Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments

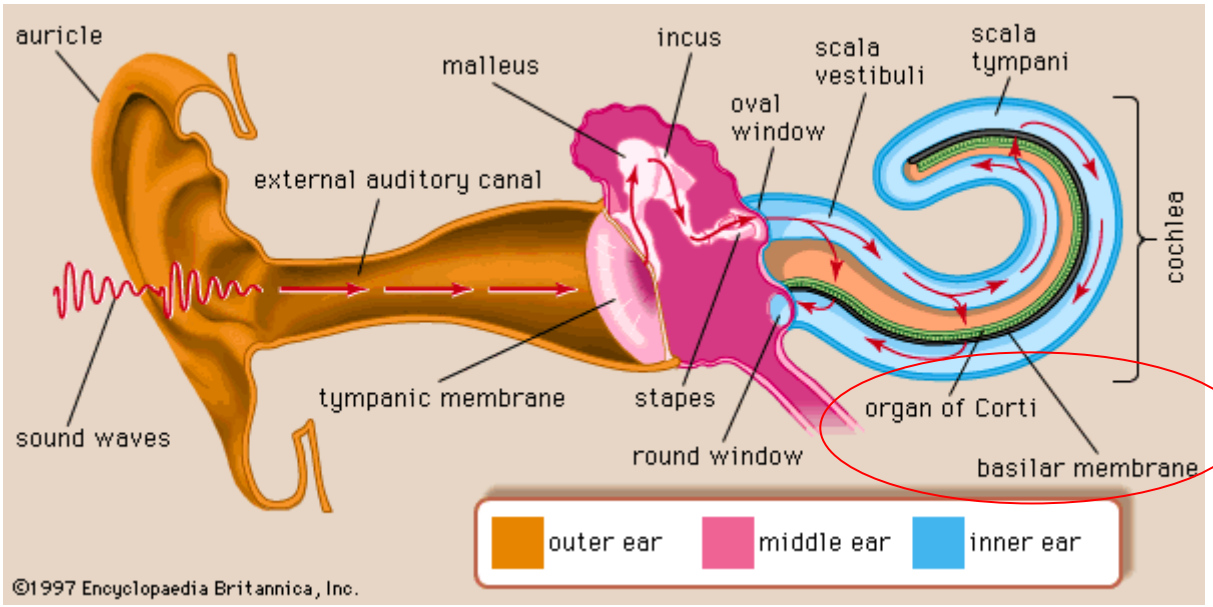
Doh-Suk Kim, Soo-Young Lee, Rhee M. Kil

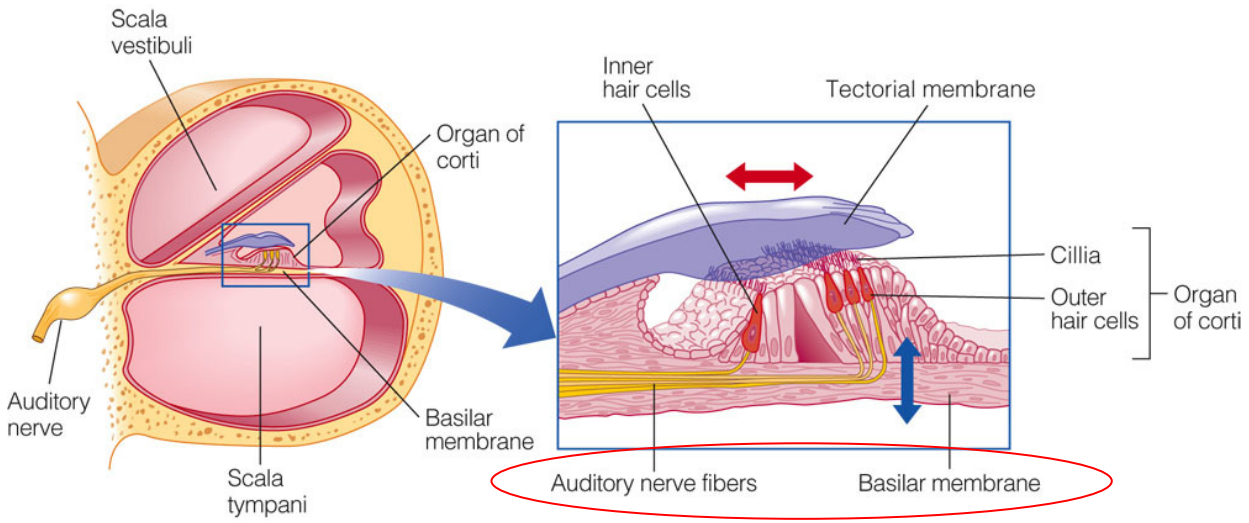
Reporter:邱聖權

Professor:陳嘉平

introduction

- This paper introduces a robust feature extraction method motivated by a mammalian auditory periphery.
- The model consists of two parts
  - cochlear band-pass filters
  - nonlinear operations which obtain frequency information and intensity information


auricle
malleus
incus
scala vestibuli
scala tympani
oval window
external auditory canal
cochlea
tympanic membrane
stapes
organ of Corti
round window
basilar membrane
sound waves
outer ear
middle ear
inner ear
©1997 Encyclopaedia Britannica, Inc.


