Auditory Nerve Representation Criteria for Speech Analysis/Synthesis

ODED GHITZA

Abstract—**Traditional speech analysis/synthesis techniques are designed to produce synthesized speech with a spectrum (or waveform) that is as close as possible to the original. It is suggested, instead, that representations of the synthetic and the original speech be matched at the auditory nerve level. This concept has been used in conjunction with the sinusoidal representation of speech analysis/synthesis suggested by McAulay and Quatieri [6]. Based on informal listening, the synthesized speech is natural, including some tonal artifact, and highly intelligible for various kinds of speech material, in both quiet and noisy environments. The inherent dominance property of the auditory nerve representation reduces the number of sinusoidal components needed for synthesis by approximately 70 percent, offering a potential for reduced data rate.**

I. Introduction

TRADITIONAL speech analysis/synthesis systems are designed to produce synthesized speech with properties similar to the original. Waveform coders attempt to reproduce the original waveform. Vocoder schemes are considered to perform well if they reproduce the short-term power spectrum. The deviation from the original depends on the quantization used to achieve a desired data rate.

This paper suggests a new speech analysis technique that exploits the limitations imposed by the human peripheral auditory mechanism on speech transduction. It is proposed that speech be synthesized so as to match the representations of the synthesized and the original speech at the auditory nerve level, that is, at the output of the auditory periphery. An analysis/synthesis system has been created to examine this approach. The analysis is based on a simple approximation of the auditory periphery (up to the auditory nerve level) followed by an heuristic nonlinear relative spectrum intensity measure that is assumed to be derived in higher stages of the nervous system. This measure uses timing–synchrony information in an attempt to exploit the in-synchrony phenomena observed in the neuron firing patterns. We call the resulting representation the "in-synchrony-bands spectrum" (SBS). In an ideal situation, if the modeling is correct, this representation should contain all the necessary information about the acoustic environment since the auditory nerve is the only link to the higher stages. It is this information that the synthesis should reproduce. Thus, an "inverse law" is required to determine the appropriate parameters that should control the synthesizer. Alternatively, one can run a synthesis-by-analysis system that matches the SBS of the synthesized speech to that of the original.

The desired inverse law has not yet been found. However, a one-step rule for synthesis by analysis is proposed. It is based on combining the SBS with the sinusoidal representation system suggested by McAulay and Quatieri [6]. In the proposed system, termed "the SBS controlled sinusoidal representation," the SBS is used to select the frequency components necessary in the sinusoidal representation to preserve the SBS. The resulting synthesized speech has an SBS spectrum identical to that of the original. Furthermore, the resulting speech is natural, with some tonal artifact, and highly intelligble for all kinds of acoustical stimuli in our database, including single male and female speakers in quiet and in noise, two overlapped speech waveforms, music waveforms, and speech in musical background. Very few frequency components suffice for an adequate representation of all the speech features. This suggests a potential for an efficient speech coding system based on this technique. Finding the appropriate coding strategies, however, is beyond the scope of this work.

The next section describes the SBS spectrum. The integration of the SBS concept into the sinusoidal representation system is described in Section III, and the resulting performance of the overall system is demonstrated in Section IV. The Appendix briefly describes the sinusoidal representation analysis/synthesis system.

II. The In-Synchrony-Bands Spectrum (SBS)

A. Motivation

All the information that is processed by the higher auditory system stages must exist in the auditory nerve firing patterns since the auditory nerve is the only link from the auditory periphery to the brain. For speech analysis purposes, it is thus sufficient to retain only properties of the speech signal that determine the auditory nerve firing patterns. Measurements of the firing patterns of cats' auditory nerve fibers in response to speech-like stimuli (Sachs and Young [7], Young and Sachs [8], Delgutte and Kiang [2]) suggest that firing rate is an insufficient carrier of speech information. Some use of temporal characteristics

Manuscript received July 23, 1985; revised January 23, 1987.

