Robust Auditory-Based Speech Processing Using the Average Localized Synchrony Detection

Ahmed M. Abdelatty Ali, *Member, IEEE*, Jan Van der Spiegel, *Fellow, IEEE*, and Paul Mueller

***Abstract*—In this paper, a new auditory-based speech processing system based on the biologically rooted property of the average localized synchrony detection (ALSD) is proposed. The system detects periodicity in the speech signal at Bark-scaled frequencies while reducing the response's spurious peaks and sensitivity to implementation mismatches, and hence presents a consistent and robust representation of the formants. The system is evaluated for its formant extraction ability while reducing spurious peaks. It is compared with other auditory-based and traditional systems in the tasks of vowel and consonant recognition on clean speech from the TIMIT database and in the presence of noise. The results illustrate the advantage of the ALSD system in extracting the formants and reducing the spurious peaks. They also indicate the superiority of the synchrony measures over the mean-rate in the presence of noise.**

***Index Terms*—ALSD, auditory, extraction, feature, formant, processing, recognition, speech, synchrony.**

I. Introduction

THE SUPERB ability of the human auditory system to process speech in the presence of noise has motivated many researchers to build auditory-based speech processing systems for automatic speech recognition (ASR) applications. The most widely used systems implement some auditory effects (such as Bark- or Mel-scale filtering and nonlinear compression) in a short-time fast (discrete) Fourier transform (FFT) framework. Examples are the Mel-frequency cepstral coefficients (MFCC) [13], [51], the perceptual linear predictive analysis (PLP) [23], [26], and the RASTA processing [24], [25]. Those systems have shown clear improvement over traditional cepstral and LPC analyzes in speech recognition applications. However, despite their usefulness, those systems still suffer from the fixed-window limitation of short-time FFT systems, which causes the frequency-time resolution tradeoff [31], [45]. Moreover, their modeling of the auditory effects is neither complete nor accurate. They, however, have the advantage of fast processing.

Manuscript received April 24, 2001; revised March 17, 2002. This work was supported by a grant from Catalyst Foundation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jerome Bellegarda.

A. M. A. Ali is with Research and Development, Texas Instruments, Inc., Warren, NJ 07059 USA and also with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104-6390 USA (e-mail: ahm@ee.upenn.edu).

J. Van der Spiegel is with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia, PA 19104-6390 USA (e-mail: jan@ee.upenn.edu).

P. Mueller is with Corticon, Inc., King of Prussia, PA 19406 USA (e-mail: corticon@aol.com).

Publisher Item Identifier 10.1109/TSA.2002.800556.

Some researchers, on the other hand, worked on more accurate auditory modeling [12], [17]–[19], [32], [44]–[47]. In those models, the auditory effects are emulated as accurately as possible according to the current understanding. Such trials have yielded systems that outperformed the MFCC, PLP and RASTA systems in speech recognition applications especially in the presence of noise and other adverse conditions [10], [17], [28]–[31], [39], [43], [50]. They, however, suffer from very slow processing that makes real-time software implementation, for the overall ASR system, difficult and uneconomic with the current state-of-the-art computation and storage powers [31]. For example, their speed was found to be between 40 and 120 times real-time on a Sparc-2 workstation [30], with 35–600 times the number of operations required for the traditional LPC processing [31], [34].

This work concentrates on auditory-based systems of the latter type. The superior ability of the human auditory system to handle and recognize speech in the presence of noise makes the understanding and modeling of such capability a necessity. The slow processing time could be economically overcome by relying on hardware analog VLSI implementation of the system, which will enable parallel real-time processing [37], [38]. In the next sections, we investigate some of the auditory-based systems that proved to yield relatively good and robust performance, and are readily implementable in analog VLSI technology. Those include the Bark-scaled filter bank mean-rate output, the lateral inhibitory network (LIN) output [46], [47], and the generalized synchrony detector (GSD) output [44], [45]. A new system is developed by the authors as a modification to the GSD. It is called the average localized synchrony detector (ALSD) [3], [7] and is designed to alleviate some of the limitations of the GSD. The ALSD is evaluated and compared with the other three systems in their formant extraction ability as indicated by vowel recognition experiment for multiple speakers with seven different dialects of the American English from the TIMIT database.

II. Auditory-Based Processing

The general structure of the auditory-based front-end processing systems used in this work is shown in Fig. 1. It consists of a Bark-scaled filter bank of 36 bandpass filters with a spacing of half a Bark between neighboring filters. The filter bank used is a software simulation of an actual analog cochlea that was implemented in VLSI [35], [36]. This choice of the filter bank is made in order to ensure its practicality from the hardware implemention standpoint.

The Bark-scale filter distribution preserves the temporal structure of the output waveforms and avoids the frequency-time

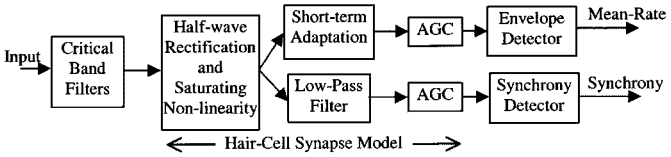

Fig. 1. Auditory-based front-end system.

resolution tradeoff encountered in fixed-window short-time FFT-based systems. The low-frequency region is characterized by relatively sharp narrow-band filters. The high-frequency region has wider-band, and hence faster, filters. This achieves high-frequency resolution in low-frequency regions where it is needed to distinguish sonorants, which are usually static in nature and hence do not need high time resolution. On the other hand, the high-frequency (wideband) filters will have low-frequency resolution and high time resolution. This will be useful for dynamic sounds that are characterized by high-frequency energy and usually require high time resolution. Thus, this filter bank approach, as opposed to the fixed-window approach, invests the appropriate resolution where it is needed and is therefore compatible with the characteristics of the speech signal and the requirements of speech processing. The amplitude responses of the filter bank are shown in Fig. 2.

The filter bank is followed by a nonlinear stage that performs half-wave rectification with a compressive and saturating nonlinearity. The description of this stage is given by [44]

$$\begin{aligned} y &= 1 + A\tan^{-1} Bx \qquad x > 0 \\ &= e^{ABx} \qquad x \le 0 \end{aligned} \tag{1}$$

where x is the input, y is the output, A and B are constants (10 and 65 respectively). It is clear that the function is exponential for negative inputs, linear for small input values and compressive for larger signals.

The system is then divided into two branches. One branch gives the mean-rate response and the other gives the synchrony (phase-locked) response. The mean-rate response path begins with a short-term adaptation and forward masking (STA) module, followed by an automatic gain control (AGC) module and finally ending with an envelope detector. The synchrony path, on the other hand, has a low-pass filter (LPF), an AGC and a synchrony detector.

The STA system models the short-term adaptation and forward masking effects that take place in the cochlear response [22], [49]. The model describes two separate mechanisms that influence the concentration of neurotransmitters. A membrane allows the flow of a supply from a source region at a rate proportional to the concentration gradient across the membrane, with a proportionality constant of μ_a. When the concentration gradient is negative (i.e., the concentration in the supply region is too small), the channels in the membrane close and the neurotransmitters are lost by natural decay at a rate that is proportional to its concentration within the region, with a proportionality constant of μ_b. This is shown mathematically as follows [44]:

$$\begin{aligned} dC(t)/dt &= \mu_a[S(t) - C(t)] - \mu_b C(t) \qquad C(t) < S(t) \\ &= -\mu_b C(t) \qquad C(t) \ge S(t) \end{aligned} \tag{2}$$

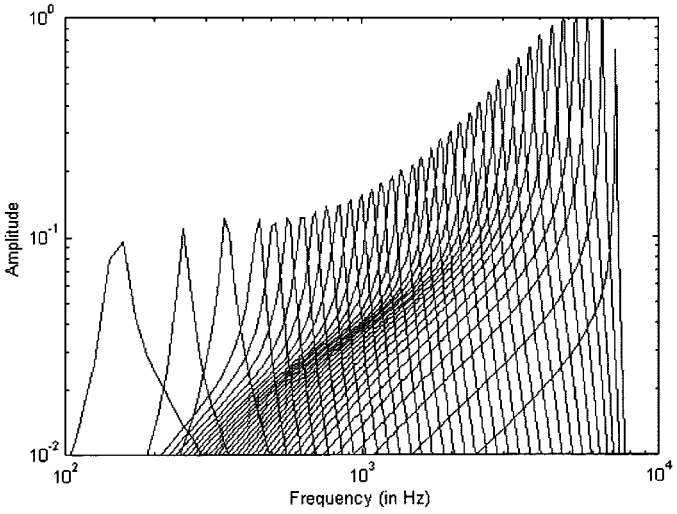

Fig. 2. Frequency (amplitude) responses of the filter bank.