Scala vestibuli
Organ of corti
Inner hair cells
Tectorial membrane
Cillia
Outer hair cells
Organ of corti
Auditory nerve
Basilar membrane
Scala tympani
Auditory nerve fibers
Basilar membrane
(a)
(b)
© 2007 Thomson Higher Education

The ZCPA model

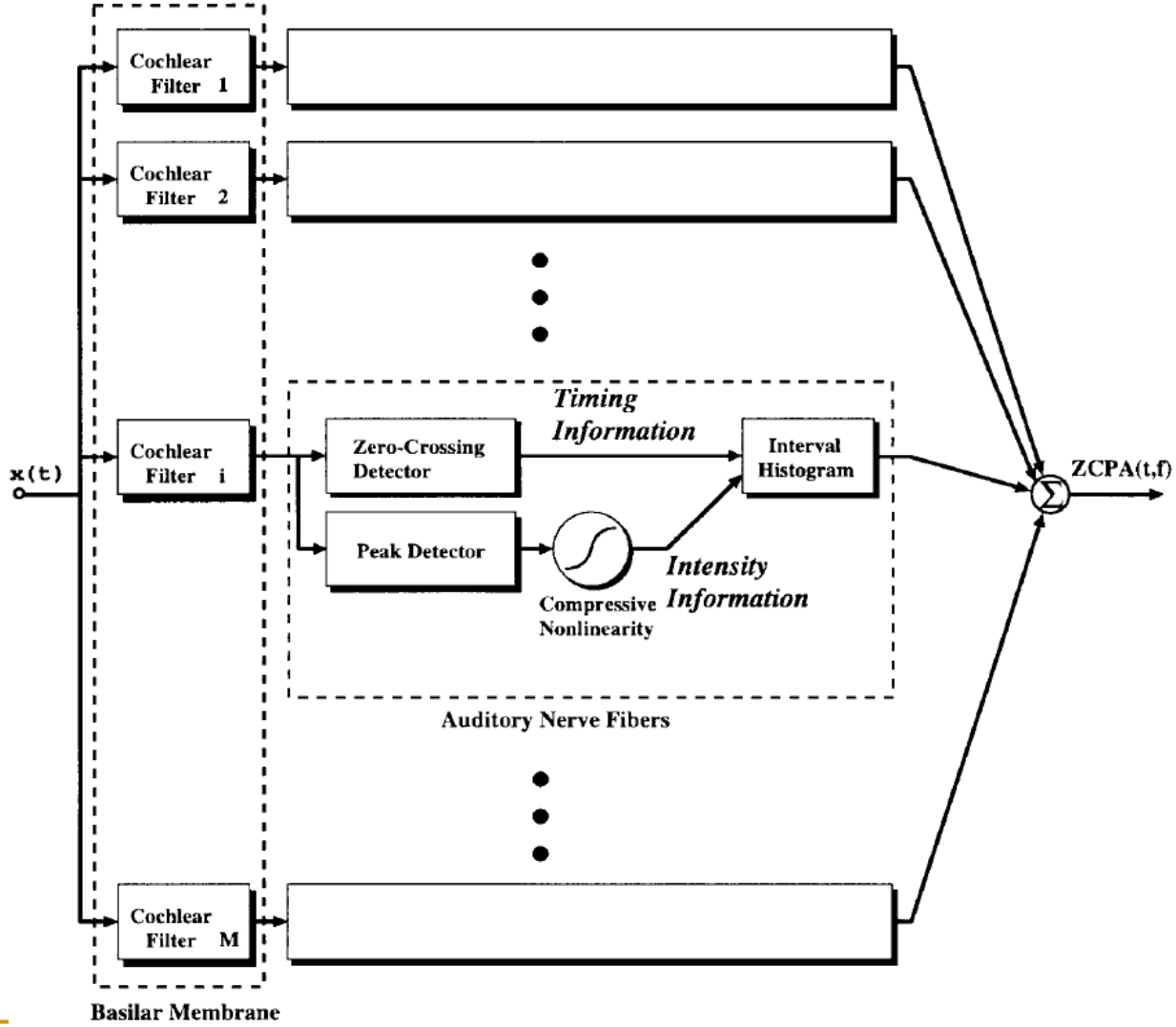


Fig. 1. Block diagram of the zero-crossings with peak-amplitudes (ZCPA) model.

