Feature Extraction Based On Zero-Crossings With Peak Amplitudes For Robust Speech Recognition In Noisy Environments

Conference Paper *in* Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing · June 1996

DOI: 10.1109/ICASSP.1996.540290 · Source: IEEE Xplore

Doh-Suk Kim
Dolby Laboratories, Inc.
31 PUBLICATIONS **710** CITATIONS

Soo-Young Lee
Korea Advanced Institute of Science and Technology
1,039 PUBLICATIONS **18,252** CITATIONS

FEATURE EXTRACTION BASED ON ZERO-CROSSINGS WITH PEAK AMPLITUDES FOR ROBUST SPEECH RECOGNITION IN NOISY ENVIRONMENTS

Doh-Suk Kim *Jae-Hoon Jeong* *Jae-Weon Kim* *Soo-Young Lee*

Department of Electrical Engineering
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

The Ensemble Interval Histogram (EIH) is an auditory model which can be used as a robust "front-end" for speech recognition systems. The utilization of multiple level-crossing detectors in the EIH provides frequency and intensity information, which may be useful for speech processing. Proper determination of the number of levels and the level values is very important for reliable performance of the system. In this paper, an analytic relationship is developed for variance and SNR of the level-crossing intervals as a function of the crossing level value, and a new feature extraction method based on zero-crossings with peak amplitudes is proposed for robust speech recognition in noisy environments. The proposed method not only can preserve intensity information, but also is robust to noise in estimating frequency information without the efforts to determine the level values and the number of levels. Experimental results show the robustness of the proposed method.

1. INTRODUCTION

Several auditory models have been investigated for robust speech recognitions in noisy environments [1, 2, 3]. Among them EIH (Ensemble Interval Histogram), proposed by Ghitza [1], is computationally efficient and robust enough to be used as a "front-end" for speech recognition systems. The EIH is composed of a bank of bandpass cochlear filters and an array of level-crossing detectors at the output of each cochlear filter. The filter bank models frequency selectivity at various points along a basilar membrane in a cochlea, and the array of level-crossing detectors models the ensemble of nerve fibers innervating a single inner hair cell. Each level represents a fiber of specific threshold connected to an inner hair cell, and the level values are uniformly distributed on a log scale over the positive amplitude range of the signal. This corresponds to the half-wave rectification observed in cilia attached to inner hair cells. Neural firings are simulated as the level-crossing events of the signal at the output of each bandpass filter. Inverse of time intervals between adjacent neural firings at given levels is coded as a frequency histogram, and the histograms for every level and filter channel are combined together to represent outputs of the EIH. From the viewpoint of signal processing, the utilization of multiple level-crossing detectors can provide intensity information, which may be useful for speech processing. However proper determination of the number of levels and the level values is very important to the performance. Unfortunately there is no theory available to determine those values.

In this paper, an analytic formula is developed for variance and SNR of the level crossing intervals due to additive white Gaussian noise as a function of the level-crossing value, and a new feature extraction method, zero-crossings with peak amplitudes (ZCPA), is developed for robust feature extractions in noisy environments.

2. STATISTICAL ANALYSIS OF THE LEVEL VALUES

Let's consider an input signal of the form

$$x(t) = \sum_{i=0}^{M-1} A_i \cos(\omega_i t + \theta_i) + gv(t), \tag{1}$$

where $v(t)$ is white Gaussian noise with zero mean and unit variance, and SNR (Signal-to-Noise Ratio) is determined by the parameter g. Let's assume that the filter characteristics of the filterbank is ideal bandpass, and the bandwidth of a filter is B. Suppose that each sinusoidal component in the input signal is separated by the filter bank, and the output of the k-th filter consists of a single sinusoid and bandpass noise as

$$x_k(t) = A_i \cos(\omega_i t + \phi_i) + gv_k(t). \tag{2}$$

As shown in Fig. 1, let's denote the upward level-crossing locations by t_n, i.e. $x_k(t_n) = l$, $n = 1, 2, ...$, the successive level-crossing intervals by $\tau_n = t_{n+1} - t_n$, and the perturbation in the level-crossing positions by r_n. The mean of the upward level-crossing interval can be approximated by $2\pi/\omega_i$ for $A_i \gg g$, using the dominant frequency principle [4]. From Fig. 1, one obtains

