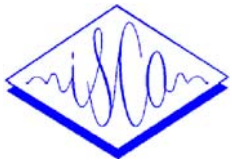


EUROSPEECH '97
5th European Conference on Speech
Communication and Technology
Rhodes, Greece, September 22-25, 1997

COMPARATIVE EVALUATIONS OF SEVERAL FRONT-ENDS FOR ROBUST SPEECH RECOGNITION

Doh-Suk Kim†, *Jae-Hoon Jeong*†, *Soo-Young Lee*†, *Rhee M. Kil*‡

†Department of Electrical Engineering/ ‡Division of Basic Science
Korea Advanced Institute of Science and Technology
373-1 Kusong-dong, Yusong-gu, Taejon 305-701, Korea
E-mail: dsk@eekaist.kaist.ac.kr

ABSTRACT

Zero-crossings with peak amplitudes (ZCPA) model motivated by human auditory periphery is simple compared with other auditory models, but powerful speech analysis tool for robust speech recognition in noisy environments. In this paper, improvement in recognition rate of ZCPA model is addressed by incorporating time-derivative features with several different time-derivative window lengths. Experimental results show that ZCPA has relatively higher sensitivity to derivative window length than conventional feature extraction algorithms. Also, experimental comparisons with several front-ends including some auditory-like schemes in real-world noisy environments demonstrate the robustness of ZCPA model. ZCPA model shows superior performance compared with other front-ends especially in noisy condition corrupted by white Gaussian noise.

1. INTRODUCTION

Automatic speech recognition (ASR) is the leading technology as a human-computer interface for real-world applications. However there are various types of background noises in real environments which degrade the performance of ASR systems, and there have been many researches on modeling functional roles of the peripheral auditory systems to design robust front-ends of ASR systems. ZCPA model, which is relatively simple and efficient, was proposed as a robust front-end for ASR in noisy environments and was shown to be robust to additive white Gaussian noise than both LPC-derived cepstrum and ensemble interval histogram (EIH) model [1] in our previous work [2]. In this paper, the performance of ZCPA is evaluated as the time-derivative window length is varied for improved recognition accuracy, and is compared with other front-ends including some auditory-like schemes as well as conventional ones in several real-world noisy environments.

2. ZCPA ANALYSIS

The ZCPA model consists of a bank of bandpass cochlear filters and nonlinear stages at the output of each cochlear filter as shown in Fig. 1. The cochlear filterbank represents frequency selectivity at various locations along a basilar membrane in the cochlea, and is implemented with a bank of hamming band pass filters. Auditory nerve fibers tend to fire in synchrony with the stimulus, and the synchronous neural firing is simulated as the upward-going zero-crossing event of the signal at the output of each bandpass filter, and the inverse of time interval between adjacent neural firings is represented as a frequency histogram. Further, each peak amplitude between successive zero-crossings is detected, and this peak amplitude is used as a nonlinear weighting factor to a frequency bin to simulate the relationship between the stimulus intensity and the degree of phase-locking of auditory nerve fibers. The histograms across all filter channels are combined to represent the pseudo-spectrum of the auditory model. As a result, frequency information of the signal is obtained by zero-crossing intervals of subband signals, and intensity information is also incorporated by a peak detector followed by a saturating nonlinearity.

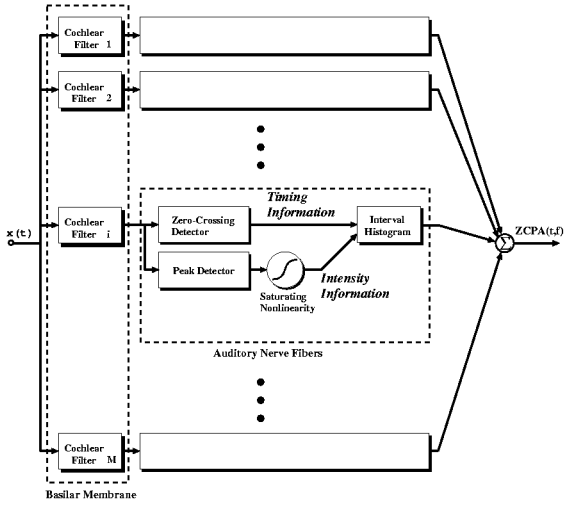


Figure 1: Block diagram of the zero-crossings with peak amplitudes (ZCPA) model.