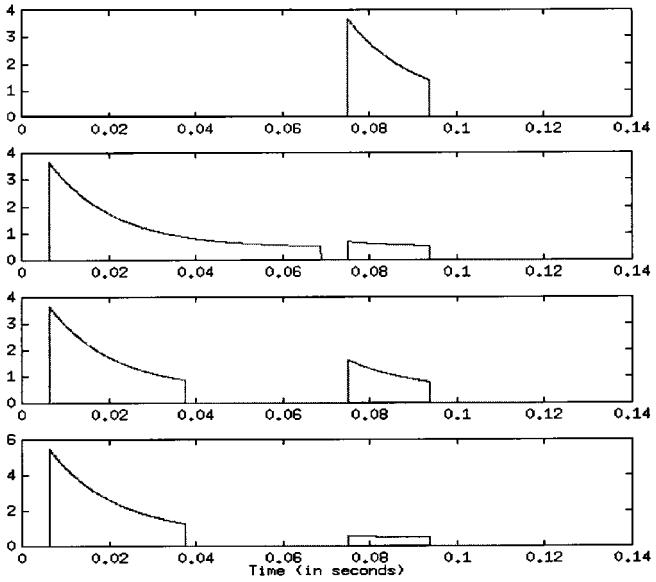

Fig. 3. Short-term adaptation and forward-masking modeling. Responses of a probe tone presented alone (in the top curve) and with a preceding masker of various durations and amplitudes.

where $C(t)$ is the concentration of neurotransmitters within the region, and $S(t)$ is the concentration in the source region (i.e., the input). The output of the system is represented by the flow rate across the membrane: $\mu_a[S(t) - C(t)]$. The constants μ_a and μ_b are 8.3 s^{-1} and 58.3 s^{-1}, respectively. The discrete-time realization is achieved by approximating d/dt by a first difference in time and normalizing with respect to the sampling frequency. The response of this block is shown in Fig. 3.

This module is included in the mean-rate response only since it was found that, biologically, the synchrony response is only marginally affected by such effects [14], [41], [42], [53]. On the other hand, from a practical viewpoint, including such an effect in the synchrony response strongly obliterates the formant structure for long steady sounds (like vowels). Short-term adaptation was found to improve the immunity of the system to noise [39]. By enhancing the changes it helps attenuate stationary noise. This is obvious in its high-pass filter effect that

eliminates any stationary source of distortion. This principle is used in the RASTA processing and shown to improve the robustness of the system [24], [25]. Moreover, the adaptation and masking effects help significantly in marking and emphasizing the boundaries between different speech segments. This proved to be useful in the segmentation of speech into acoustically-compatible segments [7].

The AGC is used for dynamic range compression, to accommodate inputs with various amplitudes. using the relation [44]

$$y[n] = \frac{x[n]}{1 + K_{AGC}\langle x[n]\rangle} \tag{3}$$

where K_{AGC} is a constant, and $\langle\rangle$ represents the "expected value of" obtained by passing $x[n]$ through a low-pass filter with a time constant of about 3 ms.

The envelope detector is a simple low-pass filter with a cutoff frequency of 50 Hz. On the other hand, the low pass filter in the synchrony path is used to model the synchrony suppression that occurs at high frequencies due to the neural latencies and response jitters. This attenuates the phase-locking capability above 4 kHz.

The synchrony detector is used to detect the temporal phase-locking characteristics of the response. Examples of synchrony detectors are the lateral inhibitory network (LIN) and the GSD.

The LIN approach has different forms and has been used by several researchers [15], [16], [46], [47]. It is based on inhibiting each filters output by one or more of the neighboring filters. This helps enhance the spectral peaks and improves the frequency resolution. This could be as simple as subtracting the neighboring filter output, or it could be more involved like using a feedforward or feedback inhibitory network. The approach used in this work is using a feedforward lateral inhibitory network similar to that used by Shamma and described in [47]. The output of each unit is computed by subtracting a weighted sum of its neighbors, followed by a threshold operation and a time-window average. In this way, the peaks (formants) are enhanced by detecting the filters that have strong phase differences with their neighbors. This is a simple, fast and effective approach for detecting the synchrony and producing a robust formant representation. It is interesting to know that the order of operations did not cause any noticeable difference in the output. Thus performing the lateral inhibition on the AGC output followed by the averaging (filtering) gave nearly the same results as averaging before the lateral inhibition.

The GSD is designed by Seneff [45] to enhance the prominent peaks at the formant resonances, improve the spectral resolution, reduce features of the spectrograms associated with the glottal excitation, and normalize for amplitude. It detects the periodicity in the temporal response (instead of the envelope mean-rate) by computing an auto-correlation-like output. It generates a soft-limited ratio of the expected (averaged) magnitude value of the sum and difference of the output of each filter and a delayed version of it, as shown in Fig. 4. The delay of each GSD must match its corresponding filters center frequency (i.e., the

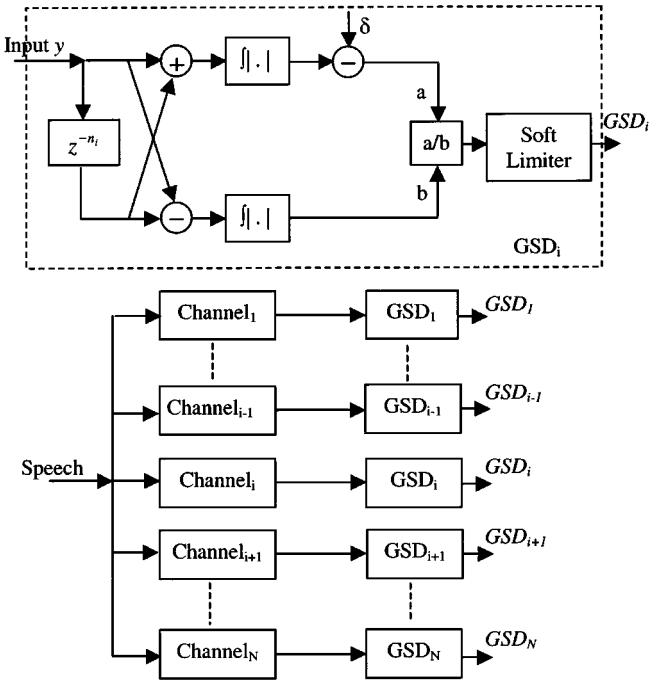

Fig. 4. GSD block diagram.

delay is equal to the inverse of the center frequency). This operation is expressed as follows:

$$GSD_i(y) = A_s \tan^{-1}\left[\frac{1}{A_s}\left(\frac{\langle|y[n] + y[n-n_i]|\rangle - \delta}{\langle|y[n] - \beta^{n_i} y[n-n_i]|\rangle}\right)\right] \tag{4}$$

where $y[n]$ is the input to the GSD (output from the AGC stage) at time sample n, GSD_i is the synchrony output of the ith channel [i.e., tuned to the ith filter by setting $n_i = f_s/f_i$, where f_i is the ith filter center frequency (CF) and f_s is the sampling frequency], $\langle\rangle$ represents envelope detection, and A_s, β, and δ are constants.

The sum and difference waveforms are constructed from the output of the AGC stage, full-wave rectified and low-pass filtered to obtain the envelope response in (4). The constant β is set to a value slightly less than 1.0 in order to position the zero of the denominator slightly inside the unit circle and hence reduce the sharpness of the nulls at multiples of n_i. This was found to be useful for low-frequency filters in order to decrease the preciseness of the tuning [45]. A small threshold δ is subtracted from the numerator in order to suppress the response to small amplitude signals. Its value is chosen to be slightly larger than the spontaneous rate. The saturating nonlinearity is used to soft-limit the output and prevent infinite responses. At small amplitudes, the response is nearly linear and then saturates for large input amplitudes. The linear range of the input is controlled by the value of A_s.

As shown in Figs. 1 and 4, the GSDs are used to compute the synchrony from the AGC outputs. Each AGC output is applied to a GSD tuned to the center frequency of the corresponding auditory filter. Thus if there is a prominent peak in the signal at a particular frequency, f, it will show up as periodicity in the AGC waveforms. The channel whose CF is closest to f will

detect the correct periodicity by generating a response that is significantly larger than its neighbors.

The GSD has many advantages over other systems. First, because it measures periodicity rather than frequency, it avoids the problem of detecting synchrony to the second harmonic of a strong peak. It could, however, detect synchrony at half the frequency especially if the high-frequency slopes of the filters are not very steep. Detecting periodicity also makes it more immune to noise. Moreover, taking the ratio between the sum and difference waveforms performs an energy-normalization that reduces the temporal fluctuations in the response due to the envelope of the glottal excitation [45].

Nevertheless, the GSD response contains significant spurious peaks that are due to individual harmonics of the fundamental frequency (F0), noise, and other artifacts, especially for female speakers below the first formant region (F1). This was described by Seneff [45] as a major problem that limits the effectiveness of the GSD in ASR applications. Moreover, it requires accurate matching between the filters CF and the GSD delay time (tuning frequency). This matching may need to be as tight as 0.1% for the low-frequency filters [45]. Such requirement is not easily achieved in practical analog VLSI implementations due to technology limitations. If not achieved, the spurious peaks increase significantly.

A. Average Localized Synchrony Detector

The aforementioned limitations of the GSD are mainly due to the sharp tuning of the GSD that is desired to improve the resolution and enhance the formant peaks. The way to reduce those problems is by increasing the filter bandwidth [45]. This could be the input bandpass filter or the GSD filter (by decreasing β for example). The smoothing effect of such process will decrease the sharpness of the GSD and hence reduce the above problems. Unfortunately, such *blind* smoothing will also deteriorate the resolution of the GSD in a way that defeats its original purpose. In other words, we are faced with a resolution-accuracy tradeoff that is increasingly manifested in the absence of accurate matching. To get rid of spurious peaks that affect the formant extraction accuracy, we need to smooth the spectrum by using wider-band filters, which would deteriorate the resolution.