$$A_i \cos(\omega_i t_n + \phi_i) = l - V_n \tag{3}$$

$$A_i \cos(\omega_i (t_n - r_n) + \phi_i) = l, \tag{4}$$

where V_n is the instantaneous value of the bandpass noise at t_n. Now one substitutes $\alpha = \omega_i t_n + \phi_i$ and $\beta = \cos^{-1}(l/A_i)$, and obtains

$$\omega_i r_n = \alpha - \beta. \tag{5}$$

Since only the upward level crossings at positive level values are considered and $\omega_i r_n$ is assumed to be small, so $3\pi/2 \leq$

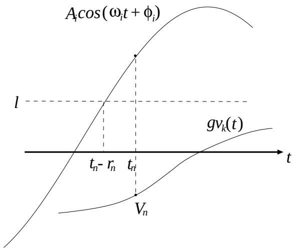

Figure 1. Signals and noise components at the output of a bandpass filter

$\beta \leq 2\pi$. By taking cosine function on Eq.(5), one obtains

$$\cos(\omega_i r_n) = \frac{l - V_n}{A_i} \cdot \frac{l}{A_i} + \left[\left(1 - \left(\frac{l - V_n}{A_i}\right)^2\right)\left(1 - \left(\frac{l}{A_i}\right)^2\right)\right]^{1/2}. \quad (6)$$

When $\omega_i r_n$ is small, Eq.(6) can be approximated further as

$$\begin{aligned} r_n^2 \;\; &\approx \;\; \frac{2}{\omega_i^2}\left[1 - R\left(R - \frac{V_n}{A_i}\right)\right] \\ &- \frac{2}{\omega_i^2}\left[\left(1 - \left(R - \frac{V_n}{A_i}\right)^2\right)\left(1 - R^2\right)\right]^{1/2} \quad (7) \end{aligned}$$

where $R = l/A_i$, $\sin\alpha = -[1 - ((l - V_n)/A_i)^2]^{1/2}$, and $\sin\beta = -[1 - (l/A_i)^2]^{1/2}$ are utilized.

Now, let's consider two successive level crossing perturbations r_n and r_{n+1}, and the perturbation of the corresponding level-crossing interval $|r_n - r_{n+1}|$. Variance of the interval perturbations is given as

$$\begin{aligned} \sigma_i^2 \;\; &= \;\; E\left\{|r_n - r_{n+1}|^2\right\} \\ &= \;\; E\left\{r_n^2\right\} + E\left\{r_{n+1}^2\right\}, \quad (8) \end{aligned}$$

where the random variables r_n and r_{n+1} are assumed to have zero mean and negligible correlation. From Eq.(7), one obtains

$$\begin{aligned} E\left\{r_n^2\right\} &\approx \frac{2}{\omega_i^2}\left(1 - R^2\right) \\ &- \frac{2}{\omega_i^2} E\left\{\left[\left(1 - R^2\right)\left(1 - \left(R - \frac{V_n}{A_i}\right)^2\right)\right]^{1/2}\right\} \quad (9) \end{aligned}$$

with $E\{V_n\} = 0$. The second term in the right side of Eq.(9) is approximated as

$$\begin{aligned} &E\left\{\left[\left(1 - R^2\right)\left(1 - \left(R - \frac{V_n}{A_i}\right)^2\right)\right]^{1/2}\right\} \approx \\ &\left[1 - R^2\right] - \left[\frac{1}{A_i^2}\left(1 + \frac{R^2}{1 - R^2}\right) \cdot \frac{1}{2}\left(\frac{B}{\pi} g^2\right)\right], \quad (10) \end{aligned}$$

where $E\left\{V_n^2\right\} = Bg^2/\pi$ and the 2nd order Taylor series expansion

