Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments

Doh-Suk Kim, *Associate Member, IEEE,* Soo-Young Lee, *Member, IEEE,* and Rhee M. Kil, *Member, IEEE*

***Abstract*—This paper presents a new approach to an auditory model for robust speech recognition in noisy environments. The proposed model consists of cochlear bandpass filters and nonlinear operations in which frequency information of the signal is obtained by zero-crossing intervals. Intensity information is also incorporated by a peak detector and a compressive nonlinearity. The robustness of the zero-crossings in spectral estimation is verified by analyzing the variance of the level-crossing intervals as a function of the crossing level values. Compared with other auditory models, the proposed auditory model is computationally efficient, free from many unknown parameters, and able to serve as a robust front-end for speech recognition in noisy environments. Experimental results of speech recognition demonstrate the robustness of the proposed method in various types of noisy environments.**

***Index Terms*—Auditory model, noise robustness, speech recognition, zero-crossing.**

I. Introduction

AUTOMATIC speech recognition (ASR) is one of the leading technologies serving as a man-machine interface for real-world applications. In general, the performance of an ASR system is usually degraded when there exist environmental mismatches between training and test phases. One type of mismatch in real environments is the various kinds of background noises that affect the feature extraction stage in an ASR system. In this sense, the front-end for robust speech recognition requires to reduce redundancy and variability as well as the ability to capture important cues of speech signals, even in noisy environments. One of the most widely used feature representations are cepstral coefficients derived from linear predictive coding (LPC) in which the speech signal is assumed to be the output of the all-pole linear filter simulating the vocal tract of a human being. The ASR systems with LPC-derived cepstrum work well in clean environments, but speech recognition performance is severely degraded in noisy environments.

Manuscript received January 8, 1997; revised December 18, 1997. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Kuldip K. Paliwal.

D.-S. Kim was with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA. He is now with Samsung Advanced Institute of Technology, Suwon 440-600Korea.

S.-Y. Lee is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Taejon 305-701, Korea.

R. M. Kil is with the Division of Basic Sciences, Korea Advanced Institute of Science and Technology, Taejon 305-701, Korea.

Publisher Item Identifier S 1063-6676(99)00174-1.

There has been considerable research devoted to the modeling of the functional roles of peripheral auditory systems. Seneff [1], [2] suggested a generalized synchrony detector (GSD) to identify formant peaks and periodicities of the speech signal. Hunt and Lefebvre [3] performed recognition experiments on noisy speech using a dynamic time warping (DTW) recognizer, and showed noise-robustness of the GSD. Perceptual linear prediction (PLP) analysis method [4], [5] is a perception-based technique in which the speech spectrum is transformed to the auditory spectrum by several perceptually motivated relationships before performing conventional linear prediction (LP) analysis. The robustness of the PLP analysis to additive noise was reported in [6]. Subband-autocorrelation (SBCOR) analysis technique [7], [8] was suggested to extract periodicities present in speech signals by computing autocorrelation coefficients of subband signals at specific time-lags, and was shown to outperform the smoothed group delay spectrum for speech recognition tasks under noisy environments.

Although computational auditory models have been shown to outperform conventional signal processing techniques, especially in noisy environments, modeling peripheral auditory systems is still a difficult problem. First, studying an auditory model requires interdisciplinary research, including physiology, psychoacoustics, physics, and electrical engineering. Second, little is known about the exact mechanism of the auditory periphery for detailed construction of the model. Since the auditory model usually involves multistage nonlinear transformations, analytical treatments are intractable, and most auditory models rely heavily on experiments, even though there have been some efforts to analyze auditory models [9]–[12]. Furthermore, auditory models require careful determination of many free parameters and much computation time, which makes it difficult for them to be widely used in speech recognition systems.

In this paper, a robust feature extraction method motivated by a mammalian auditory periphery is introduced to extract reliable features from speech signals, even in noisy environments. The developed auditory model is computationally efficient and free from many unknown parameters compared with other auditory models. In addition, it is shown both analytically and experimentally that the proposed method can maintain reliable features of speech signals. We also provide further performance improvements by incorporating conventional techniques and performance comparisons with other front-ends. This paper is organized as follows. Section II

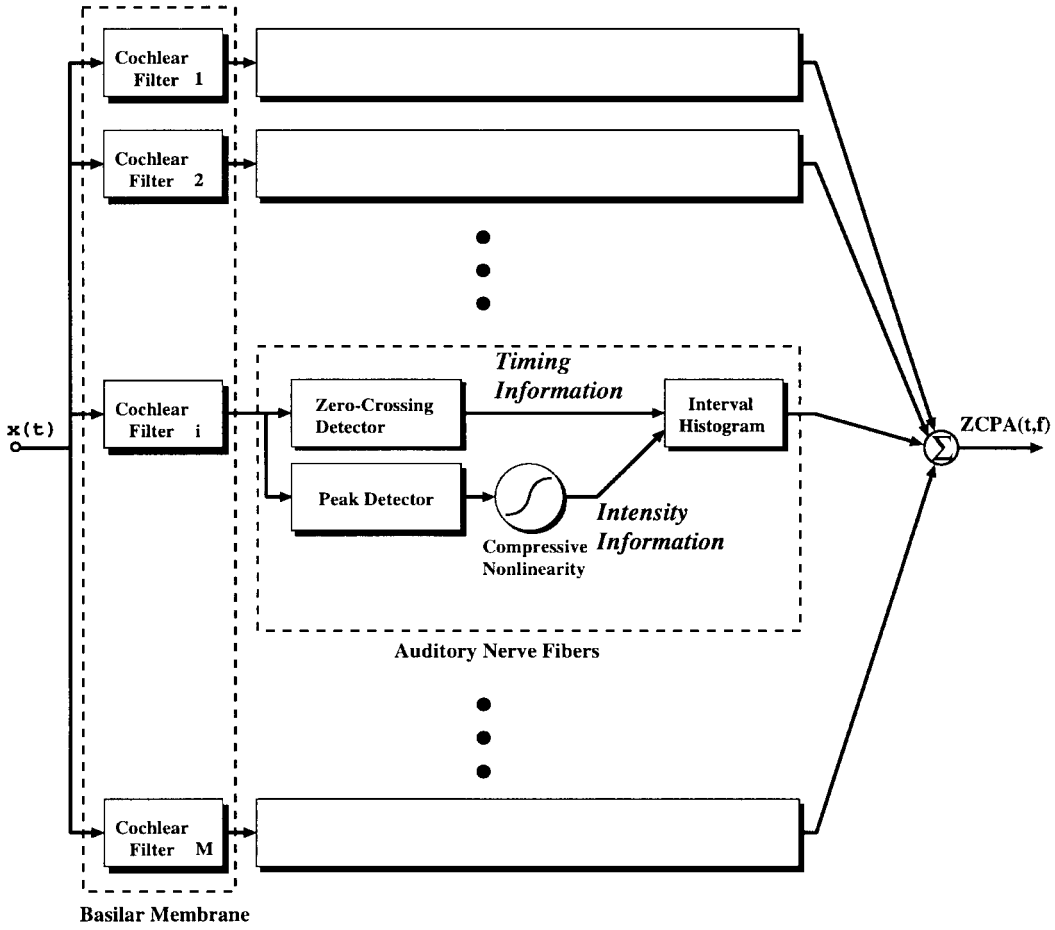


Fig. 1. Block diagram of the zero-crossings with peak-amplitudes (ZCPA) model.

presents the developed auditory model for robust feature extraction. A comparative study of the developed model with the EIH's, as well as several properties of the model, are provided in Section III. Also, a statistical analysis of the effect of level values on level-crossing intervals is provided. This analysis verifies the superiority of the zero-crossings to the level-crossings in estimating frequency when the signal is corrupted by noise. Experimental results in Section IV demonstrate the robustness of the proposed model. In Section V, improved performance of the auditory model is presented, and comparative evaluations with other front-ends are provided. Finally, conclusions are given in Section VI.

II. Auditory Processing of Acoustic Signals

Acoustic signals are transformed into perceptual representations in auditory systems [13]. Speech signals are transformed into mechanical vibrations of the eardrum at the outer ear, and then transmitted to the cochlea at the inner ear through the middle ear, which contains a complicated linkage of bones. The role of the middle ear is known as impedance matching between the outer and inner ear. Speech signals transmitted through the oval window at the base of the cochlea are converted into traveling waves of the basilar membrane as a response to liquid pressure inside the cochlea. The site of maximum excursion of the traveling wave on the basilar membrane is dependent on the frequency. The mechanical vibrations of the basilar membrane are then transduced into neural firings of the auditory nerve fibers as a response to the bending movement of the cilia attached to inner hair cells, which are organized along the length of the basilar membrane. The transduced neural firings are transferred to the auditory cortex via auditory pathways, which are composed of many cell groups. It is suspected that there are complex feature detectors in these higher stages of the auditory system, but little is known of their functional roles.

Fig. 1 represents a block diagram of the suggested auditory signal processing, zero-crossings with peak amplitudes (ZCPA), which consists of a bank of bandpass cochlear filters and nonlinear stages [14]. The bank of cochlear filters simulates the basilar membrane, common in most auditory models. The nonlinear stage performs a series of nonlinear signal processings to simulate the transformation of the mechanical vibrations of the basilar membrane into neural firings of auditory nerve fibers.