The author was with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139. He is now with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

IEEE Log Number 8613861.

of the firing patterns seems necessary. It is also evident from these measurements that as the stimulus intensity increases, more fibers fire in synchrony with the stimulus periodicity. It is thus proposed to consider the width of the region in which all the fibers fire in synchrony with the stimulus periodicity as a measure of the stimulus intensity. Specifically, we adopt the temporal nonplace approach suggested by Carlson *et al.* [1]. It is assumed that information about the relative intensity of differnt spectral portions of the signal is in the number of fibers that fire synchronously, regardless of the fiber's characteristic frequency. Furthermore, the phase response of the fiber's firing are assumed to be irrelevant. The basis for this assumption is the finding of Goldstein and Srulovicz [4] that the interspike interval statistic is adequate to explain the psychophysics of the perception of the pitch of complex tones. The in-synchrony idea was also applied to unvoiced speech with its energy located at the high portion of the spectral band (up to 4–5 kHz). Although synchrony drops as fiber CF increases (Johnson [5]), it is assumed that the amount of synchrony in these fibers is still useful.

B. Implementation

Based on these observations, a speech analyzer composed of two stages is suggested [Fig. 1(a)]. The first stage models the peripheral auditory processing structure up to the level of the auditory nerve. The second stage is an heuristic nonlinear relative spectrum intensity measure. It operates on the output of the first stage and plays the role of the higher nervous system. The design of the first stage is based on the general overall behavior of the peripheral auditory system. A simple model of the cochlear filters is used. The simulated cochlear filter bank consists of 100 highly overlapping filters equally spaced on a logarithmic scale with a 3 percent frequency step. Their frequency responses are similar to the tuning curves of the auditory nerve fibers and have a simple description in the log-amplitude log-frequency scale [Fig. 1(b)]. Each filter is identified by its characteristic frequency CF. The filters with CF up to 1000 Hz have a frequency response that is symmetric around CF on a log-frequency scale, with a +18 dB per octave incline on the low-frequency side and a −18 dB per octave rolloff on the high-frequency side. The filters with CF above 1000 Hz have a +18 dB per octave incline on the low-frequency side, but a very sharp rolloff on the high-frequency side. Group delay characteristics of the cochlear filters have been ignored. Thus, the filtering can be easily performed in the log-amplitude log-frequency domain by adding the input log spectrum with the log-frequency response of the filter.

The second stage of the proposed speech analyzer is based on the assumptions made in Section II-A. These assumptions lead to the following definitions.

Definition 1: An in-synchrony-band is a region of L_n successive filters having the same dominant frequency f_n, where the "dominant frequency" is the frequency of the strongest component in the filter's output signal. For a region to be declared as an in-synchrony band, it is necessary that L_n be greater or equal to a threshold M.

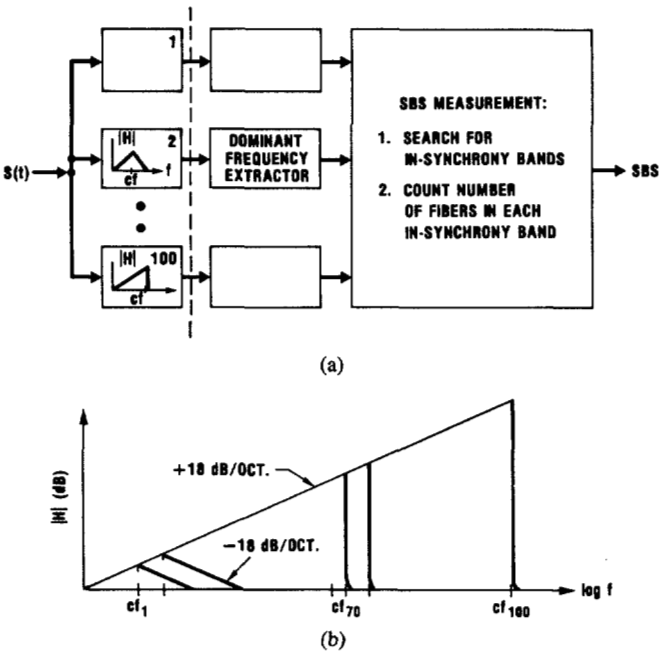

Fig. 1. (a) The in-synchrony-bands spectrum (SBS) analyzer. The first stage models the peripheral auditory processing structure up to the level of the auditory nerve. The second stage, an heuristic nonlinear relative spectrum intensity measure, operates on the output of the first stage and plays the role of the higher nervous system. (b) Frequency response of the simplified cochlear filters in a log-amplitude log-frequency scale. The gain of the filters is arbitrary since details of the spectral energy distribution are not considered and only the *frequency* of the strongest (dominant) component is measured.