$$\begin{aligned} E\{h(x)\} \;\; &= \;\; \int_{-\infty}^{\infty} h(x) f_X(x) dx \\ &\approx \;\; h(\eta_X) + h''(\eta_X)\frac{\sigma_X^2}{2}, \quad (11) \end{aligned}$$

is utilized for $V_n/A_i \ll 1$, where η_X and σ_X^2 are the mean and variance of the random variables respectively. From Eqs.(8),(9) and (10), the variance of the time intervals between two adjacent level-crossings are represented as

$$\begin{aligned} \sigma_i^2 \;\; &= \;\; \frac{(2B/\pi)g^2}{(\omega_i A_i)^2} \cdot \frac{1}{1 - (l/A_i)^2} \\ &= \;\; \sigma_{i_0}^2 \frac{1}{1 - (l/A_i)^2} \quad (12) \end{aligned}$$

where $\sigma_{i_0}^2$ is the variance in the case of zero-crossings, and

$$\begin{aligned} SNR_i \;\; &= \;\; \frac{2\pi/\omega_i}{\sigma_i} \\ &= \;\; \left(\frac{A_i \pi}{g}\right)\left[\frac{2\pi}{B}\left(1 - (l/A_i)^2\right)\right]^{1/2} \quad (13) \end{aligned}$$

is the SNR of the time intervals. The variance of the time interval between two adjacent level-crossings is minimum when l is zero. As the level value l increases for given A_i and g, the variance increases and SNR of the time intervals decreases. Therefore results with higher level values are more sensitive to additive noise.

3. ZERO-CROSSINGS WITH PEAK AMPLITUDES

Even though the higher value of level is sensitive to noise, a pilot experiment shows that the performance of the EIH with multiple level-crossing detectors is somewhat superior to that of the EIH with single level-crossing detector provided the level values were determined properly. This may come from the intensity information in the multiple level-crossing detectors. However the frequency information in higher levels may be incorrect in noisy condition as shown in section 2. Thus, a robust method which can estimate frequency information as well as intensity information even in noisy conditions is required. We propose a method to incorporate intensity information in the zero-crossing data.

The developed zero-crossings with peak amplitudes (ZCPA) method utilizes zero-crossing only, but peak amplitude between the two zero-crossing times is used as a weighting factor for the frequency component. The output of the ZCPA at time t is given as

$$y(t,i) = \sum_{channel} \sum_{k=1}^{K-1} \delta_{i j_k} f(A_k), \quad 1 \leq i \leq N, \quad (14)$$

where K is the number of upward zero-crossings at each filter channel, N is the number of frequency bins, j_k is the index of frequency bin computed using the k-th and $(k+1)$-th zero crossings, A_k is the peak amplitude between the k-th and $(k+1)$-th zero- crossings, and δ_{ij} is a Kronecker delta. The firing rate of auditory nerve fibers saturates above a certain stimulus intensity [5]. $f()$ is a monotonic function which implements this saturating nonlinearity. In connection with human auditory system, log function is used in our experiments. The frequency component is found by zero-crossing intervals only, and the frequency bin of histogram is increased by an amount of $f(A_k)$. The use of zero-crossings in finding frequency components makes it more robust to noise, and the spectral intensity information is also incorporated. It is also free from complications to determine the level values and the number of levels in the EIH.

From the signal processing viewpoints the ZCPA utilizes zero-crossings of the signal, and a zero-crossing based signal representation is valid for band-limited signals. In the case of periodic band-limited signal, they can be recovered within a scale factor from their real zeros, and ratios between any of two DFT coefficients can be computed. In the case of aperiodic signals, they can only be recovered approximately [6]. Sreenivas and Niederjohn [7] proposed an algorithm to analyze spectrum based on the noise threshold for the detectability of a sinusoid which was derived from the statistical properties of the zero-crossing intervals at the output of a filter bank, and showed that their algorithm was robust to noise. Comparing the ZCPA with Sreenivas' method, Sreenivas' method utilizes the first and second order statistics of zero-crossing intervals while the ZCPA is the probability density function of the inverse of zero-crossing intervals and incorporates intensity information in a nonlinear manner.

