ROBUSTNESS AGAINST NOISE: THE ROLE OF TIMING-SYNCHRONY MEASUREMENT

Oded Ghitza

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

ABSTRACT

In a previous report (Ghitza, 1987, [1]) we described a computational model based upon the temporal characteristics of the information in the auditory nerve fiber firing patterns, which produced an "auditory" spectral representation (the EIH) of the input signal. We also demonstrated that for speech recognition purposes, the EIH is more robust against noise compared to the traditional Fourier power spectrum. This paper reports on the first step towards understanding the role of different parameters in the EIH in achieving this performance. Both, the Fourier power spectrum measurement and the EIH measurement can be partitioned into two parts, a filter-bank followed by feature analyzer. In the Fourier power spectrum, the filter bank consists of uniformly shaped Hamming filters and the analyzer is based on power measurements. In the EIH, the filter bank consists of the cochlear filters and the analyzer is based on timing-synchrony measurements. The present study examines the relative importance of the filter-bank properties as compared to the analysis principle. For this purpose a modified EIH model has been created in which the cochlear filters have been replaced by the uniformly shaped Hamming filters. The output of the filter bank is processed by the timing-synchrony analyzer, as with the original EIH. The modified EIH and the Fourier power spectrum differs, therefor, only in the kind of analysis performed on the filter bank output. The modified EIH has been used as a front-end to a Dynamic Time Warp (DTW), using the same set-up as in Ghitza, 1987, [1]. A speaker dependent, isolated word recognition test has been conducted, on a database consisted of a 39 word alpha-digits vocabulary spoken by two male and two female speakers, in different levels of additive white noise. Compared to the Fourier-based front-end, the recognition scores have been slightly improved in clean environment but significantly improved in noisy environments. Furthermore, compared to the original EIH, the recognition scores have also been improved, both in clean and in noisy environments. These results demonstrate that the timing-synchrony measurement is significantly more robust against noise compared to the power measurement. They also show that the robustness is due to the timing-synchrony analyzer and not to the unique shape of the cochlear filters.

1. INTRODUCTION

In a previous report (Ghitza, 1987, [1]) we described a computational model based upon the temporal characteristics of the information in the auditory nerve fiber firing patterns. The model is shown in Fig. 1. It produces a frequency domain representation of the input signal in terms of the ensemble histogram of the inverse of the interspike intervals, measured from firing patterns generated by a simulated nerve fiber array. The nerve fiber firing mechanism has been modeled by a multi-level crossing detector at the output of each cochlear filter. We used 85 cochlear filters, equally spaced on a log-frequency scale from 200 Hz to 3200 Hz, and the level crossings have been measured at positive threshold levels which were uniformly distributed on a log scale. The resulting Ensemble Interval Histogram (EIH) "spectrum" has two main properties: (1) fine spectral details are well preserved in the low frequency region but become fuzzy at the high frequency end, (2) the EIH is more robust in noise, compared to the traditional Fourier power spectrum. The

