A Cochlea Filter Bank for Speech Analysis

Elaine Tsiang
et@monowave.com

ABSTRACT

The filter bank design derives from a three-dimensional solution to the equation of motion of the basilar membrane. In particular, the inter-filter coupling constants depend quadratically on the breadth of the basilar membrane, which may vary along its length.

We describe the implementation of a design based on the physical properties of the average human cochlea. Decreasing inter-filter feedback significantly attenuates frequency modulations in the range of formant movements in speech, while leaving stationary tones relatively intact. We demonstrate this effect with cochleagrams for chirps, tones and speech.

This filter bank design is currently used as a front-end processor to a neural-network based continuous-speech recognizer. Results are encouraging. Because formant dynamics encode important perceptual information, this filter bank may be useful as initial analysis for automatic speech recognition.

1. Introduction

The proposed filter bank is motivated by the need to obtain an initial spectral representation of speech signals that bears sufficient similarity to that performed by the human ear to support a neural model of speech perception[1][2][3]. The model is to be used for acoustic-phonetic decoding in speech recognition. The aim is not to match cochlea tuning curves, or to reproduce its various nonlinear responses, but to capture critical speech dynamics.

Of the many and intricate structures in the cochlea, the basilar membrane and the hair cells have been most extensively modeled (see reviews [4],[5]). This paper deals only with the basilar membrane.

Many simplifying assumptions about the hydrodynamics of the fluid reduce the nonlinear Navier-Stokes equations to the linear homogeneous Laplace equation for the fluid pressure[6]. We further simplify the shape of the cochlea from a spiral to two stacked parallelopipeds separated by the basilar membrane(Figure 1). We can then formulate the fluid pressure as a solution to the Laplace equation with appropriate boundary conditions.The membrane itself is assumed to consist of members with significant transverse stiffness, and negligible longitudinal stiffness, making it a system of vibrating beams, coupled only via the surrounding fluid. We assume the difference in the fluid pressure across the membrane alone drives its vibration, the displacement of which reacts on the fluid pressure (the point impedance assumption).

Many filter banks have been proposed based on such an underlying cochlea model. The proposed filter bank differs from others(e.g., [7],[8])in its substantial inter-filter feedback. Each filter inputs from all other filters, and outputs to all other filters. We show the utility of this feedback in capturing formant dynamics.

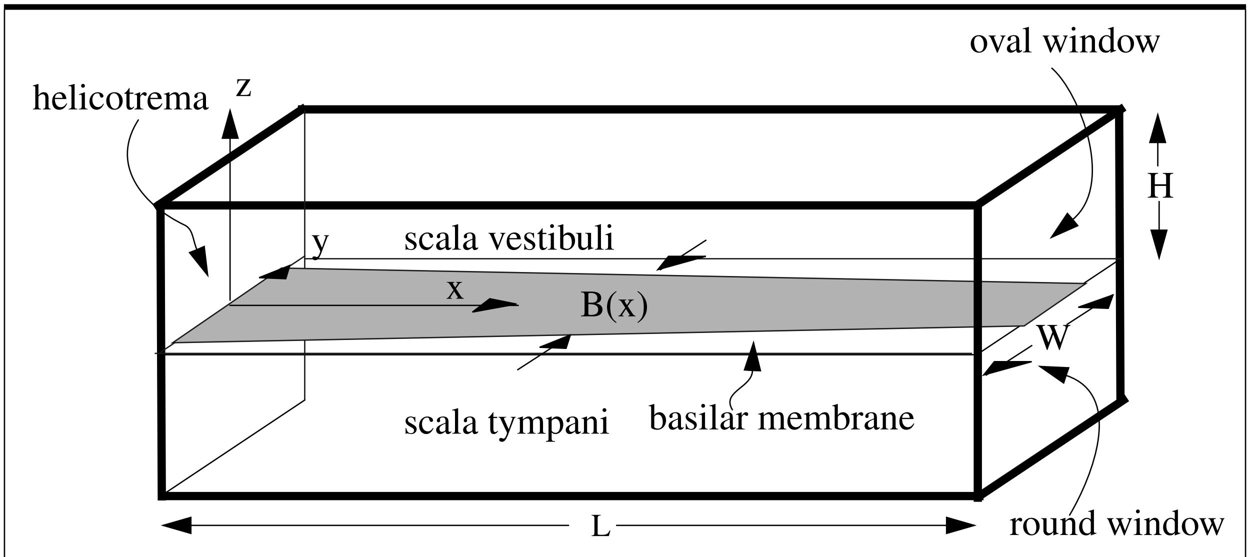


Figure 1 Cochlea Model

2. The Cochlea Filter Bank

The equation of motion of the basilar membrane as point impedances is:

$$-2p(x, y = 0, z = 0)B(x) = m(x)\frac{d^2z}{dt^2} + r(x)\frac{dz}{dt} + k(x) \qquad \text{(eq. 1)}$$

where p is the pressure in the scala vestibuli, B is the effective breadth of the basilar membrane, m the mass, r the damping and k the stiffness of the membrane sections. p(x,y,z) is the solution to the Laplace equation which results from the simplifications of the hydrodynamics of the cochlear chamber as incompressible fluid in inviscid flow within rigid walls:

$$p(x, y, z) = \int_{\delta V} g\frac{\partial p}{\partial n}dS \qquad \text{(eq. 2)}$$

where the integral is over the boundary δV of the enclosed volume, including the upper and lower surfaces of the basilar membrane itself but excluding the helicotrema. g is the Green's functions with boundary condition

$$\left.\frac{\partial g}{\partial n}\right|_{\delta V} = 0 \qquad \text{(eq. 3)}$$

where $\frac{\partial}{\partial n}$ is the partial derivative normal to the boundary. At the helicotrema, both g and p = 0. The boundary condition for p(x,y,z) is

$$\left.\frac{\partial p}{\partial n}\right|_{\delta V} = 0 \qquad \text{(eq. 4)}$$

except at the windows:

$$\left.\frac{\partial p}{\partial n}\right|_{oval} = -\left.\frac{\partial}{\partial n}(p)\right|_{round} = \rho\frac{du_x}{dt} \qquad \text{(eq. 5)}$$

and the membrane:

$$\left.\frac{\partial p}{\partial n}\right|_{vestibuli} = -\left.\frac{\partial}{\partial n}(p)\right|_{tympani} = -\rho\frac{\partial u_z}{\partial t} \qquad \text{(eq. 6)}$$

ρ is the fluid density, u_x is the stapes velocity and u_z is the fluid velocity at the membrane. The integral over the boundary therefore reduces to one over the windows and the membrane only. The Green's functions is

$$g(x, y, z|\xi, \eta, \zeta) = f(\xi - x, y, z|\eta, \zeta) - f(\xi + x, y, z|\eta, \zeta) \qquad \text{(eq. 7)}$$

$$f(\chi, y, z|\eta, \zeta) = \frac{1}{4\pi}\sum_{k=-\infty}^{\infty}\sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty}\frac{(-1)^k}{\left((2kL+\chi)^2 + (mW + (-1)^m\eta - y)^2 + (nH + (-1)^n\zeta - y)^2\right)^{\frac{1}{2}}}$$

Defining the effective breadth B(x) as

$$B(x)g(x, 0, 0|\xi, 0, 0)\frac{d^2 z}{dt^2} = \int_{-\frac{W}{2}}^{\frac{W}{2}}\left(g(x, y, z|\xi, \eta, \zeta)\cdot\frac{\partial^2 u_z}{\partial t^2}\right)d\eta \qquad \text{(eq. 8)}$$

and taking the first approximation of (eq. 7), the equation of motion,(eq. 1), when discretized along x at a rate of $\frac{1}{D}=\frac{M}{L}$, for a total of M filter sections, becomes:

$$\left[m^i + 2\rho\frac{D^2(B^i)^2}{A}i\right]\frac{d^2 z^i}{dt^2} + r^i\frac{dz^i}{dt} + k^i z^i = -2\rho DB^i i\frac{du_x}{dt} - 2\rho D^2 B^i\sum_{j=0}^{M}\Theta_j^i\frac{B^j}{A}\frac{d^2 z^j}{dt^2} \qquad \text{(eq. 9)}$$

where the cross-sectional area $A = W\cdot H$ and

$$\Theta_j^i = \begin{Bmatrix} j, j<i \\ 0, j = i \\ i, i>j \end{Bmatrix} \qquad \text{(eq. 10)}$$