On the other hand, EIH model utilizes level crossings for frequency information. However, unlike ZCPA model, multiple level-crossing detectors with different

level values are utilized both for frequency and intensity information in EIH model. In implementing EIH, one has to determine several parameters such as the number of levels and level values, which are extremely critical for reliable performance. However, there is no elegant method to determine these values, except by trial-and-error. The utilization of zero-crossings in frequency estimation makes ZCPA model free from unknown parameters associated with the level, more efficient for calculations, and more robust to noise than EIH model.

3. DATA BASE AND RECOGNITION SYSTEMS

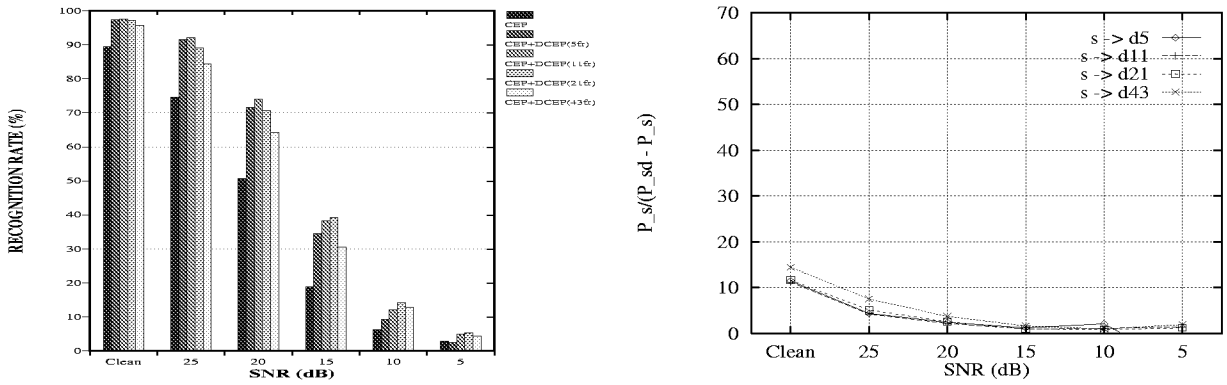
In consideration of practical applications of ASR, 50 Korean words for control of electric home appliances including TV and VCR were chosen. The utterances from 16 male speakers were sampled at 11.025 kHz sampling rate with 12 bit precision via SONY ECM-220T condenser microphone. 900 tokens of 9 speakers were used as training of recognizers, and 1050 tokens of the other speakers as test evaluations. There are many kinds of noises in real environments which are not stationary in general, and the performance evaluation in real situations may be very important for practical applications of ASR. Factory noise and military operations room noise, contained in NOISEX-92 CD ROMS [3], were added to the test data sets for test evaluations in real situations. To evaluate the recognizer-independent reliability of the front-ends, both discrete hidden Markov model (HMM) and multilayer perceptron (MLP) are used as speech recognizers. Each HMM is iteratively trained with Baum-Welch algorithm based on maximum likelihood estimation (MLE). The codebook of size 256 is trained with training data in iterative manner. Although there have been a lot of schemes proposed to apply neural networks to speech recognition, the static MLP recognizer with trace-segmentation algorithm [4] is used for normalization of time scale without serious computation time.