Cochlear filters

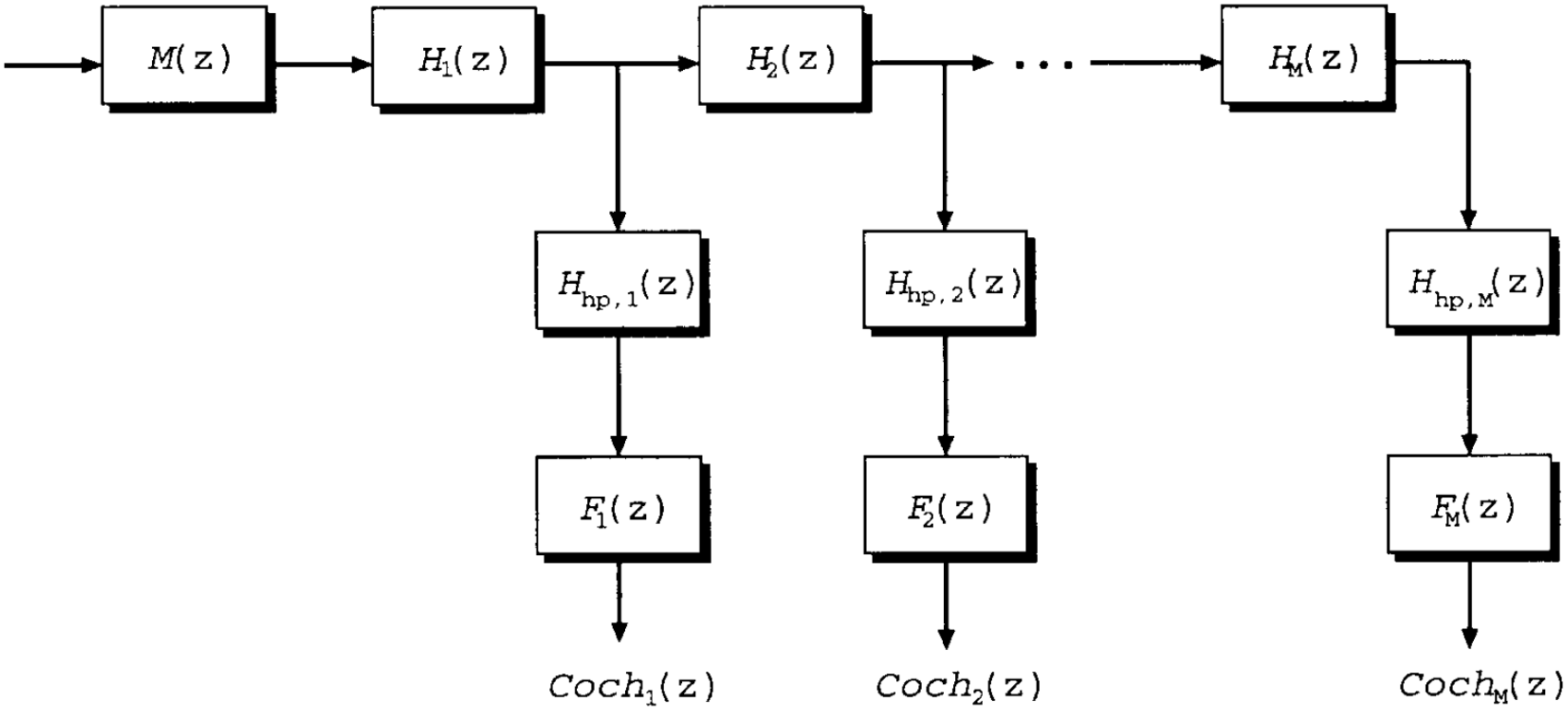


Fig. 2. Block diagram of Kates' cochlear filters.

Middle ear

- The role of the middle ear is known as impedance matching between the outer and inner ear.
- *M(z)* is a second-order high-pass filter having a resonance frequency of 350 *Hz* and a *Q* of 0.7, and simulates the behavior of the human middle ear.

Traveling waves

- $H_k(z)$ is a second-order low-pass filter which provides
  - a gain for frequencies near the resonance frequency of the filter
  - attenuation for frequencies above the resonance frequency
  - and unity gain for frequencies below the resonance frequency

Traveling waves(2)

$$H_k(s) = \frac{1 + (\mu + 1/Q_k)(s/\omega k) + b(\mu/Q_k)(s/\omega_k)^2}{(1 + \mu s/\omega_k)(1 + \frac{1}{Q_k}\frac{s}{\omega_k} + \frac{s^2}{\omega_k^2})}$$

$\mu = 0.5, b = 0.5$

Q_k ranges from 0.28 at 100 Hz to 0.45 at 10 kHz

$0 < \omega_k < \pi$, and $\omega_{k-1} > \omega_k$

The resonance frequency decreases as the index k increses, and can be describes as $F = A(10^{ax} - 1)$,

$A = 165.4, a = 2.1, x$ is the normalized distance along the basilar membrane.