Discretized in time by

$$z(n+1) = z(n) + \left(\frac{T}{2}\right)\left(\frac{d}{dt}z(n+1) + \frac{d}{dt}z(n)\right) \qquad \text{(eq. 11)}$$

(eq. 9) yields the following filter bank equations

$$\tilde{z}^i(n+1) - \tilde{z}^i(n) + a_1^i z_1^i(n) + a_2^i z_2^i(n) = b^i[u_x(n+1) - u_x(n)] \qquad \text{(eq. 12)}$$

$$z_2^i(n+1) = z_2^i(n) + \left(\frac{T}{2}\right)\left(z_1^i(n+1) + z_1^i(n)\right)$$

$$\tilde{z}^i(n+1) = \sum_{j=0}^{M}\left(\delta_j^i + c^i d^j\Theta_j^i\right)\left(z_1^i(n+1)\right)$$

The filter bank consists of resonators with all-to-all inter-filter feedback. T is the sampling period, and the coefficients are given by

$$\mu^i = m^i + 2\rho\frac{D^2(B^i)^2}{A}i + \left(\frac{T}{2}\right)r^i + \left(\frac{T}{2}\right)^2 k^i$$

$$a_1^i = \left(\frac{2}{\mu^i}\right)\left[\left(\frac{T}{2}\right)r^i + \left(\frac{T}{2}\right)^2 k^i\right] \qquad \text{(eq. 13)}$$

$$a_2^i = \left(\frac{2}{\mu^i}\right)\left(\frac{T}{2}\right)k^i$$

$$b = -\left(\frac{2\rho}{\mu^i}\right)DB^i i$$

$$c^i = \left(\frac{2\rho}{\mu^i}\right)DB^i$$

$$d^i = \frac{DB^i}{A}$$

We have implemented this filter bank in the time domain using the physical properties of the average human cochlea as given by [9]. A = 0.01 cm^2. B^i tapers from 0.0075 cm at the oval window at a rate of 0.0012(cm/cm). 105 filter sections span a frequency range from 187Hz to 6Khz. The mass, m(x), damping, r(x) and stiffness, k(x) vary exponentially with x in such a way as to keep the quality factor

$$Q = \frac{\sqrt{mk}}{r}$$

constant. This implies the filters are narrowband in the lower frequencies, and become increasingly broadband towards the higher frequencies.

Input signals are sampled at 13.3Khz. The filter bank output is rectified, leakily integrated, compressed and decimated to 1Khz.

3. Inter-filter Feedback and Formant Dynamics

The coupling coefficients $c^i d^j\Theta_j^i$ are quadratic in the breadth, $B^i B^j$ and determine the strength of the inter-filter feedback. The matrix $\delta_j^i + c^i d^j\Theta_j^i$ is invertible:

$$z_1^0(n) = \sum_{j=0}^{M}\Gamma_j\tilde{z}^j(n)$$

$$z_1^i(n) = \left(\alpha^i z_1^0(n) + \sum_{j=0}^{M}\beta_j^i\tilde{z}^i(n)\right) \qquad \text{(eq. 14)}$$

where $i \geq 1$.The matrix β_j^i is a lower-triangular matrix, in which the first column β_0^i is comparable in magnitude to the diagonal elements. The off-diagonal elements in each row decrease to $O(10^{-3})$ relative to the diagonal elements within about 60 sections. Because the matrix multiplication is the most expensive computation in the implementation, we looked at the effects of reducing the inter-filter feedback by truncating the off-diagonal elements to the first 63 in each row. The number 63 is chosen because any additional truncation results in response instability.

Figure 2 shows the effects of truncation. The multiple tones are spaced at intervals of 1/4 octave, with the top tone at 6KHz. Feedback reduction leaves the tones more or less intact; in fact, it boosts them around the 63rd channel. However, feedback reduction also broadens the filter response to frequencies higher than their resonant frequencies, resulting in an undesirable saturation artifact in a range of channels around channel 63. In the case of the up-down chirp, whose rates of frequency change are typical of glides, the reduction in feedback severely attenuates the lower half of the frequency modulations. The rectangles in the cochleagrams for "greasy wash" indicate the areas of severe attenuation of formant movements. In addition, in those portions of both the chirp and the speech signal where high frequencies predominate, the feedback reduction again results in the saturation artifact.

