AUDITORY SPEECH PREPROCESSORS

George Zweig
Signition, Inc.
P.O. Box 1020
Los Alamos, New Mexico 87544

ABSTRACT

A nonlinear transmission line model of the cochlea (Zweig 1988) is proposed as the basis for a novel speech preprocessor. Sounds of different intensities, such as voiced and unvoiced speech, are preprocessed in radically different ways. The Q's of the preprocessor's nonlinear filters vary with input amplitude, higher Q's (longer integration times) corresponding to quieter sounds. Like the cochlea, the preprocessor acts as a "subthreshold laser" that traps and amplifies low level signals, thereby aiding in their detection and analysis.

• *Speech preprocessors are important.* Small improvements at the beginning of the recognition process can lead to substantial improvements by the end. Resolving acoustic ambiguities decreases the number of possibilities that must resolved by higher level linguistic processing.

• The past: *Much has been learned about speech preprocessing from the inner ear of vertebrates.* Historically, this approach dates back to Ohm (of Ohm's Law fame), Helmholtz (1863), and more recently to work at Bell Laboratories (Flanagan 1965). Even information about hearing mechanisms in lower vertebrates is of interest because the sounds they analyze are qualitatively similar to speech. Many natural sounds, like speech, are created by exciting resonant systems either periodically or chaotically. Presumably speech sounds evolved to take advantage of preexisting signal processing mechanisms in hearing. Past research has shown:

1. The external and middle ears of humans act together as a linear acoustic filter that boosts high frequency sound by 6 dB/octave. The long-time average spectrum of speech is approximately the inverse of the product of the external and middle ear transfer functions (Zweig 1987). Correspondingly, speech preprocessors routinely boost the high frequencies in speech by differentiating the acoustic signal.
2. Ohm's acoustic law states that the cochlea analyzes sound by decomposing it into different frequency components. Current speech preprocessors all extract frequency information contained in sound by one of several methods — moving Fourier transforms, filter banks, or LPC analysis.
3. The frequency-position map within the cochlea (Liberman 1982), like the psychoacoustically derived mel scale, is approximately linear at low frequencies and exponential at high frequencies. Speech preprocessors based on filter banks have the center frequencies of their filters equally spaced on the mel scale. LPC based preprocessors may also use the mel scale (e.g. the BBN system uses a "mel-cepstrum" analysis).

4. Cochlear filters are approximately constant bandwidth at low frequencies and constant Q at high frequencies (Kiang et al. 1986). The bandwidths of filters in speech preprocessing filter banks follow this same arrangement.

• The present: *Vertebrate inner ears analyze sounds of differing intensities with different analysis systems.* As a corollary, voiced and unvoiced speech are processed differently by the human inner ear. Correspondingly, differences in the acoustic structure of voiced and unvoiced speech are matched to differences in their respective analysis systems. In particular, since the vocal tract is excited chaotically in unvoiced speech, unvoiced speech must by processed by each auditory filter for a longer time than voiced speech if the resonant modes (formants) of the vocal tract are to be extracted. The nonlinear auditory filters of the inner ear of both lower vertebrates and mammals have Q's that vary with input amplitude, with higher Q's (longer integration times) being used for quieter sounds. Current speech preprocessors are linear and use a single algorithm for all sounds, independent of their amplitude. Perhaps they shouldn't.

Although there are many qualitative similarities in the way in which lower vertebrates and mammals hear, there are important differences in the functioning of their inner ears. The inner ear of lower vertebrates works as a nonlinear filter bank with approximately the same sound stimulating all hair cells which act as independent filters. In the mammalian inner ear, the individual cellular components are strongly coupled, sound exciting them into collective traveling wave oscillations that deliver different stimuli to different hair cells (Zweig 1988). The sensitivity, resolution, and dynamic range of hearing in mammals are much greater than that in lower vertebrates. Although collective phenomena can give rise to qualitative differences in signal processing for these two classes of systems, similarities exist and a study of the simpler lower vertebrate inner ear is also informative.

Lower vertebrates: The hair cell output voltage $O(t)$ (relative to its resting voltage) satisfies the second order nonlinear inhomogeneous differential equation (Lackner and Zweig 1988):

$$\ddot{O}(t) + \omega^2 O(t) = \epsilon\, \alpha\Big(I(t), \dot{I}(t), O(t), \dot{O}(t)\Big)\,, \tag{1}$$

where the diacritical dot denotes differentiation with respect to time, ω is the angular frequency of the freely oscillating hair cell in the small amplitude linear limit, $I(t)$ the input signal (transducer conductance), ϵ a small constant, and α a nonlinear function of I, $\dot{I}$, O, and $\dot{O}$. A linear second order filter, by comparison, has output $O(t)$ satisfying:

$$\ddot{O}(t) + \omega^2 O(t) = \alpha_0 I(t) + \alpha_1 \dot{I}(t) + \alpha_2 O(t) + \alpha_3 \dot{O}(t)\,, \tag{2}$$

where the α_i are constants. For sounds near the threshold of hearing, hair cell responses are linear and Eq. 1 reduces to Eq. 2.

