A TEMPORAL AUDITORY MODEL WITH ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION

Serajul Haque, Roberto Togneri, Anthony Zaknich

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

ABSTRACT

Rapid and short-term adaptation are dynamic mechanisms of human auditory system. 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 environment. 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.

Index Terms— Auditory system, speech recognition, feature extraction, adaptive system, hidden Markov model.

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. 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 single auditory-nerve fibres is higher 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. The decrease in response rate, referred to as adaptation, has been determined by psychophysical experiments as observed responses to pure tones [2, 11].

Perceptual features including adaptation have been employed in ASR with favorable results. Holmberg *et. al* [3] incorporated a simplified model of synaptic adaptation into MFCC features as a competitive strategy to RASTA [7] and the cepstral mean subtraction (CMS). It showed improved ASR performance compared to baseline MFCC. Strope and Alwan [4] described a dynamic model with a 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 *et. al* [10] 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 cochlear models duplicating the functionalities of the inner ear and the cochlea [5, 6], but the ASR performance of these models are not well documented.

An auditory model which utilizes the zero-crossing principle is the ZCPA proposed by Kim *et. al* [1]. It is an enhancement to the Ensemble Interval Histogram (EIH) model where it replaces multiple levels of EIH with a single zero level for feature extraction. It utilizes a zero-crossing detector for frequency estimation. In addition, it utilizes a peak amplitude detector to extract the intensity information. The model utilizes the dominant frequency principle i.e. the zero-crossings are dominated by the dominant component in the waveform that contains more power than others. This contributes to the noise robustness of the ZCPA. The disadvantage of the model is the increased processing time since the processing is done in the temporal domain.

The base ZCPA auditory model does not utilize the perceptual property of adaptation in the interval histogram construction. In this paper we have proposed to extend the ZCPA auditory model with a simplified adaptation scheme which is consistent with psychoacoustic observations. It was observed that with the adaptation strategy, the ASR performance of the auditory model may be improved. To reduce computational time of the ZCPA, we implemented a simplified model with fewer number of filters with some optimization of the parameters and feature extraction algorithm. The performance of the ZCPA with adaptation was 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 added to it. In section 3, the ASR performance of isolated digits in clean and four types of additive noise are presented. Section 4 describes the performance enhancement obtained with dynamic features added to the adaptation. Finally, section 5 presents an overall discussion and the conclusions drawn from the experiments.

2. THE ADAPTATION MODEL

Speech transitions and 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) [2].

Holmberg *et. al* [3] 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, but differed 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 was not included in the Holmberg filter because this fine structure is removed by the MFCC feature extraction. Since in the ZCPA the temporal structures are preserved, rapid adaptation may be implemented in the ZCPA model. We defined the first order high pass IIR filter function as

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

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

Figure 1 shows the rapid and the short-term adaptation responses with a time constant of 40 ms to a 2 kHz tone of 625 ms duration followed by a tone of 150 ms duration. Adaptation accentuates signal onsets by following a high initial firing rate. A rapid adaptation component enhanced voicing and short-term adaptation was found to improve the immunity of the system to stationary noise. This principle is used in RASTA processing and shown to improve the robustness of the system [7].

Forward masking can be viewed as a consequence of auditory adaptation. Ghulam *et. al* [10] 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 [6]. Moreover, level values higher than the zero level result in higher sensitivity in the estimated intervals and frequencies for the ZCPA [1]. We have also related the process of adaptation with a time constant as observed by psychoacoustic experiments.

Longer time constants are important for speech processing which may give better recognition. Forward and simultaneous masking can last up to 200 ms and the best time constant for ASR lies between 200 and 300 ms [3]. 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

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

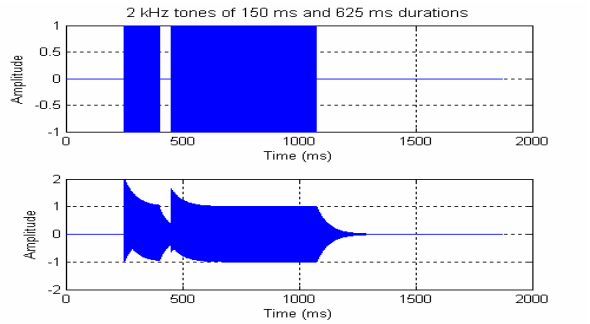


Fig. 1. Rapid and short-term adaptation to 2-kHz tone bursts of 150 and 625 ms durations. The short-term adaptation time constant is 40 ms.

This gave a corner frequency of 0.636, Hz which is below 1-16 Hz which is the modulation spectrum, considered important for human speech intelligibility. In the ZCPA, the adaptation filtering was implemented by summing a temporally highpass filtered version of the filter output with the original filter output.