To alleviate this problem, we modified the GSD in order to represent the average localized synchrony [14], [42], [53]. The output of each ALSD is the average of n GSDs tuned to the *same* frequency but applied to several filters in the neighborhood of the corresponding filter. The value of n is decided empirically based on the resolution and bandwidth of the filters used. This can be expressed as follows:

$$ALSD_i = \frac{1}{n} \sum_{k=i-n1}^{i+n2} GSD_i(y_k) \tag{5}$$

where $ALSD_i$ is the ALSD output of the ith channel (filter); GSD_i is the output of the GSD which is tuned to the ith filter; y_k is the output of the kth filter (after the AGC stage); $GSD_i(y_k)$ is the output of the ith GSD (i.e., the GSD tuned to the ith filter) when applied to the kth filter. The constants $n1$ and $n2$ add up to n. (i.e., $n = n1 + n2$). We chose n to be equal to 3, with one filter on each side of the center filter (i.e., $n1 = n2 = 1$ and k ranges from $i-1$ to $i+1$). The ALSD block diagram and connection are shown in Fig. 5.

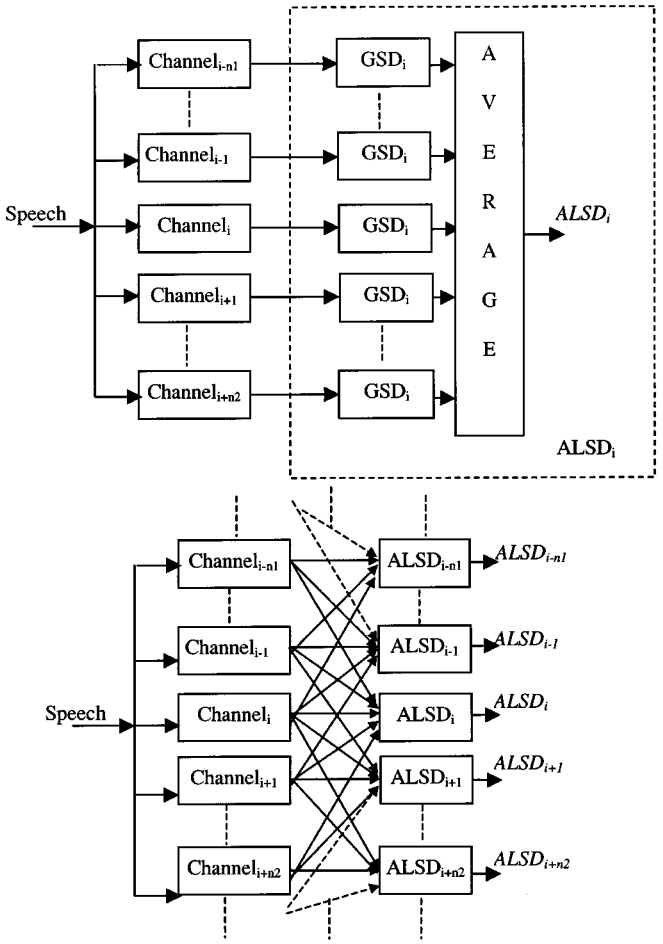

Fig. 5. ALSD block diagram.

We need to emphasize that the operation described in (5) is not equivalent to simply averaging the inputs of neighboring filters and applying them to the same GSD. It is also different from averaging the outputs of neighboring GSDs. The nonlinearity of the GSD and its tuning characteristics make the ALSD output substantially different from those two averaging operations as will be shown later.

The ALSD provides an extra degree of freedom that enables us to achieve *selective* smoothing while preserving the resolution and formant structure. It also decreases the system response to individual harmonics (compared to formants). It is interesting to have a close look and investigate why it has such a selective-smoothing effect. The following remarks can be readily observed.

1) Formants extend to neighboring filters while spurious peaks and individual harmonics usually do not. This enhances the response of the ALSD to formants, relative to nonformants or using wider filters, while preserving the resolution.
2) Smoothing performed by widening the filters will inevitably deteriorate the filter's resolution. This is not

necessarily the case with the ALSD since the original sharp filters are still preserved.

3) Wider filters cause loss (or reduction) in the synchrony due to the decrease in the signal's periodicity. This causes significant deterioration in the output and hence weakens the response. On the other hand, the ALSD preserves the sharp filters. Strong periodicity will cause a strong response from the central GSD that is tuned to the corresponding filter. Such strong response, (though averaged with weaker responses from GSDs tuned to the same filter but applied to neighboring filters), will lead to a strong overall response due to the nonlinearity of the divider.

This last point needs more elaboration. The reason the ALSD gives better performance than using wider filters is because the GSD is a nonlinear processor that is based on taking the inverse of a certain measure of periodicity. Using wider filters is equivalent to an averaging process of neighboring filters which is followed by taking the inverse. On the other hand, the ALSD operation is equivalent to averaging *after* taking the inverse. The interesting properties of the ALSD could be illustrated by investigating the relationship between the mean of the inverses (call it y_{mi}) and the inverse of the means (call it y_{im}).

Equation (4) shows that the difference term in the denominator is the main periodicity-indicator, while the sum term in the numerator is mainly for normalization. We will denote the difference term as x. The more periodicity the signal exhibits, the smaller x will be. If we ignore the limiter, we can represent the relationship between the output y and the difference term x by an inverse relationship, i.e., $y = K/x$, where x is nonnegative. This relationship is shown in Fig. 6. It is clear that strongly periodic signals have very small values of x compared to aperiodic signals. At $x = 0$, the signal is perfectly periodic and the output is infinite (before being limited by the saturating nonlinearity).

In Fig. 6, assume that we have two inputs x_1 and x_2 to the *same* GSD with corresponding outputs y_1 and y_2. Averaging the input filters' outputs is equivalent to averaging the periodic-indicators (difference terms). The mean of x_1 and x_2 is x_m whose output is y_{im}. It can be easily proved that the mean of y_1 and y_2 lies at the midpoint of the straight line connecting the two points as shown in the figure. Therefore, it is clear that the mean of the outputs, y_{mi}, is larger than the output of the mean y_{im}. This could be represented as follows:

$$y_1 = K/x_1, \quad y_2 = K/x_2, \quad x_m = (x_1 + x_2)/2 \tag{6}$$

$$y_{im} = K/x_m = 2K/(x_1 + x_2) \tag{7}$$

$$y_{mi} = (y_1 + y_2)/2 = K/2x_1 + K/2x_2 \tag{8}$$

$$\delta y = y_{mi} - y_{im} = \frac{(x_1 - x_2)^2}{2x_1x_2(x_1 + x_2)} \geq 0. \tag{9}$$

This argument can be extended for averaging more than two terms. If we have three inputs x_1, x_2 and x_3 with the corresponding outputs y_1, y_2 and y_3. Then

$$\delta y = y_{mi} - y_{im} = \frac{x_3(x_1-x_2)^2 + x_2(x_1-x_3)^2 + x_1(x_3-x_2)^2}{3x_1x_2x_3(x_1 + x_2 + x_3)} \geq 0. \tag{10}$$

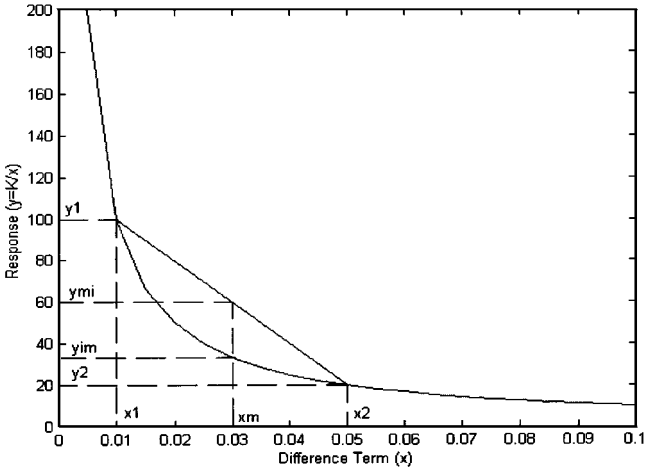

Fig. 6. Illustration of the inverse-mean relation.

Using mathematical induction, we can prove that for n inputs $x_1, x_2, \ldots, x_n$, whose corresponding outputs are: $y_1, y_2, \ldots, y_n$, we have

$$\delta y = \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \left[(x_i - x_j)^2 \prod_{\substack{k \neq i \\ k \neq j}} x_k \right]}{n \prod_{i=1}^{n} x_i \sum_{j=1}^{n} x_j} \geq 0. \tag{11}$$

Equation (11) shows that the mean of the inverses is always greater than (or equal to) the inverse of the means. Therefore, the ALSD output is greater than the output of the same GSD tuned to a wider filter. It is important to note that this argument is not valid for the mean of the outputs from GSDs tuned to *different* frequencies. Moreover, careful examination of (11) leads to interesting conclusions regarding the difference δy.