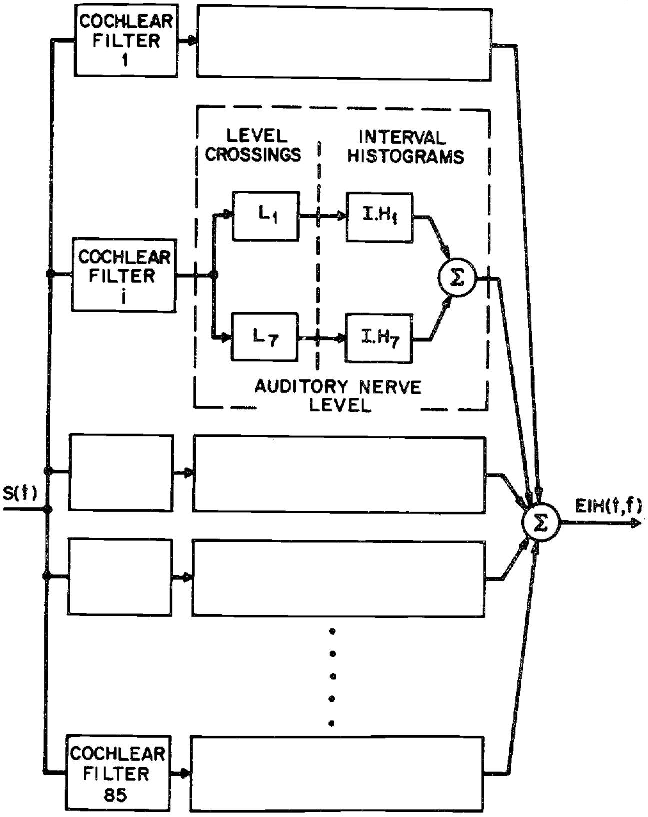

Figure 1.

robustness of the EIH for speech recognition tasks was measured quantitatively. This was done by using, alternately, the EIH and the Fourier power spectrum as front-ends to a Dynamic Time Warping (DTW), speaker dependent, isolated word recognizer. The database consisted of a 39 word alpha-digits vocabulary spoken by two male and two female speakers, in different levels of additive white noise. In the noise-free case, the performance of the EIH front-end was comparable to a conventional Fourier-based front-end. In the presence of noise, however, the EIH front-end was more robust. In view of these results, the following question arises: What is the structural difference between a Fourier power spectrum measurement and the EIH measurement in providing the robustness against noise ?

Fig. 2(a) and 2(b) shows a block diagram description of a Fourier power spectrum measurement and an EIH measurement, respectively. An N-point Fourier power analysis on a Hamming windowed signal can be viewed as a two step spectral estimation. At first, the signal is filtered by N identically shaped filters, their frequency response is the Fourier transform of the Hamming window. This filter will be called the "Hamming shaped filter". The Hamming filters are equally spaced in a frequency band which is determined by the sampling frequency and are highly overlapped. The power spectrum estimate is achieved by computing the short-term power at the output of each filter.

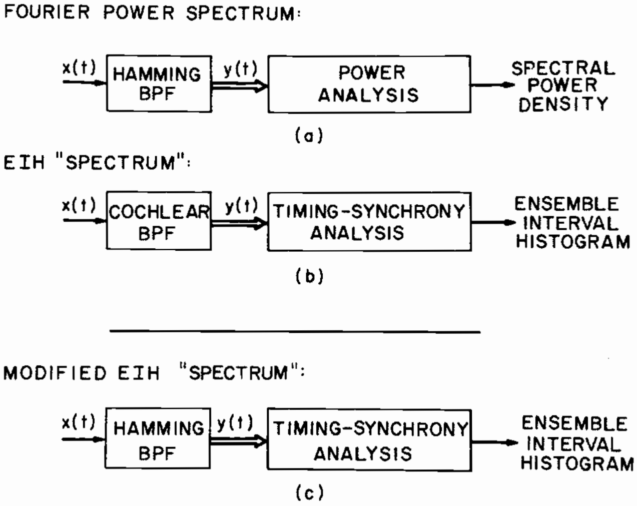

Figure 2.

Similarly, the EIH model is also a two stage analyzer. However, the first stage is a cochlear filter-bank which represents the auditory periphery, and the second stage is a timing-synchrony analyzer which processes the output signals of the cochlear filter-bank. The EIH measurement, thus, differs from the Fourier power spectrum measurement both in the filter-bank part and in the analysis part. The filter banks are different in two ways. First, the cochlear filters are equally spaced on a logarithmic scale as opposed to the Hamming filters which are equally spaced on a linear scale. Second, the shape of the cochlear filters is unique. Roughly, the filters with CF up to 500 Hz have a frequency response which is symmetric around CF on a log-frequency scale, with an +18 dB per octave incline on the low frequency side and a -18 dB per octave roll-off on the high frequency side. The filters with CF above 500 Hz have a +18 dB per octave incline on the low frequency side but a very sharp roll-off on the high frequency end (e.g. -120 dB per octave). As for the differences in the analysis part, in contrast to the power measurement in the Fourier power analysis, the timing-synchrony measurement consists of three steps, each of which is non-linear. First, a multi-level crossing (MLC) analysis is applied to the output signal produced by each filter in the filter bank. This provide a N-dimensional point process, where N equals the number of filters in the filter-bank multiplied by the number of levels in the MLC. In the next step, N independent interval histograms are constructed, one for each dimension of the N-dimensional point process. The interval histogram should be viewed as an estimate of the short term probability density function of the intervals at each level. The final step is to locate regions in the N-dimensional point process which fire synchronously and to use the width of the region as an estimate of the relative spectral intensity of the underlying frequency components.