A. Basilar Membrane

Mechanical displacement of the basilar membrane is simulated by a bank of cochlear filters. The cochlear filterbank represents frequency selectivity at various locations along the basilar membrane in a cochlea. For implementation, Kates'

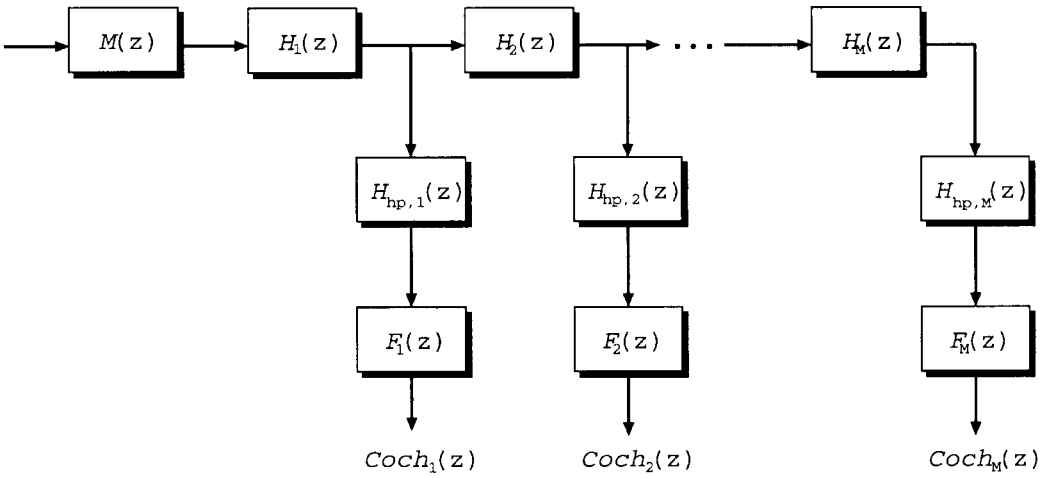


Fig. 2. Block diagram of Kates' cochlear filters.

traveling wave filters [15] without adaptive Q adjustments are used in this paper. A block diagram of the cochlear filters composed of traveling wave filters is depicted in Fig. 2. This filter section, implemented by a cascade of linear filters, simulates a combination of traveling waves progressing through the cochlea and additional filtering caused by the coupling of the tectorial and basilar membranes. This was designed to reproduce the basic mechanical and neural behaviors of the cochlea, and the intended application is the study of auditory impairment. His model is based on the analog model developed by Lyon and Mead [16] that represents wave propagation along an active cochlear partition. $M(z)$ is a second-order highpass filter having a resonance frequency of 350 Hz and a Q of 0.7, and simulates the behavior of the human middle ear. $H_{hp,i}(z)$ is a one-pole highpass filter that models the pressure-to-velocity transformation, and $F_i(z)$ is a notch filter by which the total response shows two resonance frequencies, which coincides with biological observations. $H_k(z)$ is a single section of the traveling wave filter, which provides a gain for frequencies near the resonance frequency of the filter, attenuation for frequencies above the resonance frequency, and unity gain for frequencies below the resonance frequency. Details of the form of these transfer functions and design procedures are described in [15].

Wave propagates from the base to the apex of the cochlea, and high frequencies show maximum excursion near the base while low frequencies near the apex. Thus, the resonance frequency of $H_k(z)$ decreases as the index k is increased. Resonance frequencies of $H_k(z)$'s are distributed according to the frequency-position relationship,

$$F = A(10^{ax} - 1) \tag{1}$$

where F is frequency in Hz, and x is the normalized distance along the basilar membrane with a value of from zero to one [17]. The appropriate constants for the human cochlea, $A = 165.4$ and $a = 2.1$, are used in this work.

As a result, the transfer function of each cochlear filter is expressed as

$$\text{Coch}_i(z) = M(z)H_{hp,i}(z)F_i(z)\prod_{k=1}^{i} H_k(z) \tag{2}$$

and the frequency responses of the 20 filters are shown in Fig. 3. The magnitude response shows an asymmetric property: each response shows a long tail on the lower frequency side, while the slope on the higher frequency side is very steep. Also, the higher frequency filter has a sharper resonance than does the lower frequency filter.

B. Neural Transduction—Auditory Nerve Fibers

How the acoustic signals are represented and coded in the mammalian auditory system has been an important issue for several decades, and there are two parallel theories of the auditory nerve representations of acoustic stimulus.

- *Rate Representation:* Since each ascending fiber innervates an inner hair cell located at a specific position on the basilar membrane, which is tuned to a certain frequency, the auditory nerve fibers can be characterized by their best frequencies. That is, the auditory nerve fibers are tonotopically organized, and they convey the spectral contents of the stimulus by an average firing rate.
- *Temporal Representation:* This interpretation is based on the observation that the auditory nerve fibers are capable of locking, or synchronizing, to harmonics of stimuli that correspond to formants of speech signals [18].

Many research papers suggest that the average firing rate is insufficient to represent speech information, and that the temporal information of the firing patterns should be included [19]–[21]. According to the temporal representation, auditory nerve fibers tend to fire in synchrony with the stimulus, and this synchronous firing pattern contains useful frequency information. In the proposed model, a 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 the time interval between adjacent neural firings is collected

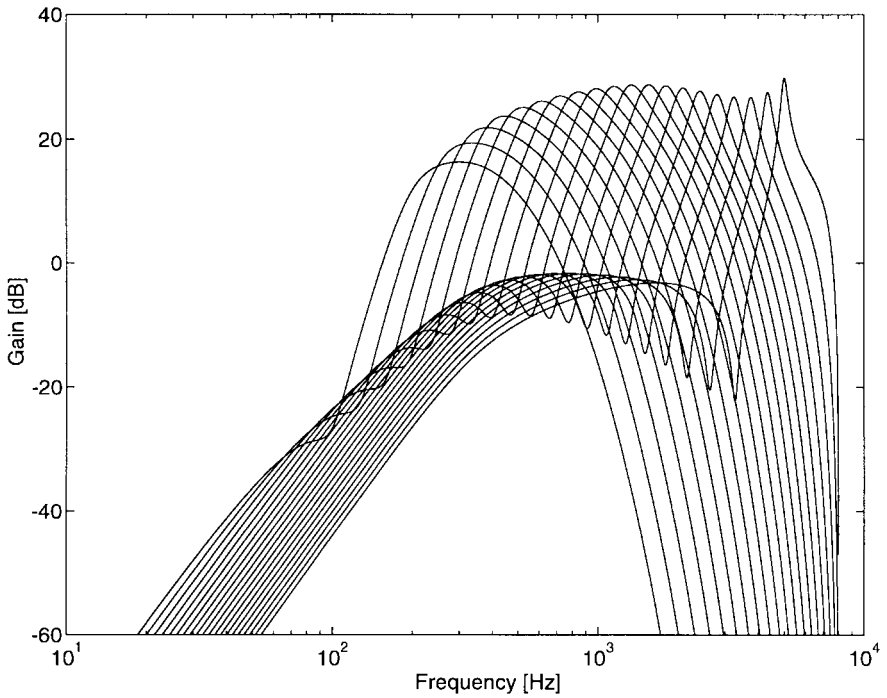


Fig. 3. Frequency response of 20 cochlear filters.

and 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 firing rate. The histograms across all filter channels are combined to represent the output of the auditory model.

Let us denote the output signal of the kth bandpass filter by $x_k(n)$ and the frame of $x_k(n)$ at time m by $x_k(n; m)$ described as

$$x_k(n; m) = x_k(n)w_k(m-n), \qquad k = 1, \cdots, N_{ch} \quad (3)$$

where $w_k(n)$ is a window function of finite length, and N_{ch} is the number of the channel, i.e., the number of cochlear filters in Fig. 1. Further, let us denote Z_k by the number of upward-going zero-crossings of $x_k(n; m)$, and P_{kl} by the peak amplitude between the lth and $(l+1)$th zero-crossings of $x_k(n; m)$, respectively. Then the output of the ZCPA at time m is described as

$$y(m, i) = \sum_{k=1}^{N_{ch}} \sum_{l=1}^{Z_k-1} \delta_{ij_l} g(P_{kl}), \qquad 1 \le i \le N \quad (4)$$

where N is the number of frequency bins, and δ_{ij} is the Kronecker delta. For each channel, the index of frequency bin, j_l, is computed by taking the inverse of the time interval between the lth and $(l+1)$th zero crossings, for $l = 1, \cdots, Z_k-1$. Then the value of the frequency histogram at the frequency bin, j_l, is increased by $g(P_{kl})$. The histograms across all channels are combined to obtain the output of the ZCPA. $g(\,)$ is a monotonic function simulating the relation between the stimulus intensity and the degree of phase-locking of the auditory nerve fibers. As a candidate of $g(\,)$, the logarithmic function of the form,

$$g(x) = \log(1 + x) \quad (5)$$

is used in connection with the human auditory system, where the constant term is added to prevent negative contribution to the frequency bin. As for the frequency bin allocation, we used a linear bin allocation for speech analysis tasks in Section III, or a bin allocation according to the bark scale [22] for recognition tasks.

In the suggested auditory model, the length of window function, L_k, is set to $10/F_k$ to capture about ten periods of the signal at each channel, provided that the signal is a sinusoid with a frequency equal to the characteristic frequency of the channel, F_k [23]. Thus, the window lengths become long for low frequencies, and short for high frequencies. As a result, frequency resolutions are finer, while time resolutions are poorer at lower frequencies, and vice versa at higher frequencies. This property is consistent with psychoacoustic observations.

The operation of the ZCPA is significantly different from conventional signal processing schemes in that the temporal frequency and intensity information of one period of the signal is measured, and then an accumulation of the temporal information is carried out to obtain the final output. During this operation, the bandpass signal is first converted to point process by the zero-crossing detector in the ZCPA. If the bandlimited signal can be approximated as an algebraic polynomial function of order n, then this can be further decomposed as a product expansion in which each zero component appears in each term. Assuming all the zeros of the polynomial are real, the signal can be represented by their zero-crossings and a coefficient of the highest order term of the polynomial. In the case of periodic bandlimited signals, they can be recovered within a scale factor from their real zeros, and ratios between any of two discrete Fourier transform (DFT) coefficients can be calculated. In the case of aperiodic signals, they can only be recovered approximately [24]. Sreenivas and

Niederjohn [25] proposed an algorithm to analyze the spectrum based on the statistics of the zero-crossings.