- The difference δy is inversely proportional to $x_1, x_2, \ldots, x_n$. Thus, the response enhancement (relative to smoothing using wider filters) is stronger for small xs which indicate higher periodicity (peaks). This indicates that the ALSD enhances peaks more than flat regions or valleys. It also enhances formants (which tend to extend to neighboring filters causing the neighboring xs to be small) more than spurious peaks.
- The difference δy is directly proportional to the differences between the xs and inversely proportional to their product. Thus δy increases with dissimilar xs. This indicates that the ALSD works better with sharp filters and strong peaks. This is an intuitive result that agrees with our experiments. It indicates that when using sharp filters, the ALSD will smooth the response while preserving the peaks by enhancing them relative to other regions, therefore preserving resolution.
- The difference δy is inversely proportional to the number of averaged inputs n. This clearly indicates that using too many filters in the averaging process is not desirable since it destroys the ALSD advantage and hence deteriorates the resolution. It emphasizes the importance of *localizing* the averaging process in order to preserve the "place" infor-

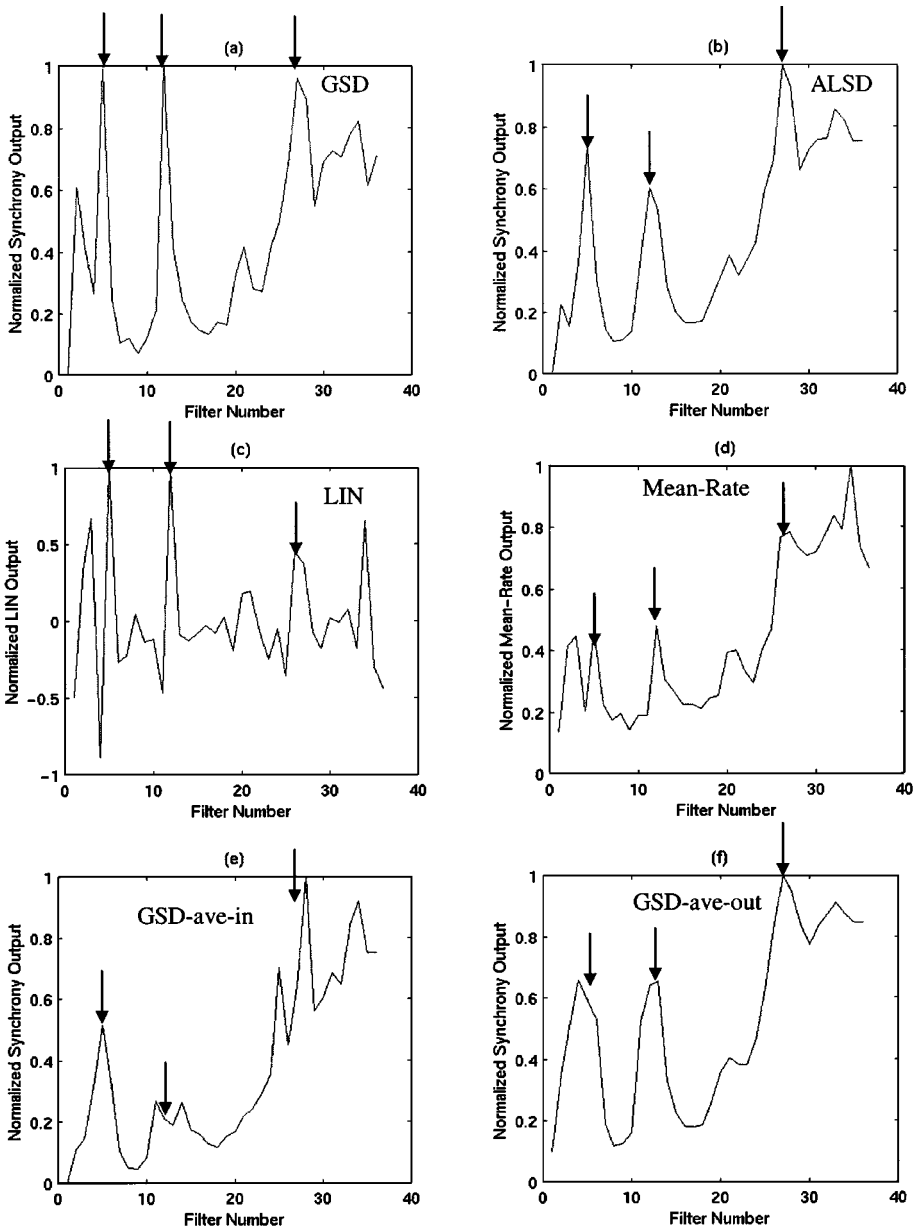

Fig. 7. System response for three noisy sinusoidal signals with frequencies 500 Hz, 1000 Hz, and 3000 Hz and SNR $=$ 0 dB. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs.

mation. A similar effect was found in the human auditory system [41].

These comments lead to the conclusion that the ALSD has the ability to smooth the response and hence decrease the spurious peaks and the sensitivity of the system to mismatches. Nevertheless, it still relatively preserves the resolution by enhancing the true formants relative to spurious peaks, flat regions and valleys. Therefore, its operation is a combination of averaging (smoothing) and lateral-inhibition (sharpening) simultaneously. Whereby, it selectively smooths out the undesired peaks (spurious peaks and harmonics) while sharpening the desired peaks (formants).

It is also clear how the ALSD provides an additional degree of freedom. Choosing the number of channels n to include in the averaging depends on the resolution and sharpness of the filters. When we have a large number of sharply tuned filters (i.e., high resolution), we can use larger value of n than with a smaller number of loosely tuned filters since in the former case we would have large differences between, and small products of, the xs. The ALSD would not be needed in the latter case anyway since the response is smooth enough.

In general, the choice of n does not have to be the same for all channels. Thus different channels could average different number of GSDs. This is especially useful for high-frequency channels that tend to be wider in band, less affected by harmonics and less sensitive to mismatches than low-frequency channels. In some cases, smoothing for such channels may not be needed at all. The choice of n is therefore decided by experimentation according to the resolution, accuracy, and weight required for each channel.

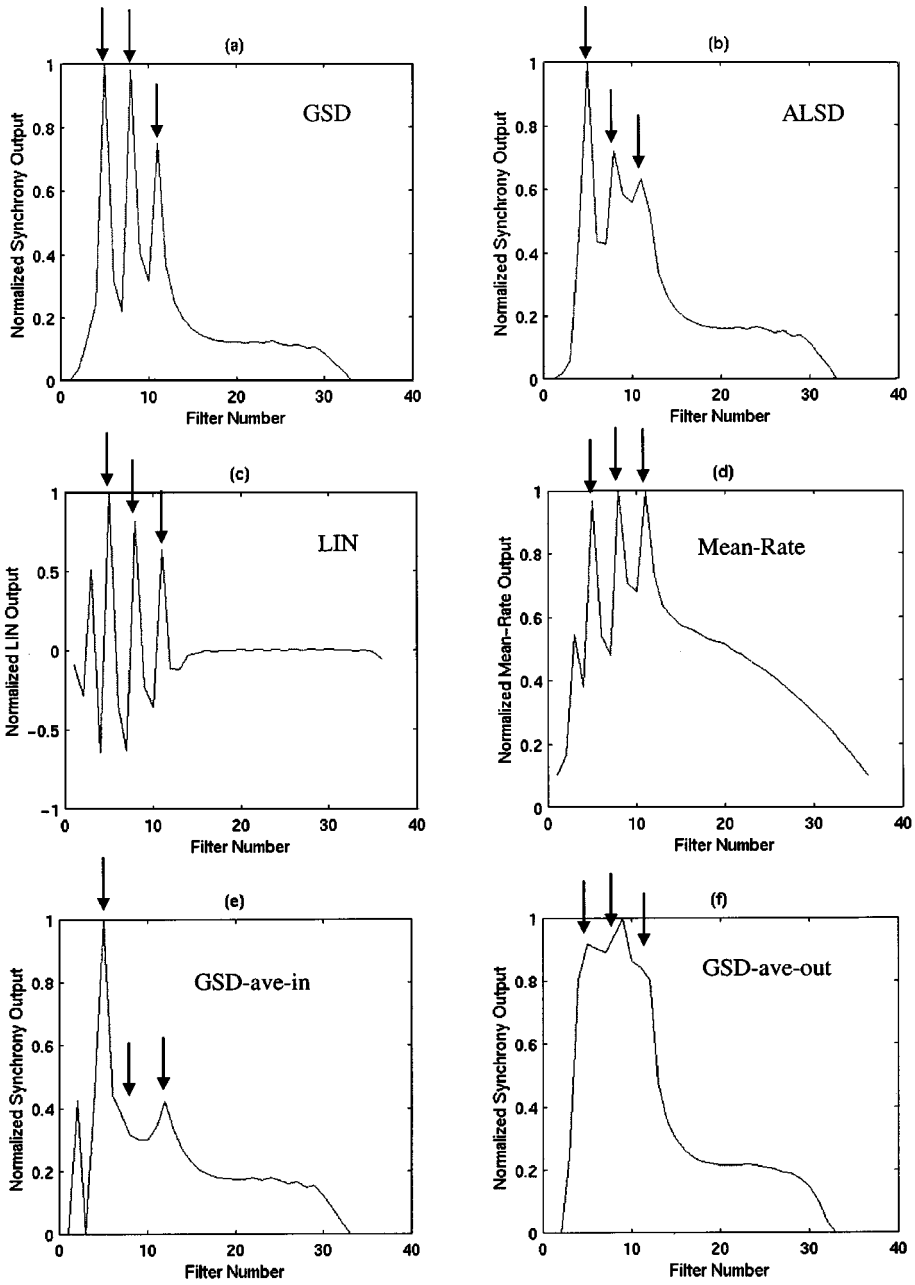

Fig. 8. System response for three sinusoidal signals with frequencies 500 Hz, 700 Hz, and 900 Hz. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs.

III. Performance Evaluation

A. Formant Representation

The spectral responses of the various auditory-based systems discussed previously to sine waves and speech are shown in Figs. 7–12. Each response represents a snapshot of the spectrum in the middle of the sound. Each figure shows the responses of six different systems. The first four parts, (a)–(d), in each figure are the GSD (generalized synchrony detector, developed by Seneff [44]), ALSD (average localized synchrony detector, developed by the authors [3], [7]), LIN (lateral inhibitory network, developed by Shamma [47]), and the mean-rate [7], [44], respectively. Part (e) shows the GSD response using wider filters (i.e., averaged inputs) and part (f) shows the GSD response when the outputs from neighboring GSDs are averaged (i.e., averaged outputs).