Velocity transformation

- $H_{hp,i}(z)$ is a one-pole high-pass filter that models the pressure-to-velocity transformation

Second filter

- Tuning-curve measurements show a notch in the frequency response approximately one octave below the center frequency.
- Phase measurements show a π phase shift occurring at the same frequency as the notch.
- $F_i(z)$ is a notch filter by which the total response shows two resonance frequencies, which coincides with biological observations

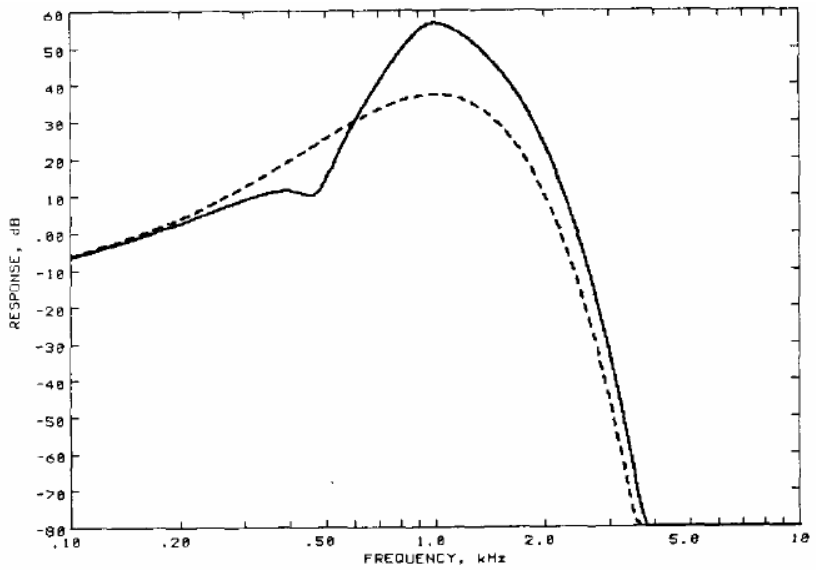
Second filter(2)

$$F(s) = \frac{\omega_0^2 + (\omega_0 / Q_0)s + s^2}{\omega_P^2 + (\omega_P / Q_P)s + s^2}$$

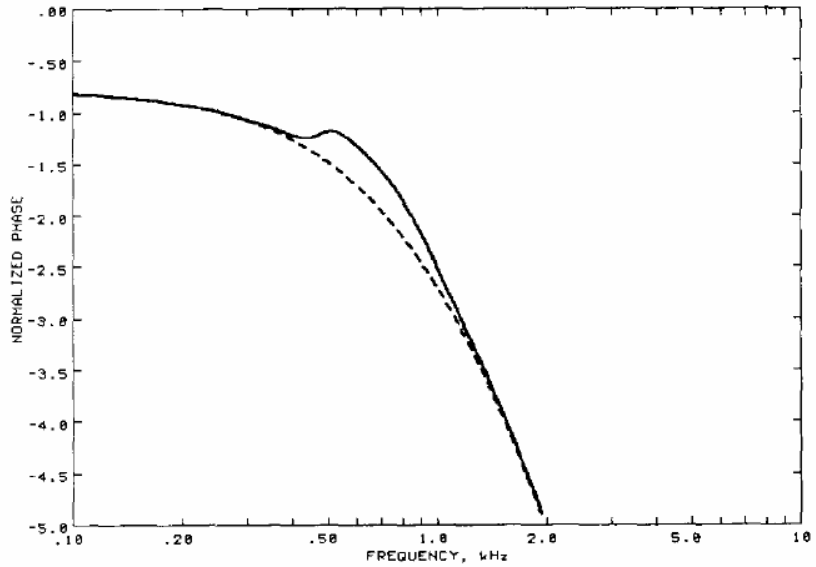
ω_0 and Q_0 are thr resonance frequency and Q for the zero,

ω_P and Q_P are thr resonance frequency and Q for the pole

$$Coch_i(z) = M(z)H_{hp,i}(z)F_i(z)\prod_{k=1}^{i} H_k(z)$$



(a)



(b)

Fig. 3. Basilar-membrane velocity (a) magnitude and (b) phase response at the 1-kHz tap for the traveling-wave filters alone (dashed line) and with the addition of the second filter (solid line). The phase is normalized by a factor of 2π.

