AUDITORY NEURAL FEEDBACK AS A BASIS FOR SPEECH PROCESSING

Oded Ghitza

AT&T Bell Laboratories
Acoustics Research Department
Murray Hill, New Jersey 07974

ABSTRACT

This paper describes the closed-loop Ensemble-Interval-Histogram (EIH) model. It is constructed by adding a feedback system to the former, open-loop, EIH model (Ghitza, *Computer, speech and Language*, 1(2), pp. 109-130, Dec. 1986). While the open-loop EIH is a computational model based upon the ascending path of the auditory periphery, the feedback system is motivated by the descending path and attempts to capture the functional contribution of the neural feedback mechanism in the auditory periphery.
The capability of the resulting closed-loop EIH to preserve relevant phonetic information in quiet and in noisy acoustic environments was measured *quantitatively* using the model as a front-end to a Dynamic Time Warping (DTW), speaker-dependent, isolated-word recognizer. The database consisted of a 39 word alpha-digit vocabulary spoken by two male speakers, in different levels of additive white noise. In the absence of noise the recognition scores based on the closed-loop EIH are comparable to those based on the open-loop EIH. However, recognition performance based on the closed-loop EIH does not decline as much as with the open-loop EIH at low signal-to-noise ratios. At SNR of 6 dB, the average correct-recognition score with the closed-loop EIH is 82 %. This is equivalent to the recognition score obtained with the open-loop EIH at 10 dB SNR, a gain of 4 dB.

1. INTRODUCTION

Little is known, at present, about the functional operation of the auditory brain-stem and auditory cortex. However, studies of the population response of single auditory-nerve fibers in the cat to speech-like stimuli provide a rich source of information concerning the principles by which such sounds are encoded in the auditory nerve. As the sole afferent path from the cochlea to the cochlear nucleus, the auditory nerve contains, by necessity, all the information available to the more central regions of the auditory pathway. Thus, from the perspective of speech analysis it is sufficient to retain only those properties of the speech signal encoded in the firing patterns of the auditory-nerve fibers.

In the past (Ghitza, [7] and [8]), we have used this approach to examine the representation of speech in quiet and in noisy acoustic environments. In [7], detailed physiological modeling of the mechanical motion of the basilar membrane has been used (adopting Allen's model [1]) in combination with a functional representation of the inner-hair cell and a plausible non-linear processing, to functionally simulate the type of processing which might occur at higher levels of the auditory pathway. The capability of the resulting Ensemble-Interval-Histogram (EIH) representation to preserve relevant phonetic information in quiet and in noisy acoustic environments was measured *quantitatively* using the model as a front-end to a speech recognition system. It was shown that the EIH representation preserves the relevant phonetic information in quiet and outperforms the Fourier power spectrum representation in the presence of high levels of additive background noise. (Note that the EIH is one among several possible representations that can be constructed from the information available at the auditory-nerve level. For a review of current auditory models see [9]. See also Delgutte, [3], and Lyons, [11]).

While the EIH model was based upon physiological studies at the auditory nerve level, the present study aims at incorporating recent physiological studies concerning the role of higher stages in the auditory brain-stem. In particular, we will make use of on-going physiological studies (in the cat) on the effect of activity in the medial olivocochlear (MOC) nerve bundle on the neural firing activity in the afferent path of the auditory nerve. The origin of the MOC nerve bundle is in the medial region of the superior olive, which is the second neural processing center in the auditory brain-stem, following the cochlear nucleus. The MOC nerve fibers project back to different places along the cochlea partition, making synapse connections to the outer-hair cells. As such, the MOC nerve bundle serves as a frequency-dependent feedback mechanism which conveys neural signals (which are the outcome of some neural processing taking place in the medial superior olive region) from the superior olive back to the cochlea.