Definition 2: The *in-synchrony-bands spectrum* (*SBS*) is a discrete function of frequency, consisting of a set of lines located at frequencies f_n with magnitudes L_n where f_n and L_n are as in Definition 1.

Note that in the proposed measure, each filter contributes at most one dominant frequency to the SBS. That is, it is assumed that the higher-stage apparatus can lock to the dominant component of the filter's output. Note also that the processing of the second stage is based on timing–synchrony. Details of the spectral energy distribution are not considered, and only the *frequency* of the strongest (dominant) component is measured. In the implementation, the dominant frequency is extracted simply by picking the frequency of the strongest component in the filter's output power spectrum.

A frame-oriented SBS analyzer was implemented since speech signal characteristics remain steady over approximately 20 ms and can be tracked adequately in 100 frame/s analysis. Fig. 2(b) shows the SBS lines for a sample high-resolution speech power spectrum plotted in Fig. 2(a). The analysis conditions are described in Section IV. The left-hand side plots are samples of male speech, while the right-hand side shows samples of female speech. The magnitude of the SBS lines represents the relative importance of each activity region. The figures demonstrate the *dominance effect*, which is a basic property of the SBS measure. Since each filter contributes

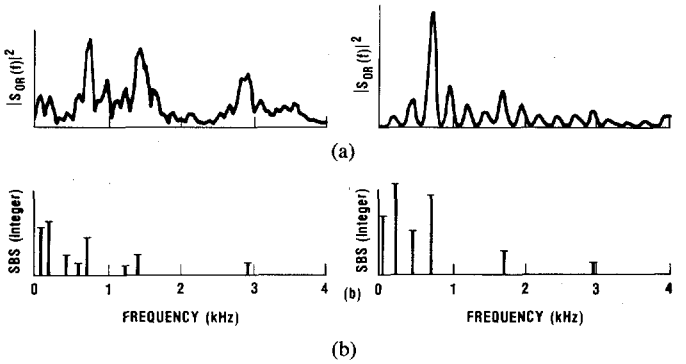

Fig. 2. (a) High-resolution FFT power spectra of a female (right-hand side) and male (left-hand side) speech samples. (b) SBS displays of the speech samples of Fig. 2(a).

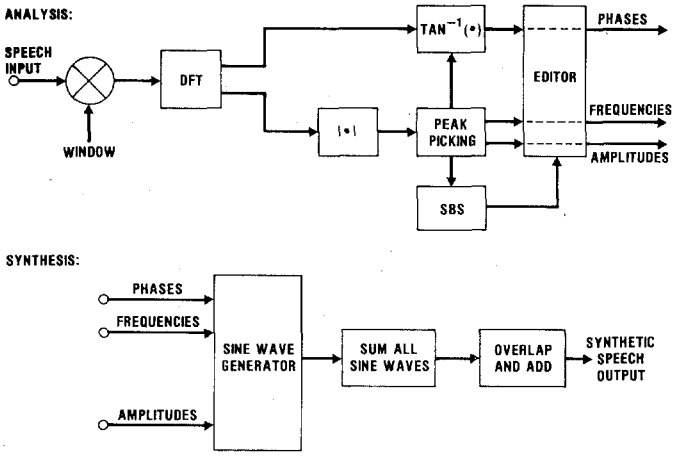

Fig. 3. The SBS controlled sinusoidal representation analysis/synthesis system.

only one dominant frequency, a strong frequency component f_0 dominates the activity of filters with CF greater than f_0. The number of filters that are dominated by the f_0 component depends on its relative intensity compared to the intensity of the next component. Note the intentional use the concept *dominance* instead of *masking*. The relation between the dominance property of the SBS and the psychophysical meaning of masking is yet to be studied.

Fig. 2 also shows another characteristic of the SBS. Because of the nonuniform distribution of fibers along the Basilar membrane, most of the frequency components in the first formant region are usually represented in the SBS display. The presence of a region with successive frequency components is necessary for the central pitch estimation mechanisms. The other formants are represented with many fewer SBS lines.