In view of these differences between the Fourier power spectrum and the EIH, it is possible to relate the robustness of the EIH against noise to the unique shape of the cochlear filters, to the properties of the timing-synchrony analysis, or to both. In order to examine the relative contribution of the cochlear filters and the timing-synchrony analysis separately, the *modified EIH* system in Fig. 2(c) has been constructed. Here, the timing-synchrony analyzer operates on the output signals produced by the uniformly shaped and uniformly spaced Hamming filters. Thus, the Fourier power spectrum measurement and the modified EIH measurement differ only in the analysis part. To obtain a quantitative comparison between the two, the modified EIH model was examined as a front-end to a DTW recognizer under the same conditions as with the original EIH and with the Fourier based front-ends (Ghitza, 1987, [1]). The experiments are discussed in Section 2. The results demonstrate that the timing-synchrony measurement is significantly more robust against noise compared with power measurement. The main contribution to the robustness is by the timing-synchrony analyzer and not by the unique shape of the cochlear filters.

2. THE MODIFIED EIH : RECOGNITION RESULTS

To obtain a quantitative comparison to the Fourier power spectrum, the modified EIH model was examined as a front-end to a DTW recognizer using the same set-up as in Ghitza, 1987, [1]. The system that was used for the previous experiments as well as the analysis conditions, the database, and the signal conditions are summarized in Appendix A. Exactly the same procedures have been used to study the performance of the modified EIH front-end. The filter bank for the modified EIH is shown in Fig. 3. It consists of 85 Hamming filters equally spaced in linear frequency scale, 40 Hz apart. The -3 dB points bandwidth of each filter is 420 Hz. Five noise conditions have been tested: clean speech, +18 dB, +15 dB, +12 dB and +6 dB global-SNR (see Appendix A.3 for our definition of global-SNR). The noisy data was created by adding white noise to the clean database.

The recognition scores for the modified EIH front-end are presented in Fig. 4, combined with the previous results for the Fourier based and the original EIH front-ends. Under noise-free conditions, the Fourier power spectrum and the modified EIH are identical (in a statistical sense); both are slightly superior to the original EIH. However, with increasing noise, the recognition scores with the Fourier based front-end drop much faster than with the original EIH and the modified EIH. In addition, the modified EIH is significantly more robust than the original EIH using cochlear filters. Note that the improvement in the recognition scores by the modified EIH is much bigger for the females database.

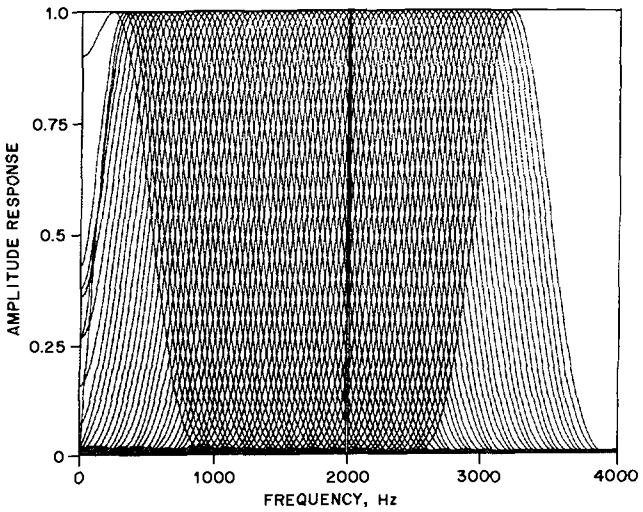

Figure 3.