In view of these neurophysiological observations, we modify the former (open-loop) EIH model by constructing a feedback connection that controls the gain of the peripheral filters in a frequency-dependent way, determined by the EIH profiles which still serve as the output of the model. As opposed to the open-loop EIH, the closed-loop EIH is a time varying system, which changes its parameters (the gain of the filters at present) as a function of the information it processes. As a result, stationary signals that are presented to the system are represented in a non-stationary way.

This paper is organized as follows. In Section 2 we describe the closed-loop EIH model. As we shall see there, we do not attempt to draw a detailed physiological model. Instead, we propose a system that attempts to capture the basic functional behavior that seems to occur in the auditory system in order to examine the possible advantages that might result by applying such processing principles to the speech processing problem. Furthermore, to study the role of the feedback mechanism in isolation from the intra-cochlea processing, we replaced the cochlear filters in the EIH with constant-bandwidth filters. This is a simpler set of filters which was found to be suitable for studying the representation of the gross spectral structure of speech (Ghitza, [8]).

In Section 3, two key questions are addressed with respect to this novel form of speech representation: (1) Does the closed-loop EIH representation preserve all of the phonetically relevant information of the speech signal? (2) Is the closed-loop EIH representation more resistant to noise than the open-loop EIH? In order to answer these questions, the open-loop EIH and the closed-loop EIH representations both served as the input to a speech recognition system. The comparison between the open-loop EIH and the closed-loop EIH was made as a function of signal-to-noise ratio. The speech database consisted of a 39-word alpha-digit vocabulary, spoken by two males and presented over a range of signal-to-noise ratios. In the absence of noise the recognition scores based on the closed-loop EIH are comparable to those based on the open-loop

CH2561-9/88/0000-0091 $1.00 © 1988 IEEE

EIH. This result demonstrate that the feedback mechanism does not distort the phonetically relevant information of the speech signal. However, recognition performance based on the closed-loop EIH does not decline as much as with the open-loop EIH at low signal-to-noise ratios. At SNR of 6 dB, the average correct-recognition score with the closed-loop EIH is 82 %. This is equivalent to the recognition score obtained with the open-loop EIH at 10 dB SNR, a gain of 4 dB.

2. THE CLOSED-LOOP EIH

The closed-loop EIH is constructed by adding a feedback system to the former, open-loop, EIH system. While the open-loop EIH is a computational model based upon the ascending path of the auditory periphery, the feedback system is motivated by the descending path and attempts to capture the functional contribution of the feedback mechanism in the auditory periphery as discussed by Kiang *et al*, [6], Liberman, [10], and Winslow and Sachs, [13].

The open-loop system that was used is the constant-bandwidth EIH system (Ghitza, [8]). By using the simpler constant-bandwidth filter-bank instead of the cochlear filter-bank we isolate the complex effects of the intra-cochlea processing and concentrate on the contribution of the other parts of the overall system. However, the use of the constant-bandwidth filters restrict our study only to the representation properties of the gross spectral structure of speech. The open-loop EIH is shown in Fig. 1, inside the dashed box.

The feedback system uses different kinds of partial-information that are derived from the EIH function (Fig. 1). The integral over the EIH provides an estimate of the loudness of the input signal (Chien, [2]). The ratio of the maximum value to the minimum value of the EIH reflects the degree of "flatness" of the EIH function and is used to construct a synchronization-index flag which is either 1 (if there exist a certain amount of coherent activity across the simulated fiber-array) or 0. As we shall see later, the synchronization index is a soft flag. The smoothing of the EIH is done by convolving the EIH function with an Hamming-window function. The width of the window is determined by the current pitch value, estimated from the EIH (the underlying assumption here is the possible existence of a feedback link between the pitch neural-processing center and the center that provides a representation of the gross spectral structure).

