A Measure of In-Synchrony Regions in the Auditory Nerve Firing Patterns as a Basis for Speech Vocoding

Oded Ghitza

Research Laboratory of Electronics and Lincoln Laboratory
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

A speech spectrum intensity measure based on temporal non-place modeling of the cat's auditory nerve firing patterns is introduced where the spectrum intensity values are estimated using timing-synchrony measurements only. The ability of this measure to serve as a speech information carrier was tested psychoacoustically, by integrating the proposed measure into a buzz-hiss type analysis-synthesis vocoder. Informal listening suggests that considerable speech information is preserved when this new measure is used to replace the spectrum analyzer portion of the vocoder.

INTRODUCTION

It seems to be true that humans perceive speech much better than any hardware processor designed for similar tasks. For example, when a person is listening to speech in the presence of any enviromental noise, the speech intelligibility remains relatively high; on the other hand, the intelligibility score of vocoded noisy speech drastically drops, for any usual kind of vocoder (Gold and Tierney, 1983, [7]). It is thus reasonable to adopt the human auditory system's structure in order to get better processing of speech. Our goal here is restricted to the design of a physiologically based speech spectral envelope estimator and to its experimental application as part of a analysis-synthesis vocoder which separately handles excitation (pitch and buzz-hiss decision) and spectral envelope information. We believe that such a measure may do well with noisy speech and also produce parameters capable of low bit rate encoding.

Measurements of the firing responses of cat auditory nerve fibers to speech-like stimuli (Sachs and Young, 1979, [12]; Young and Sachs, 1979, [15]; Sachs and Young, 1980, [13]; Delgutte and Kiang, 1984, [1]-[5]) suggest that firing rate is an insufficient carrier of speech information. Some use of temporal characteristics of the firing pattern seems necessary. It is also evident from these measurements that as the stimulus intensity increases, more fibers fire in synchrony with the stimulus frequency. It is thus possible to consider the width of the region in which all the fibers are in-synchrony with the stimulus frequency as a measure of the stimulus intensity.

In this work, a speech analyzer was implemented in two stages. The first stage models the auditory periphery processing structure up to the level of the auditory nerve. A heuristic, non-linear spectrum intensity measure is applied to the output of the first stage. This measure uses timing-synchrony measurements in an attempt to exploit the in-synchrony phenomena observed in the neuron firing patterns. We term the output of this measure "the in-Synchrony Bands Spectrum" (SBS). To verify its ability to convey speech information we used a channel vocoder as a test analysis synthesis system. Only the spectrum analysis path of the vocoder was replaced by the SBS analyser. DRT tests are still to be performed; however, informal listening to utterances of several male and female talkers suggests that considerable speech information is preserved.

ANALYSIS

The analyzer (Fig. 1-a) is comprised of two stages. The first is related to peripheral auditory physiology while the second plays the role of higher level processing. The design of the first stage is based on the general overall behavior of the peripheral auditory system. Fine details in realizing the different blocks were ignored. The first stage consists of a 100 filter filter-bank, where the filters are highly overlapped and equally spaced in the logarithmic scale with a 3% frequency step. Their frequency responses are similar to the tuning curves of auditory nerve fibers: each filter is identified by its characteristic frequency and has a frequency response with a -18 dB per octave rolloff in its low end and a very sharp rolloff at the high frequency end (Fig. 1-b). The shape of each filter is obtained by weighting appropriate FFT spectral points around the characteristic frequency. In the following text, we sometimes refer to such a filter as a "fiber".

The second stage of the analyzer processes the filter bank's outputs to provide the estimated spectral envelope information. The estimation principles are based on conclusions, yet to be psychophysically verified, evoked from the gross characteristics of the cat's auditory nerve firing patterns. Specifically, we adopt a temporal non-place model. It is assumed that information about the intensity of different spectral portions of the signal is in the number of fibers which fire synchronously, regardless of the fiber's characteristic frequency. Furthermore, the phase