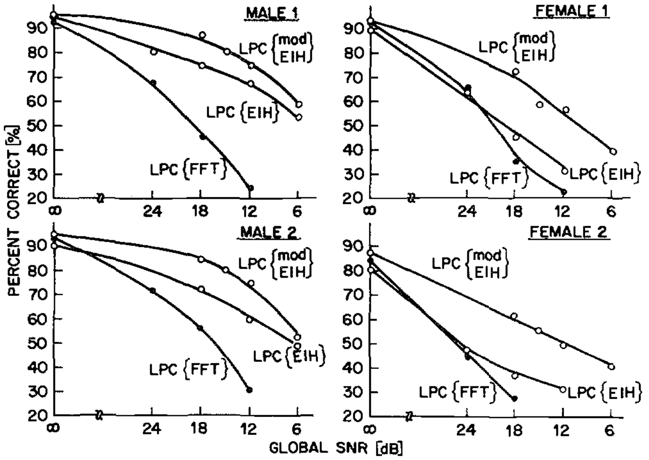

Figure 4.

3. CONCLUSIONS

The timing-synchrony concept, which is motivated by the temporal characteristics of the information in the auditory nerve fiber firing patterns, was studied as a means of estimating the relative spectral distribution of an acoustic sound instead of the traditional spectral density power measurement. This principle has been applied previously as part of a detailed peripheral auditory model and experiments have shown that the auditory model is more robust to noise relative to an Fourier power analysis. However, it was unclear what were the relative contributions of the cochlear filters versus the timing-synchrony analysis in the overall performance of the auditory model. By replacing the cochlear filter-bank with a Hamming filter-bank we have demonstrated the following: (1) timing-synchrony analysis maintains all the speech information which is relevant for speech recognition tasks, (2) timing-synchrony measurement is significantly more robust against noise as compared to power measurement, and (3) the main contribution to the robustness is by the timing-synchrony analyzer and not by the unique shape of the cochlear filters.

ACKNOWLEDGMENT

I wish to thank J. B. Allen, B. S. Atal, D. A. Berkley and M. M. Sondhi for stimulating discussions throughout this work.

APPENDIX A

This appendix summarizes the system, the analysis conditions, the database and the signal conditions that were used in this and in the previous experiment (Ghitza, 1987, [1]).

A.1 The system

For the recognition experiments, we used the DTW recognizer described by Wilpon and Rabiner, 1985, [2]. In this version, the distance $D(x_i,x_j)$ between patterns (words) x_i and x_j is defined as

$$D(x_i,x_j) = \sum_k d(k,w(k),i,j) \qquad (1)$$

where the local frame distance $d(k,w(k),i,j)$ is the log likelihood distance between the kth frame of x_i and the $w(k)$th frame of x_j, i.e.,

$$d(k,w(k),i,j) = \log \left[\frac{(\mathbf{a}^j_{w(k)})'\mathbf{R}^i_k(\mathbf{a}^j_{w(k)})}{(\mathbf{a}^i_k)'\mathbf{R}^i_k(\mathbf{a}^i_k)} \right] \qquad (2)$$

where $\mathbf{a}^i_k$ is the vector of linear prediction coefficients of the kth frame of pattern i, $\mathbf{R}^i_k$ is the autocorrelation matrix of the kth frame of pattern i, and $'$ denotes vector transpose. The function $w(k)$ is the warping function obtained from a dynamic time alignment of pattern j to pattern i which minimizes $d(k,w(k),i,j)$ over a constrained set of possible $w(k)$. Equation (2) implies that in this configuration, recognition is based on spectral envelope variations only while energy clues have been omitted.

Fig. A-1 describes the system that was used for the experiments. The inputs to the existing alignment procedure were the first P autocorrelation coefficients, representing the spectral envelope. For the Fourier based front-end, these coefficients were computed directly from the input signal with P equal to 9 (Wilpon and Rabiner, 1985, [2]). To separate the EIH (or the modified EIH) spectral fine structure from the spectral envelope, we used the source-filter model, treating the EIH as the output log spectrum of a linear system excited by a given source. We further assumed that the hypothetical system can be modeled by an all-pole filter. Thus, the EIH envelope was estimated by using the conventional frequency domain linear prediction approximation methods. Since the EIH is measured in logarithmic units, the exponent of the EIH was first computed to obtain the EIH on a linear scale. An inverse DFT was then applied, to obtain the autocorrelation function. The degree of smoothing depends on the number of autocorrelation coefficients to be included. We found that for the EIH (or the modified EIH), an appropriate envelope fit was achieved by applying a 19*th* order all-pole polynomial fit.