4. EXPERIMENTS

4.1. Experimental Conditions

Speaker-independent word recognition experiments were conducted to evaluate the robustness of performances using word utterances made by 20 speakers. The vocabulary consists of 75 phonetically-balanced Korean words which are mutually very confusable. Each speaker uttered the words once in a quiet office environment via a Sennheiser HMD224X headset. The utterances were sampled by 16 kHz sampling rate with 16 bit resolution. The data were divided into 4 sets, 5 speakers each. Three sets are used as references, and the other set is used as test patterns. By changing the combination of the sets, one obtains 4 different results for each experiment, and recognition accuracy is averaged over the 4 experiments to normalize the sensitivity of the results to the data sets. To evaluate noise robustness of the features, white Gaussian noise is added to isolated word utterances to be used as test patterns at various SNR's. The gain of the noise is adjusted to make the desired SNR, where SNR is the energy ratio of the whole utterance to noise. The filterbank used in the EIH and the ZCPA is the cochlear filter [8] with 20 bands where center frequencies are distributed from 200 to 5000 Hz according to the frequency-position relationship [9]

$$F = A\left(10^{ax} - 1\right), \qquad (15)$$

where F is frequency in Hz, x is the normalized distance along the basilar membrane with value from 0 to 1. The appropriate constants for the human cochlea are $A = 165.4$ and $a = 2.1$. The length of analysis windows is ten times of the inverse of the center frequency to get fine frequency resolution in lower frequency side and fine time resolution in higher frequency side. Frequency bins divide the frequency range [0, 5000] Hz into 18 regions according to the bark scale. Nearest neighbor classifier with trace-segmentation [10] is used for the test evaluation.

4.2. Results of the EIH

Fig. 2 compares recognition rates of the EIH with several different number of levels and different level values. Level values of the EIH are uniformly distributed on a log2 scale over the positive amplitude range of the signal, and the first digit following the "L" denotes the number of levels used in the EIH. The second digit represents the range of the crossing level values. Higher values of the second digit mean that the crossing level values are distributed in lower range. For example, the highest level value of L3.1 is four times higher than that of L3.3. And the same value of the second digit for the EIH's with different number of levels means that the highest level values are same. For example the highest level value of L5.7 is same as that of L3.7 and L7.7, and L5.7 has additional 2 lower levels compared with L3.7.

Recognition rates of the EIH tend to increase as the level values are lowered. However if the level values are too low, the information obtained from some lower levels will be duplicated, and recognition rates decrease to some extent.

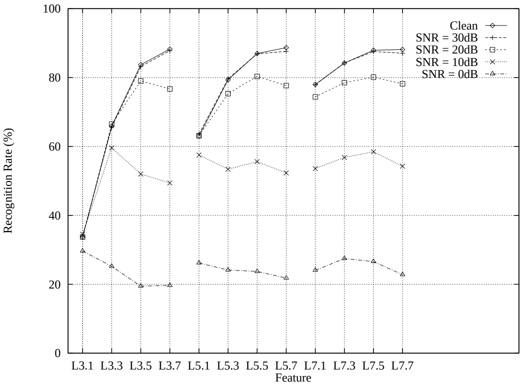

Figure 2. Recognition rates of the EIH's with multiple levels at various SNR's. The first digit following "L" denotes the number of levels used in the EIH, and higher values of the second digit mean that the crossing level values are distributed in lower range.

4.3. Results of the ZCPA and Comparison with Other Features

Fig. 3 compares recognition rates of the ZCPA with the EIH's and LPC-derived cepstrum at various SNR's. Comparison of recognition rates of the ZCPA and other features at various SNR's is shown in Fig. 3. LPC-derived cepstrum is one of the most widely used features in speech recognition tasks. The number of cepstral coefficients was varied to be 12 and 18, and we chose 18 cepstral coefficients which produced higher recognition rate. Recognition rate of the LPC-derived cepstrum decreases severely as noise level is increased. If we compare the ZCPA with the EIH(L7.1) of which the highest level value is set to be at 6.4% of the possible maximum value of the signal at the output of each channel, recognition rates of the ZCPA are higher by 10.3%, 8.9%, 7.2%, and 11.1% than that of the EIH at clean, 30dB SNR, 20dB SNR, and 10dB SNR, respectively. Comparing the ZCPA with the EIH(L7.5) of which the level values are set to proper values (the highest level value is set to be at 0.4% of the possible maximum value of the signal), the differences in recognition rates between the EIH and the ZCPA are less than 2% above 20dB SNR. Recognition rate of the ZCPA is 6.2% higher than that of the EIH(L7.5) at 10dB SNR. This demonstrates low sensitivity of the ZCPA to additive random noise.

