A SPECTRAL-TEMPORAL SUPPRESSION MODEL FOR SPEECH RECOGNITION

P. L. Divenyi

Speech and Hearing Research Facility, Veterans Administration Medical Center, Martinez, California, 94553, and Department of Speech and Hearing, University of California, Santa Barbara, California, 93106, U.S.A.

INTRODUCTION

Speech recognition systems, however heterogeneous in their conceptions and schemes, share at least one basic feature: the inclusion of a vocoder-type front-end. While many of the early, and some of the contemporary, systems adopted a pragmatic design for their front-end filter bank, there were some efforts (e.g., Chistovich et al., 1975; Searle et al., 1979) toward providing the recognizer with an input stage that was modeled after the human ear. The motivation for such a design was the desire to optimize the recognition process from the very first stage on. However, work by auditory physiologists on auditory nerve responses to speech (Young and Sachs, 1979; Delgutte, 1980) signaled a welcome convergence of interests by two groups of scientists on the problem of speech processing in the auditory system. More recent work by several investigators, some of which is included in the present symposium, has been directed toward designing recognizer front-ends that resembled the ear more-and-more closely, and toward examining effects of model parameter modifications on recognition performance.

Computational models of the auditory system fall into two major classes, depending on whether the calculations are performed in the time or in the spectral domain. The advantage of time-domain algorithms lies mainly in their speed, whereas spectrally-based algorithms may more closely approximate the actual auditory processes because they are able to deal more directly with non-linear filtering operations. The present model is spectral in the sense that the filtering computations are executed in the frequency domain.

DESCRIPTION OF THE MODEL

The present model has been built around the physiologically-based and fine-tuned spectral model proposed by Shannon (1979). That work stands out in that it computes the magnitude of peripheral auditory activity across all frequency-specific channels, taking into account passive and active cochlear filtering, compressive nonlinearity, and suppression on both sides of a given channel. It is, however, restricted to spectral processing. The present modeling work was undertaken in an effort to see how time-varying signals can benefit from spectral suppression, i.e., an enhancement of the contrast between channels differing in their activity level, as offered by the Shannon model. The five stages of this model are connected in a strict sequential order, i.e., without feedback loops.

1. The Spectral Estimator Stage.
The physical continuum of frequency was mapped into 120 discrete channels between 50 and 10kHz using the frequency-to-basilar membrane distance transformation proposed by Greenwood (1961). The purpose of the spectral estimator was to provide the inner ear simulator (that operated in the spectral domain) with an estimate of the input magnitude that excited each channel. This input magnitude had to reflect the duration of the assumed equivalent impulse response of the corresponding inner-ear filter, i.e., it had to be gated using a window whose length was a function of the inner-ear filter width. Thus, a separate magnitude estimate had to be made for the narrow active- and the wider passive filters of each channel (see Stage 3). We adopted a Hamming window with a skew that emphasized more recent events. We arbitrarily assigned a 10-Hz maximum frequency resolution to our 50-Hz channel and calculated the window length for each channel assuming linear impulse response and applying the Greenwood mapping. We also limited the minimum window length to 2 ms, in order to account for an indelible neural refractoriness. The actual estimation was represented by Direct Fourier Transform coefficients of the windowed input at the frequency corresponding to a given channel.

2. The Outer- and Middle-Ear Response Simulator.
To account for ear canal resonance and middle ear attenuation, we included a spectral shaping algorithm gradually falling off below 2.5 and above 4 kHz. The attenuation (in dB) was a linear function of basilar membrane distance.

3. The Inner-Ear Spectral Response Simulator.
This stage, the actual Shannon model, is characterized by two concurrently working filter banks. One of the banks consists of passive, broadly-tuned, linear filters having a hhigh (30-dB SPL) threshold. Filters in the other bank are active, sharply tuned, low-threshold filters with a nonlinear compressive response that makes any activity increment beyond 40 dB SPL negligible. The active filters are followed by a sub-stage representing the suppression of high tones by low tones. The output of this sub-stage is linearly added, channel-by-channel, to that of the passive filter bank. The output of the mixer is followed by the sub-stage of suppression of low tones by high tones. In sum, the output of the inner-ear simulator represents the magnitude of the activity in the auditory nerve across tonotopically organized channels. This output compresses a 120-dB dynamic range in the input into a 20-to-25-dB range in the output.