2.1. The ZCPA auditory model with adaptation

According to the temporal representation of auditory processing, a single auditory nerve fibre tends to fire in synchrony with the stimulus periods corresponding to the formant frequencies and their harmonics. This synchronous firing, also known as phase lock phenomena, contains useful frequency information. In the ZCPA, a synchronous neural firing is simulated as the upward going zero-crossing event of the signal. The inverse of the time interval between adjacent zero-crossings is collected in a frequency histogram. The expected value of the zero-crossings count, D, of a zero mean Gaussian process $\{Z_t\}$, $t \in T$, is related to its spectral representation by

$$\cos\left(\frac{\pi E\{D\}}{N-1}\right) = \frac{\int_0^{\pi} \cos(\omega) dF(\omega)}{\int_0^{\pi} dF(\omega)} \qquad (3)$$

where $f(\lambda) = F'(\lambda)$, $-\pi \le \lambda \le \pi$, is the spectral density of Z_t [9].

Figure 2 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. The number of filters were kept low

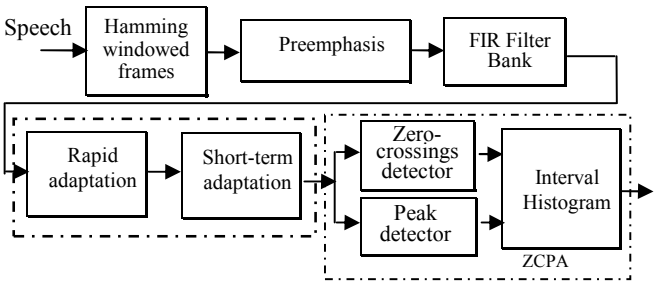


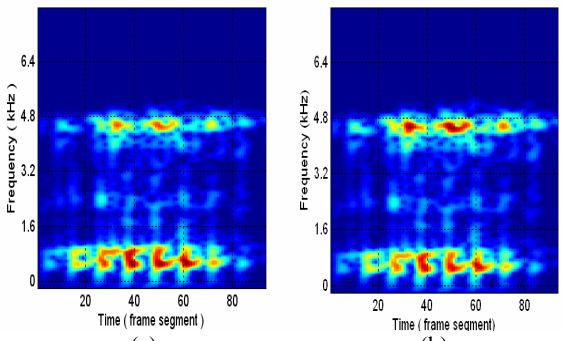
Fig. 2. Schematic of the ZCPA with adaptation (ZCPA_ADP)

Table 1. Comparison of recognition rates (%) of the base ZCPA, ZCPA with adaptation (ZCPA_ADP) and ZCPA with adaptation and dynamic features (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	59.1	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	68.8	72.7	31.8	59.1	22.7	81.8	77.3	81.8

to reduce the computational cost. However, fewer number of filters result in greater frequency overlap of adjacent frequency channels. This introduces a histogram bias in the extracted features at the cost of faster processing. In each filter output subband, the inverse of zero-crossing intervals were collected in 26 frequency bins uniformly spaced at the ERB scale between 10 Hz and 4 kHz. The interval histogram was weighted by the logarithm of the peak value within the subband. At the filter outputs, the zero crossing intervals were collected over an analysis window of length $10/f_k$ for lower frequencies and $60/f_k$ for higher frequencies, where f_k was the filter centre frequency. The histogram was normalized to reduce effects of biasing. Although longer window lengths give better parameter estimates, the window size should not be too large to violate the stationarity assumption. Considering these, a maximum window size of 80 ms was used. One ZCPA frame was obtained every 10 ms. Thirteen cepstra without C0 were generated and retained from each speech frame using inverse FFT of log FFT of the extracted features.

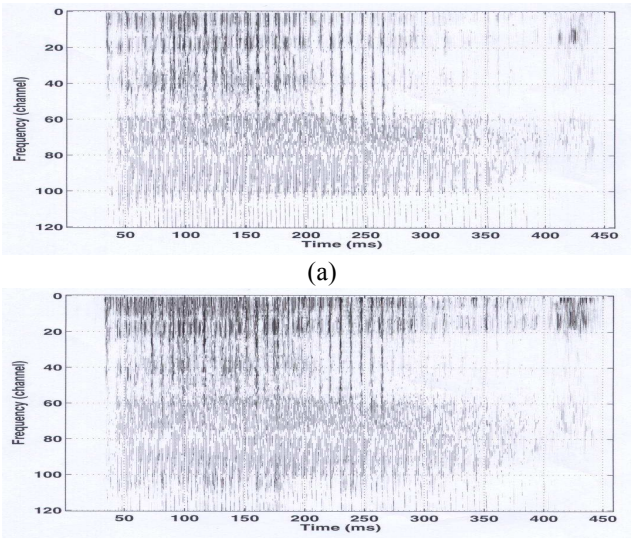
Figure 3 shows the spectrogram of the 35-th frame for the male utterance of the CV (consonant followed by a vowel sound) /ba/. It is observed in 3 (b) that high frequency segments are enhanced by the application of the adaptation strategy.



(a) (b)

Fig. 3. Spectrogram showing the effects of adaptation in the high frequency segments for the 35-th frame of the male utterance /ba/ in clean (a) without adaptation, (b) with adaptation with a time constant of 250 ms.

Figure 4 shows a time-frequency plot of the point process obtained from the simulated firing pattern for the voiced plosive /ba/ in clean conditions. 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. It is seen in 4 (b) that the burst after the closure is emphasized with adaptation in the high frequency regions (lower channels), indicating increased neural activity at higher frequencies.