4. INCORPORATION OF DYNAMIC FEATURES

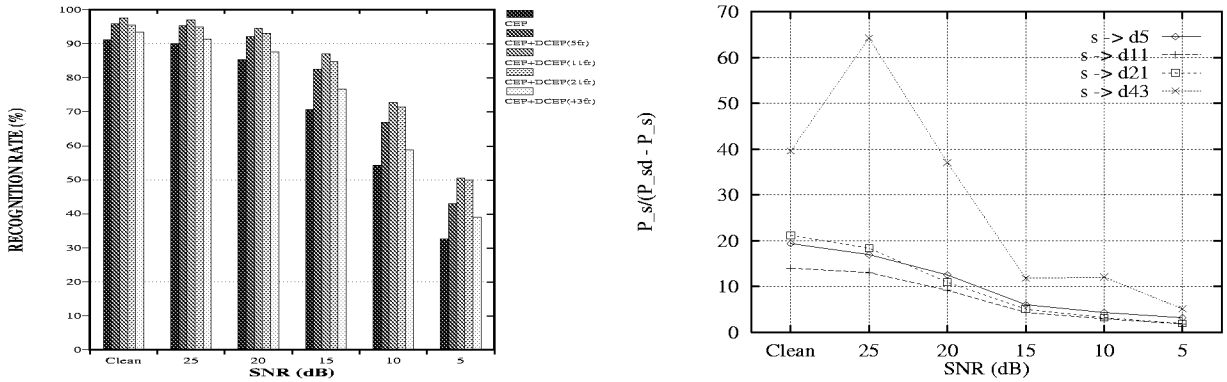
It is well known that the utilization of time-derivative features improves recognition accuracy not only in clean condition but also in noisy conditions since the time-derivative features are less affected than static features in noisy environments and the time-derivative features provide extra information over several frames which cannot be handled by HMM-based recognizers. However, unlike conventional feature processing techniques, the length of time window is dependent on the channel index in both EIH and ZCPA, i.e., it varies inversely with characteristic frequency of the channel. For example, the time window length of the channel with lowest characteristic frequency spans up to 50 msec, which is quite long compared with the frame rate of about 10 msec. Thus, it is highly recommended to investigate the performance of the auditory model when time-derivative features are incorporated to static features for practical applications where higher recognition performance is required.

Fig. 2 summarizes recognition results of HMM recognizer as the derivative window length is varied when speech data is corrupted by white Gaussian noise. Sixteen hamming bandpass filters are used as the cochlear filterbank of ZCPA, and the frequency range between 1.5 bark and 17.5 bark is divided into 16 frequency bins which are equally spaced by one bark according to the critical-band rate. Twelve cepstral coefficients of ZCPA and regression implementation of delta-cepstrum are computed every 10.15 msec to constitute the feature vector. Two independent codebooks are constructed for cepstrum and delta-cepstrum respectively, under the assumption that the static and the time-derivative features are statistically independent each other. Recognition rates obtained by using static feature only (CEP) and by using both static and dynamic features are shown in the left column, and the right column represents the inverse of performance improvement rate, $\gamma = P_s/(P_{sd} - P_s)$, where P_s and P_{sd} denote recognition rate obtained by using static features and by using both static and dynamic features, respectively. The improvements incurred by time-derivative features are more eminent for noisy data than for clean data for both MFCC and ZCPA. However the performance of ZCPA is much more sensitive to the derivative window length compared with MFCC as shown in right plots of Fig. 2. It is clear that the contribution of time-derivative features is poor if the time-derivative window length is too long (43 frames), and the window length of 11 frames shows the best performance on average for ZCPA while the difference in performance improvement is negligible for MFCC.

On the other hand the results of MLP recognizer is significantly different from that of HMM recognizer, even though the detailed results are not shown in this paper. There is no performance improvement by utilizing time-derivative features for MLP recognizer. Further, different time-derivative window lengths do not make much differences in recognition rates. Since the time-derivative features are obtained by linear combination of static features over several frames, appropriate time-derivative features, or the beyond of those, can be represented internally in hidden representations of MLP network with static features only. Thus, it is sufficient to use only static features for MLP recognition systems.



(a) MFCC



(b) ZCPA

Figure 2: The effect of augmenting time derivative features with several different derivative window lengths to static features.

