The School of Engineering

THE UNIVERSITY OF AUCKLAND
NEW ZEALAND

Auditory Based Feature Vectors for Speech Recognition Systems

Dr. Waleed H. Abdulla

Electrical & Computer Engineering Department

The University of Auckland, New Zealand

[w.abdulla@auckland.ac.nz]

Outlines

- Introduction
- ASR Systems and Signal Modelling
- The Human Ears
- Equivalent rectangular band (ERB)
- The Gammatone Filterbank (GTF)
- Speech Signal Analysis based on GTF
- Classification Evaluation
- Conclusions

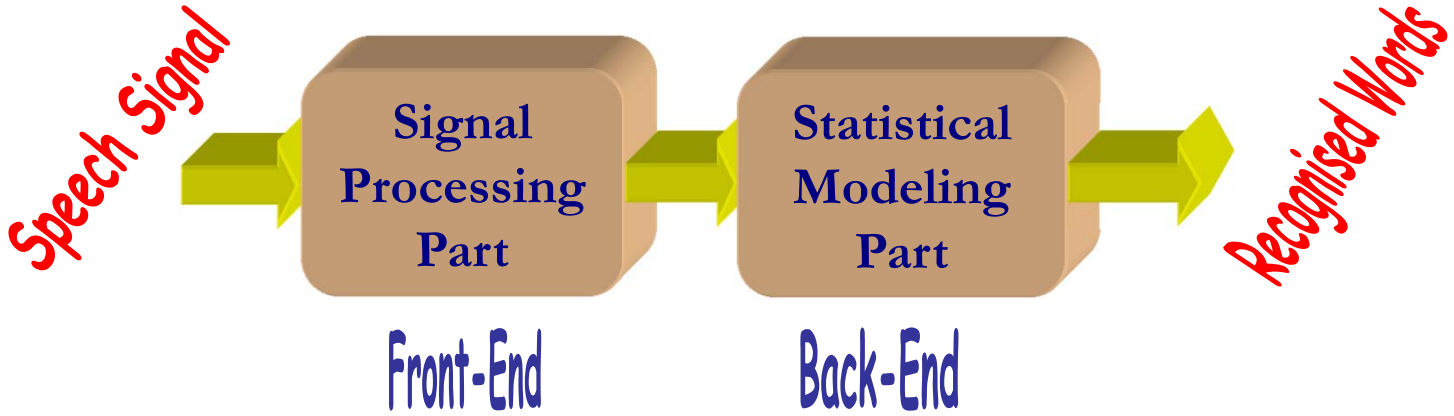
Introduction

- Automatic speech recognition (ASR) is the process of converting an incoming acoustic signal to its corresponding stream of words.

- ASR systems can be:
  - Speaker Dependent **OR** Speaker Independent
  - Isolated Words **OR** Continuous
  - Limited vocabulary **OR** Large vocabulary
  - Resticted Domain **OR** Unrestricted Domain