All these partial-information paths are used to construct a gain-control function (as a function of frequency) that, after an appropriate time-delay d, determines the gain of every individual filter in the filter-bank. The dependence of the gain-control function

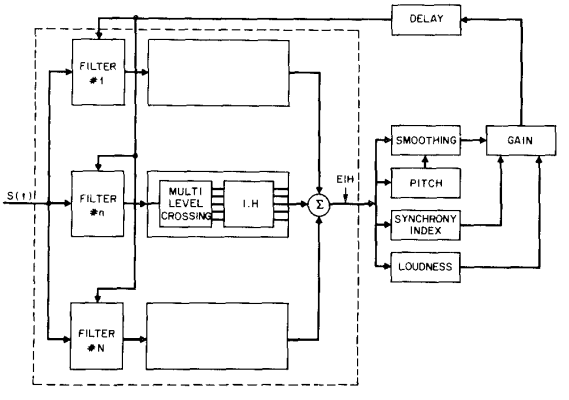

Figure 1.

in frequency reflects the anatomical arrangement of the MOC efferent system, which shows projections of the MOC effector-neurons to corresponding parts along the cochlear partition.

Let $g(\cdot)$ be the gain-control function. We define $g(\cdot)$ at time t_0 as:

$$g_{t_0}(f) = \frac{1}{w} \sum_{t=t_0-d-w}^{t_0-d} c_t^s(f) \qquad (1)$$

Where d is the time-delay introduced by the feedback-system, w is the time-duration of a smoothing window beginning at d, and $c_t^s(\cdot)$ is defined as follows, depending on the value of the synchronization index, s. For $s=1$:

$$c_t^1(f) = K_t^1 \, f(\Xi_t(f)) \qquad (2)$$

Where $\Xi_t(\cdot)$ is the smoothed-EIH at time t and $f(\cdot)$ is a non-linear compressive function such that the high-value peaks of Ξ_t are more attenuated. This is roughly in accordance with the response of the auditory feedback system to narrow-band sounds, where the amount of attenuation in the afferent firing activity (which is determined by the firing-rate of the MOC effector-neurons) is proportional to the signal intensity. K_t^1 is a normalization factor, such that:

$$\int c_t^1(f)df = \int \Xi_t(f)df \qquad (3)$$

Where the right-hand-side of (3) is the estimated loudness at time t. For $s=0$:

$$c_t^0(f) = \frac{K_u^0}{M_t} \qquad (4)$$

Where K_u^0 is a normalization factor, such that:

$$\int c_u^0(f)df = \int \Xi_u(f)df \qquad (5)$$

and u denotes the starting-time of a new string of successive frames with $s=0$. M_t is the number of accumulated successive frames with $s=0$, from u to time t. Thus, $c_t^0(\cdot)$ decreases with time, to further attenuate the filter-bank output signals. This is roughly in accordance with the effect of the MOC effector-neurons on the afferent pathway in the presence of broad-band signals, which results an on-going relaxation in the afferent neural activity.

It should be noted that the gain-control function is derived from observations on the EIH function. Hence, the gain-control is based upon information which was obtained by some degree of processing of the firing activity in the simulated fiber-array. This is in accordance with the observation that the origin of the MOC efferent system is in the region of the medial superior olive (which is the second relay center in the brain-stem). However, choosing a gain-control function which is based upon an EIH-type of representation should be considered as heuristic until further physiological studies.

The resulting closed-loop EIH system is a time-varying system. Hence, it represents the temporal properties of speech in a non-stationary manner. A systematic study of the representation of speech sounds by the closed-loop EIH is beyond the scope of this paper. However, some illustrations can be made. Consonant-vowel transition, for example, last about 40 milliseconds, while the following vowel lasts 100-200 milliseconds. The overall effect of the closed-loop system is to emphasize the transitions (which are short duration high-frequency signals) over the higher intensity, long duration, low-frequency vowel signals.