4. Conclusions

This design is currently used as a front-end processor to a speech recognizer[2][3]. We attribute the high phoneme accuracy (87%) we obtained partly to the ability of the filter bank at capturing speech dynamics.

Acknowledgment

This research was supported by Monowave Partners, L.P. and Steven A. Reisler.

References

[1] Tsiang, E., "Multiresolution Elementary Tonotopic Features for Speech Perception", *Proc. International Conference on Neural Networks,* 1997, pp. 575-579.

[2] Tsiang, E., "A Neural Architecture for Computing Acoustic-phonetic Invariants", *Monowave Technical Report*, 1997.

[3] Tsiang, E., "System for recognizing speech", *U.S. Patent* 5377302, 1994.

[4] Allen, J.B. and Neely, S.T., "Micromechanical models of the cochlea", *Physics Today*, Vol. 45, No. 7, 1992, pp. 40-47.

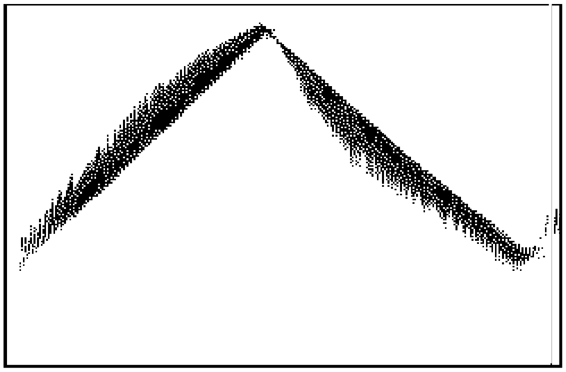
[5] Hudspeth, A.J. and Markin, V.S., "The ear's gears: mechanoelectrical transduction by hair cells", *Physics Today*, vol. 47, no. 2, 1994, pp 22-37.

[6] Monderer, B., *Exploring the space-time structure at the output of a cochlear model*, Columbia University dissertation, 1988.

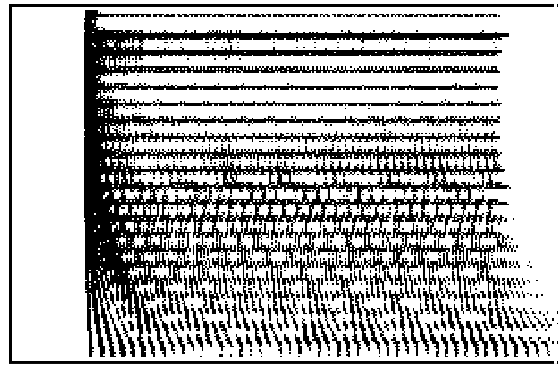
[7] Kates, J.M., "Accurate Tuning Curves in a Cochlear Model", *IEEE Trans. Speech and Audio Processing*, vol. 1, no. 4, 1993, pp. 453-462.

[8] Giguere, C. and Woodland, P.C., "A computational model of the auditory periphery for speech and hearing research", *J. Acoust. Soc. Am.*, vol. 95, no. 1, 1994, pp. 331-342

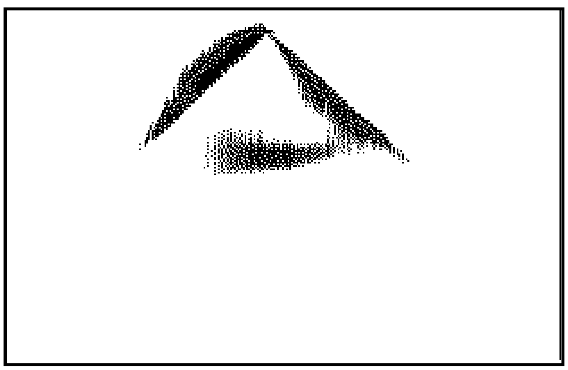
[9] Viergever, M.A., *Mechanics of the Inner Ear: A Mathematical Approach*, Delft University Press, 1980.