Introduction

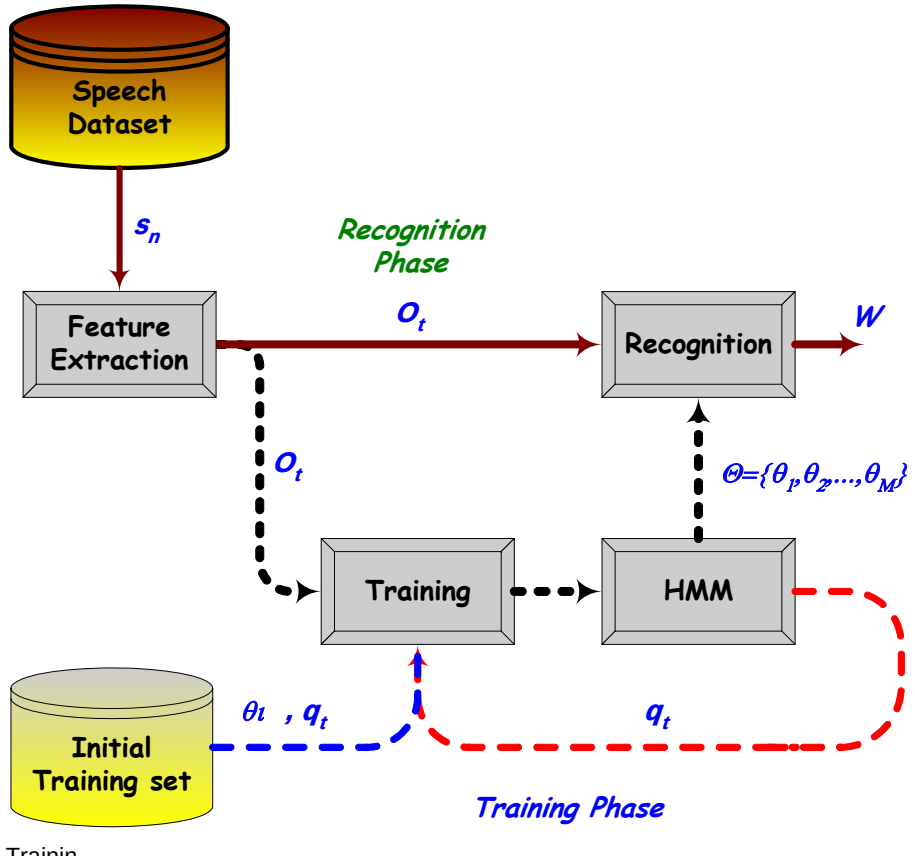
- The general paradigm of speech recognition systems comprises two main parts: front-end and back-end