Frequency response of 20 cochlear filters

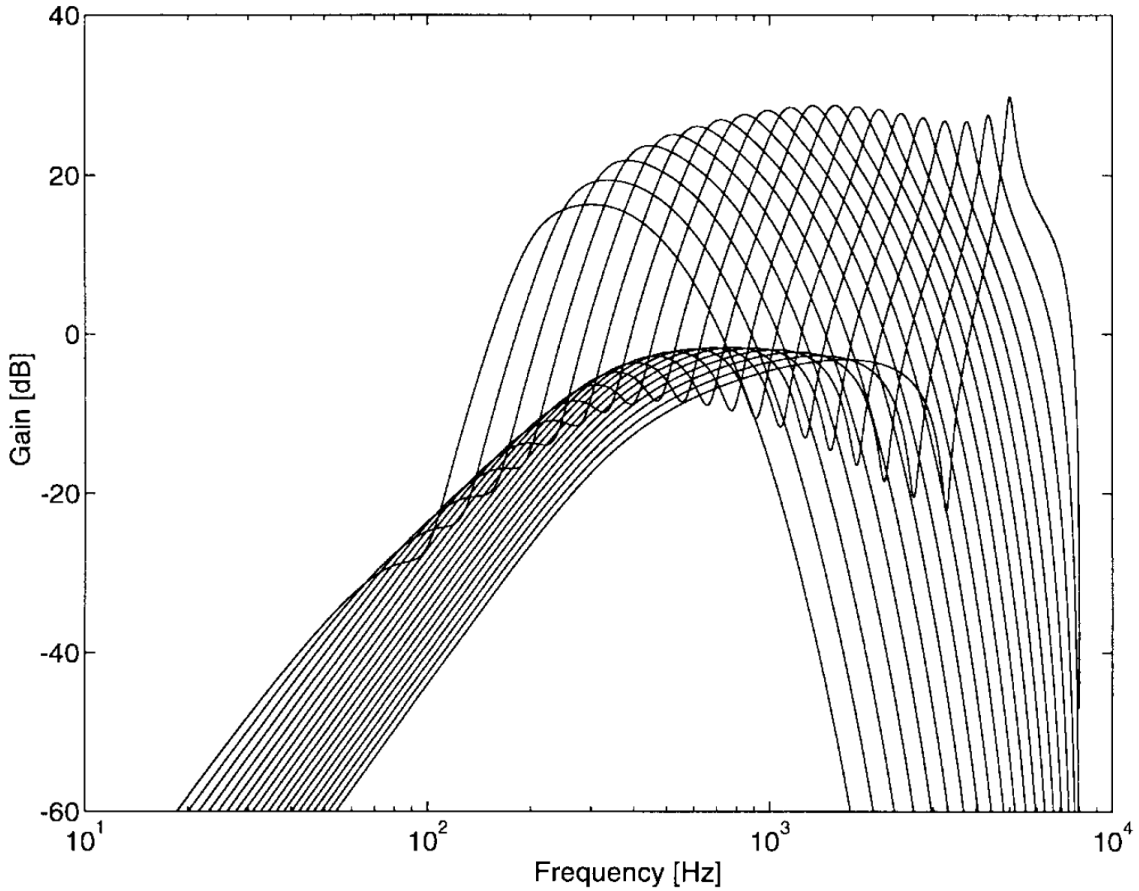
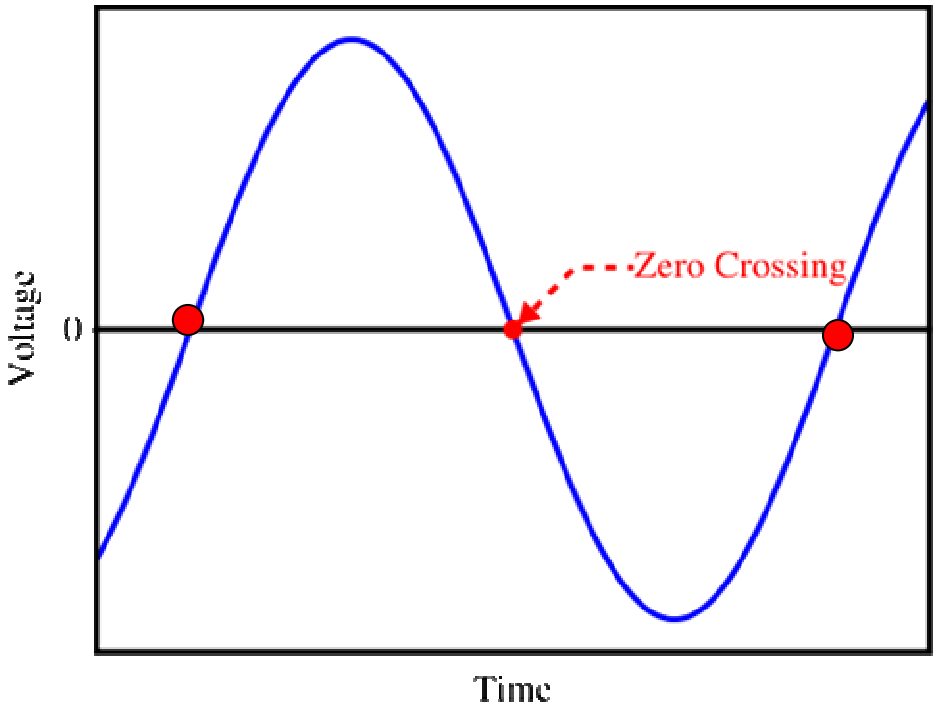


Fig. 3. Frequency response of 20 cochlear filters.