III. SBS Controlled Sinusoidal Representation

In order to use the SBS information directly for synthesis, an "inverse law" is required, to compute the appropriate parameters that should control the synthesizer. Another possibility is to run a synthesis-by-analysis system that matches the SBS of the synthesized speech to that of the original.

The desired inverse law has not yet been found. However, a one-step rule synthesis-by-analysis system is proposed. In the system, termed "the SBS controlled sinusoidal representation" (Fig. 3), the SBS analyzer is integrated into the analysis side of the sinusoidal representation analysis/synthesis system suggested by McAulay and Quatieri [6]. Only the SBS dominance information is used, to indicate which of the speech frequency components should be transmitted. The transmitted parameters are the SBS line frequencies and the *original* amplitudes and phases at those frequencies, computed from the speech input spectrum (as in the full sinusoidal representation system). In the synthesizer, only those frequency components are included in the summation of (1) of the Appendix.

In this one-step rule synthesis-by-analysis system, the SBS displays of the original and the synthesized speech must be identical since the suppressed frequency components do not change the dominance relations and hence do not affect the SBS. Fig. 4(c) shows the SBS displays of both the original speech spectra and the synthesized speech spectra (Fig. 4(a) and (b), respectively). The spectra in Fig. 4(a) are the decibel versions of the speech power spectra of Fig. 2(a).

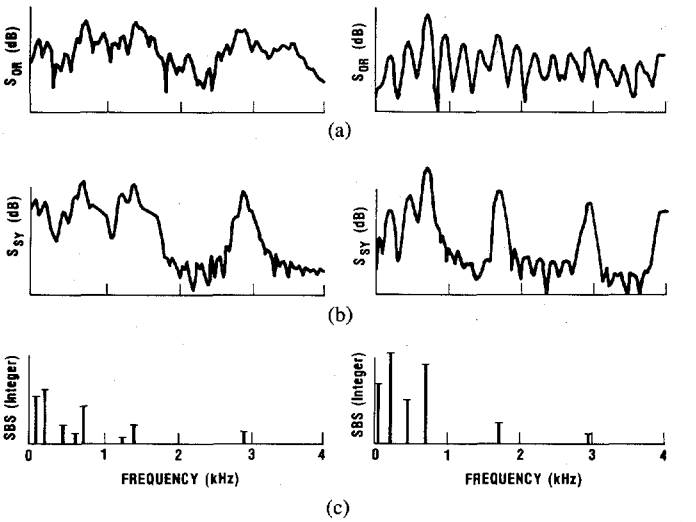

Fig. 4. (a) Power spectra, in decibels, of the same speech samples as in Fig. 2(a). (b) Power spectra, in decibels, of the SBS controlled synthesized speech of the same speech samples as in Fig. 4(a). (c) SBS displays of the original and the synthesized speech samples of Fig. 4(a) and (b), respectively.

IV. Experimental Results

The database comprises of a variety of male and female single speaker utterances, two overlapping superposed speech waveforms, music waveforms, and speech in musical background. The speech was low-pass filtered to 5 kHz, preemphasized by a 6 dB/octave preemphasis analog network, digitized at 10 kHz, and analyzed at 100 frames/s. The analysis was performed using a Hamming-window-weighted, 20 ms frame. The synthesizer uses a simplified version of the analysis/synthesis system suggested by McAulay and Quatieri [6], termed by the au-

thors as the "overlap-and-add system." The simplified version is described in the Appendix.

The synthesized speech is natural and highly intelligible, but includes some tonal artifact. The distortion is noticeable, especially in very-low-pitched male utterances where about 80 percent of the frequency components are removed. Two observations have been made. First, compared to the full sinusoidal representation, the average number of frequency components is about 40 percent for unvoiced frames and about 30 percent for voiced frames. This is due to the fact that in the unvoiced case there are no clear dominance regions since the spectral structure is inharmonic. Second, there is about 10 percent more saving for a male talker compared to a female talker. This is due to the lower pitch of male talker, which causes a larger density of spectral frequency components to be removed. These results hold for all kinds of input stimuli, including the two overlapping speech waveforms and speech in musical background.