5. COMPARISON WITH OTHER FRONT-ENDS

In this section, the performance of ZCPA is compared with other front-ends including LPC cepstrum (LPCC), MFCC, subband autocorrelation (SBCOR) [5], perceptual linear prediction (PLP) [6], and EIH under various types of noisy environments. Table 1 summarizes word error rate of both (a) HMM recognizer and (b) MLP recognizer for the data corrupted by several kinds of real-world noises. The gain of the noise is adjusted to make the global SNR of 15 dB. The feature vector consists of static features and their time-derivative coefficients for each front-end, and the window length of time-derivative features is set 11 frames. For LPCC, speech signal is first multiplied by hamming window of 20.3 msec duration every 10.15 msec, and 8 LPC coefficients and 12 cepstral coefficients are obtained successively. For MFCC, 16 mel-scale triangular bandpass filters are used to obtain 12 coefficients. To calculate SBCOR, the same 16 hamming bandpass filters used in both ZCPA and EIH are utilized. In PLP processing, 16 critical-band filters are used and LPC order is 8. Performance of several EIHs with 7 level crossing detectors were evaluated by varying the number of levels and level values, and only the best case among them is shown. In clean condition PLP performs the best, but the difference between all front-ends in error rate is quite small. And the error rate of SBCOR is lower than that of PLP when speech data is corrupted by white Gaussian noise, but is higher under real-world noisy environments on the contrary. The superiority of ZCPA over the other front-ends is prominent in all kinds of noisy conditions, and the difference in error rate between ZCPA and the other front-ends is maximum (ZCPA - 13 % and LPCC - 88 %) when speech data is corrupted by white Gaussian noise. However, for speech data corrupted by military operations room noise, the difference is reduced compared with the other kinds of noisy environments. This may be due to the characteristics of the military operations room noise in which frequency band of speech noise is overlapped with the original input signal.

Since the incorporation of time-derivative features does not improve recognition accuracy in the MLP recognizer, results obtained with only static features are shown in Table 1 (b). ZCPA outperforms all the other features, as does in the HMM recognizer.

Table 1: Comparison of word error rates (%) of (a) HMM recognizer and (b) MLP recognizer obtained by using several features augmented by time-derivative features under various types of noisy environments. WGN, FAC, and MOP denote white Gaussian noise, factory noise, and military operations room noise, respectively. Only the results obtained with static features are shown in (b).

(a) HMM results

	Clean	WGN	FAC	MOP
LPCC	5.5	88.0	47.4	46.5
MFCC	2.5	61.7	32.7	29.1
SBCOR	3.6	27.3	22.2	23.1
PLP	1.8	44.5	18.4	17.6
EIH	2.6	15.7	13.3	20.2
ZCPA	2.4	13.0	9.7	14.2

(b) MLP results

	Clean	WGN	FAC	MOP
LPCC	5.0	75.1	43.7	38.7
MFCC	3.0	58.2	35.2	26.7
SBCOR	4.0	30.2	25.2	28.7
PLP	1.6	38.3	21.3	20.2
EIH	1.7	11.9	9.5	9.2
ZCPA	2.2	8.3	6.6	6.8

6. CONCLUSIONS

The ZCPA model based on human auditory periphery was proposed as a robust front-end for speech recognition systems in noisy environments in our previous work. In this paper, several different lengths of time have been tried to both MFCC and ZCPA, and result in higher sensitivity of ZCPA to time-derivative window length. And the time-derivative window length of 11 frames shows better recognition accuracy with HMM classifier, but does not make much differences with MLP classifier. MLP classifier shows better recognition rates than HMM classifier in most of all the cases. Since different lengths of bandpass signals are considered in computing ZCPA output according to the characteristic frequency of the channel while the frame rate is fixed, it may be possible to apply different lengths of time-derivative windows according to the characteristic frequency of the channel.

Also, comparative evaluations of ZCPA model with several feature extraction methods demonstrate the robustness of ZCPA model in several real-world noisy environments. The superiority of ZCPA over the other front-ends is prominent in all kinds of noisy conditions, especially when speech data is corrupted by white Gaussian noise.

7. REFERENCES

[1] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing* (S. Furui and M. M. Sondhi, eds.), pp. 453–485, New York: Marcel Dekker, 1992.

[2] D. S. Kim, J. H. Jeong, J. W. Kim, and S. Y. Lee, "Feature extraction based on zero-crossings with peak amplitudes for robust speech recognition in noisy environments," in *Proc. ICASSP*, (Atlanta, USA), pp. 61–64, May 1996.

[3] A. Varga and H. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[4] H. F. Silverman and N. R. Dixon, "State constrained dynamic programming (SCDP) for discrete utterance recognition," in *Proc. ICASSP*, pp. 169–172, 1980.

[5] S. Kajita and F. Itakura, "Speech analysis and speech recognition using subband-autocorrelation analysis," *J. Acoust. Soc. Japan*, vol. 15, no. 5, pp. 329–338, 1994.

[6] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. America*, vol. 87, no. 4, pp. 1738–1752, 1990.