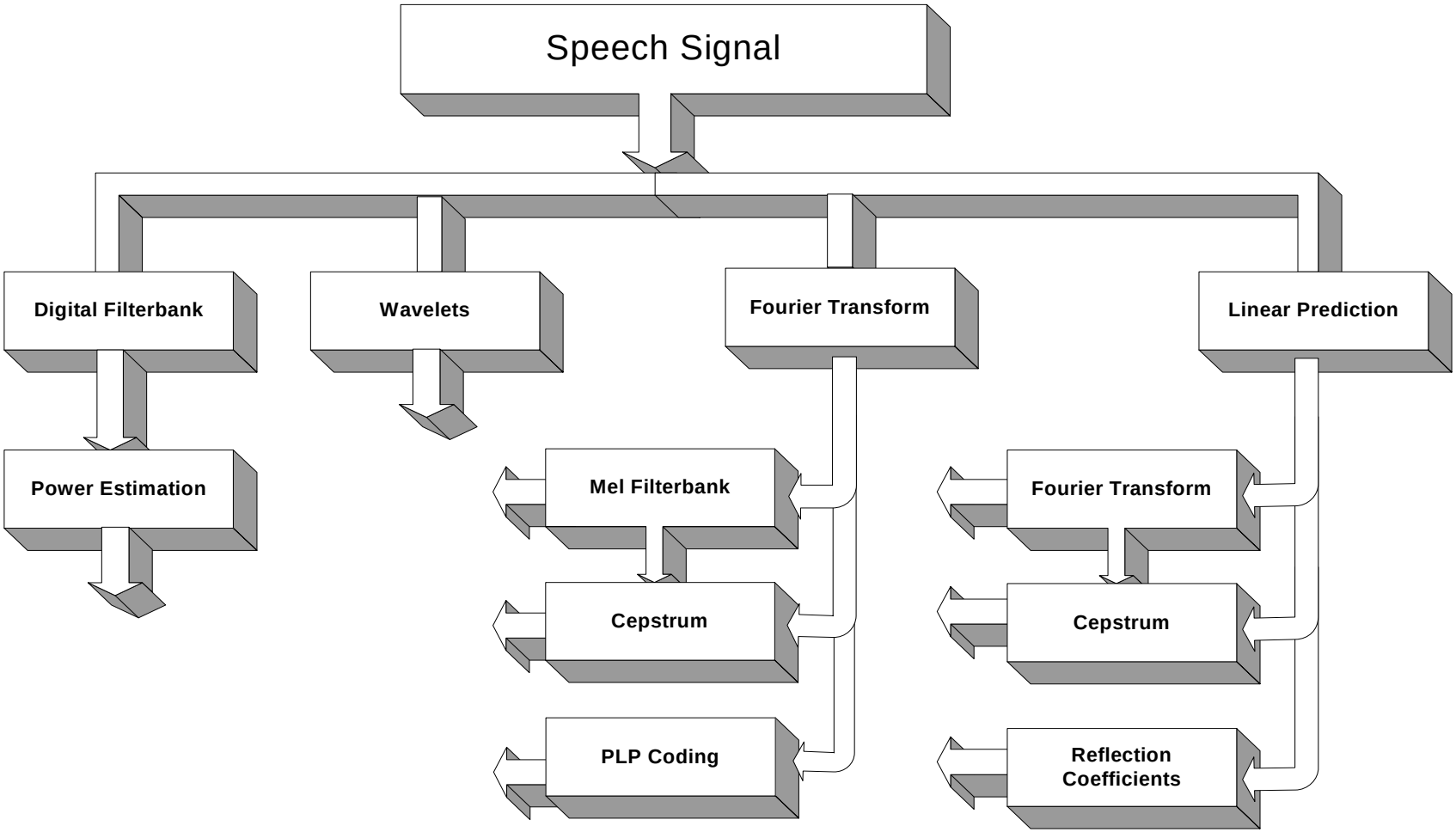
Block diagram of the ASR systems

ASR systems comprises:

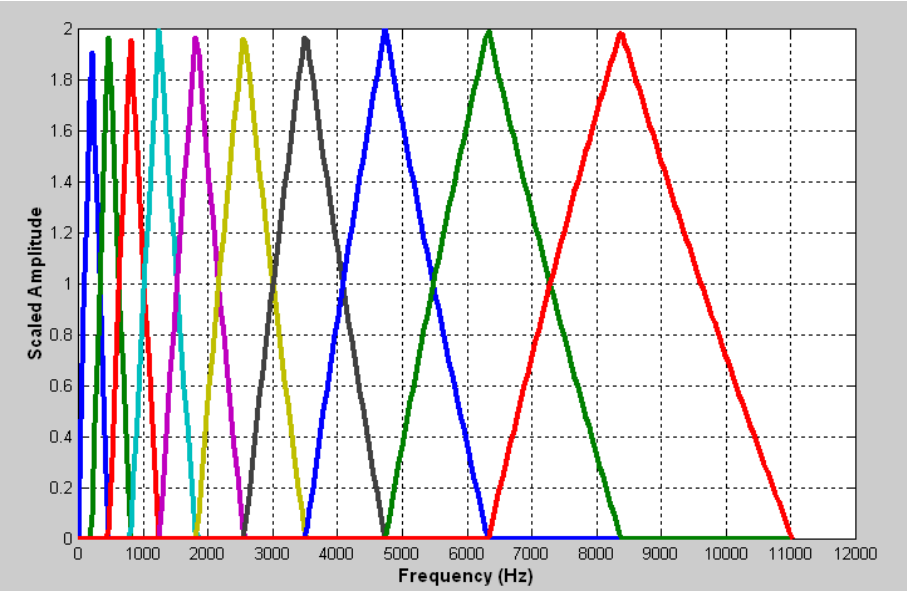
- **Training**
- **Recognition**



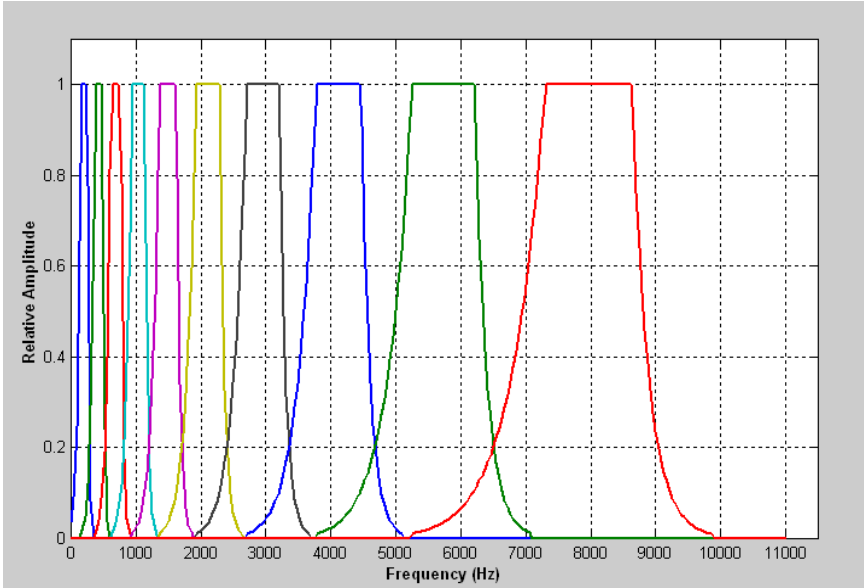
Speech Signal Processing



MFCC & PLP Filterbanks

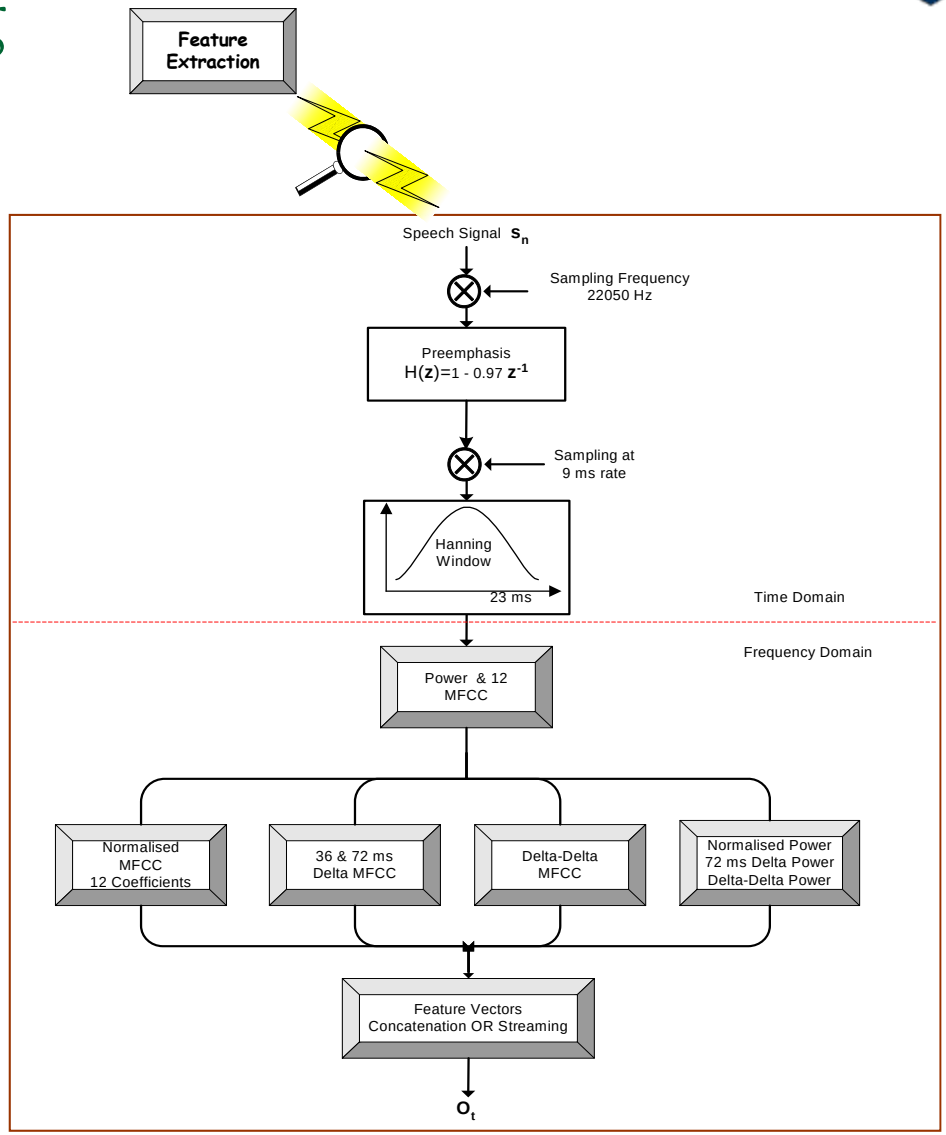


MFCC Filterbank

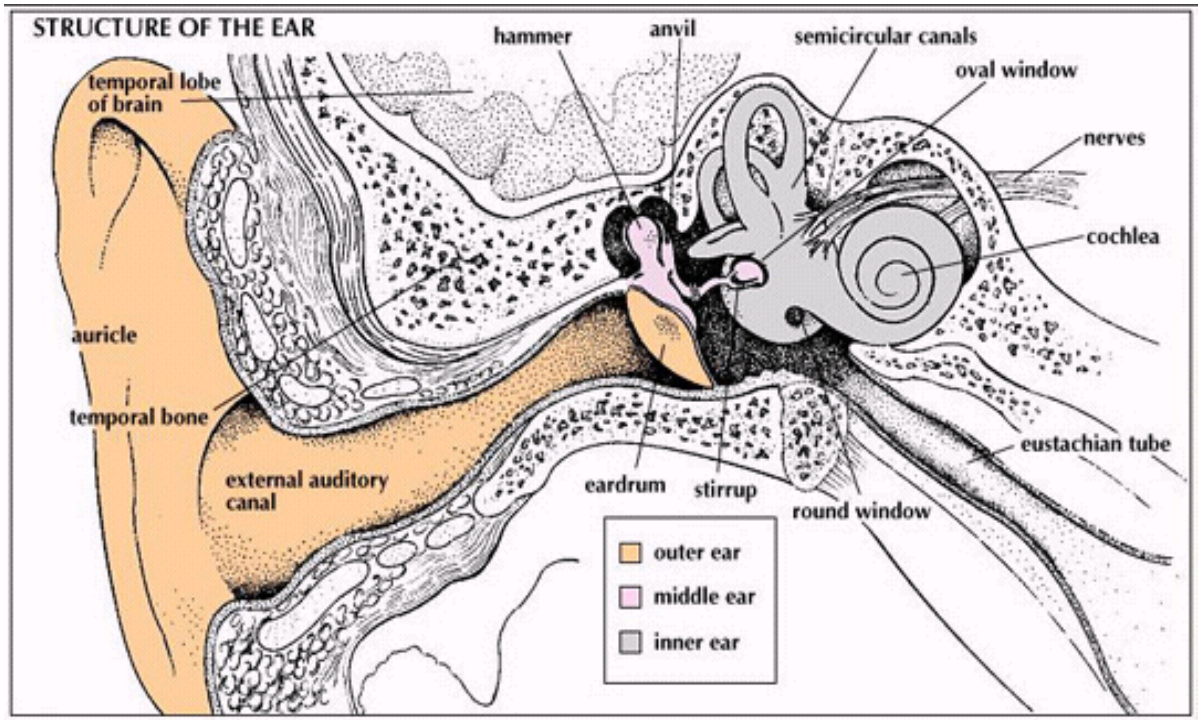