Fig. 7 shows the responses for three sine waves at 500 Hz, 1000 Hz, and 3000 Hz. All systems gave strong peaks at the three frequencies. The GSD peaks are sharper than those of the mean-rate. It suffers, however, from strong spurious peaks. The ALSD managed to significantly suppress the spurious peaks while preserving the sharpness of the peaks. The LIN and the averaged-input (wider-filter) GSD had significant spurious peaks at several frequency locations. The noise added to the signals shows the GSD and ALSD systems to yield better responses than the mean-rate. The LIN suffers from significant spurious peaks but the original peaks are still dominant.

Fig. 8 shows the responses to three sine waves at 500 Hz, 700 Hz, and 900 Hz. This represents a worst-case scenario, for the ALSD, from the resolution standpoint since those sine waves

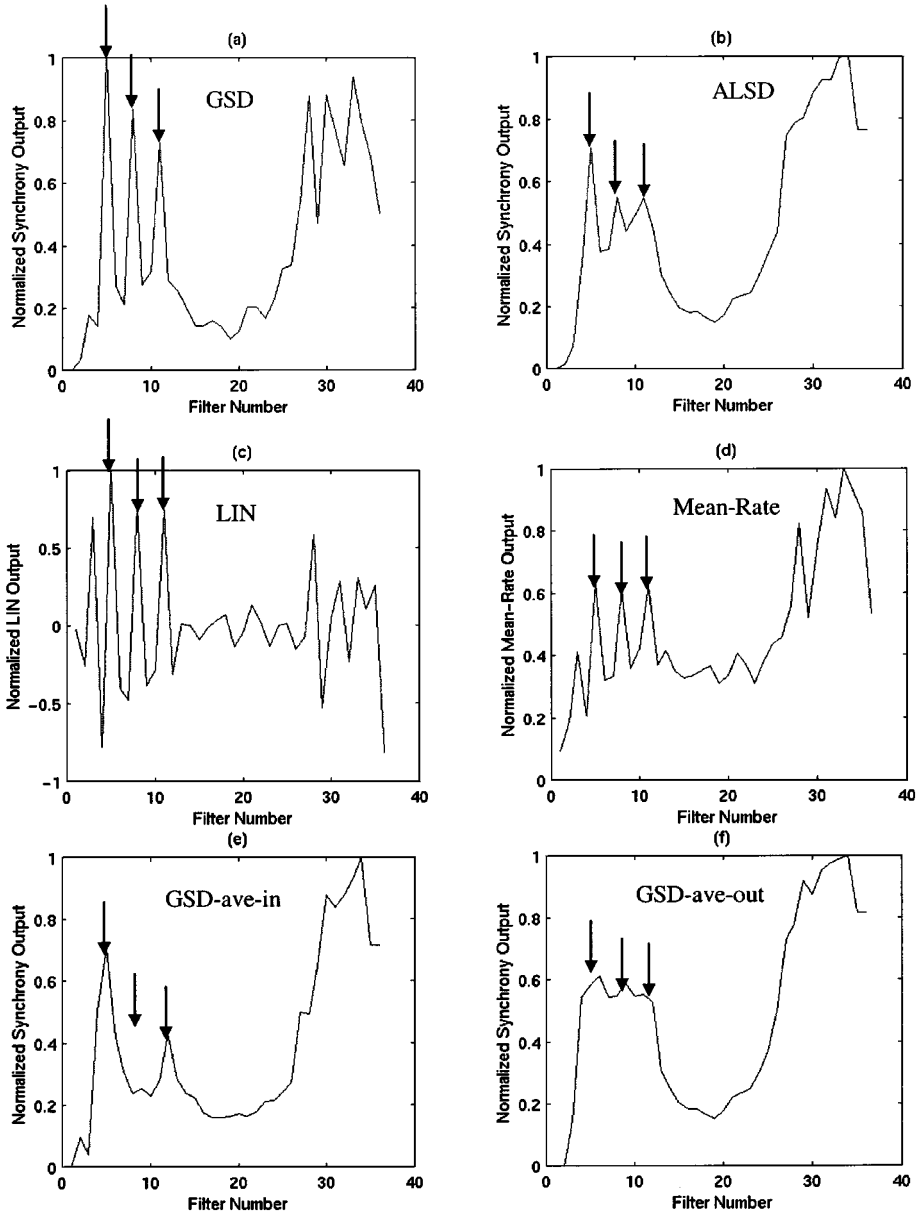

Fig. 9. System response for three noisy sinusoidal signals with frequencies 500 Hz, 700 Hz, and 900 Hz and SNR = 0 dB. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs.

are spaced about one Bark apart and hence represent the maximum resolution of the system. It is clear that the ALSD is still capable of resolving the three signals (though with considerable smoothing). On the other hand, the averaged-input (wider-filter) and averaged-output GSDs were unable to resolve the three signals. This clearly illustrates the selective-smoothing effect of the ALSD, that was mentioned earlier, and its ability to smooth the response while preserving the resolution. Fig. 9 shows that the ALSD is still able to resolve the three signals in the presence of significant noise (SNR = 0 dB).

Fig. 10 shows the responses to the back vowel /ao/ spoken by a female speaker (from the TIMIT database). The GSD, LIN, and mean-rate outputs suffer from significant peaks below the first formant (F1), which are due to the fundamental frequency individual harmonics. Those peaks are worst for the GSD. The ALSD, however, has significantly attenuated such harmonics. It is instructive to compare the performance of the ALSD with that of the averaged-input (wider-filter) and averaged-output GSDs. Though the three systems have smoothed the response, the selective smoothing of the ALSD could be easily contrasted with the blind smoothing of the other two systems. The ALSD attenuated the spurious peaks while preserving F1 almost unaffected. On the other hand, the wider-filter GSD attenuated F1, while strongly enhancing one of the harmonics. The averaged-output GSD was not able to remove one of the harmonics and it affected the F1 peak by shifting it from its correct position.

Fig. 11 illustrates, strikingly, the benefit of the ALSD system. The responses are for the vowel /aa/ spoken by a female speaker. This is a low-back vowel that is characterized by high F1 and

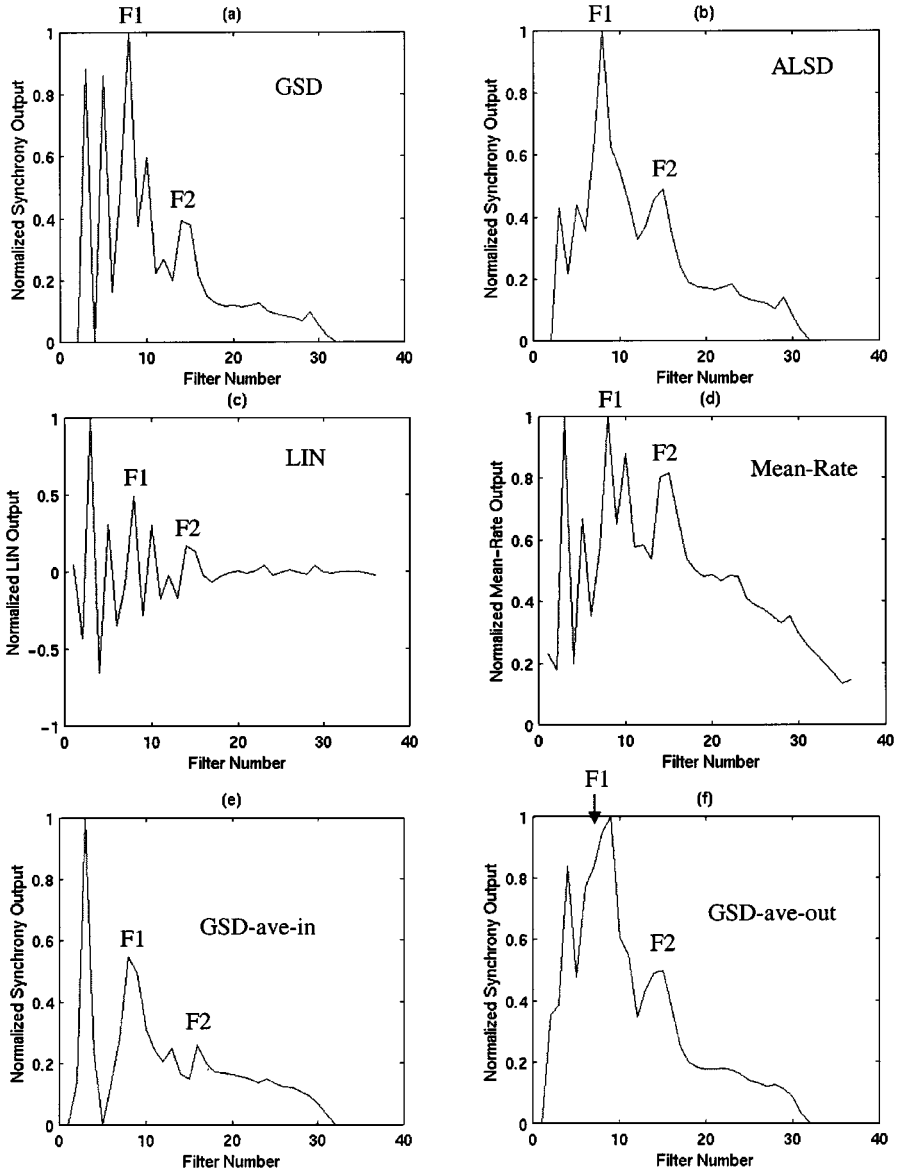

Fig. 10. System response for the vowel /ao/ spoken by a female speaker. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs. In (f), F1 is shifted.

low F2. Thus it has a relatively tight requirement on the resolution to resolve the two close formants. It is clear that not only was the ALSD able to resolve the formants, but it also significantly reduced the spurious peaks that existed in the GSD, LIN, and mean-rate outputs. Those spurious peaks were located below F1 and between F1 and F2. They were so strong that it was difficult to tell the real formants from the spurious peaks. The wider-filter GSD, on the other hand, totally failed in suppressing the harmonics or extracting the formants. The averaged-output response was also unacceptable due to the significant peak-splitting at the formant positions. This was another example of the superiority of the selective-smoothing approach of the ALSD over traditional smoothing techniques.