Fig. 2 demonstrates the response characteristics of the system to speech in the presence of wide-band noise. The utterance is the emphasized part of the sentence "line up at *the screen door*", spoken by a male speaker. The speech was low-pass filtered to 5 kHz, preemphasized by a 6 dB/octave preemphasis analog network, digitized at 10 kHz and analyzed every 3.2 msec. The upper panels are wide-band Fourier spectrograms. In creating the spectrograms no additional preemphasis was used, the duration of the Hamming window was set to 10 milliseconds and the number of points in the FFT was 128. The center panels are the open-loop EIH representations and the lower panels are the closed-loop EIH representations, with d equals 6.4 msec and w equals 16 msec. The left-hand-side column is the representation in quiet and the right-

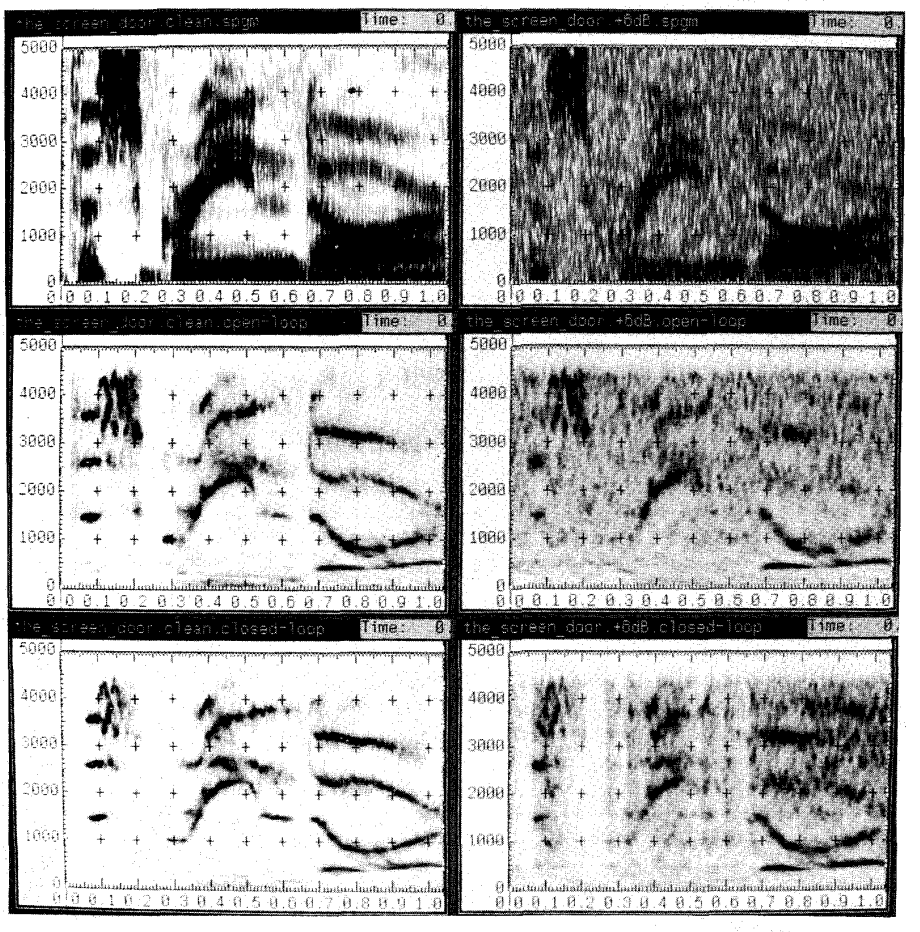

Figure 2.

hand-side column is the representation in the presence of additive, white Gaussian noise with global signal-to-noise ratio of +6 dB. The feedback system causes an emphasis of the gross temporal structure of the speech, via the reduction of the response to the noisy parts. The contribution of the frequency-shaped gain-control function is in the EIH representation during high-energy regions of the input signal.

3. THE CLOSED-LOOP EIH : RECOGNITION RESULTS

In this section we quantitatively evaluate the capability of the closed-loop EIH to preserve relevant phonetic information of speech. In the experiment, the closed-loop EIH was compared to the open-loop EIH by using them, alternately, as front-ends to a speech recognition system. The speech recognition system used was the Dynamic Time Warping (DTW) speech recognizer described by Wilpon and Rabiner, [12]. The EIH was treated as the output log-spectrum of an all-pole filter and although the underlying assumptions for this may not be entierly appropriate, they provide a simple mechanism for adapting the EIH to the DTW recognizer. If these assumptions are valid, the EIH "spectrum" envelope can be determined using conventional, frequency-domain, linear prediction approximation methods. For the EIH function, an appropriate envelope fit is achieved by applying a 16th-order LPC fit. This order of the fit is required because of the larger dynamic range of the EIH (compared to the Fourier representation).