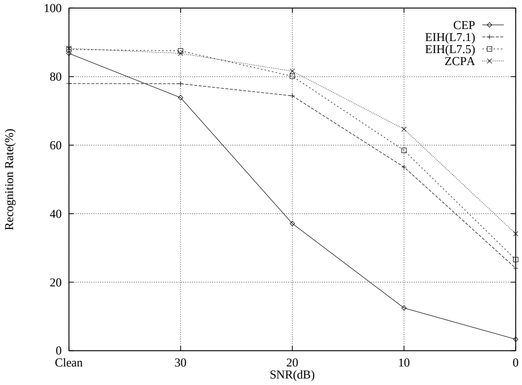

Figure 3. Comparison of recognition rates of the ZCPA with other features at various SNR's

5. CONCLUSION

The EIH is an auditory model which can be used as a robust "front-end" for speech recognition systems. The utilization of multiple level-crossing detectors in the EIH provides frequency and intensity information of input signal which may be useful features, and proper determination of the number of levels and the level values is very important. However there is no theory available to determine those values. In this paper it is shown theoretically that the variance of the level-crossing intervals increases as the level value is increased in presence of additive white Gaussian noise. Also a new feature extraction method based on zero-crossings with peak amplitudes (ZCPA) is introduced in which the intensity information of the stimulus is incorporated by the peak detection and saturating nonlinearity, and the utilization of zero-crossings in estimating frequency makes it more robust to noise without complications of determining level-crossing values. Speaker-independent word recognition experiment demonstrates the robustness of the proposed feature extraction method over the conventional one.

6. ACKNOWLEDGMENT

Speech database used in this work is distributed by Communications Research Laboratory, Department of Electrical Engineering, KAIST. Also, the authors would like to thank Professor Xuelong Zhu of Tsinghua University for his useful comments.

REFERENCES

[1] O. Ghitza, "Auditory models and human performances in tasks related to speech coding and speech recognition," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 1, part II, pp. 115–132, 1994.

[2] S. Seneff, "Pitch and spectral estimation of speech based on auditory synchrony model," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp. 36.2.1–36.2.4, 1984.

[3] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech and Audio Processing*, vol. 2, no. 3, pp. 421–435, 1994.

[4] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, pp. 1477–1493, November 1986.

[5] M. B. Sachs and P. J. Abbas, "Rate versus level functions for auditory nerve fibers in cats: Tone burst stimuli," *J. Acoust. Soc. America*, vol. 56, no. 6, pp. 1835–1847, 1974.

[6] S. M. Kay and R. Sudhaker, "A zero crossing-based spectrum analyzer," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 96–104, Feb. 1986.

[7] T. V. Sreenivas and R. J. Niederjohn, "Zero-crossing based spectral analysis and SVD spectral analysis for formant frequency estimation in noise," *IEEE Trans. Signal Processing*, vol. 40, no. 2, pp. 282–293, 1992.

[8] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. Signal Processing*, vol. 39, no. 12, pp. 2573–2592, 1991.

[9] D. Greenwood, "A cochlear frequency-position function for several species–29 years later," *J. Acoust. Soc. America*, vol. 87, no. 6, pp. 2592–2650, 1990.

[10] H. F. Silverman and N. R. Dixon, "State constrained dynamic programming (SCDP) for discrete utterance recognition," in *Proc. IEEE Int. Conf. on Acoust., Speech, and Signal Processing*, pp. 169–172, 1980.