The performance of the system is not seriously affected by the presence of noise, even as high as 0 dB peak-vowel-to-average-noise ratio. In fact, there may be some noise reduction due to the coherence property of the Fourier transform and the dominance effect of the SBS. Obviously, each frequency component in the speech frame spectrum is filtered by the narrow-band DFT filter, resulting in an accurate amplitude estimate, especially for the high-energy components. Since the low-energy components are dropped by the dominance effect and the synthesizer uses only high-energy components, the overall signal-to-noise ratio may be improved. Fig. 5 shows the original and the synthesized power spectra and their SBS (Fig. 5(a)–(c), respectively) for a female speech frame in quiet (left-hand plots) and noisy (with a 3 dB peak-vowel-to-average-noise ratio) environments. Note that the SBS display is hardly affected by the noise.

In order to get a crude estimate of the upper bound for the number of lines needed per frame, the maximum number of sine wave components to be superposed by the synthesizer was set to 10. Thus, in each frame, the ten SBS components with the largest DFT amplitudes were retained. The resulting synthesized speech was perceptually indistinguishable from the unconstrained SBS controlled synthesized speech. This result holds for all the stimuli in the database described above, in both quiet and noisy environments.

Finally, a hypothesis was made for the cause of the tonal artifact. It was suggested that the SBS representation might not be complete for voiceless sound since it introduces a discrete spectrum structure to a smoothed input spectrum. The hypothesis was based on the result that in the cases of two overlapping superposed speech waveforms and speech in musical background (where the voiceless parts of one speaker are masked by the voiced speech of the other or by the music), the synthesized speech barely suffers from any tonal artifact. To further examine the hypothesis, speech was generated by hybridizing the *original* voiceless frames and the SBS controlled

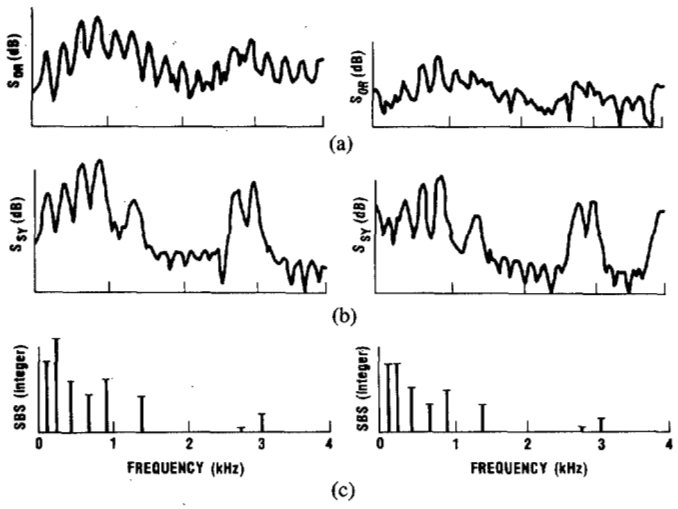

Fig. 5. (a) Power spectra, in decibels, of an original female speech sample in quiet (left-hand side) and noisy (right-hand side) environments. The noise level is 3 dB peak-vowel-to-average-noise ratio. (b) Power spectra, in decibels, of the SBS controlled synthesized speech of the speech samples in Fig. 5(a). (c) SBS displays of the original and the synthesized speech samples of Fig. 5(a) and (b), respectively.

synthesized voiced frames. The buzz–hiss classification was made by using the Gold–Rabiner pitch detector. Only the single-talker utterances were examined. An informal A–B comparison showed that the tonal artifact was eliminated almost completely. This suggests that tonal artifact is created within voiceless frames.

V. Summary

Motivated by the in-synchrony characteristics of the timing information in the auditory nerve firing patterns, we have used the in-synchrony-bands spectrum (SBS) measure to define the perceptually meaningful frequency regions of the stimulus spectra. It seems that for analysis/synthesis purposes, the SBS profiles are a sufficient display of speech as represented at the auditory nerve level. For all kinds of acoustical stimuli in our database, including single male and female speakers in clear and in noise, two overlapped speech waveforms, music waveforms, and speech in musical background, the very few frequency components indicated by the SBS are sufficient for an adequate representation of all the speech features. More than half of the SBS lines used are in the region of the first formant, representing the first formant and the intonation (pitch) information. The other formants are represented with many fewer lines. The SBS representation of voiceless sounds, however, seems to be incomplete since it introduces tonal artifacts into the synthesized speech.