4. The Auditory Nerve Temporal Response Simulator.
Single unit studies have demonstrated that there is a sizable temporal adaptation effect in the response of single auditory nerve fibers (Smith and Zwislocki, 1975). This effect is characterized by a strong burst of activity at the onset of the stimulus followed by a gradual decrease, and by a moment of sudden decrease of the activity at stimulus offset, followed by a gradual recovery. We used Smith's theoretical expression for this temporal process, noting that the effect is independent in each channel and that the adapted output is affected only by the magnitude of the present and the immediately preceding output epoch, rather than by the input. Thus, the effect is not unlike that of a high-pass filter with a floor (i.e., the spontaneous activity level). It was implemented in our model as simple exponential differentiators having different time constants for adaptation (18 ms) and recovery (36 ms). This stage enhances temporal contrasts in the input.

5. The Temporal Integrator Stage.
Auditory psychophysical data, however, depict the auditory system as one with memory: Detection of signals at threshold and ~~detection of~~ envelope

fluctuations, for example, clearly speak for the existence of a low-pass process, i.e., of a leaky integrator. We implemented this stage as an exponential integrator placed on each channel at the output of the temporal adaptation stage. The time constant we chose was short (1.5 ms) -- in agreement with other workers (Penner, 1978). We also noted that, because this integrator operates on the compressed output rather than on the input, a single, short time constant must be capable of accounting for both temporal integration at threshold and envelope discrimination at suprathreshold levels.

EXAMPLES

We have completed several tests with simple, easily definable input signals, in order to obtain an optimized set of model parameters. The output of two simple signals, a 100-dB SPL, 2-ms click and a 50-dB SPL 50-ms Gaussian white noise burst, are shown in Fig. 1. We have also examined the behavior of the model in response to natural speech sounds. One example, the beginning of the phonetically-balanced sentence "The goose was brought straight from the old market" is shown as a spectrogram in Fig. 2 and as a "neurogram", or time-frequency channel model output, in Fig. 3. In addition, we have also examined a large number of natural CV utterances, in an attempt to search for invariant cues (not shown here).

SPEECH RECOGNITION TESTS

In order to see whether the model could embody an improved front-end to a cepstrum-based recognizer, we conducted a series of experiments on a natural sentence data base. Recognition performance with the raw output of the model as input to the recognizer was significantly poorer than when the front-end was a simple vocoder. Much of the performance degradation could be attributed to the presence of individual low harmonics that dominated the model output. It seems, therefore, that some type of feature detection would be necessary before the model could become a useful tool in automatic speech recognition.

ACKNOWLEDGMENTS

This research was conducted when the author was visiting at the Institut National de la Recherche Scientifique, Université de Québec, and has been supported by the Veterans Administration.

REFERENCES

Chistovich, L., Fyodorova, N., Lissenko, D., & Zhukova, M. (1975). Auditory segmentation of acoustic flow and its possible role in speech processing. In G. F. a. M. Tatham (Ed.), Auditory analysis and perception of speech (pp. 221-232). London: Academic.

Delgutte, B. (1980). Representation of speech-like sounds in the discharge patterns of auditory nerve fibers. J. Acoust. Soc Amer., 68, 843-857.

Greenwood, D. D. (1961). Auditory masking and the combination band. J. Acoust. Soc. Amer., 33, 484-502.

Penner, M. (1978). A power-law transformation resulting in a class of short-term integrators that produce time-intensity trades for noise bursts. J. Acoust. Soc. Amer., 63, 193-201.

Sachs, M. B., & Young, E. D. (1979). Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. J. Acoust. Soc. Amer., 66, 470-479.

Searle, C. L., Jacobson, J. Z., & Rayment, S. G. (1979). Stop consonant discriminatin based on human audition. J. Acoust. Soc. Amer., 65, 799-809.

Shannon, R. (1979). A model for psychophysical suppression. J. Acoust. Soc. Amer., 65, S56.

Smith, R. L., & Zwislocki, J. J. (1975). Short-term adaptation and incremental response of single auditory nerve fibers. Biol. Cybern., 17, 169-182.

FIGURE LEGENDS

1. a: 3-D picture of the model's response to a 2-ms click presented at 100 dB SPL. Frame size: .25 ms. Only the first 10 ms of the response are shown. b: 3-D picture of the model's response to a 50-ms burst of white noise presented at 50 dB SPL. Frame size: 2 ms. Only the first 80 ms of the response are shown.
2. Conventional spectrogram of the utterance "The goose wa(s)..." by a male talker.
3. Model output ("neurogram") of the same utterance. Difference between the darkest and the lightest parts of the output is 13 dB. Frame size: 2 ms.