CH2118-8/85/0000-0505 $1.00

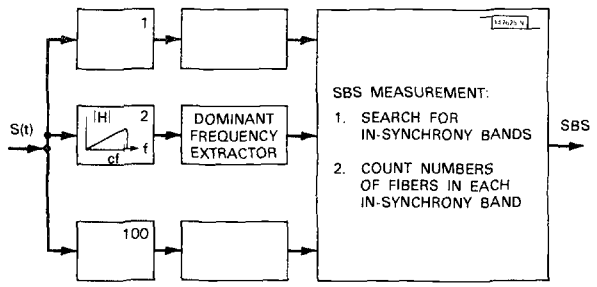

FIGURE 1.a

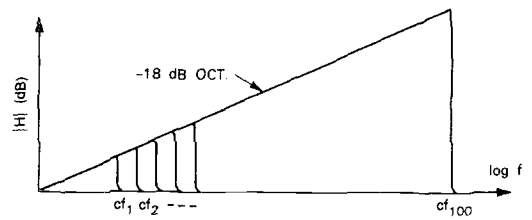

FIGURE 1.b

response properties of the fibers are assumed to be irrelevant. The basis for this assumption is the findings of Goldstein and Srulovicz, 1977, [9] and Srulovicz and Goldstein, 1984, [14], that the interspike interval statistic is adequate to explain the psychophysics of the perception of the pitch of complex tones. We also apply the in-synchrony idea 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 characteristic frequency increases (Johnson, 1980, [10]), we assume that the amount of synchrony in these fibers is still useful.

The following definitions are made to assist in describing the analysis technique:

Definition 1: An **in-synchrony band** is a region of L_n successive filters all 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 an in-synchrony band, it is necessary that L_n be greater or equal to a threshold, M.

Definition 2: The **in-Synchrony Band Spectra (SBS)** is a discrete function in 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.

The implementation of the SBS measurement is shown in Fig. 1-a. An array of dominant frequency extractors determines the strongest frequency component in each filter output signal. This dominant frequency is not necessarily the characteristic frequency of that filter. For the dominant frequency extraction we used the zero-crossing counting technique, which can be shown to be a consistent estimator of a hidden period in a given random process (Kedem, 1984, [11]). Now, the in-synchrony bands are detected by considering only the frequency regions where at least M (we set M=6) successive fibers possess the same dominant frequency. Finally, the number of fibers in those regions are counted, to obtain the SBS estimate. Fig. 2-b shows the SBS lines for a sample high resolution speech spectra (computed via FFT) plotted in Fig. 2-a. The left side plots are samples of female speech, while the right side show samples of male speech. The figures demonstrate the dominance effect which is the basic property of the SBS measure. The strongest harmonics (usually at the formant-peak regions for the voiced frames and at the high energy portion of the band for the unvoiced) are dominating the activity of fibers with higher characteristic frequencies. The magnitudes of the SBS lines (obtained by using a highly non-linear operation namely counting the fibers in synchrony with the dominant harmonic frequency) represents the relative importance of each activity region.

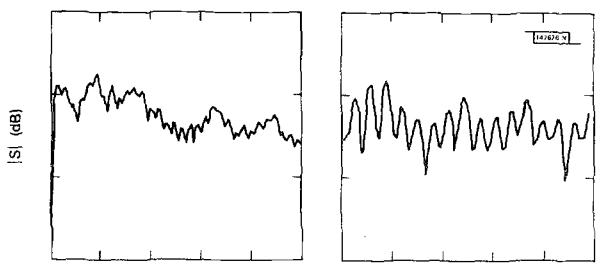

FIGURE 2.a

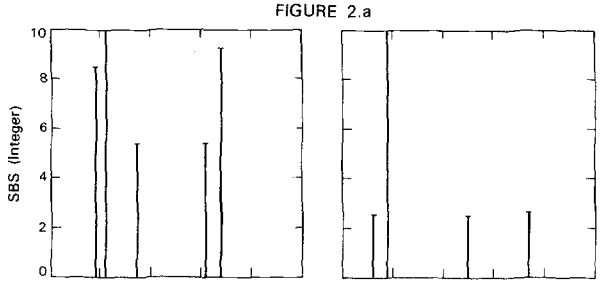

FIGURE 2.b

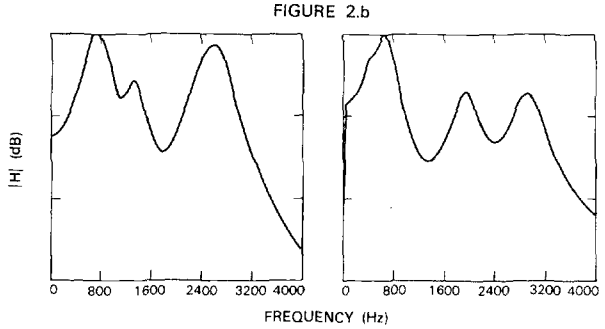

FIGURE 2.c

SYNTHESIS

One could examine the SBS reliability as an adequate speech information carrier either by visual or aural means. In the first approach, a kind of spectrogram can be created for a given utterance upon which a decision has to be made about the accuracy of the representation. However, no quantitative criterion exists for this judgment. In the psychoacoustic approach, an appropriate transformation should be suggested, to

create a spectral envelope out of the SBS lines. This spectral envelope is to be presented to some vocoder synthesizer, to create the synthesized output for a given input speech. The advantage in using the aural dimension is the existence of standard intelligibility and quality measures which can be applied to the synthetic speech.