3.1 Database and analysis conditions

The database for the recognition experiments was a 39-word alpha-digit vocabulary spoken by two males. The set of 39 words includes the letters of the alphabet, the digits, and the control words STOP, ERROR and REPEAT. Each of the words was manually edited, to mark the endpoints.

The analog speech signal was recorded using a standard telephone line, bandpass filtered from 100 Hz to 3200 Hz, and sampled at a 6.67 kHz rate. Four signal conditions have been tested: clean speech, +12 dB, +6 dB, and +0 dB global-SNR. For the noisy conditions, a zero-mean, white Gaussian noise was added to the test word, at each specified SNR.

3.2 The recognition system

The recognizer used two reference templates for each word, created from an uncorrupted training set. No clustering techniques had been used in creating the templates. Hence, the recognition scores reported here are conservative estimates of the system's capability. For the purposes of the present comparison, however, the templates used are satisfactory since the study is concerned with *relative* performance of the DTW system using both EIH front-ends.

Given the template for each word, the recognition was performed using standard dynamic time warping techniques. A weighted cepstral distance measure was used, where the cepstral coefficients

were derived from the LPC parameters that represent the EIH envelope and the weighting coefficients were determined by the following expression (Juang *et al*, [5]):

$$w_k = 1 + 0.5L\sin(\pi\frac{k}{L}), \quad k=1,2,...,L \qquad (6)$$

The word corresponding to the template that produces the smallest distance from the test word was chosen as the recognized word.

3.3 Recognition results

The experimental results are shown in Fig. 3. In the absence of noise the recognition scores based on the closed-loop EIH are comparable to those based on the open-loop EIH. This result demonstrates that the feedback mechanism does not distort the phonetically relevant information of the speech signal. However, recognition performance based on the closed-loop EIH does not decline as much as with the open-loop EIH at low signal-to-noise ratios. At SNR of 6 dB, the recognition scores with the closed-loop EIH are equivalent to those obtained with the open-loop EIH at 10 dB SNR, a gain of 4 dB. The recognition performance for the Fourier power-spectrum front-end is quoted from a study by Ephraim *et al*, [4]. The recognition system for that experiment was the same as described here, with two exceptions: (1) The order of the LPC fit was 8 (to achieve optimum performance), and (2) The two reference templates were designed using advanced clustering techniques (Wilpon and Rabiner, [12]).

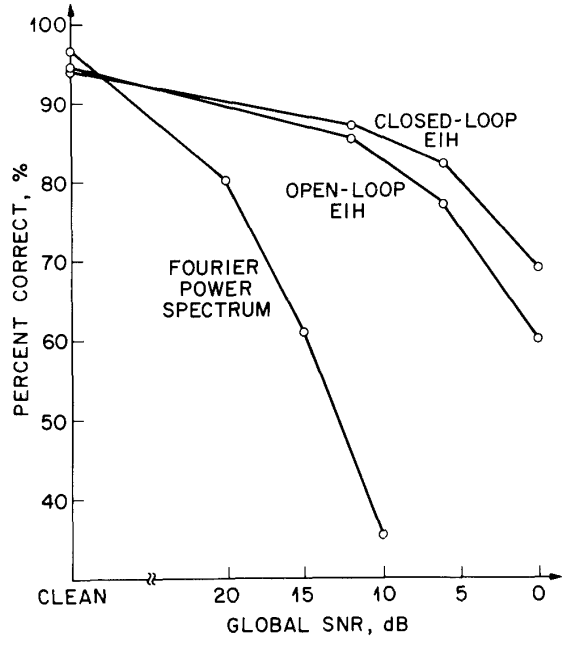

Figure 3.