Zero Crossing



ZCPA model of auditory nerve fibers

- A synchronous neural firing is simulated as the upward-going zero-crossing event of the signal at the output of each band-pass filter.
- The inverse of the time interval between adjacent neural firings is collected and represented as a frequency histogram.
- Each peak amplitude between successive zero-crossings is detected, and this peak amplitude is used as a nonlinear weighting factor to a frequency bin to simulate the firing rate.

Frame of filtered signal

- Denoting the output signal of the *k*th band-pass filter by $x_k(n)$, and it's frame at time *m* by $x_k(n;m)$

$$x_k(n;m) = x_k(n)w_k(m-n), \qquad k = 1,\ldots,N_{ch}$$

w_k is the window function, and N_{ch} is the number of cochlear filters.

Intensity information

- The output of the ZCPA at time *m* is

$$y(m,i) = \sum_{k=1}^{N_{ch}} \sum_{l=1}^{Z_k - 1} \delta_{ijl}\, g(P_{kl}), \quad 1 \le i \le N$$

Z_k is the number of upward - going zero - crossings of $x_k(n;m)$,

P_{kl} is peak amplitude between the lth and $(l+1)$th zero - crossings of $x_k(n;m)$,

N is number of frequency bins (linear or bark scale),

δ_{ij} is Kronecker delta,

jl is computed by taking the inverse of the time interval between the lth and $(l+1)$th zero - crossings

ZCPA output

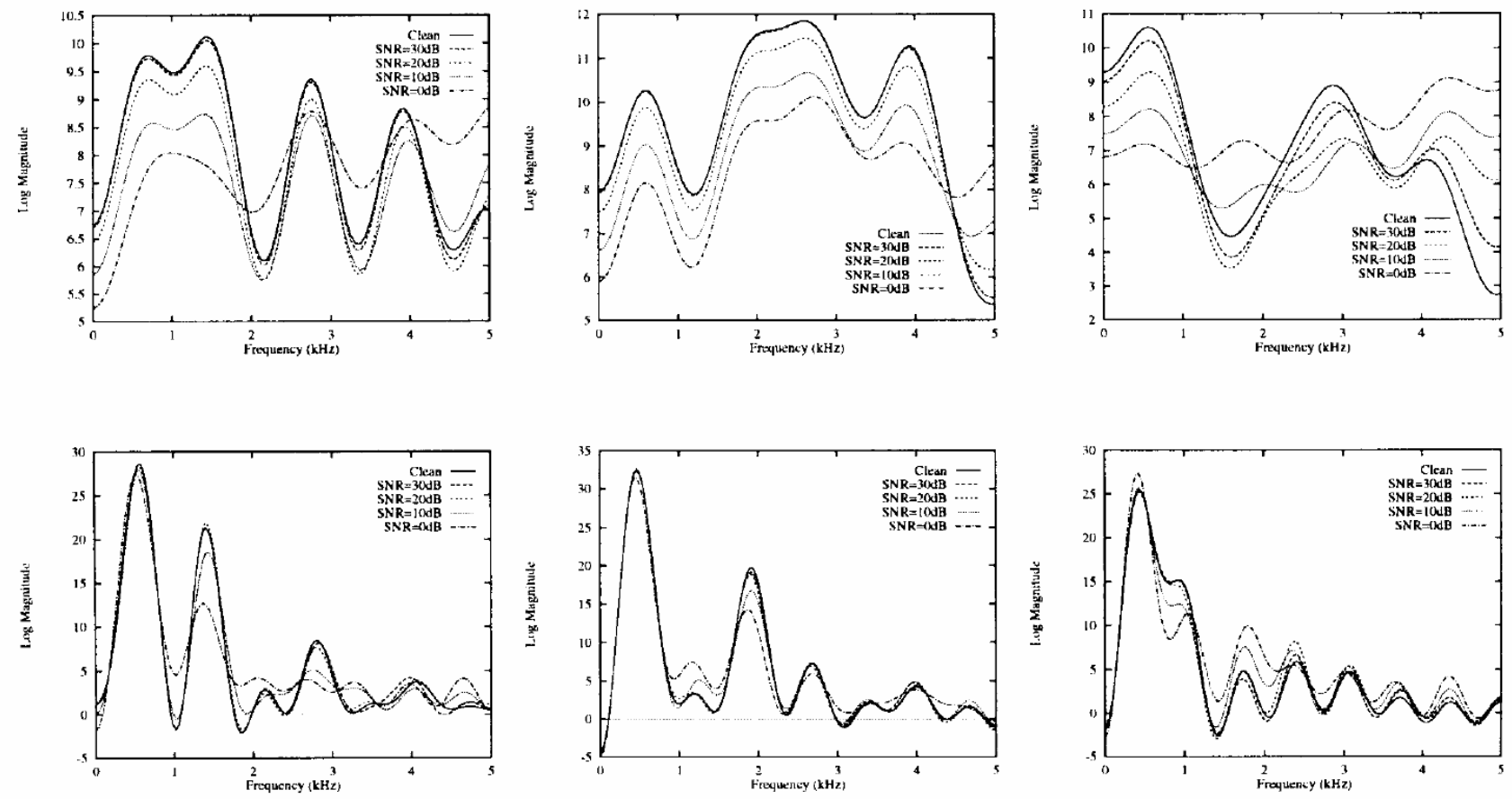
- The value of the frequency histogram at the frequency bin, jl , is increased by $g(P_{kl})$.