Using the psychoacoustic approach, it is necessary to ensure that no side information from the spectral envelope is transferred to the synthesizer. The ideal speech synthesizer, thus, is the buzz-hiss type analysis-synthesis vocoder which uses the classical speech production model of a buzz-hiss type of excitation combined with spectral envelope representation of the vocal tract transfer function. We used the channel vocoder of Gold et al, 1981, [8], which is implemented at Lincoln Laboratory on a high speed signal processing computer, as a host vocoder. No change was made in the synthesizer. On the analysis side, only the buzz-hiss/pitch path of the vocoder was used while the spectrum envelope estimator was replaced. The synthesizer was driven by appropriate spectral values that were computed from SBS information. The SBS-to-spectral envelope transformation is of great importance since it determines the intelligibility as well as the quality of the synthesized speech. The transformation technique we used is a parallel formant spectral envelope, where the formant frequencies are the f_n's and their amplitudes are the L_n's of the SBS. The bandwidth of each formant is computed from an expression motivated by Fant's findings, 1960 [6], using speech production considerations. Finally, the resulting spectrum energy is adjusted, to fit the input frame energy. Fig. 2-c shows the spectral envelope functions derived from the SBS patterns in Fig. 2-b.

CONCLUSIONS

An SBS based analysis-synthesis vocoding was applied to several utterances of male and female talkers. Informal listening to the synthesized speech suggests that the in-synchrony-bands spectrum measure can be considered to retain speech information. However, the degree of accuracy in the SBS representation remains to be quantitatively confirmed by DRT tests. In addition, work aimed at improving the quality of the SBS-based speech synthesis is in progress. Other issues which are now under investigation are the robustness of the SBS measurement and its use as a low bit-rate encoder. For the first issue, the use of a periodicity estimate (as opposed to an energy estimate) and of the in-synchrony-band condition (Definition 1) seems to be appropriate for reducing the effect of noise. As for the low rate coding, the observation that very few SBS lines are needed to represent a speech frame and the restricted values each SBS-line magnitude can possess (as a result of a counting procedure) suggest a possibility for an efficient encoding scheme.

ACKNOWLEDGEMENT

I wish to thank W. M. Siebert for his share in formulating the framework of this study, and to B. Gold and J. Tierney for stimulating discussions throughout this work.

REFERENCES

[1-4] B. Delgutte and N.Y.S. Kiang, "Speech Coding in Auditory Nerve: I, III, IV, V," J. Acoust. Soc. Am., 75, No. 3, pp. 866-918, March 1983.

[5] B. Delgutte, "Speech Coding in Auditory Nerve: II," J. Acoust. Soc. Am., 75, No. 3, March 1983.

[6] G. Fant, Speech Sounds and Features (MIT Press, Cambridge, MA).

[7] B. Gold and J. Tierney, "Vocoder Analysis Based on Properties of the Human Auditory System," Technical Report 670, M.I.T. Lincoln Laboratory, December 1983.

[8] B. Gold, P.E. Blankenship, and R.J. McAulay, "New Applications of Channel Vocoders," IEEE Trans. Acoust., Speech, and Signal Processing, ASSP-29, No. 1, pp.13-23, February 1981.

[9] J.L. Goldstein and P. Srulovicz, "Auditory-nerve Spike Intervals as an Adequate Basis for Aural Spectrum Analysis," in Psychophysics and Physiology of Hearing (Academic Press, London), p. 337.

[10] D.H. Johnson, "The Relationship Between Spike Rate and Synchrony in Responses of Auditory-Nerve Fibers to Single Tones," J. Acoust. Soc. Am., 68, No. 4, p. 1115, October 1980.

[11] B. Kedem, "Detection of Hidden Periodicities by Means of Higher Order Crossings," to be published.

[12] 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. Am., 66, No.2, p. 470, August 1979.

[13] M.B. Sachs and E.D. Young, "Effects of Nonlinearities of Speech Encoding in the Auditory Nerve," J. Acoust. Soc. Am., 68, No. 3, p. 858, September 1980.

[14] P. Srulovicz and J.L. Goldstein, "A Central Spectrum Model: A Synthesis of Auditory-Nerve Timing and Place Cues in Monaural Communication of Frequency Spectrum," J. Acoust. Soc. Am., 73, No. 4, p. 1266, April 1984.

[15] 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. Am., 66, No. 5, p. 1381, November 1979.

This work was supported in part by the Department of the Air Force. The experimental work described here was conducted at Lincoln Laboratory, where Dr. Ghitza is a consultant. During this work Dr. Ghitza has been supported as a post-doctoral fellow at the Research Laboratory of Electronics under the Myron A. Bantrell Charitable Trust.

The U.S. Government assumes no responsibility for the information presented.