4. CONCLUSIONS

A closed-loop system was constructed by adding feedback mechanism to the former (open-loop) EIH system. While the open-loop EIH is based upon the ascending path of the auditory periphery, the feedback system is motivated by the descending path and operates along lines that were motivated by recent studies on the anatomy and the physiology of the reflex feedback mechanisms in the auditory system. Only part of the reflex system was adopted, concerning the medial olivocochlear (MOC) efferent system. Since the origin of the MOC bundle is in the medial region of the superior olive (which is the second neural processing center in the auditory brain-stem following the cochlear nucleus), the resulting auditory-based processing system is motivated by larger and deeper parts of the auditory periphery. We did not attempt to model the available neurophysiological data in detail. Instead, we have proposed a system that attempts to capture the basic functional behavior that seems to occur and examined its behavior in response to speech. The quantitatively-based conclusions of this study concern only a narrow aspect of this novel representation of speech. Since the recognition experiment was conducted using a database of manually-endpointed isolated words, only the representation *within* the token has been examined. Within this framework, the closed-loop EIH representation preserves the relevant phonetic information of speech and proves to be less sensitive to wide-band noise as compared to the open-loop EIH representation.

ACKNOWLEDGMENT

I wish to thank D. A. Berkley for stimulating discussions throughout this work and for reviewing an early version of the manuscript.

REFERENCES

[1] Allen, J. B. (1985)."Cochlear modeling", *IEEE ASSP magazine*, p. 3, January.

[2] Chien, P. S. (1987). Personnel communication.

[3] Delgutte, B. (1986)."Analysis of stop consonants using a model of the peripheral auditory system", in *Invariance and variability in speech processes* Perkell and Klatt, Eds. New-Jersey: Lawrence-Erlbaum.

[4] Ephraim, Y., Wilpon, J. G. and Rabiner, L. R. (1987)."A linear predictive front-end processor for speech recognition in noisy environments", *Inter. Conf. on Acoust. Speech and Signal Proc., ICASSP'87*, Vol. 4, Dallas, Texas, p. 1324, April.

[5] Juang, B. H., Rabiner, L. R. and Wilpon, J. G. (1987)."On the use of bandpass liftering in speech recognition", *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. ASSP-35, p. 947, July.

[6] Kiang, N. Y. S., Guinan, Jr., Liberman, M. C., Brown, M. C. and Eddington, D. K. (1988)."Feedback control mechanisms of the auditory periphery: implications for cochlear implants", to appear.

[7] Ghitza, O. (1986)."Auditory nerve representation as a front-end for speech recognition in a noisy environment", *Computer Speech and Language*, Vol. 1., No. 2, p. 109, December.

[8] Ghitza, O. (1987)."Robustness against noise: the role of timing-synchrony measurement", *Inter. Conf. on Acoust. Speech and Signal Proc., ICASSP'87*, Vol. 4, Dallas, Texas, p. 2372, April.

[9] Greenberg, S., *Editor* (1988). A special issue on the "Representation of speech in the auditory periphery", *Journal of Phonetics*, Vol. 15, No. 4, January.

[10] Liberman, M. C. (1987). Personnel communication.

[11] Lyons, R. F. (1986)."Experiments with a computational model of the cochlea", *Inter. Conf. on Acoust. Speech and Signal Proc., ICASSP'86*, Vol. 3, Tokyo, Japan, p. 1975, April.

[12] Wilpon, J. G. and Rabiner, L. R. (1985)."A modified K-means clustering algorithm for use in isolated word recognition", *IEEE Trans. Acoust. Speech and Signal Proc.*, Vol. ASSP-33 (3), p. 587, June.

[13] Winslow, R. L. and Sachs, M. B. (1987)."Effect of electrical stimulation of crossolivocochlear bundle on auditory nerve response to tones in noise", *Journal of Neurophysiology*, Vol. 57, p. 1002.