Fig. 12 shows the response to the vowel /aa/ spoken by a male speaker. Though the harmonics below F1 were less severe than the female case, similar conclusions to those of the female speakers were obvious. The ALSD was consistently capable of resolving the formants and reducing the spurious peaks that plagued the GSD, LIN, and mean-rate. The wider-filter and averaged-output GSDs, on the other hand, repeatedly showed failures and errors in their responses.

It is important to note that the aforementioned figures were shown for illustration purposes and they represent cases were the problems of the traditional systems were evident and strong. There are many cases in which the GSD, LIN, and mean-rate responses were quite similar to the ALSD output. The ALSD response, however, showed consistent ability (with various sounds and speakers) to correctly extract the formants, selectively smooth the response and significantly reduce the spurious peaks.

B. Recognition Experiments

The other auditory-based systems mentioned in this work (i.e., the mean-rate, LIN, and GSD) have been previously eval-

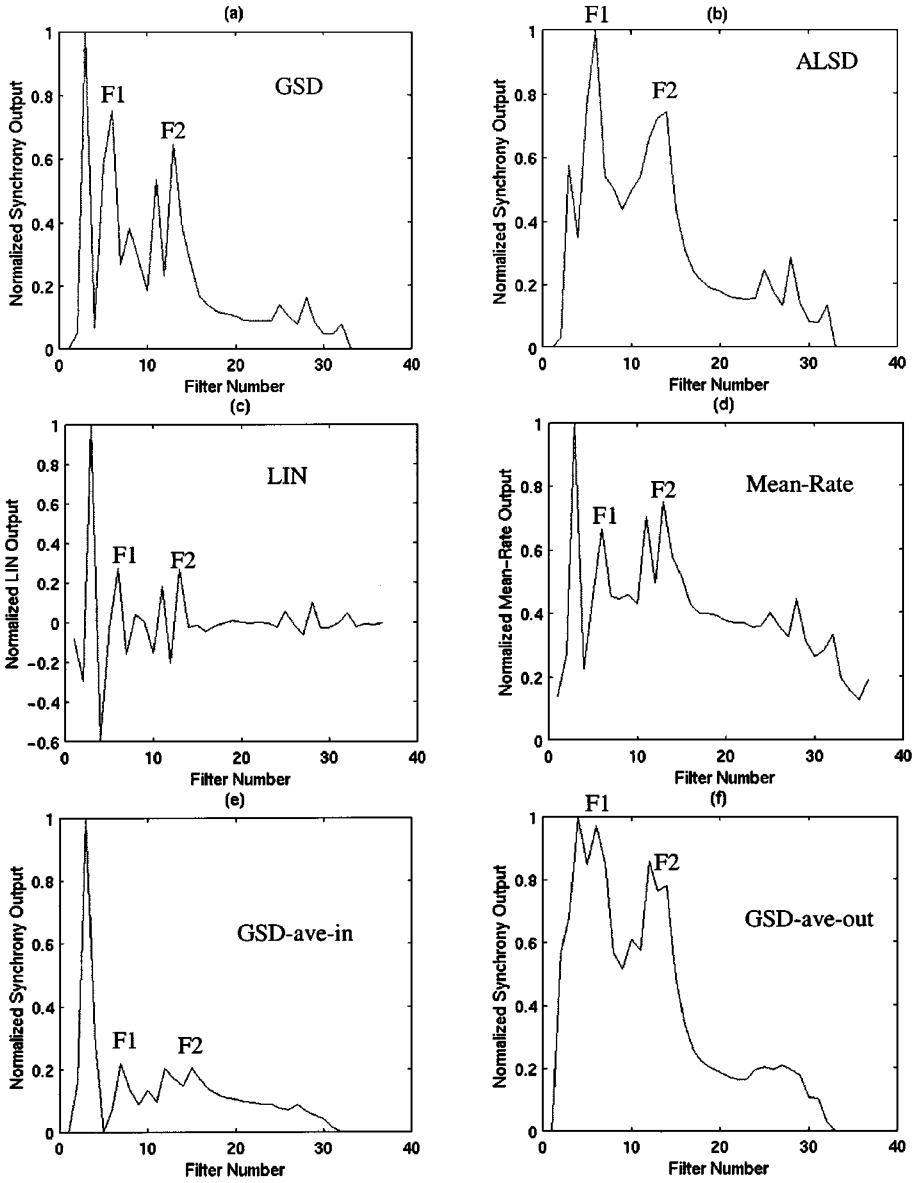

Fig. 11. System response for the vowel /aa/ spoken by a female speaker. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs.

uated in various ASR applications in the literature. The ALSD front-end system was also tested and evaluated in numerous phoneme recognition experiments by the authors [1]–[7]. In this work, our goal is to evaluate the ALSD system's ability to extract the formants and compare its performance with other auditory-based systems.

Formant extraction has been one of the most challenging tasks in speech processing [45], [52]. Though extremely desirable and useful, an accurate formant tracker is something that is yet to be built. The reason behind this is the peak-splitting and peak-merging phenomena that often happen in the spectrum and cause significant inconsistencies in the formant structure. Some researchers tried to build sophisticated formant-tracking algorithms to deal with this problem, while others opted to use alternative approaches like using the "effective" formants in lieu of regular formants, where "effective" usually refers to a simpler and more consistent version of the formants [11], [27], [40].

The ALSD was tested in some ASR experiments that aim at evaluating its formant extraction and representation ability for both clean and noisy speech. The outputs tested were the traditional Mel-scaled cepstral analysis [8], [30], the mean-rate, the LIN, the GSD, and the ALSD. They were evaluated in a vowel classification experiment that classifies four vowels: /aa/, /uw/, /ae/, and /iy/. The vowels were extracted from different contexts of continuous speech of multiple speakers with seven different dialects of American English from the TIMIT database. The vowels are chosen to represent the four extremes of the vowel chart and the four main tongue positions. Therefore, they could be classified using the first two formants. The experiments are performed on clean speech and on speech distorted by white Gaussian noise with different signal-to-noise ratios.

The first two formants are extracted using a relatively simple formant tracking algorithm. The algorithm picks the peaks that