Information about the functional form of $\alpha(I, \dot{I}, O, \dot{O})$ for the turtle has been determined from experiment (Lackner and Zweig 1988). A turtle hair cell acts as a nonlinear harmonic oscillator with the useful properties that its integration time, Q/ω, increases with decreasing amplitude, and its output is compressed to lie within a limited range.

Speech in noise should be preprocessed with a nonlinear filter bank modeled after the turtle inner ear to see if the formants of both voiced and unvoiced speech are clearer than formants obtained from a conventional filter bank. A preprocessor with two linear filter banks operating in parallel, and separately optimized for voiced and unvoiced speech, would be simpler but less effective than a preprocessor based on Eq. 1. (Two linear filter banks would not provide the formant sharpening created by the two tone suppression (Sachs and Kiang 1968) implicit in Eq. 1.)

Hair cells of lower vertebrates and mammals are active (energy producing) as well as nonlinear. The implications of these active and nonlinear elements for signal processing in lower vertebrates may be determined from Eq. 1. The situation for mammals is both more complex and surprising.

Mammals: The cochlea acts as an active nonlinear one-dimensional mechanical transmission line with time delayed feedback (Zweig 1988). The parameters defining the circuit elements vary gradually along the line. Each section of the line contains a negatively damped nonlinear harmonic oscillator stabilized by a negative feedback force whose strength is proportional to the displacement of the oscillator at a previous time. The time delay is proportional to the oscillator's period (with the proportionality constant approximately $1\frac{3}{4}$).

Nonlinearities arise through the dependence of damping and feedback strength on oscillator displacement. The damping increases and the feedback strength decreases with increasing oscillator displacement. The precise functional form of these nonlinearities, which become important at intermediate and high sound pressure levels (levels of voiced speech), is currently under investigation. Preliminary results indicate that the nonlinearities provide at least three benefits: automatic gain control necessary for the analysis of speech whose components differ markedly in amplitude; sharpening of formants, making them easier to recognize in the presence of noise; and adjustment of filter bandwidths with amplitude so that quieter unvoiced portions of speech will have longer time windows (narrower filters) for their analysis than the louder voiced portions.

In addition, *the transmission line model possesses a remarkable mechanism for the amplification and analysis of sound near threshold.* This mechanism is related to another unexpected feature of the model: standing waves are generated in the cochlea at low sound pressure levels, not just traveling waves as first observed by von Békésey at high sound pressure levels.

For example, a pure tone sets up a conventional forward traveling wave which moves from the stapes along the organ of Corti to its point of maximum displacement, the response to lower frequency tones peaking further down the cochlea. The active elements increase the amplitude of the forward traveling wave. This wave is partially reflected by spatial variations in the mechanical properties of the cochlea. The amplitude and phase of the backward traveling wave depend on the nature and spatial distribution of the mechanical inhomogeneities and the amplitude and phase of the forward traveling wave at its points of partial reflection. Contributions to the backward traveling wave are largest when they originate in the peak region of the forward traveling wave. The phase of the backward traveling wave changes monotonically as it propagates toward the stapes, and the active elements increase its amplitude. At the stapes the backward

traveling wave is partially reflected and then interferes with the incident forward traveling wave. If the interference is constructive, the forward traveling wave is reinforced, leading, in turn, to a larger backward traveling wave. The process of backward, and then forward, wave creation and amplification builds in this case a large standing wave, as first conjectured on general grounds by Kemp (1980).

Thus, mechanical inhomogeneities and "stimulated emission" from the active regions of the cochlea lead to coherent amplification of the incident wave at stimulus frequencies where there is constructive interference at the stapes. If the incident sound is a pure tone, then increasing its frequency results in shifting the position of the maximum displacement towards the stapes, and the model predicts a concomitant alternation of constructive and destructive interference. The mechanical inhomogeneities and active elements are responsible for amplifying low level signals at certain frequencies, much like a "subthreshold laser", thereby increasing the overall sensitivity of hearing. The sound pressure level necessary for the detection of a pure tone is expected to vary in a cyclical fashion with frequency. Such microstructure in the hearing threshold curve is easily observed (Elliot 1958; Zweig 1973; Thomas 1975; Kemp 1979; Zwicker and Schloth 1984).

The energy in the backward traveling wave not reflected at the stapes vibrates the middle ear bones and ultimately appears in the external ear canal as sound. Thus the model predicts the existence of "otoacoustic emissions," also in agreement with experiment (Kemp 1978). In fact, the maxima in the spectra of otoacoustic emissions correlate strongly with the periodic minima in the threshold hearing curve (Kemp 1979; Zwicker and Schloth 1984). As predicted by the model, the ear emits most loudly at those frequencies to which it is most sensitive.