III. Properties of the Auditory Model

A. Comparison with the EIH Model

The ensemble interval histogram (EIH), an auditory model proposed by Ghitza [23], [26], is computationally efficient and robust for use 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 attached to the output of each cochlear filter. Both the EIH and ZCPA utilize zero or level crossings for frequency information. However, unlike the ZCPA model, the EIH model utilizes multiple level-crossing detectors with different level values for intensity information. The array of level-crossing detectors represents the ensemble of nerve fibers innervating a single inner hair cell. Each level is associated with a fiber of a specific threshold attached to an inner hair cell, and the level values are uniformly distributed on the log scale over the positive amplitude range of the signal. Each level-crossing detector detects upward-going level-crossing points of the signal, and intervals between successive crossing points are contributed to a frequency histogram. Thus, there are multiple numbers of timing information obtained from the independent level-crossing detectors, even in the same channel. As the amplitude of the signal is increased, more levels will detect their own crossing points, and the amount of contribution to the frequency histogram will also be increased. Thus, the utilization of multiple level-crossing detectors provides intensity information of the signal, which may be one of the useful cues for automatic speech recognition.

In implementing the EIH, one has to consider carefully to determine such parameters as level values and number of levels. Qualitatively speaking, if level values are near zero, the intensity information of the signal will not be well represented. On the other hand, some of level-crossing detectors with high level values may be useless if level values are too high when compared with the signal level. Further, even if a level-crossing detector with a high level value captures the crossing points at the high amplitude range of the signal, timing information at the higher level becomes incorrect in noisy environments, as will be shown in Section III-B. In this case inaccurate spectral contents are contributed to the frequency histogram, which make the EIH model sensitive to noise. Thus, proper determination of the number of levels and the level values is very important for reliable performance of the EIH model especially in noisy environments. However, there is no method available to determine those values, except by trial-and-error.

On the other hand, the proposed ZCPA model utilizes only zero-crossings for frequency information. The use of zero-crossings in estimating frequency makes it more robust to noise without serious efforts to determine free parameters associated with the level. This property is described in Section III-B. Furthermore, intensity information is also incorporated by using the peak amplitude as a weighting factor to the frequency bin estimated from zero-crossings. The sensitivity of ZCPA model to the peak amplitude estimation in noisy condition is addressed in the Appendix.

B. The Effect of the Level Value on Level-Crossing Intervals

In our approach, the spectrum of the signal is estimated from the intervals of the point process generated by level-crossing events. This method is severely influenced by the level value, especially when the signal is corrupted by additive noise. As the level value is increased, spectral estimation becomes sensitive to the additive noise. To verify the above statement, it is sufficient to show that the variance of the time interval perturbation between two adjacent level-crossing points increases as the level value is increased.

Consider an input signal of the form

$$x(t) = \sum_{i=0}^{M-1} A_i \cos(\omega_i t + \theta_i) + A_g v(t) \tag{6}$$

where $v(t)$ is bandlimited white Gaussian noise with a rectangular power spectrum of bandwidth W [rad/s] and has zero mean and unit variance. Signal-to-noise ratio (SNR) is determined by the parameter A_g for fixed A_i's. Suppose that $x(t)$ is filtered by a bank of ideal bandpass filters, of which the bandwidths are all B, and each sinusoidal component in the input signal is separated by the filterbank. The output of each filter can be considered to be of one of two types: 1) a signal with bandpass noise only and 2) a signal with a single sinusoid in addition to bandpass noise. If one considers the latter case, the output of the kth filter can be represented as

$$x_k(t) = A_i \cos(\omega_i t + \phi_i) + A_g v_k(t). \tag{7}$$

As shown in Fig. 4, let us denote the upward-going level-crossing locations of the bandpass signal by t_n, i.e., $x_k(t_n) = l$ for $n = 1, 2, \cdots$, the successive level-crossing intervals by $\tau_n = t_{n+1} - t_n$, and the perturbation in the level-crossing positions by r_n. From Fig. 4, one obtains

$$A_i \cos(\omega_i t_n + \phi_i) = l - V_n \tag{8}$$

and

$$A_i \cos(\omega_i (t_n - r_n) + \phi_i) = l \tag{9}$$

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

$$\omega_i r_n = \alpha - \beta. \tag{10}$$

In our case, only the upward level crossings at positive level values are considered. In other words, $3\pi/2 \le \beta \le 2\pi$. Then from (10)

$$\cos(\omega_i r_n) = \frac{l - V_n}{A_i} \cdot \frac{l}{A_i} + \left[\left(1 - \left(\frac{l - V_n}{A_i} \right)^2 \right) \left(1 - \left(\frac{l}{A_i} \right)^2 \right) \right]^{1/2}. \tag{11}$$

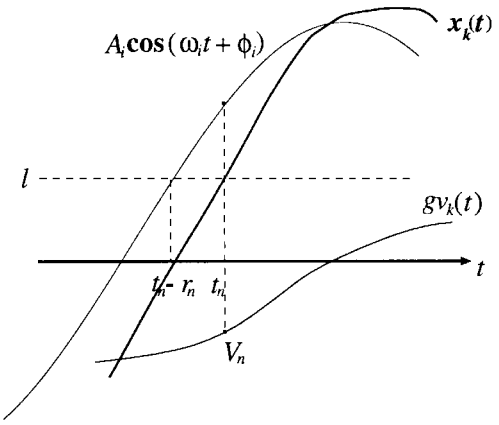


Fig. 4. Signal and noise component at the output of a bandpass filter.

When $\omega_i r_n$ is small, (11) can be approximated as

$$r_n^2 \approx \frac{2}{\omega_i^2}\left[1-R\left(R-\frac{V_n}{A_i}\right)\right] - \frac{2}{\omega_i^2}\left[\left(1-\left(R-\frac{V_n}{A_i}\right)^2\right)(1-R^2)\right]^{1/2} \quad (12)$$

where $R = l/A_i$.

Now, let us 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

$$\begin{aligned}\sigma_i^2 &= E\{|r_n - r_{n+1}|^2\} \\ &= E\{r_n^2\} + E\{r_{n+1}^2\}\end{aligned} \quad (13)$$

where the random variables r_n and r_{n+1} are assumed to have zero mean and negligible correlation. From (12) and $E\{V_n\} = 0$

$$E\{r_n^2\} \approx \frac{2}{\omega_i^2}(1-R^2) - \frac{2}{\omega_i^2}E\left\{\left[(1-R^2)\left(1-\left(R-\frac{V_n}{A_i}\right)^2\right)\right]^{1/2}\right\}. \quad (14)$$

The second term in the right side of (14) is approximated as

$$E\left\{\left[(1-R^2)\left(1-\left(R-\frac{V_n}{A_i}\right)^2\right)\right]^{1/2}\right\} \approx [1-R^2] - \left[\frac{1}{A_i^2}\left(1+\frac{R^2}{1-R^2}\right)\cdot\frac{1}{2}\left(\frac{B}{W}A_g^2\right)\right] \quad (15)$$

where $E\{V_n^2\} = BA_g^2/W$ and the second-order Taylor series expansion

$$\begin{aligned}E\{h(x)\} &= \int_{-\infty}^{\infty} h(x) f_X(x)\,dx \\ &\approx h(\eta_X) + h''(\eta_X)\frac{\sigma_X^2}{2}\end{aligned} \quad (16)$$

is utilized for $V_n/A_i \ll 1$, where $f_X(x)$, η_X, and σ_X^2 are the probability density function, mean, and variance of the random variable X, respectively. From (13)–(15), the variance of the time interval perturbation between two adjacent level-crossings is represented as

$$\begin{aligned}\sigma_i^2 &= \frac{2A_g^2 B/W}{(\omega_i A_i)^2}\cdot\frac{1}{1-(l/A_i)^2} \\ &= \sigma_{i_0}^2\frac{1}{1-(l/A_i)^2}\end{aligned} \quad (17)$$

where $\sigma_{i_0}^2$ is the variance in the case of zero-crossings [14]. The variance of the time interval perturbation between two adjacent level-crossings has a minimum value of σ_{i_0} for $l = 0$, which is in agreement with the result of [25]. As the level value l increases for fixed A_i and A_g, the variance increases. This implies that higher level values result in higher sensitivity in the estimated intervals and frequencies.

Fig. 5 illustrates the effects of the level value of a single level crossing detector on the result of spectral estimation. The input signal is composed of a sinusoid at 0.2π and additive white Gaussian noise, and 20 statistical realizations of noise are used for each superimposed spectra. The input signal is filtered by a bandpass filter, of which both the center frequency and the bandwidth are 0.2π, and successive upward level crossing intervals are measured and coded as a frequency histogram, which is composed of 129 linearly spaced frequency bins. The left three columns in Fig. 5 are generated by a level-crossing detector with different level values, and the rightmost plots are for magnitudes of the DFT. The ratio of the level value to the amplitude of the signal is denoted by R. Level value does not affect the result of spectral estimation at relatively clean conditions, e.g., SNR $= 20$ dB. As the ratio of the level value to the amplitude of the signal increases, the spectral estimation becomes rough at noisy conditions. This simulation result is consistent with the result derived in (17). In addition, Fig. 5 shows the usefulness of the zero-crossing, even at heavily noisy conditions, even though (17) is derived for $A_i \gg g$. Also, the spectrum based on zero-crossings has a tendency to enhance a dominant signal component, and to suppress the noise components adjacent to the signal components. This property may be explained by the dominant frequency principle [27], which states that the property of zero-crossings is dominated by the dominant component that contains more power than others, and contributes to the noise-robust property of the developed auditory model.