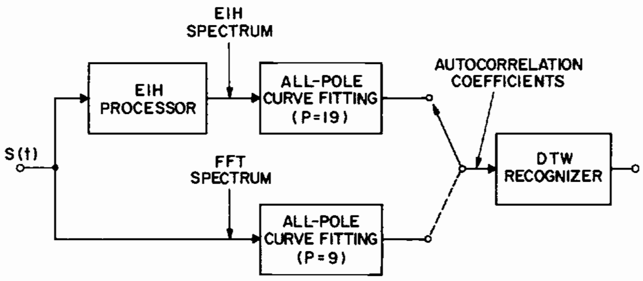
EIH SPECTRUM
EIH PROCESSOR
ALL-POLE CURVE FITTING (P=19)
AUTOCORRELATION COEFFICIENTS
S(t)
FFT SPECTRUM
DTW RECOGNIZER
ALL-POLE CURVE FITTING (P=9)

Figure A-1.

A.2 Analysis conditions

The speech signal was recorded using a standard telephone line, bandpass filtered from 100 to 3200 Hz, and sampled at a 6.67 kHz rate. In order to increase the measurement accuracy in the high frequency Hamming filters (recall that the timing-synchrony analysis is based on measuring the time interval between successive upgoing level crossings) the original digital samples were upsampled by a factor of 6, to 40 kHz sampling rate. When an upwardgoing level crossing was found, a linear interpolation was used to compute the precise time of crossing. Timing-synchrony analysis was performed every 5 milliseconds, taking into account the last, at most, twenty intervals with a maximum of 40 mS memory.

A.3 Database and signal conditions

The database for the recognition experiments was the 39 word alpha-digit vocabulary. It included the letters of the alphabet, the digits, and the control words STOP, ERROR and REPEAT. Two male and two female speakers repeated the vocabulary a total of 13 times. Recordings were made in a soundproof booth over a standard telephone line. Each of the words was manually edited, to mark the endpoints.

The noisy database was created by adding white noise to the clean database. The global-SNR was defined using *global* measurements. First, the total energy E_T of the clean word was computed. Then, the average energy per sample E was determined, dividing E_T by the number of samples in the word signal. Finally, E was used to set the variance of a white noise generator, according to the required global-SNR. Note that our definition of global-SNR is very severe; in most of the utterances in the alpha-digit vocabulary, a large portion of the signal is a vowel. This implies that the global-SNR conditions in the consonant regions is higher than the labeled global-SNR.

A.4 Experiment procedure and previous results

Data were first processed with the Fourier based front-end, producing autocorrelation files. The same data were then processed by the EIH front-end, producing a similar set of autocorrelation files. The autocorrelation files from both front-ends were then processed by the same DTW recognizer, described in Section A.1 of this appendix. A speaker dependent recognition test was then performed. For a given speaker, 5 of the repetitions were marked as the training set and the other 8 repetitions were used as the test set. The distance between every pair of repetitions belonging to the training set was computed and the most mismatched pair was chosen as the template set for the DTW. Note that the recognition scores obtained by this training procedure represent a lower bound on performance, since a better template set can be designed with more advanced techniques (Wilpon and Rabiner, 1985, [2]). For this comparative study, however, the above template set was satisfactory, since the same pair of repetitions was used for both front-ends. The recognition results for the 8 repetitions in the test set are summarized in Fig. 4.

REFERENCES

[1] Ghitza, O. (1987)." Auditory nerve representation as a front-end for speech recognition in a noisy environment", Computer Speech and Language, Vol. 2., to appear.

[2] 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., ASSP-33 (3), June, p.587.