This approach can be considered as an efficient method for extracting a minimal set of DFT maxima needed to retain the main speech features. In fact, ten frequency components, at most, are sufficient to obtain natural (with some tonal artifact) and highly intelligible synthesized speech. This suggests a potential for an efficient speech coding system based on this technique.

Finally, the in-synchrony characteristics of the timing information in the auditory nerve firing patterns can be

applied to the speech recognition problem. The use of these principles as a front end for speech recognition in a noisy environment is described in [3].

Appendix

This Appendix describes a simplified version of the sinusoidal representation analysis/synthesis system suggested by McAulay and Quatieri [6] termed by the authors as the "overlap-and-add system" [6, p. 753]. This version was the basic system from which the SBS controlled sinusoidal representation system was derived.

In the overlap-and-add system, the speech waveform is modeled as a sum of sine waves. The sine wave parameters are kept constant over the frame. If $s_i(n)$ represents the speech waveform samples in the ith frame, then

$$s_i(n) = \sum_k A_{ki} \sin (2\pi f_{ki} nT + \theta_{ki}) \qquad (1)$$

where A_{ki} and f_{ki} are the amplitude and frequency of the kth component in the ith frame and θ_{ki} is the phase of the kth frequency component referenced to the center of the ith frame. In order to determine these parameters, a high-resolution FFT is computed every frame, and a set of sine wave frequencies is generated by applying simple peak picking to the interpolated magnitude function. The amplitudes and the phases are measured from the interpolated FFT at the location of the peaks. In the synthesizer, the summation (1) is used to produce the frame output signal on which a triangular window is applied. Since the analysis is performed every 10 ms on a 20 ms window, the current and the preceding 20 ms time-weighted frame output signals are overlapped by 10 ms delay and added to obtain the output synthesized speech. There is no voicing decision, and the same procedures are applied for the voiced and the unvoiced frames.

Acknowledgment

The author wishes to thank B. Gold, R. J. McAulay, W. M. Rabinowitz, J. Tierney, and T. F. Quatieri for stimulating discussions throughout this work and to thank D. A. Berkley and M. M. Sondhi for reviewing an earlier version of the paper.

References

[1] R. Carlson, G. Fant, and B. Granstrom, "Two formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, G. Fant and M. A. A. Tatham, Eds. London, England: Academic, 1975, pp. 55–82.

[2] B. Delgutte and N. Y. S. Kiang, "Speech coding in auditory nerve: I–V," *J. Acoust. Soc. Amer.*, vol. 75, no. 3, pp. 866–918, Mar. 1984.

[3] O. Ghitza, "Auditory nerve representation as a front-end for speech recognition in a noisy environment," in *Comput. Speech Lang.*, vol. 1, no. 2, to appear.

[4] J. L. Goldstein and P. Srulovicz, "Auditory-nerve spike intervals as an adequate basis for aural spectrum analysis," in *Psychophysics and Physiology of Hearing*, E. F. Evans and J. P. Wilson, Eds. London, England: Academic, 1977, p. 337.

[5] D. H. Johnson, "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," *J. Acoust. Soc. Amer.*, vol. 68, no. 4, p. 1115, Oct. 1980.

[6] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, p. 744, Aug. 1986.

[7] 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, no. 2, p. 470, Aug. 1979.

[8] 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, no. 5, p. 1381, Nov. 1979.

Oded Ghitza was born in Haifa, Israel, on September 24, 1948. He received the B.Sc., M.Sc., and Ph.D. degrees in electrical engineering from Tel-Aviv University, Tel-Aviv, Israel, in 1975, 1977, and 1983, respectively.

From 1980 to 1984 he worked at the Signal Corps Research Laboratory of the Israeli Defence Forces. During 1984–1985 he was a Bantrell postdoctoral Fellow at the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, and a Consultant at the Lincoln Laboratory Speech System Technology Group, Lexington, MA. Since 1985 he has been with the Department of Acoustic Research, AT&T Bell Laboratories, Murray Hill, NJ, where he is studying auditory-based processing techniques for speech coding and speech recognition.