Fig. 6 compares the smoothed spectrum of LPC and the ZCPA for three different vowels, /a/, /e/, and /o/, uttered by a male speaker. White Gaussian noise is added to the utterances to make the desired SNR levels. In every panel, the smoothing is performed by a cepstral curve fitting by which spectral fine structures from a pseudoperiodic glottal source are eliminated and only the spectral envelope representing resonant characteristics of the vocal tract is shown. The upper part of Fig. 6 shows the plots from the LPC-derived cepstrum with an LPC and cepstral order of 20, while the lower part of Fig. 6 shows the plots from the 20th-order cepstral curve fit of the ZCPA. Table I shows the log spectral distance between clean and noisy spectra, normalized by the dynamic range of the spectrum, as a function of SNR. Both formant

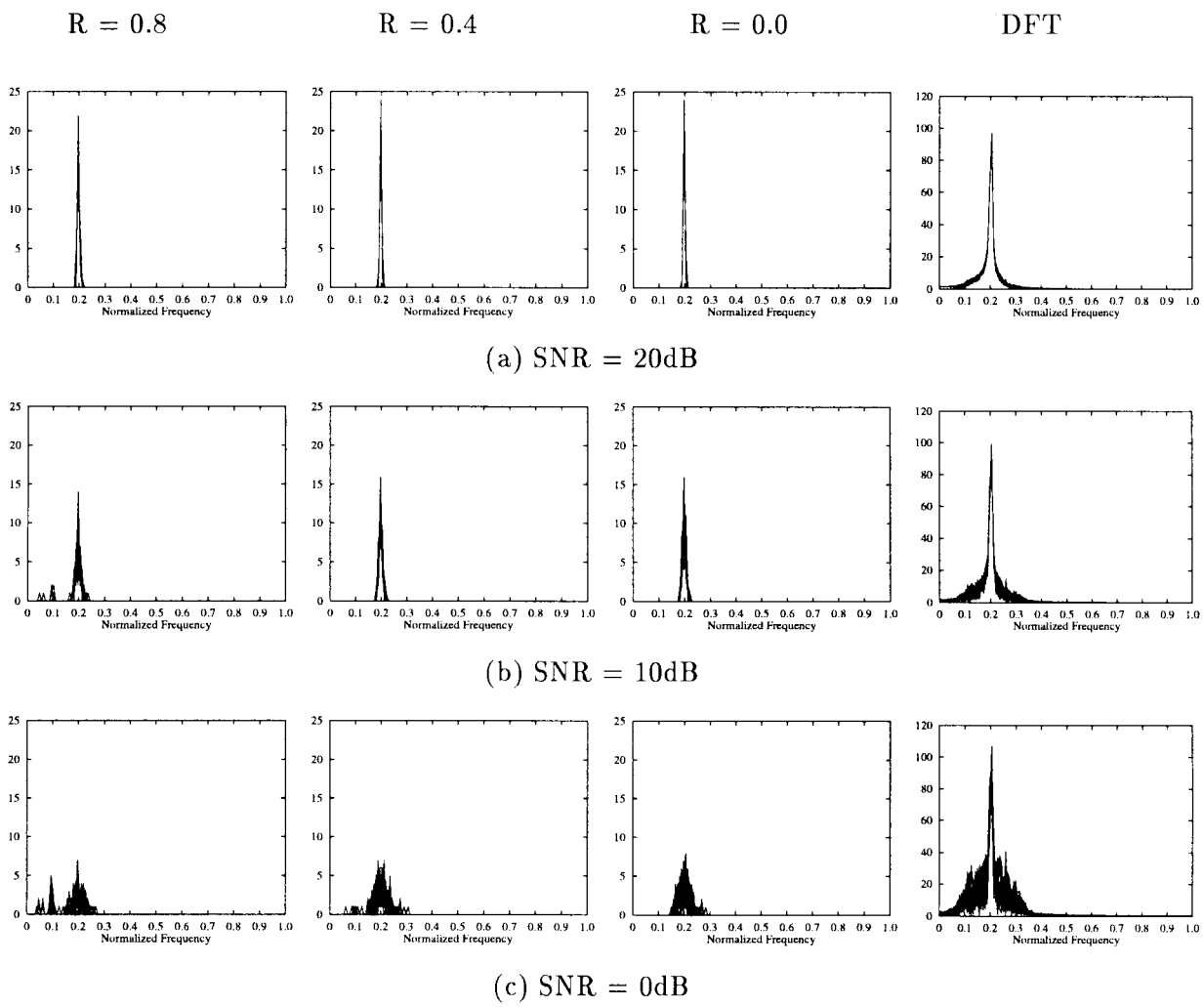


Fig. 5. Superimposed spectral plots with different SNR's. Left three columns are generated by a level-crossing detector with different level values, and plots in the rightmost column are the magnitude of DFT. The signal is composed of a sinusoid at 0.2π and additive white Gaussian noise, and 20 statistical realizations of noise are used for each superimposed plot. $R = l/A_i$ is the ratio of the level value to the amplitude of the signal.

frequencies and the detailed characteristics of each vowel are well preserved even in severely noisy conditions for the ZCPA, while the LPC spectrum deviates heavily from the original one when SNR's is below 20 dB. These results show that the ZCPA representation is much more robust to noise than the LPC spectrum.

IV. Recognition Results

A. Experimental Conditions

Speaker-independent word recognition experiments were conducted to evaluate the robustness of the proposed feature extraction algorithm under noisy environments. The speech data consist of 75 phonetically balanced Korean words made by 20 male speakers [28]. Each speaker uttered the words once in a quiet office environment via a Sennheiser HMD224X lip microphone. The utterances were sampled at 16 kHz sampling rate with 16-b resolution. The data were divided into four sets of five speakers each. Three sets were used as references, and the other set was used as test patterns. By changing the combination of the sets, four different recognition rates were obtained and averaged to reduce the sensitivity of the results to data sets. To evaluate noise robustness of the features, white Gaussian noise was 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. This SNR measurement coincides with the global SNR [26]. For the computation of the ZCPA features every 10 ms, 20 cochlear filters are used, and the frequency range between 0–5000 Hz is divided into 18 frequency bins equally spaced by one bark according to the critical-band rate [22].

For time alignment, we adopted a simple trace-segmentation scheme [29], which provides reasonable performance without serious computation time [30]. For each isolated word, cumulative distances between adjacent feature vectors are calculated each time frame, and an overall trace of the feature is defined as the cumulative distance at the end point. To generate normalized features with N time frames, the trace is then divided by $(N-1)$, representing an uniform feature changes between each normalized time interval. This simple normalization procedure removes variations in speech periods,

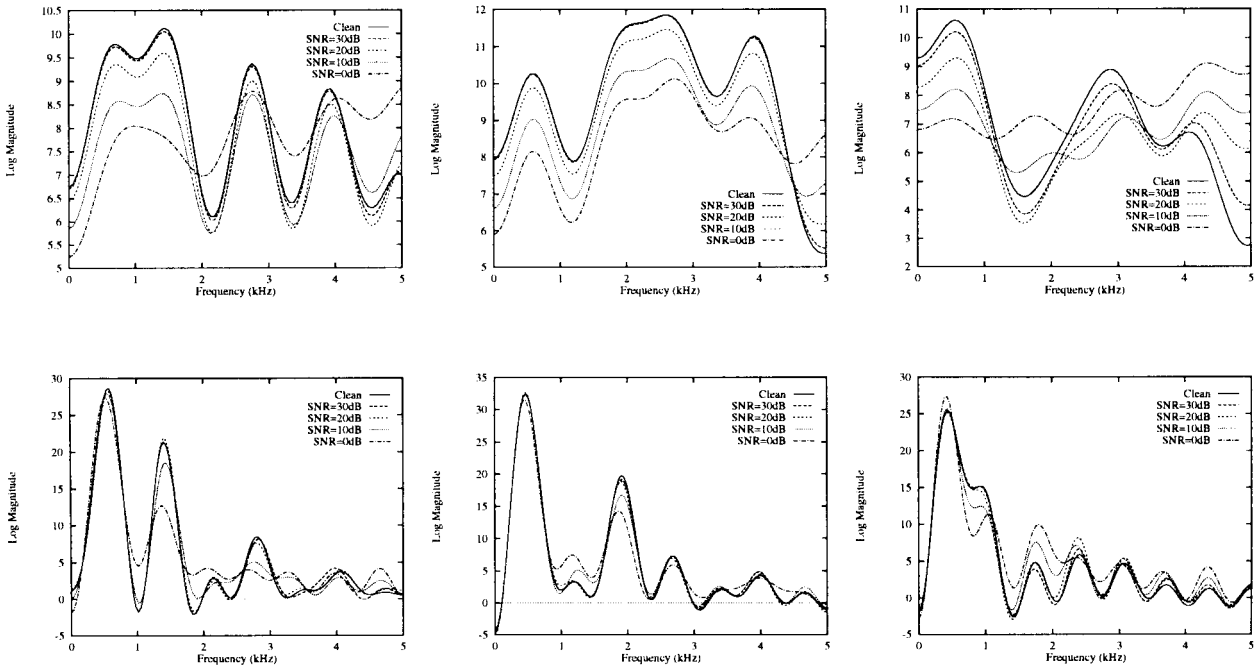


Fig. 6. Cepstral curve fitting of the vowels: (a) vowel $/a/$, (b) vowel $/e/$, and (c) vowel $/o/$ at various SNR levels, uttered by a male speaker. Each upper plot is the cepstral curve fit from the LPC-derived cepstrum, where the order of the LPC and cepstrum is 20. Each lower plot is the cepstral curve fit of the ZCPA output where 20th order cepstral coefficients are used.