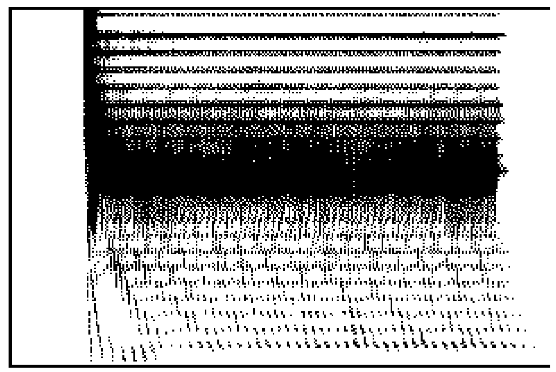
Chirp with full feedback



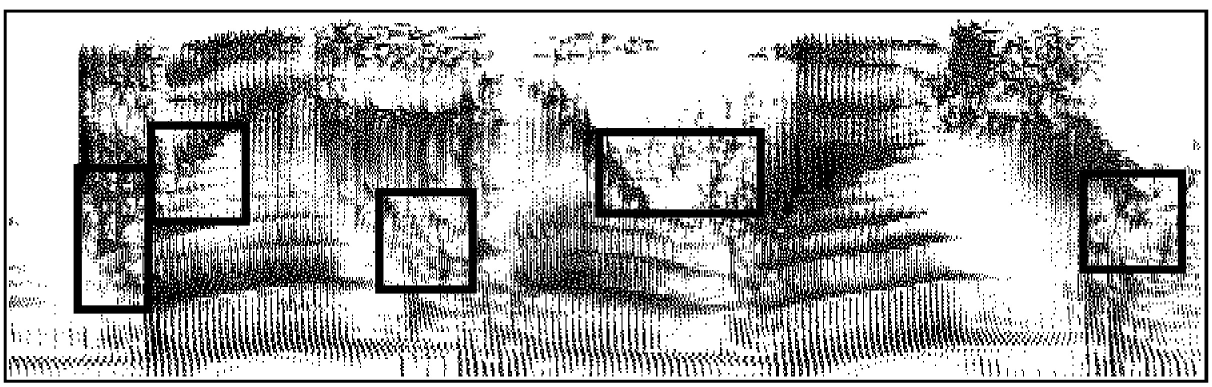
Multiple tones with full feedback



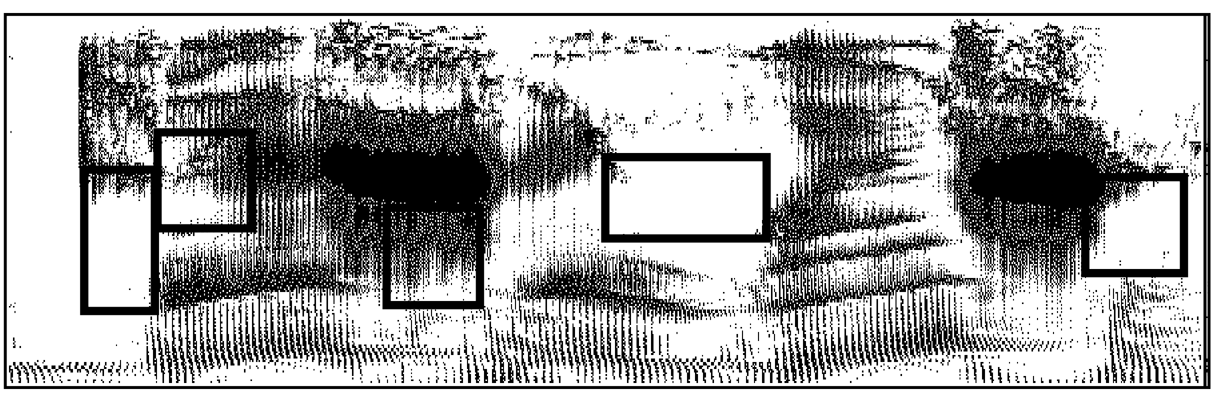
Chirp with reduced feedback



Multiple tones with reduced feedback



"greasy wash" with full feedback



"greasy wash" with reduced feedback

Figure 2