Spontaneous emissions, corresponding to certain mild cases of tinnitus (Kemp 1981), arise when the product of the magnitudes of the reflection and amplification factors exceeds unity and the traveling waves add coherently at the stapes. Spontaneous emissions are therefore expected to occur at frequencies to which the ear is most sensitive, and in fact experimental measurements indicate that this is true (Zwicker and Schloth 1984). These emissions, once initiated, require no external sound for their maintenance; they are created by an oscillating biological "hydromechanical laser." The stapes acts as a partially reflecting mirror that feeds back energy to the cochlear amplifier in the form of a forward traveling wave. Both spontaneous and externally stimulated emissions are limited in amplitude by mechanical nonlinearities which prevent the formation of large standing waves at high sound pressure levels.

Finally, at low amplitudes the energy of transients (e.g., clicks or the initial burst in unvoiced plosives) is also trapped and amplified within the cochlea. The energy that does leak out into the ear canal (an "echo" of the incident sound) has a frequency spectrum that peaks at those frequencies to which the ear is most sensitive (Zwicker and Schloth 1984).

In contrast to mammals, lower vertebrates utilize different (and also remarkable) mechanisms that do not depend on collective phenomena to increase their sensitivity to sound, but their thresholds of hearing are much higher than those of mammals.

• The future: *A number of nonlinear signal processing principles remain to be abstracted from the peripheral hearing process and applied to the design of speech preprocessors.* The intensity-dependent induced emission of mechanical energy, and its coherent amplification at certain regularly spaced frequencies through multiple reflections, provide a dramatic example of how the detection and analysis of sound in the cochlea depend on sound intensity. It also shows that the cochlea processes information in a surprising fashion through the *collective* action of many components. Enough experimental evidence already exists to indicate that nonlinear signal processing in the cochlea at intermediate and high sound pressure levels is also the product of collective hair cell action. Its mechanisms and effects promise to be both novel and effective in resolving acoustic ambiguities that currently plague conventional speech preprocessors.

Acknowledgements

Scott Konishi, Klaus Lackner, and Christopher Shera have contributed to the research reported here. DARPA and Los Alamos National Laboratory have provided financial support.

References

Elliot, E. (1958). "A ripple effect in the audiogram," *Nature* **181**, 1076.

Flanagan, J. L. (1965). *Speech Analysis – Synthesis and Perception.* Berlin: Springer Verlag.

Helmholtz, H. L. F (1863). *Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik.* Braunschweig: Vieweg. Trans. by A. J. Ellis, *On the Sensations of Tone as a Physiological Basis for the Theory of Music.* New York: Dover, 1954.

Kemp, D. T. (1978). "Stimulated acoustic emissions from within the human auditory system," *J. Acoust. Soc. Am.* **64**, 1386–1391.

Kemp, D. T. (1979). "The evoked cochlear mechanical response and the auditory microstructure – evidence for a new element in cochlear mechanics," *Scand. Audiol. Suppl.* , 35–47.

Kemp, D. T. (1979). "Evidence of mechanical nonlinearity and frequency selective wave amplification in the cochlea," *Arch. Otorhinolaryngol.* **224**, 37–45.

Kemp, D. T. (1980). "Towards a model for the origin of cochlear echoes," *Hearing Res.* **2**, 533–548.

Kemp, D. T. (1981). "Physiologically active cochlear micromechanics – one source of tinnitus" in *Tinnitus*, edited by D. Evered and G. Lawrenson, 54–81. London: Pitman.

Kiang, N. Y. S., M. C. Liberman, W. F. Sewell, and J. J. Guinan (1986). "Single unit clues to cochlear mechanisms," *Hearing Res.* **22**, 171–182.

Lackner, K. S. and G. Zweig (1988). "A nonlinear oscillator model of hair cells," paper presented at the International Symposium on Current Concepts of Hair Cell Function, Ann Arbor, Michigan.

Liberman, M. C. (1982). "The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristic frequency," *J. Acoust. Soc. Am.* **72**, 1441–1449.

Sachs, M. B. and N. Y. S. Kiang (1968). "Two-tone inhibition in auditory-nerve fibers," *J. Acoust. Soc. Am.* **43(5)**, 1120–1128.

Thomas, I. B. (1975). "Microstructure of the pure-tone threshold," *J. Acoust. Soc. Am.* **57** Suppl. 1, S26–S27.

Zweig, G. (1973), unpublished observation of auditory threshold microstructure.

Zweig, G. (1987), DARPA Progress Report.

Zweig, G. (1988). "Cochlear mechanics," submitted to *J. Acoust. Soc. Am.*

Zwicker, E. and E. Schloth (1984). "Interrelation of different oto-acoustic emissions," *J. Acoust. Soc. Am.* **75**, 1148–1154.