TABLE I
NORMALIZED LOG SPECTRAL DISTANCE OF THE NOISY VECTORS TO THE CLEAN VECTOR FOR THE VOWELS SHOWN IN FIG. 6

($\times 10^{-3}$)

SNR (dB)	/a/		/e/		/o/	
	LPC	ZCPA	LPC	ZCPA	LPC	ZCPA
30	0.27	0.25	0.07	0.03	5.61	0.25
20	6.76	0.18	3.93	0.17	31.80	0.68
10	28.82	2.66	31.00	0.77	62.21	2.70
0	82.69	10.54	79.65	3.39	110.53	9.19

especially for steady long-pronounced vowels. The normalized word pattern is fed into the nearest neighbor classifier to obtain recognition results.

B. Results of the EIH's

Table II shows recognition rates of the EIH with a single level-crossing detector at various SNR's as a function of the level value. The ratio, R, of level value of the level-crossing detector to the possible maximum value of the signal, i.e., 2^{15} for the 16-b quantization, is represented in decibels. Overall recognition rates decrease as the level value increases. When $R \leq -60$ dB, there exists no big difference in recognition rates at relatively clean conditions, while the recognition rate decreases to some extent at noisy conditions as the level value increases. This is due to the fact that the incorrect spectral estimation with a high level value becomes more evident when speech is corrupted by additive noise.

On the contrary however, the recognition rate increases at low SNR's as the level value increases in the range above $R > -60$ dB. In such cases, the amplitudes of the weak parts of the signal become smaller than the level value, and only the higher amplitude parts of the signal can be represented as nonzero feature vectors. This causes an increase in recognition rates at noisy conditions as the level value increases, whereas the recognition rate at clean conditions decreases due to information loss. Thus the lower level value is preferred for reliable performance.

Table III compares recognition rates of the EIH's with several different numbers of levels and different level values. The ratio, H, of the highest level value to the maximum possible signal amplitude is represented in decibels. The level values of the EIH are uniformly distributed on the logarithmic scale over the positive amplitude range of the signal. Thus, for the three-level EIH with the highest level value of -30 dB, the three levels are located at -42, -36, and -30 dB, respectively. The recognition rate of the EIH's tends to increase as the level values decrease. However, if the level values are too low, the information obtained from some of lower levels will become redundant, and the recognition rate decreases to some extent. Also, comparing Table II with Table III, the recognition rate of the seven-level EIH's with $H = -24$ dB is worse than that of the zero-crossings. However, recognition rate of the zero-crossing is worse than that of the seven-level EIH's with $H = -54$ dB at clean, 30, and 20 dB SNR by 2.4, 2.1, and 3.6%, respectively. These results show that the utilization of multiple levels is preferable to a single zero-crossing, provided level values are properly determined. However, it is not so trivial to determine the proper level values.

C. Results of the ZCPA

A comparison of recognition rates of the ZCPA and other features is given in Table IV. The number of coefficients of the LPC-derived cepstrum (LPCC) was varied to be 12 and 18. We chose 18 cepstral coefficients, which produced

TABLE II
RECOGNITION RATE (%) OF THE EIH WITH SINGLE LEVEL AT VARIOUS SNR CONDITIONS AS THE LEVEL VALUE IS CHANGED

SNR (dB)	R, Relative level value (dB)												
	$-\infty$	-102	-96	-90	-84	-78	-72	-66	-60	-54	-48	-42	-36
Clean	85.9	86.2	86.1	86.6	86.7	87.7	88.5	88.7	88.3	83.8	76.2	59.7	37.9
30	85.3	86.0	84.5	85.1	84.7	86.6	86.9	87.3	87.6	83.4	76.0	60.6	38.3
20	75.7	75.4	75.5	75.4	74.8	73.3	73.2	72.3	75.7	79.8	76.4	61.5	38.5
10	53.7	54.0	53.7	54.9	52.3	50.2	49.3	46.7	48.4	48.9	59.3	58.5	39.4
0	23.6	23.1	22.3	22.6	22.3	20.7	20.2	19.1	19.1	18.8	21.3	23.1	30.1

TABLE III
RECOGNITION RATE (%) OF THE EIH WITH MULTIPLE LEVELS AT VARIOUS SNR CONDITIONS AS THE RANGE OF LEVEL VALUES IS VARIED: (a) THREE LEVELS; (b) FIVE LEVELS; (c) SEVEN LEVELS

SNR (dB)	H, Relative value of the highest level (dB)						
	-60	-54	-48	-42	-36	-30	-24
(a) 3-Level EIH							
Clean	88.3	87.9	83.7	78.3	65.9	49.3	33.7
30	87.8	87.5	83.2	78.5	65.6	49.0	33.8
20	76.7	79.0	79.0	76.4	66.5	49.4	33.7
10	49.4	50.7	52.0	57.7	59.6	49.3	34.5
0	19.7	20.4	19.5	21.5	25.2	29.3	29.7
(b) 5-Level EIH							
Clean	88.7	88.1	87.0	83.7	79.3	72.6	63.0
30	87.6	87.8	86.9	84.1	79.7	72.2	63.6
20	77.7	79.3	80.3	79.4	75.3	72.1	63.1
10	52.3	53.9	55.6	54.5	53.4	55.6	57.5
0	21.8	22.5	23.7	23.7	24.1	24.1	26.2
(c) 7-Level EIH							
Clean	88.1	88.3	88.0	86.8	84.2	81.9	78.0
30	87.1	87.4	87.5	85.7	84.3	82.1	77.9
20	78.2	79.3	80.1	80.1	78.5	77.3	74.4
10	54.3	55.9	58.5	58.3	56.8	54.3	53.6
0	22.8	23.8	26.6	27.5	27.5	25.8	24.1

TABLE IV
COMPARISON OF RECOGNITION RATES OF THE ZCPA WITH OTHER FEATURES AT VARIOUS SNR'S

SNR (dB)	Feature			
	LPCC	ZC	EIH	ZCPA
Clean	86.8	85.9	88.3	88.3
30	73.9	85.3	87.4	86.8
20	37.1	75.7	79.3	81.6
10	12.5	53.7	55.9	64.7
0	3.3	23.6	23.8	34.1

higher recognition rate. ZC is equivalent to the ZCPA without intensity information. Among several choices of EIH's, only the best case, i.e., the seven-level EIH with $H = -54$ dB, is shown in Table IV. The recognition rate of the LPCC decreases severely as the noise level increases. Comparing the ZCPA with the EIH's, the difference in recognition rates between the EIH's and the ZCPA is less than 1.5% above 20 dB SNR, and the recognition rate of the ZCPA is higher than that of the EIH's by 6.2% at 10 dB SNR. This demonstrates the low sensitivity of the ZCPA to additive random noise.

V. TOWARD PERFORMANCE IMPROVEMENTS AND COMPARATIVE EVALUATIONS WITH OTHER FRONT-ENDS

A. Preliminaries

There are many stand-alone applications of automatic speech recognition technology in which the whole platforms, such as workstations and personal computers, may not be available. Also, even though the ZCPA is a very simplified auditory model and the computational complexity is much less than other auditory models, the required computation time is still greater than conventional feature extraction algorithms. Therefore, it may be necessary to develop stand-alone hardware such as application specific integrated circuits (ASIC's), and several factors including computational complexity of the auditory model should be reconsidered.

In this section, the choice of a cochlear filterbank is considered for the efficient digital hardware implementations of the ZCPA model. In addition, the performance of the ZCPA is compared with other front-ends, including several auditory-like schemes, as well as conventional ones in various types of noisy environments.

B. Experimental Conditions

In consideration of practical applications of robust speech recognition, 50 Korean words, including ten digits ($0 \sim 9$) and 40 command words for control of electric home appliances, were chosen. Sixteen speakers uttered the words three times each in a quiet office environment via a SONY ECM-220T condenser microphone. The utterances were sampled at 11.025 kHz sampling rate with 12-b resolution. This quality, somewhat lower than the data base used in Section IV, is from the consideration of the cost and speed of hardware under development [31].

An isolated-word discrete density hidden Markov model (HMM) recognizer was used to model each word as a sequence of either five states for a one-syllable word or eight states for a multisyllable word. For training of the recognizer, 900 tokens of nine speakers were used in the construction of the codebook, where the number of codewords is 256, and in the conventional Baum–Welch procedure [32]. For test evaluations, 1050 tokens of the other speakers were used.

To evaluate the performances of front-ends in real environments, factory noise, military operations room noise, and car noise, contained in NOISEX-92 CD ROMS [33], as well as white Gaussian noise, were added to the test data sets at various SNR's.

