. Continuous Gaussian density HMM with 15 states per digit, 5 mixture components per state and diagonal covariances were used to define each model. A 3-state silence model was inserted at the beginning of each utterance. The Baum-Welch re-estimation using a flat-start



(a)



(b)

Figure 5: Time-frequency (channel) plot of the point process obtained from simulated firing pattern for the utterance /ba/ (a) without adaptation and (b) with adaptation (time constant of 250 ms), showing enhanced onsets at higher frequencies (lower channels).

scheme and 15 estimation iterations was used for training under clean conditions. Test speech was corrupted with Gaussian white noise, factory noise, babble noise and car noise from the NOISEX 92 database. A left-to-right Viterbi recognizer was used for testing word accuracy. Recognition results are shown in Table 1. It was observed that in clean condition and in white Gaussian noise at high SNR, the recognition rate was improved with adaptation over the base ZCPA. The adaptation accentuated signal onsets and had an effect of shifting the global mean towards zero, resulting in improved estimation of the Gaussian model parameters. However, at low SNRs, there was a degradation in the recognition rate. This is due to the fact that the high pass adaptation filter gain accentuates the high frequency white noise components. This further increases the variance of the zero-crossing perturbation and increases low frequency formant contrast.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 2: Comparison of recognition rates (in %) of the MFCC with ZCPA, and with ZCPA with adaptation and dynamic processing (ZCPA_ADP_DEL) in clean (shaded) and in white Gaussian and factory noise.

SNR dB	White					Factory				
	MFCC	ZCPA	MFCC_DEL	ZCPA_DEL	ZCPA_ADP_DEL	MFCC	ZCPA	MFCC_DEL	ZCPA_DEL	ZCPA_ADP_DEL
Clean	100.0	95.4	100.0	95.4	100.00					
40	100.0	90.9	100.0	95.4	95.4	100.0	95.4	100.0	95.4	100.0
30	81.8	81.8	95.4	86.3	90.9	100.0	86.3	100.0	95.4	95.4
15	18.1	77.3	27.2	77.3	59.0	95.4	81.8	100.0	81.8	81.8
10	18.1	68.8	22.2	50.0	54.5	86.4	68.8	100.0	77.3	77.3
5	9.0	50.0	13.6	31.4	22.7	54.5	50.0	95.4	59.0	72.7

For nonstationary factory noise, the improvements with adaptation were observed at lower SNRs, which is consistent with the results of Kim et al. (1999). It is expected that larger frame lengths would give poor estimates in time domain processing in presence of nonstationary factory noise. This may be a reason for poor recognition results at higher SNRs. For babble and car noise, there were no significant changes of recognition rates with adaptation. We observed that the performance of the ZCPA with CMS did not show any improvements over the baseline performance. Therefore we did not use CMS in our experiments.

3. Performance enhancement by dynamic features

Dynamic behavior or time varying features of speech are important for ASR. The ZCPA features with adaptation was further tested with the dynamic (delta and delta-delta) features. Regression analysis was applied to each time function of the cepstrum coefficients over three frame intervals to the left and three frame intervals to the right of the centered frame, every 10 ms (Furui, 1986). The dynamic features were concatenated to the ZCPA_ADP with a time constant of 250 ms to generate a 39-dimensional feature vector (ZCPA_ADP_DEL). The recognition results are shown in Table 1 for the four types of additive noise. We observed that the performance of the ZCPA_ADP was further enhanced with the addition of dynamic features at clean and high SNRs. The ZCPA_ADP_DEL consistently performed better in factory, babble and car noise over a wide range of SNR.

4. Performance comparison of ZCPA-ADP and MFCC features

For speech recognition purposes and in research, MFCC is widely used for speech parameterization and is accepted as the baseline. We further compared the performance of the baseline MFCC with the ZCPA_ADP. The MFCC features were generated with 40 (13 linear and 27 log) triangular filters using a 25 ms frame size. A 512-point FFT was used and one frame was generated every 10 ms. The log energy output was computed as the product of the filter coefficients and input magnitude spectrum. From each frame, 13 cepstra including C0 were computed using a DCT.

The ZCPA features were generated on the ERB scale using the same frame parameters, but with a 80 ms frame size. Dynamic features for MFCC (MFCC_DEL) were generated from the static features using regression analysis as stated in Section 3. The recognition results are stated in Table 2. We observed that the performance of the ZCPA_ADP_DEL were significantly improved over the MFCC_DEL in white noise below 30 db SNR. For example, at 15 dB SNR the recognition rate was 59.0% for ZCPA_ADP_DEL compared to 27.2% for the MFCC_DEL. The ZCP_ADP_DEL also performed better than ZCPA_DEL for a wide range of SNR.

In the usable recognition rate (>95%) the standard MFCC_DEL is still the best feature and it is less computationally expensive than the ZCPA based techniques. This may raise the question as to why we need the proposed algorithm. It can be said that, the computational complexity of ZCPA is more than the MFCC because it duplicates the functions of the auditory periphery in greater details than the MFCC (dominant frequency principle and the phase-lock phenomenon). The main benefit of ZCPA over MFCC lies in its performance in additive noise conditions, attributed to the dominant frequency principle. In real life situations (in presence of additive and convolutive noise), it is not possible to achieve a high recognition rate above 95% as MFCC performance rapidly degrades. Although, it can be alleviated very simply by CMS to some extent, and more efficiently by appending dynamic (delta and delta-delta) features but at the cost of increased feature space dimensionality. Moreover, fine temporal structures which are not preserved or lost during feature extraction in the frequency domain can be preserved in the ZCPA. The objective of this study is to demonstrate that perceptual features may also be employed to enhance the machine recognition of speech in clean and noisy conditions.