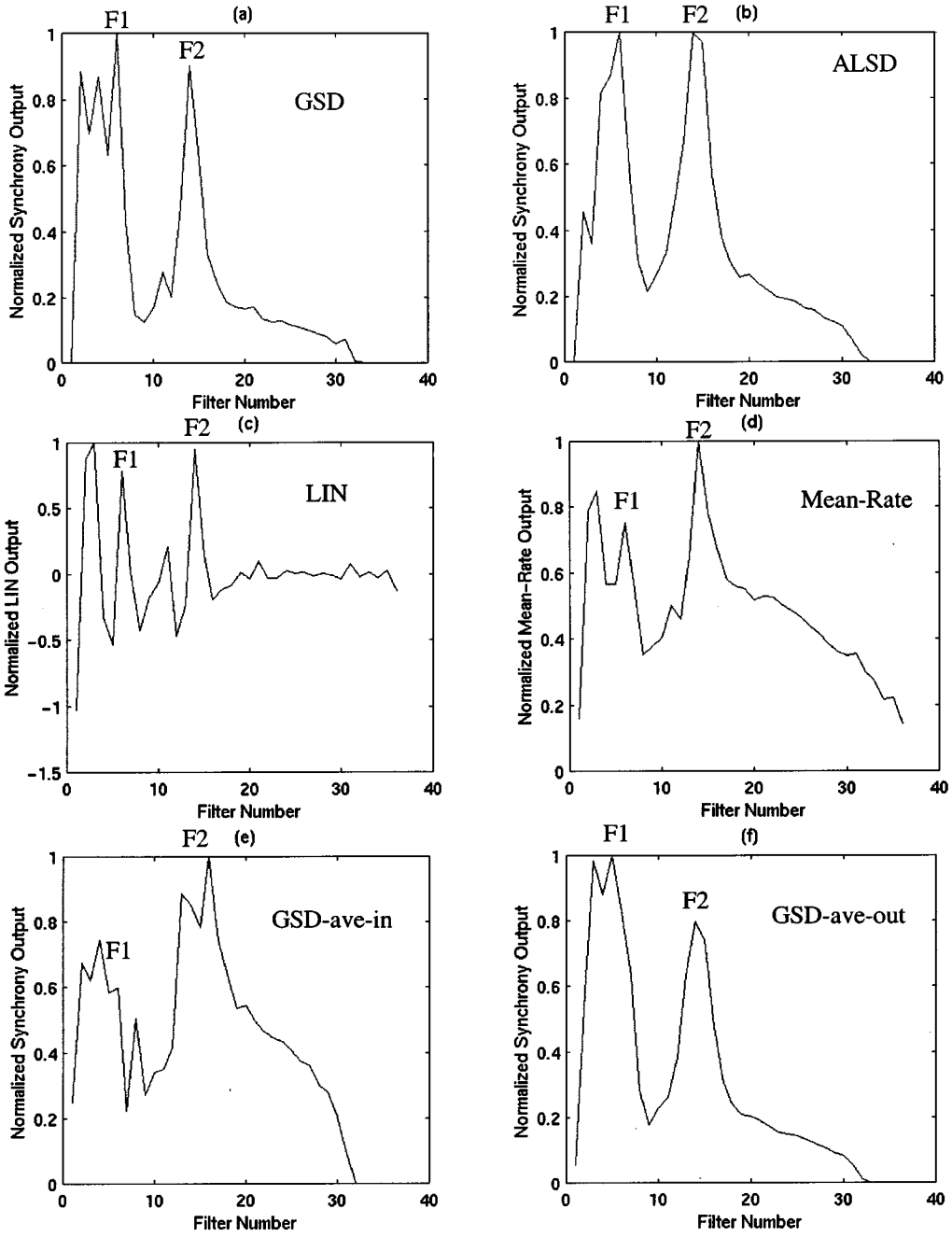

Fig. 12. System response for the vowel /aa/ spoken by a male speaker. (a) GSD, (b) ALSD, (c) LIN, (d) mean-rate, (e) GSD with averaged inputs (wider filters), and (f) GSD with averaged outputs.

satisfy certain constraints in location, amplitude and continuity. The four vowels are classified using two threshold values (for the two formants) that divide the two-dimensional space into four regions using a Bayesian classification maximum posterior probability criterion. The system is trained using 120 tokens from six speakers (three males and three females) and tested on 30 different speakers with more than 1000 of the aforementioned vowels. Three measurements are taken in the middle third of each vowel and a majority rule is used for the decision.

The choice of this classification method is motivated by the purpose of the experiments. Since we are interested in evaluating the systems' abilities to accurately extract the formants in the presence of noise, it is necessary to ensure that the classification decision is based on the formant positions and not any other spectral artifacts. Moreover, we need to evaluate the ability of the system to reduce spurious peaks and hence enable us to use a relatively simple formant tracking algorithm.

The results of the experiments are summarized in Fig. 13. The mean-rate and traditional cepstral analysis give almost identical results. For clean speech, we see that the ALSD gives the best performance, followed by the cepstral/mean-rate, the LIN, and, finally, the GSD. The relatively bad performance of the LIN and the GSD is attributed to the presence of spurious peaks that cause errors in formant extraction. The ALSD smooths the response, while preserving the formants, and hence improves the

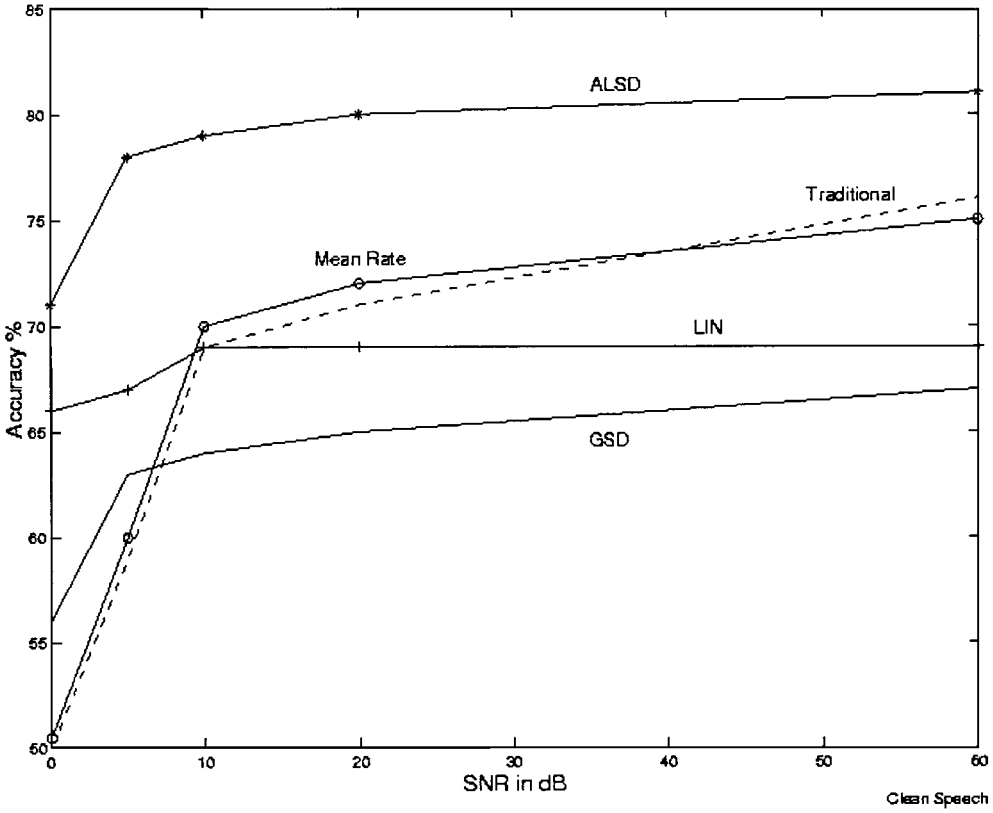

Fig. 13. Classification accuracy, in the presence of different levels of white Gaussian noise, for the different front-end processing systems in vowel recognition experiment on 30 speakers with seven different dialects of American English from the TIMIT database. The ALSD gives the best performance. The mean-rate and traditional outputs deteriorate more sharply with noise than the synchrony measures.

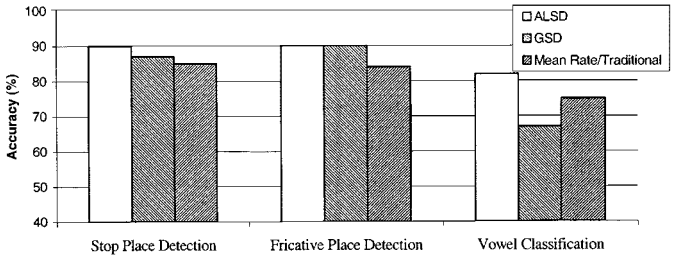

Fig. 14. Comparison between the ALSD, GSD, and mean-rate in different phoneme recognition experiments.

performance. When noise is added to the system, the performance deteriorates. The deterioration is worst for the mean-rate, which falls sharply. This is in agreement with previous findings which demonstrated that synchrony measures are usually more robust than mean-rate. We can also see that the deterioration of the ALSD response with noise is almost identical to that of the GSD, which indicates that the ALSD preserved the robustness of the GSD while improving the performance by decreasing the spurious and individual harmonic peaks.

Other large-scale phoneme recognition experiments in continuous speech have verified the usefulness of the ALSD [4]–[7]. Those include fricative, stop, and vowel recognition experiments on speaker-independent continuous speech from the TIMIT database. Statistically guided knowledge-based algorithms were used for the various recognition tasks. The results are summarized in the chart shown in Fig. 14. In the place of articulation detection of stop consonants, the ALSD showed a consistent improvement of 3% with respect to the GSD in clean and noisy speech [4], [7]. This is due to the partial reliance of such detection on the formant of the neighboring vowels [5], [7]. These included all classes of vowels spoken by different speakers in various contexts from the TIMIT database. Moreover, a comparison of the ALSD with the mean-rate and traditional cepstral analysis showed a consistent improvement of 5% and 6% in the stop and fricative place of articulation detection respectively.

IV. Conclusion

A new auditory-based speech processing system, namely the ALSD, is developed to alleviate some of the limitations of the GSD, such as the presence of spurious peaks, sensitivity to implementation mismatches and response to individual harmonics. The system is compared with several other systems in their formant extraction ability from clean and noisy speech. The other systems are the traditional cepstral analysis, the Bark-scaled mean-rate detector, the LIN detector, and the GSD.