C. Choice of Cochlear Filters

Since it is efficient to use powers of two as the number of parameters for digital hardware implementations, both the number of bandpass filters and the number of frequency bins were set to 16. Also, it is recommended to use finite impulse

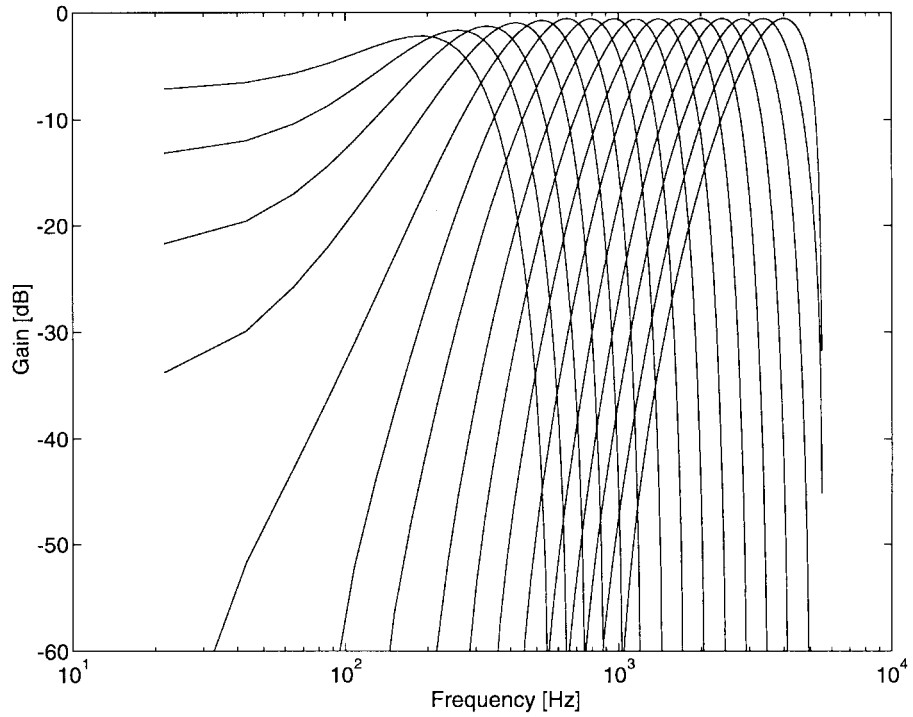


Fig. 7. Frequency response of cochlear filterbank implemented with FIR filters.

response (FIR) filters rather than infinite impulse response (IIR) filters for digital hardware implementations because roundoff noise and coefficient quantization errors are much less severe in FIR filters than in IIR filters, and the stability of IIR filters should be carefully considered. Thus, it is necessary to design the cochlear filterbank with FIR filters.

Frequency response of the filterbank consists of 16 Hamming bandpass filters (FIR filters), which are designed by the window method, is shown in Fig. 7. Even though the desired filter shape is not aimed to follow the biological neural tuning curve, center frequencies of the filterbank are determined by (1) between 200–4000 Hz, and bandwidths are set to be proportional to the equivalent rectangular bandwidth (ERB) [34]. Further, the maximum number of tabs is limited to 100 for an appropriate level of hardware implementation. The characteristics of several lower frequency channels are sacrificed by the limitation, as shown in Fig. 7.

Table V summarizes recognition rate of the ZCPA obtained using the traveling wave (TW) cochlear filters and that of the ZCPA obtained using Hamming FIR filters. In Table V, WGN, FAC, MOP, and CAR denote white Gaussian noise, factory noise, military operations room noise, and car noise, respectively. Even though cochlear filters are designed to mimic neural tuning curve shapes in detail, the recognition rate obtained by the FIR filters is higher than that of cochlear filters, regardless of the types of noise and SNR. As a result, the shape of the filter does not seem to be critical, at least for recognition performance, which is in agreement with the result of [26]. Thus it is sufficient to use the FIR filterbank if one considers digital hardware implementations of the ZCPA.

TABLE V
COMPARISON OF RECOGNITION RATE (%) OF THE ZCPA OBTAINED USING TW FILTERS AND FIR FILTERS

Noise		WGN		FAC		MOP		CAR	
Filterbank		TW	FIR	TW	FIR	TW	FIR	TW	FIR
	Clean	88.2	90.8	88.2	90.8	88.2	90.8	88.2	90.8
	25	85.8	88.1	87.5	89.1	86.4	87.5	87.8	90.0
	20	76.7	80.9	82.4	84.9	79.3	80.6	87.9	90.1
S	15	63.9	69.2	66.8	73.5	64.4	65.3	88.5	90.3
N	10	43.4	54.7	45.8	57.6	44.7	48.2	87.9	90.7
R	5	24.2	37.7	27.0	36.9	20.9	26.0	87.5	90.3
(dB)	0	–	–	12.7	18.3	8.9	10.4	84.8	88.1
	-5	–	–	–	–	–	–	76.9	81.1
	-10	–	–	–	–	–	–	63.7	64.8

D. Incorporation of Dynamic Features and Comparison with Several Front-Ends

It is well known that the transition of spectral contents through time plays an important role in human perception of speech [13], and it is common to incorporate dynamic properties of speech into speech recognition systems by augmenting dynamics such as delta and delta–delta features to static features for improved recognition accuracy, not only in clean but also in noisy conditions. Computing delta features is equivalent to an FIR filtering, which rejects lower modulation frequency variations of the speech parameters. If speech and nonspeech components occupy different ranges in the parameter domain, they can be separated by filtering in the parameter domain. Actually, the channel characteristics occupy the lower range of the modulation frequency in the logarithmic spectral domain, and lots of techniques, such as cepstral mean normalization (CMN) [35], and RASTA processing and its several variants

[36], [37], have been suggested to separate channel effects from speech parameters.

However, it was reported that the contribution of dynamic features of the EIH's to the performance improvements is much smaller than that of mel-frequency cepstral coefficients (MFCC) [38]. This may be due to the fact that the length of the time-window is channel dependent in the EIH's, i.e., it varies inversely with the characteristic frequency of the channel. For example, the length of the time-window at the channel with the lowest characteristic frequency spans up to 50 ms, which is much longer length when compared with the frame rate of about 10 ms. Thus, appropriate dynamic features cannot be obtained with the derivative window of 50 ms duration, which is used in [38]. Even though variable length of the derivative window [39], [40] may be applied to the computation of the delta features of the EIH's and ZCPA, it is beyond the scope of this study. Instead, we tried several fixed derivative window lengths.

Table VI compares recognition rates of six kinds of front-ends in various types of noisy environments, where the static and dynamic features are listed in columns. The procedure to obtain the static features of each front-end is briefly described here.

- *LPCC:* Speech signal is first multiplied by a 20.3 ms-long Hamming window every 10.15 ms. Eight LPC coefficients are obtained from the windowed segment of speech using the autocorrelation method. Then, 12 cepstral coefficients are obtained from the eight LPC coefficients.
- *MFCC:* A DFT is computed for each windowed segment of speech every 10.15 ms. Then the DFT power spectra are weighted by the magnitude frequency response of a filterbank, which consists of 16 triangular filters ranging from 100–5000 Hz, equally spaced in the mel scale. The magnitude frequency response of each filter is unity at the center frequency and linearly decreasing to zero at the center frequencies of two adjacent filters. Logarithms of the 16 output values are calculated to obtain 16 log channel energies. Twelve cepstral coefficients are then obtained by an inverse cosine transform.
- *SBCOR:* The DFT power spectrum of the windowed segment of the speech waveform is computed, and is then multiplied by the magnitude square of the frequency response of a filterbank to generate a vector of the subband power spectrum. From the subband power spectrum, the SBCOR spectrum, i.e., an array of the autocorrelation coefficient at the lag equal to the inverse of the center frequency of the corresponding filter, is computed. For the filterbank, 16 Hamming bandpass filters, which are also used in the ZCPA, are utilized in the frequency domain.
- *PLP:* The DFT power spectrum of the windowed segment is weighted by the magnitude frequency response of a filterbank, which consists of 18 critical-band filters ranging from 0 to 5000 Hz. The filtered power spectrum is preemphasized by the equal-loudness curve, and the intensity-loudness power of the hearing law is applied. The spectrum is approximated by an all-pole model with an order of eight using the autocorrelation method of all-pole spectral modeling. The eight autoregressive coefficients are then transformed into 12 cepstral coefficients.
- *ZCPA:* Due to their improved recognition accuracy with fewer coefficients than spectral representations [41], twelve cepstral coefficients are obtained by an inverse cosine transform of the ZCPA spectrum. The difference in recognition rate of the ZCPA spectrum and the ZCPA cepstrum can also be seen by comparing Tables V and VI.
- *EIH:* Twelve cepstral coefficients are obtained from the EIH output in the same manner as in the ZCPA. Performance of several of the EIH's cepstrum was evaluated by varying the number of levels and level values, as in Section IV. Only the best case (seven-level EIH with the ratio of the highest level value to the possible maxima of speech sample of -54 dB) is shown here.

As for static features, 12 cepstral coefficients ($c_1 \sim c_{12}$) without absolute log energy (c_0) were used for all front-ends except for the SBCOR. The output of the SBCOR is spectral representation in 16-dimensions ($s_1 \sim s_{16}$). For delta features, the regression coefficients of the static features over 11 time frames were computed [42]. Since the performance of the recognition system is known to be dependent on the length of time-derivative window [43], we tried four different derivative window lengths: 50.8 ms (5 frames), 111.7 ms (11 frames), 213.3 ms (21 frames), and 436.8 ms (43 frames). Finally, the length of the time-derivative window was set to be 11 frames, which yields the best performance on average. Two separate codebooks were constructed for each front-end, one for the static features and the other for the delta features, such that the number of codewords is 256 for each codebook. Recognition rates of the static and dynamic features and those of the static features were obtained with discrete density HMM by using both codebooks and by using the codebook for static features only, respectively.