5. Computation Complexity

It is also useful to consider the computational cost of the adaptation process to see whether it balances the gain in the recognition rates achieved. The adaptation filtering performs 3 multiplications, 2 additions and 1 division in each filter subband. The dynamic (delta) processing performs a total of 130 multiplications, 273 additions and 13 divisions for a regression processing with 7 frames and 13 cepstra. A comparison of improvements in recognition rates achieved vs increase in processing time using TIDIGITS connected digits (55 male and 55 females speakers each with 77 utterances of a connected digits string) is given in Table 3.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 3: Comparison of increase in recognition rates vs increase in processing time using TIDIGITS connected digits with white Gaussian noise at 30 dB SNR.

Features	% increase in recog. rate over the base	% increase in processing time over the base
ZCPA	Base	Base
ZCPA_DEL	2.39	5
ZCPA_ADP	1.58	20
ZCPA_ADP_DEL	5.80	40

It was observed that for adaptation with dynamic features, a 5.8% increase in recognition rate might be achieved with a corresponding 40% increase in processing time.

6. Discussion and conclusions

In this paper, the ASR performance of a zero-crossing temporal auditory model was investigated with the perceptually motivated property of rapid and short-term adaptation. The adaptation was implemented with a first order high pass IIR filter with a time constant of 250 ms. The ZCPA model was optimized for efficient and faster processing. The temporal adaptation was tested using a HMM recognizer on isolated digit (TIDIGITS) database in four types of additive noise. For stationary white noise, it was observed that the ASR performance of the base ZCPA might be improved with adaptation in clean conditions and upto 20 dB SNR. The adaptive feature showed enhancement in the high frequency segments of voiced plosives. Speech onsets and voicing were observed to be accentuated. However, below 20 dB, there was degradation in the recognition rate due to increased high frequency noise perturbations caused by adaptation filtering. It was found that CMS did not provide any significant improvement with the ZCPA features in clean or noise conditions.

For nonstationary factory noise, the improvements with adaptation were observed at low SNRs, which is consistent with the results of Kim et al. (1999). The ASR performance was better than in white noise since it is less affected by high frequency noise components. However, the recognition performance degraded at higher SNRs. For babble and car noise, there were no significant changes of recognition rate with adaptation.

The ASR performance of the ZCPA with adaptation may be further improved by appending dynamic (delta and delta-delta) features to it. With dynamic features, significant improvements were observed in factory, babble and car noise over a wide range of SNR.

A comparison of ASR performance of the ZCPA with baseline MFCC was made. It was observed that the ZCPA_ADP_DEL performed better than the MFCC with dynamic features (MFCC_DEL) in white noise at low SNRs. The ZCP_ADP_DEL also performed better than the ZCPA_DEL for a wide range of SNR.

The benefit of adaptation in the ZCPA for ASR is noted, especially in nonstationary noise environment with some enhancement techniques. Further investigation will be undertaken to evaluate the noise performance and to evaluate the adaptation model in other types of noise such as convolutive and reverberant noise.

7. References

Abdelatty, A. M., Spiegel, J. V. & Mueller, P. (2002). Robust auditory-based speech processing using the average localized synchrony detection. *IEEE Transactions on Speech, and Audio Processing*, 10(5), 279-292.

Dau, T. & Püschel, D. (1996). A quantative model of the "effective" signal processing in the auditory system. I. Model structure. *Journal of the Acoustical Society of America*, 99(6), 3615-3622.

Furui, S. (1986). Speaker-independent isolated word recognition using dynamic features of speech spectrum. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34(1), 52-59.

Gajić, B. and Paliwal, K. K. (2003). Robust speech recognition using features based on zero crossings with peak amplitudes. *ICASSP 2003.*

Ghitza, O. (1994). Auditory models and human performance in tasks related to speech coding and speech recognition. *IEEE Transactions on Speech and Audio Processing*, 2(1), 115-132.

Ghulam, M., Fukuda, T., Horikawa, J. & Nitta, T. (2005). Pitch-synchronous ZCPA–based feature extraction with auditory masking. *ICASSP 2005*, Philadelphia.

Hermansky, H. & Morgan, N. (1994). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing*, 2(4), 587-589.

Holmberg, M., Gelbart, D. & Hemmert, W. (2006). Automatic speech recognition with an adaptation model motivated by auditory processing. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), 44-49.

Kim, D. S, Lee, S. Y. & Kil, R. M. (1999). Auditory Processing of speech signals for robust speech recognition in real world noisy environments. *IEEE Transactions on Speech and Audio Processing*, 7(1), 55-69.

Lyon, R. F. & Mead, C. (1988). An analog electronic cochlea. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7), 1119-1134.

Seneff, S. (1988). A joint synchrony/mean-rate model of auditory processing. *Journal of Phonetics*, 85(1), 55-76.

Smith, R. & Zwislocki, J. J. (1975). Short-term adaptation and incremental responses of single auditory-nerve fibres. *Biological Cybernetics*, 17, 169-182.

Strope, B. & Alwan, A. (1997). A model of dynamic auditory perception and its application to robust word recognition. *IEEE transactions on Speech and Audio Processing*, 95(5), 451-464.

Westerman, L. & Smith, R. L. (1984). Rapid and short-term adaptation in auditory nerve responses, *Hearing Research*, 15, 249-260.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review