$$g(x) = \log(1 + x)$$

$g()$ is a monotonic function simulating the stimulus intensity.

- The histograms across all channels are combined to obtain the output of the ZCPA.
- Temporal frequency and intensity information of one period of the signal is measured, and then an accumulation of the temporal information is carried out to obtain the final output.

Comparison with LPC



Choice of cochlear filters

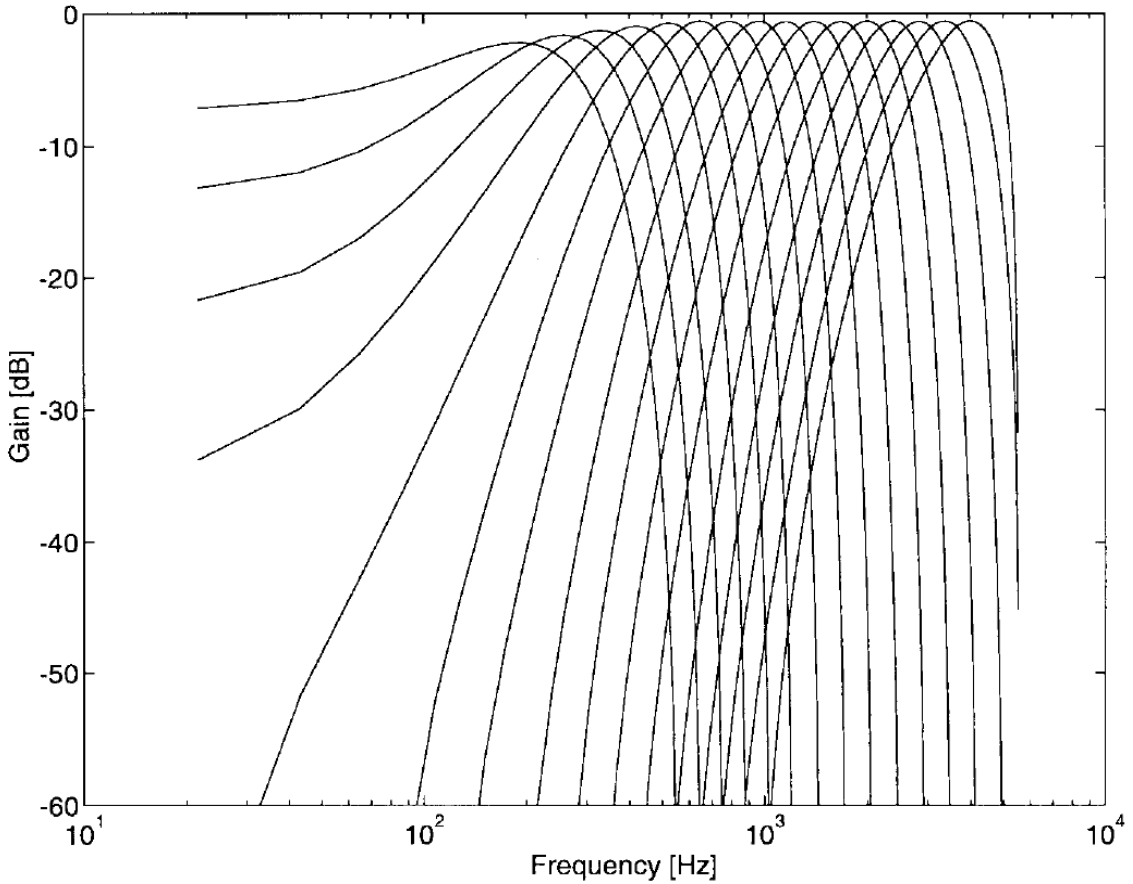


Fig. 7. Frequency response of cochlear filterbank implemented with FIR filters.