On clean speech, the recognition rates of the MFCC, PLP, EIH, and ZCPA with the static features are quite similar. As the noise level increases, the recognition rate of the ZCPA becomes higher than that of the other front-ends for all kinds of noise. And the improvements incurred by incorporating delta features are more eminent for noisy data than for clean data in all front-ends. However, the contribution of delta features of both the EIH and ZCPA to the performance improvements is smaller than that of the other front-ends. This is in agreement with the results of [38]. Nevertheless, the performance of the ZCPA with static and dynamic features is similar to that of the other front-ends on clean speech, and becomes eminent even in noisy environments, where the other front-ends are of no practical importance because of low recognition rates. When SNR is 15 dB for example, recognition rates of the LPCC, MFCC, SBCOR, PLP, and EIH on speech data corrupted by white Gaussian noise are 12.0, 38.3, 72.7, 55.5, and 84.3%, respectively, whereas the ZCPA shows 87.0%. And recognition rate of the ZCPA becomes 90.3% on speech data corrupted by factory noise, which is higher than that of the PLP and EIH's by 8.7 and 3.6%, respectively. When military operations room noise is added to test utterances, the increases in recognition rate provided by the ZCPA are 3.4

TABLE VI
COMPARISON OF RECOGNITION RATES (%) OF FRONT-ENDS IN VARIOUS TYPES OF NOISY ENVIRONMENTS. (a) WHITE GAUSSIAN NOISE. (b) FACTORY NOISE. (c) MILITARY OPERATIONS ROOM NOISE. (d) CAR NOISE.

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	56.5	74.5	82.5	86.3	90.0	90.0	74.4	92.1	94.1	96.0	97.0	96.9
20	22.3	50.6	74.6	65.4	83.0	85.3	38.5	74.0	90.0	85.9	93.5	94.6
15	6.0	18.9	51.7	37.5	68.8	70.7	12.0	38.3	72.7	55.5	84.3	87.0
10	2.9	6.3	25.6	16.8	50.7	54.3	4.2	12.1	43.2	25.7	66.3	72.7
5	3.2	2.9	7.3	6.0	30.1	32.7	2.6	4.9	16.9	7.7	45.3	50.5

(a)

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	74.4	83.2	83.5	90.8	90.0	90.3	91.2	95.9	94.3	97.1	96.6	97.1
20	54.1	70.9	75.0	83.9	85.0	85.4	79.0	90.5	91.3	95.0	94.4	95.1
15	26.8	41.4	57.7	64.2	73.9	76.0	52.6	67.3	77.8	81.6	86.7	90.3
10	11.8	16.9	31.3	41.0	55.2	58.3	20.7	33.7	50.6	52.4	70.0	75.4
5	4.1	6.3	11.6	19.5	32.3	39.3	7.9	10.5	22.2	25.7	46.3	52.3

(b)

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	74.6	84.4	83.6	91.7	88.8	89.0	91.5	96.1	95.0	96.9	96.2	96.7
20	56.3	72.1	78.0	85.3	78.0	81.9	81.4	89.2	91.3	94.5	92.2	94.0
15	31.2	47.8	58.1	66.9	60.2	66.9	53.5	70.9	76.9	82.4	79.8	85.8
10	11.4	22.0	33.1	43.7	43.8	44.8	23.8	39.4	47.6	55.3	61.8	68.0
5	4.8	6.6	15.5	23.3	19.0	20.3	7.7	16.0	23.1	29.0	30.6	37.3

(c)

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
15	77.8	86.3	84.3	92.9	91.1	90.2	93.0	96.6	95.2	98.2	96.9	97.7
10	77.0	85.1	84.3	92.3	91.8	90.6	93.6	95.7	95.5	98.1	97.2	97.6
5	75.2	85.1	83.5	90.7	90.0	90.3	93.2	95.5	95.1	97.3	97.5	97.2
0	72.4	81.0	80.2	85.1	86.1	89.1	91.5	94.6	94.5	96.3	96.1	96.6
-5	58.7	72.8	73.7	69.4	78.8	81.2	84.8	90.7	89.6	92.6	93.5	93.6
-10	39.5	54.1	57.6	37.2	59.3	64.7	68.2	78.2	75.8	72.9	84.2	86.9

(d)

and 6.0 percentage points compared with the PLP and EIH, respectively.

One noticeable thing is that the usefulness of the ZCPA in noisy environments is prominent when white Gaussian noise rather than real-world noise is added to speech data. At 20 dB SNR, for example, the difference in recognition rates between the ZCPA and MFCC is as high as 20.6 percentage points on speech data corrupted by white Gaussian noise. However, the improvement provided by the ZCPA becomes only 4.6 and 4.8 percentage points on speech data corrupted by factory noise and military operations room noise, respectively. These results can be related to the previous works provided by Ghitza [26], Stern *et al.* [44], and Jankowski *et al.* [45]. Jankowski *et al.* used speech babble noise as the background noise, and reported that two auditory models, EIH's and Seneff's model, perform *slightly better* than the MFCC in noisy conditions: the difference between the auditory models and the MFCC is from 0.6 to 4 percentage points depending on the SNR. However,

Ghitza [26] and Stern *et al.* [44] used artificially generated white noise, and showed that the EIH's and Seneff's auditory model provide much more improved robustness than the LPCC in noisy conditions. Thus, the difference in recognition rate between the conventional front-ends and auditory models is dependent on the type of noise: the difference is maximum when white Gaussian noise is used, and decreases when real-world noises are used. The reason is not clear yet, and remains as a future work. In summary, the performance of the ZCPA is superior to the other front-ends in noisy environments.

VI. Conclusion

A feature extraction method motivated by mammalian auditory periphery is proposed. The proposed ZCPA model is composed of cochlear bandpass filters and a nonlinear stage at the output of each bandpass filter. The bank of bandpass filters simulates frequency selectivity of the basilar membrane in the cochlea, and a nonlinear stage models the auditory nerve fibers, which fire in synchrony with the stimulation. The nonlinear stage consists of a zero-crossing detector, a peak detector, and a compressive nonlinearity. Frequency information is obtained by the zero-crossing detector, and intensity information is also incorporated by the peak detector followed by the compressive nonlinearity. It is shown analytically that the variance of the level-crossing interval perturbation increases as the level value increases in the presence of additive noise. Thus, the zero-crossing is more robust to noise than the level-crossing, and it offers the motivation for utilizing zero-crossings for robust speech recognition in noisy environments. The ZCPA is computationally efficient and free from many unknown parameters compared with other auditory models. Experimental comparisons of the developed ZCPA auditory model with the EIH and other popular feature extraction methods in various types of noisy environments have demonstrated greatly improved robustness of the ZCPA method, especially in noisy environments corrupted by white Gaussian noise.

Appendix
Sensitivity of the ZCPA Model to the Peak Amplitude Estimation

In this Appendix, the sensitivity of the developed ZCPA model to inaccuracies in peak amplitude estimation in noisy condition is analyzed. Consider the input signal $x(t)$ as in (6), and suppose that $x(t)$ is filtered by a bank of ideal bandpass filters, of which the bandwidths are all B, and each sinusoidal component in the input signal is separated by the filterbank. Then the output of the kth filter can be represented as

$$x_k(t) = A_i \cos(\omega_i t + \phi_i) + A_g v_k(t). \tag{18}$$

Let us assume that the zero-crossing points of $A_i \cos(\omega_i t)$ are not affected by the additive noise term, $A_g v_k(t)$, so that the index of the frequency bin estimated by zero-crossing points is retained. Further, suppose that the position of the peak amplitude is retained and only the value of the peak amplitude is changed by the noise. Then the value of the frequency histogram at the frequency bin associated with the frequency, ω_i, will be increased whenever two adjacent zero-crossing points are detected, and the only factor affected by the noise becomes the amount of contribution to the frequency bin. Let us denote the peak amplitude of the output signal at the kth channel by x_p, which is represented as

$$x_p = v_p + A_i \tag{19}$$

where v_p is a bandpass-filtered white Gaussian noise with zero-mean and variance, σ^2. If we assume that the compressive nonlinearity of the ZCPA in (5) is approximated as $g(x) \approx \log(x)$ for sufficiently large x, then

$$\begin{aligned} g(x_p) &= g(A_i + v_p) \\ &= \log(A_i + v_p). \end{aligned} \tag{20}$$

The mean of the amount of contribution to a frequency bin is given as

$$\begin{aligned} m_p &= E\{g(x_p)\} \\ &\approx \log A_i - \frac{1}{2}\frac{\sigma^2}{A_i^2} \end{aligned} \tag{21}$$

where the second-order Taylor series expansion of (16) is utilized for $\sigma/A_i \ll 1$. The mean of the amount of contribution to a frequency bin is shifted downward when the signal is corrupted by additive noise as shown in (21). However, for the moderate level of signal amplitude and SNR, the amount of shift is significantly small compared with the term, $\log A_i$.

The variance can also be approximated as

$$\begin{aligned} \sigma_p^2 &\approx |g'(\eta)|^2 \sigma^2 \\ &= \frac{\sigma^2}{A_i^2} \end{aligned} \tag{22}$$

by applying (16) to the function $g^2(x)$, and neglecting σ^4/A_i^4 term. The variance in (22) increases as SNR decreases. However, in noisy conditions with the moderate level of signal amplitude and SNR, the variance is negligible compared with the mean in (21).

If a linear function is used as the compressive nonlinearity of the ZCPA, the mean of the amount of contribution to a frequency bin is represented as $\log A_i$, which is not shifted downward. However, the variance of the amount of contribution to a frequency bin is dependent on the noise power only, which makes the ZCPA sensitive to noise severely.

Acknowledgment

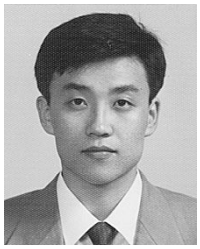
The authors would like to thank Prof. X. Zhu and J.-H. Jeong for valuable discussions on the ZCPA model, as well as O. Ghitza, the reviewers, and the associate editor for their valuable comments and suggestions improving the quality of this manuscript.

References

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

[2] ——, "A joint synchrony/mean-rate model of auditory processing," *J. Phonet.*, vol. 16, pp. 55–76, 1988.