(a)

(b)

Fig. 4. Time-frequency (channel) plot of the point process obtained from the simulated firing pattern for the CV utterance /ba/ in clean (a) without adaptation and (b) with adaptation (time constant 250 ms) showing enhanced onsets at higher frequencies (lower channels).

3. RECOGNITION RESULTS

Speaker independent isolated digits from the TIDIGITS speech database were used for recognition experiments. There were 55 male speakers in the training set and a separate set of 55 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 with 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 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 the word accuracy. Recognition results are shown in Table 1. It was observed that in clean conditions 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 HMM model parameters. However, at low SNRs, there was a degradation in recognition rate. This is due to the fact that at the output of the *k*-th adaptation filter, the variance of the input Gaussian white noise $v(n)$ with zero mean is multiplied by the square of the adaptation filter coefficient vector, **h**, which is given by

$$\sigma_k{}^2{}_{adp} = E\left\{\left|v(n)_{adp}\right|^2\right\} = \mathbf{h}^H\mathbf{R}_v\mathbf{h} = \mathbf{h}^H\mathbf{C}_v\mathbf{h} = \mathbf{C}_v\mathbf{h}^2 \qquad (4)$$

where $\mathbf{R}_v$ and $\mathbf{C}_v$ are autocorrelation matrix and the autocovariance matrix, respectively, of the input white noise. At low SNR, the high pass adaptation filter gain accentuates high frequency noise components. This further increases the variance of the zero-crossing perturbation and decreases low frequency formant contrasts.

For nonstationary factory noise, the improvements with adaptation were observed to be opposite to that with white noise with improved recognition at low SNRs, which are consistent with the results of Kim [1]. It is expected that larger window lengths would give poor estimates in time domain processing in the presence of nonstationary 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.

4. PERFORMANCE ENHANCEMENT USING DYNAMIC FEATURES

Dynamic behavior or time varying features of speech are important cues 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 [8]. 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 in clean and at high SNRs. However, the recognition rate degrades at low SNR white noise due to the high pass adaptation filtering further enhancing the high frequency noise perturbations. The ZCPA_ADP_DEL consistently performed better in factory, babble and car noise over a wide range of SNR.

5. CONCLUSIONS

In this paper, the ASR performance of a temporal auditory model based on zero-crossings was investigated with the perceptually motivated property of rapid and short-term adaptation. It was observed that with adaptation scheme the ASR performance of the base ZCPA may be improved in stationary white noise in clean conditions and upto 20 dB SNR. For highly nonstationary factory noise, the improvements with adaptation were observed to be at low SNRs. For nonstationary babble and car noise, there were no significant changes. This is because short-term adaptation by enhancing changes helps to attenuate primarily stationary noise and other sources of distortion.

The ASR performance of the ZCPA with adaptation may be further improved by appending dynamic (delta and delta-delta) features. With dynamic features, significant improvements were observed in factory, babble and car noise over a wide range of SNR. This is because delta features are able to capture instantaneous and dynamic variations with greater immunity to nonstationary noise. Further investigations will be undertaken to enhance the noise performance and to evaluate the adaptation model in other noise types, such as convolutive and reverberant noise.

6. REFERENCES

[1] D.S. Kim, S.Y. Lee, and R.M. Kil, "Auditory processing of speech signals for robust speech recognition in real world noisy environments," *IEEE Trans. Speech and Audio Proc*., vol. 7, No. 1, pp. 55-69, Jan. 1999.

[2] R. Smith and J.J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibres," *Biological Cybernetics*, 17, pp. 169-182, 1975.

[3] M.D. Holmberg, M.D. Gelbart, and W. Hemmert, "Automatic Speech Recognition with an adaptation model motivated by auditory processing," *IEEE Trans. Audio, Speech, and Lang. Proc.*, vol. 14, No. 1, pp. 44-49, Jan. 2006.

[4] B. Strope and A. Alwan, "A model of dynamic auditory perception and its application to robust word recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 95. No. 5, pp. 451-464, Sep. 1997.

[5] R.F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech and Signal Proc.,* vol. 36, pp. 1119-1134, Jul. 1988.

[6] S. Seneff, "A joint synchrony/mean-rate model of auditory processing," *J. Phonetics*, vol. 85, No.1, pp. 55-76, 1988.

[7] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech and Audio Proc*., vol. 2, pp. 587-589, Oct. 1994.

[8] S. Furui, "Speaker-Independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech and Signal Proc.*, vol. 34, No. 1, pp. 52-59, Feb. 1986.

[9] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proceedings of the IEEE*, vol. 74, No. 11, pp. 1477-1492, Nov. 1986.

[10] M. Ghulam, T. Fukuda, J. Horikawa and T. Nitta, "Pitch-synchronous ZCPA–based feature extraction with auditory masking," *ICASSP,* 2005.

[11] L. Westerman and R.L. Smith, "Rapid and short-term adaptation in auditory nerve responses," *Hearing Research,*15, pp. 249-260, 1984.