The results demonstrate the advantage of the ALSD in extracting the formants and reducing the spurious peaks. They also indicate the superiority of the synchrony measures, in the presence of noise, compared to the mean-rate and traditional systems. In spite of their superb formant extraction ability, the LIN

and GSD systems are plagued by significant spurious peaks, which complicate the formant-tracking task. The ALSD significantly reduces such spurious peaks by selectively smoothing the output response while preserving the formants and resolution of the system. It simplifies and improves the formant extraction and is more suitable for analog hardware implementation.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] A. M. A. Ali *et al.*, "An acoustic-phonetic feature-based system for the automatic recognition of fricative consonants," in *Proc. IEEE ICASSP-98*, vol. 2, 1998, pp. 961–964.
[2] A. M. A. Ali *et al.*, "Automatic detection and classification of stop consonants using an acoustic-phonetic feature-based system," in *Int. Congr. Phonetic Sciences (ICPhS 99)*, 1999.
[3] A. M. A. Ali *et al.*, "Auditory-based speech processing based on the average localized synchrony detection," in *Proc. IEEE ICASSP-2000*, vol. 3, 2000, pp. 1623–1626.
[4] A. M. A. Ali *et al.*, "Robust classification of stop consonants using auditory-based speech processing," in *Proc. IEEE ICASSP 2001*, 2001.
[5] A. M. A. Ali *et al.*, "Acoustic-phonetic features for the automatic classification of stop consonants," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 833–841, Nov. 2001.
[6] A. M. A. Ali *et al.*, "Acoustic-phonetic features for the automatic classification of Fricatives," *J. Acoust. Soc. Amer.*, vol. 109, pp. 2217–2235, May 2001.
[7] A. M. A. Ali, "Auditory-based acoustic-phonetic signal processing for robust continuous speech recognition," Ph.D. dissertation, Dept. Elect. Eng., Univ. Pennsylvania, Philadelphia, 1999.
[8] J. B. Allen, "Cochlear modeling," *IEEE ASSP Mag.*, pp. 3–29, Jan. 1985.
[9] ——, "How do humans process and recognize speech?," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 567–576, 1994.
[10] T. Anderson, "A comparison of auditory models for speaker independent continuous speech," in *Proc. IEEE ICASSP*, 1993, pp. 231–234.
[11] R. Carlson *et al.*, "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, G. Fant and M. Tatham, Eds. New York: Academic, 1975, pp. 55–82.
[12] J. R. Cohen, "Application of an auditory model to speech recognition," *J. Acoust. Soc. Amer.*, vol. 85, pp. 2623–2629, 1989.
[13] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-28, pp. 357–366, 1980.
[14] B. Delgutte, "Speech coding in the auditory nerve: Part II. Processing schemes for vowel-like sounds," *J. Acoust. Soc. Amer.*, vol. 75, no. 3, pp. 879–886, 1984.
[15] L. Deng and C. D. Geisler, "A composite auditory model for processing speech sounds," *J. Acoust. Soc. Amer.*, vol. 82, pp. 2001–2012, 1987.
[16] L. Deng *et al.*, "A composite model of the auditory periphery for the processing of speech," *J. Phonetics*, vol. 16, pp. 93–108, 1988.
[17] O. Ghitza, "Auditory models and human performance in tasks related to speech coding and speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 115–132, Apr. 1994.
[18] ——, "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.
[19] ——, "Temporal nonplace information in the auditory nerve firing patterns as a front-end for speech recognition in a noisy environment," *J. Phonetics*, vol. 16, pp. 106–123, 1988.
[20] S. Greenberg, "Acoustic transduction," *J. Phonetics*, vol. 16, pp. 3–17, 1988.
[21] ——, "The ear as a speech analyzer," *J. Phonetics*, vol. 16, pp. 139–149, 1988.
[22] D. M. Harris and P. Dallos, "Forward masking of auditory nerve fiber responses," *J. Neurophys.*, vol. 42, pp. 1083–1107, 1979.
[23] H. Hermansky and A. L. Cox, "Perceptual linear predictive (PLP) analysis–resynthesis technique," in *Proc. 2nd Eur. Conf. Speech Communication and Technology*, Genova, Italy, 1991.
[24] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, pp. 587–589, Oct. 1994.
[25] H. Hermansky *et al.*, "RASTA-PLP speech analysis technique," in *Proc. IEEE ICASSP*, vol. 1, 1992, pp. 121–124.
[26] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *J. Acoust. Soc. Amer.*, vol. 87, pp. 1738–1752, 1990.
[27] Z. Hu and Z. E. Barnard, "Efficient estimation of perceptual features for speech recognition," in *Proc. 5th Eur. Conf. Speech Communication and Technology*, Athens, Greece, May 1997.
[28] M. J. Hunt and C. Lefebvre, "Speaker dependent and independent speech recognition experiments with an auditory model," in *Proc. IEEE ICASSP*, 1988, pp. 215–218.
[29] ——, "Speech recognition using an auditory model with pitch synchronous analysis," in *Proc. IEEE ICASSP*, 1987, pp. 813–816.
[30] C. R. Jankowski *et al.*, "A comparison of signal processing front ends for automatic word recognition," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 286–293, 1995.
[31] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition, Fundamentals and Applications*. Norwell, MA: Kluwer, 1996.
[32] J. M. Kates, "A time-domain digital cochlear model," *IEEE Trans. Signal Processing*, vol. 39, pp. 2573–2592, 1991.
[33] D.-S. Kim *et al.*, "Auditory processing of speech signals for robust speech recognition in real-world noisy environments," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 55–69, Jan. 1999.
[34] C.-H. Lee *et al.*, *Automatic Speech and Speaker Recognition, Advanced Topics*. Norwell, MA: Kluwer, 1996.
[35] W. Liu *et al.*, "Voiced-speech representation by an analog silicon model of the auditory periphery," *IEEE Trans. Neural Networks*, vol. 3, pp. 477–487, 1992.
[36] R. F. Lyon and C. Mead, "An analog electronic cochlea," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1119–1134, 1988.
[37] C. Mead and M. Ismail, *Analog VLSI Implementation of Neural Systems*. Norwell, MA: Kluwer, 1989.
[38] P. Mueller *et al.*, "A programmable analog neural computer with applications to speech recognition," in *Proc. CISS'95*, 1995.
[39] Y. Ohshima, "Environmental robustness in speech recognition using physiologically-motivated signal processing," Ph.D. dissertation, Carnegie Mellon Univ., Pittsburgh, PA, 1993.
[40] K. Paliwal *et al.*, "A study of two-formant models for vowel identification," *Speech Commun.*, vol. 2, pp. 295–303, 1983.
[41] M. B. Sachs and E. D. Young, "Effects of nonlinearities on speech encoding in the auditory nerve," *J. Acoust. Soc. Amer.*, vol. 68, pp. 858–875, 1980.
[42] ——, "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.
[43] S. Sandhu and O. Ghita, "A comparative study of Mel cepstra and EIH for phone classification under adverse conditions," in *Proc. IEEE ICASSP*, 1995, pp. 409–412.
[44] S. Seneff, "A joint synchrony/mean rate model of Auditory Speech Processing," *J. Phonetics*, vol. 16, pp. 55–76, 1988.
[45] ——, "Pitch and spectral analysis of speech based on an auditory synchrony model," Ph.D. dissertation, Mass. Inst. Technol., Cambridge, 1985.
[46] S. Shamma, "Speech processing in the auditory system I: The representation of speech sounds in the responses of the auditory nerve," *J. Acoust. Soc. Amer.*, vol. 78, pp. 1612–1621, 1985.
[47] ——, "The acoustic features of speech sounds in a model of auditory processing: Vowels and voiceless fricatives," *J. Phonetics*, vol. 16, pp. 77–91, 1988.
[48] H. Sheikhzadeh and L. Deng, "Speech analysis and recognition using interval statistics generated from a composite auditory model," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 90–94, Jan. 1998.
[49] R. L. Smith and J. J. Zwislocki, "Short-term adaptation and incremental responses of single auditory-nerve fibers," *Biol. Cybern.*, vol. 17, pp. 169–182, 1975.
[50] R. M. Stern *et al.*, "Multiple approaches to robust speech recognition," in *Proc. ICSLP*, 1992.
[51] R. Vergin *et al.*, "Generalized Mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 525–532, 1999.
[52] L. Welling and H. Ney, "Formant estimation for speech recognition," *IEEE Trans. Speech Audio Processing*, pp. 36–48, 1998.
[53] 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.*, pp. 1381–1403, 1979.

Ahmed M. Abdelatty Ali (S'91–M'00) received the B.Sc. and M.Sc. degrees (with distinction and honors) in electrical engineering from Ain Shams University, Cairo, Egypt, in 1991 and 1994, respectively. He received the Ph.D. degree in electrical engineering from the University of Pennsylvania, Philadelphia, in 1999.

From 1991 to 1994, he was a Teaching Assistant with the Electronics and Communication Department, Ain Shams University, Cairo. He is currently with Texas Instruments R&D, Warren, NJ, and an adjunct Assistant Professor at the University of Pennsylvania. His research interests include mixed-signal IC design, digital signal processing, and speech processing and recognition.

Dr. Ali is the recipient of the S. J. Stein Award for his doctoral research achievements.

Jan Van der Spiegel (M'72–SM'90–F'02) received the Engineering degree in electromechanical engineering and the Ph.D. degree in electrical engineering from the University of Leuven, Leuven, Belgium, in 1974 and 1979, respectively.

From 1980 to 1981, he was a Postdoctoral Fellow at the University of Pennsylvania, after which he became Assistant Professor of electrical engineering. In 1987, he became an Associate, and, in 1995, a Full Professor of electrical engineering. He is currently the Chairman of the Department and Director of the Center for Sensor Technology. His research interests are in analog and digital integrated circuits for intelligent sensors, data acquisiton, sensory data processing systems, and acoustic–phonetic feature extraction for automatic speech recognition. He is the editor for North and South America for *Sensors and Actuators* and is on the editorial boards of the *International Journal of High Speed Electronics* and the *Journal of the Brazilian Microelectronics Society*.

Dr. Van der Spiegel is Bicentenneial Chair of the Class of 1940. He received the Presidential Young Investigator Award and the S. R. Warren and C. & M. Lindback Awards for distinguised teaching. He has served on several IEEE Program Committees, and is currently on the Program and Executive Committees of the ISSCC. He is a member of Phi Beta Delta and Tau Beta Pi.

Paul Mueller received the M.D. degree from Bonn University, Bonn, Germany.

He was formerly with the Rockefeller University, New York, and the University of Pennsylvania, Philadelphia, and is currently Chairman of Corticon, Inc., Philadelphia. Since 1953, he has worked in molecular and systems neuroscience and has been involved in theoretical studies and hardware implementation of neural networks since the early 1960s.