[3] M. Hunt and C. Lefebvre, "Speech recognition using a cochlear model," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1986, pp. 37.7.1–37.7.4.
[4] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
[5] J. C. Junqua, H. Wakita, and H. Hermansky, "Evaluation and optimization of perceptually-based ASR front-end," *IEEE Trans. Speech Audio Processing*, vol. 1, pp. 329–338, 1993.
[6] J. C. Junqua, "Toward robustness in isolated-word automatic speech recognition," Ph.D. dissertation, Univ. Nancy I, Nancy, France, 1989.
[7] S. Kajita and F. Itakura, "Speech analysis and speech recognition using subband-autocorrelation analysis," *J. Acoust. Soc. Jpn.*, vol. 15, pp. 329–338, 1994.
[8] ———, "Robust feature extraction using SBCOR analysis," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1995, pp. 421–424.
[9] M. Hunt and C. Lefebvre, "Speech recognition using an auditory model with pitch-synchronous analysis," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1987, pp. 20.5.1–20.5.4.
[10] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Trans. Inform. Theory*, vol. 38, pp. 824–839, 1992.
[11] K. Wang and S. A. Shamma, "Self-normalization and noise-robustness in early auditory representations," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 421–435, 1994.
[12] A. Morris, J. L. Schwartz, and P. Escudier, "An information theoretic investigation into the distribution of phonetic information across the auditory spectrogram," *Comput. Speech Lang.*, vol. 2, pp. 121–136, 1993.
[13] S. Handel, *Listening: An Introduction to the Perception of Auditory Events.* Cambridge, MA: MIT Press, 1993.
[14] 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. Int. Conf. Acoustics, Speech, Signal Processing*, Atlanta, GA, May 1996, pp. 61–64.
[15] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. Signal Processing*, vol. 39, pp. 2573–2592, 1991.
[16] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1119–1134, 1988.
[17] D. Greenwood, "A cochlear frequency-position function for several species—29 years later," *J. Acoust. Soc. Amer.*, vol. 87, pp. 2592–2650, 1990.
[18] M. B. Sachs, C. C. Blackburn, and E. D. Young, "Rate-place and temporal-place representations of vowels in the auditory nerve and anteroventral cochlear nucleus," *J. Phonet.*, vol. 16, pp. 37–53, 1988.
[19] M. B. Sachs and E. D. Young, "Encoding of steady state vowels in the auditory-nerve: Representation in terms of discharge rate," *J. Acoust. Soc. Amer.*, vol. 66, pp. 470–479, 1979.
[20] E. D. Young and M. B. Sachs, "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibers," *J. Acoust. Soc. Amer.*, vol. 66, pp. 1381–1403, 1979.
[21] B. Delgutte and N. Y. S. Kiang, "Speech coding in the auditory nerve: I," *J. Acoust. Soc. Amer.*, vol. 75, pp. 866–878, 1984.
[22] E. Zwicker and E. Terhart, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *J. Acoust. Soc. Amer.*, vol. 68, pp. 1523–1525, 1980.
[23] O. Ghitza, "Auditory models and human performances in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pt. II, pp. 115–132, 1994.
[24] 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.
[25] 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, pp. 282–293, 1992.
[26] O. Ghitza, "Auditory nerve representation as a basis for speech processing," in *Advances in Speech Signal Processing*, S. Furui and M. M. Sondhi, Eds. New York: Marcel Dekker, 1992, pp. 453–485.
[27] B. Kedem, "Spectral analysis and discrimination by zero-crossings," *Proc. IEEE*, vol. 74, pp. 1477–1493, Nov. 1986.
[28] I. J. Choi *et al.*, "A Korean speech database for use in automatic translation," in *Proc. 11th Workshop on Speech Communication and Signal Processing*, 1994, pp. 287–290, in Korean.
[29] H. F. Silverman and N. R. Dixon, "State constrained dynamic programming (SCDP) for discrete utterance recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1980, pp. 169–172.
[30] D.-S. Kim and S.-Y. Lee, "Intelligent judge neural network for speech recognition," *Neural Process. Lett.*, vol. 1, pp. 17–20, 1994.
[31] S.-Y. Lee *et al.*, "Voice command: A digital neuro-chip for robust speech recognition in real-world noisy environments (Invited talk)," in *Proc. Int. Conf. Neural Information Processing*, Hong Kong, Sept. 1996, pp. 283–287.
[32] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–286, 1989.
[33] 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 Commun.*, vol. 12, pp. 247–251, 1993.
[34] B. C. J. Moore and B. R. Glasberg, "Suggested formula for calculating auditory-filter bandwidth and excitation patterns," *J. Acoust. Soc. Amer.*, vol. 74, pp. 750–753, 1983.
[35] J.-C. Junqua *et al.*, "Time derivatives, cepstral normalization, and spectral parameter filtering for continuously spelled names over the telephone," in *Proc. Europ. Conf. Speech Communication and Technology*, 1995.
[36] H. Hermansky, N. Morgan, A. Bayya, and P. Kohn, "Compensation for the effect of the communication channel in auditory-like analysis of speech (RASTA-PLP)," in *Proc. Europ. Conf. Speech Communication and Technology*, 1991, pp. 1367–1370.
[37] H. Hermansky, E. Wan, and C. Avendano, "Speech enhancement based on temporal processing," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, 1995, pp. 405–408.
[38] S. Sandhu and O. Ghitza, "A comparative study of mel cepstra and EIH's for phone classification under adverse conditions," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, Detroit, MI, 1995, pp. 409–412.
[39] J. Smolders and D. V. Compernolle, "In search for the relevant parameters for speaker independent speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1993, vol. II, pp. 684–687.
[40] K. Aikawa, H. Singer, H. Kawahara, and Y. Tohkura, "A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing*, 1993, vol. II, pp. 668–671.
[41] D.-S. Kim, S.-Y. Lee, R. M. Kil, and X. Zhu, "Auditory model for robust speech recognition in real world noisy environments," *Electron. Lett.*, vol. 33, p. 12, 1997.
[42] S. Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 52–59, 1986.
[43] B. Hanson, T. Applebaum, and J.-C. Junqua, "Spectral dynamics for speech recognition under adverse conditions," in *Advanced Topics in Automatic Speech and Speaker Recognition*, C.-H. Lee, K. Paliwal, and F. Soong, Eds. Boston, MA: Kluwer, 1995.
[44] R. M. Stern *et al.*, "Multiple approaches to robust speech recognition," in *Proc. DARPA Speech V Natural Language Workshop*, Harriman, NY, 1992, pp. 274–279.
[45] C. R. Jankowski, Jr., H.-D. H. Vo, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286–293, 1995.



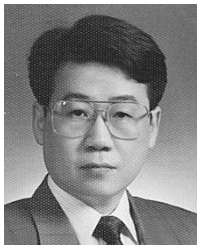
Doh-Suk Kim (A'94) was born in Seoul, Korea, in 1968. He received the B.S. degree in electronics engineering from Hanyang University, Seoul, Korea, in 1991, and the M.S. and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Taejon, Korea, in 1993 and 1997, respectively.

From 1993 to 1996, he was a part-time Researcher with the Systems Engineering Research Institute, Taejon, Korea. He served as a Postdoctoral Fellow at KAIST from March 1997 to October 1997. From November 1997 to October 1998, he was with the Acoustics and Speech Research Department, Bell Laboratories, Lucent Technologies, Murray Hill, NJ, as a Post-doctoral Member of Technical Staff. He is currently with the Human and Computer Interaction Laboratory, Samsung Advanced Institute of Technology, Suwon, Korea. His research interests include auditory psychophysics, speech recognition, speech coding, and objective speech quality assessment.

Soo-Young Lee (S'75–M'83) received the B.S. degree in electronics from Seoul National University, Seoul, Korea, in 1975, the M.S. degree in electrical engineering from the Korea Advanced Institute of Science and Technology, Taejon, in 1977, and the Ph.D. degree in electrophysics from the Polytechnic Institute of New York (PINY), Brooklyn, in 1984.

From 1977 to 1980, he was a Project Engineer with the Taihan Engineering Co., Seoul. From 1980 to 1983, he was a Senior Research Fellow at the Microwave Research Institute, PINY. From 1983 to 1985, he served as a Staff/Senior Scientist at the General Physics Corp., Columbia, MD. After a short stay at the Argonne National Laboratory, Argonne, IL, he joined the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology. His current research interests include optical computing and neural networks. He has published or presented more than 80 papers in optical inplementation of neural networks, neural network architectures and applications, and numerical simulation techniques for electromagnetics.

Dr. Lee was the Guest Editor of the Special Issue on Neural Networks of the Proceedings of the KIEE, February 1989. He organized the Korea-USA Joint Workshop on Optical Neural Networks in 1990.



Rhee M. Kil (M'94) received the B.S. degree in electrical engineering from Seoul National University, Seoul, Korea, in 1979, and the M.S. and Ph.D. degrees in computer engineering from the University of Southern California, Los Angeles, in 1985 and 1991, respectively.

From 1979 to 1983, he was with the Agency for Defense Development, Taejon, Korea, where he was involved in the development of an infrared imaging system. From 1987 to 1991, he was involved in research on the theories and applications of connectionist models. His dissertation was on the learning algorithms of connectionist models and their applications to nonlinear system control. From 1991 to 1994, he was with the Research Department, Electronics and Telecommunications Research Institute, Taejon. In 1994, he joined the Division of Basic Sciences, Korea Advanced Institute of Science and Technology, Taejon, as an Assistant Professor. His general research interests lie in the areas of pattern recognition, system identification, data coding, and nonlinear system control. His current interests focus on learning based on evolutionary computation and information representation.