FIR filter is better

TABLE V
COMPARISON OF RECOGNITION RATE (%) OF THE ZCPA
OBTAINED USING TW FILTERS AND FIR FILTERS

Noise		WGN		FAC		MOP		CAR	
Filterbank		TW	FIR	TW	FIR	TW	FIR	TW	FIR
SNR (dB)	Clean	88.2	90.8	88.2	90.8	88.2	90.8	88.2	90.8
	25	85.8	88.1	87.5	89.1	86.4	87.5	87.8	90.0
	20	76.7	80.9	82.4	84.9	79.3	80.6	87.9	90.1
	15	63.9	69.2	66.8	73.5	64.4	65.3	88.5	90.3
	10	43.4	54.7	45.8	57.6	44.7	48.2	87.9	90.7
	5	24.2	37.7	27.0	36.9	20.9	26.0	87.5	90.3
	0	–	–	12.7	18.3	8.9	10.4	84.8	88.1
	-5	–	–	–	–	–	–	76.9	81.1
	-10	–	–	–	–	–	–	63.7	64.8

Comparison with several front-ends

TABLE VI
COMPARISON OF RECOGNITION RATES (%) OF FRONT-ENDS IN VARIOUS TYPES OF NOISY ENVIRONMENTS. (a) WHITE GAUSSIAN NOISE. (b) FACTORY NOISE. (c) MILITARY OPERATIONS ROOM NOISE. (d) CAR NOISE.

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	56.5	74.5	82.5	86.3	90.0	90.0	74.4	92.1	94.1	96.0	97.0	96.9
20	22.3	50.6	74.6	65.4	83.0	85.3	38.5	74.0	90.0	85.9	93.5	94.6
15	6.0	18.9	51.7	37.5	68.8	70.7	12.0	38.3	72.7	55.5	84.3	87.0
10	2.9	6.3	25.6	16.8	50.7	54.3	4.2	12.1	43.2	25.7	66.3	72.7
5	3.2	2.9	7.3	6.0	30.1	32.7	2.6	4.9	16.9	7.7	45.3	50.5

(a)

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	74.4	83.2	83.5	90.8	90.0	90.3	91.2	95.9	94.3	97.1	96.6	97.1
20	54.1	70.9	75.0	83.9	85.0	85.4	79.0	90.5	91.3	95.0	94.4	95.1
15	26.8	41.4	57.7	64.2	73.9	76.0	52.6	67.3	77.8	81.6	86.7	90.3
10	11.8	16.9	31.3	41.0	55.2	58.3	20.7	33.7	50.6	52.4	70.0	75.4
5	4.1	6.3	11.6	19.5	32.3	39.3	7.9	10.5	22.2	25.7	46.3	52.3

(b)

Comparison with several front-ends

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
25	74.6	84.4	83.6	91.7	88.8	89.0	91.5	96.1	95.0	96.9	96.2	96.7
20	56.3	72.1	78.0	85.3	78.0	81.9	81.4	89.2	91.3	94.5	92.2	94.0
15	31.2	47.8	58.1	66.9	60.2	66.9	53.5	70.9	76.9	82.4	79.8	85.8
10	11.4	22.0	33.1	43.7	43.8	44.8	23.8	39.4	47.6	55.3	61.8	68.0
5	4.8	6.6	15.5	23.3	19.0	20.3	7.7	16.0	23.1	29.0	30.6	37.3

(c)

SNR (dB)	Static Features						Static and Dynamic Features					
	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC	LPCC	MFCC	SBCOR	PLP	EIHC	ZCPAC
Clean	83.9	89.5	85.6	92.7	91.1	91.1	94.4	97.5	96.4	98.2	97.4	97.6
15	77.8	86.3	84.3	92.9	91.1	90.2	93.0	96.6	95.2	98.2	96.9	97.7
10	77.0	85.1	84.3	92.3	91.8	90.6	93.6	95.7	95.5	98.1	97.2	97.6
5	75.2	85.1	83.5	90.7	90.0	90.3	93.2	95.5	95.1	97.3	97.5	97.2
0	72.4	81.0	80.2	85.1	86.1	89.1	91.5	94.6	94.5	96.3	96.1	96.6
-5	58.7	72.8	73.7	69.4	78.8	81.2	84.8	90.7	89.6	92.6	93.5	93.6
-10	39.5	54.1	57.6	37.2	59.3	64.7	68.2	78.2	75.8	72.9	84.2	86.9