PLP Filterbank

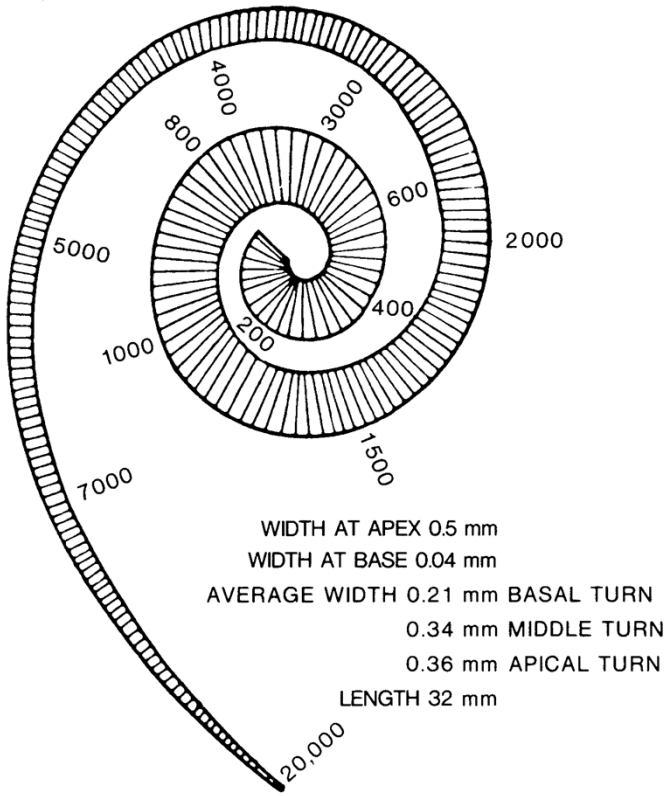
Signal Modelling



The Structure of the Human Ears



Human Basilar Membrane

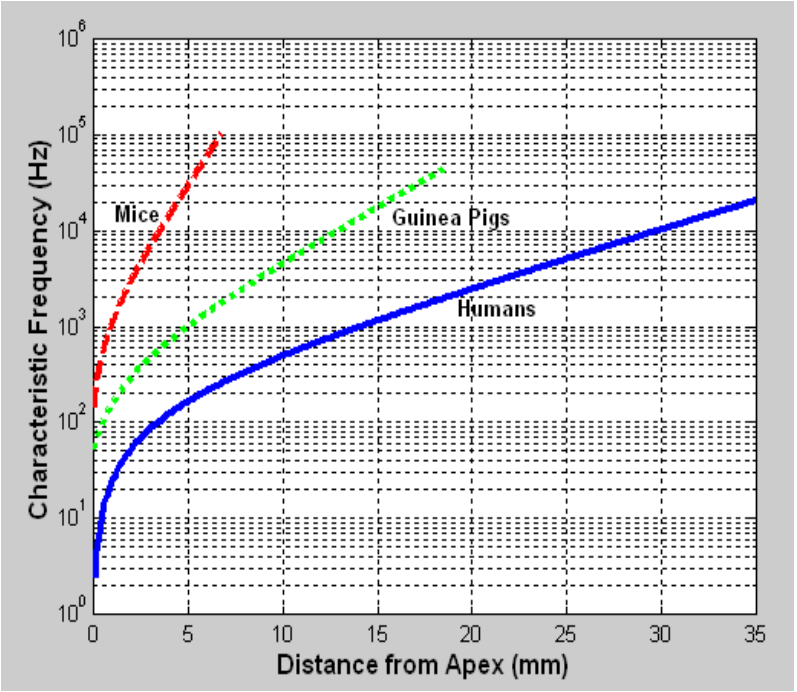


Cochlea characteristic frequency for different species

- In 1961, Don Greenwood developed a mathematical function relating the characteristic frequency, fc, at any location along the length of the cochlea to the distance, x, from the apex (Greenwood 1961).

  The function is:

$$f_c = A(10^{ax/L} - K)$$

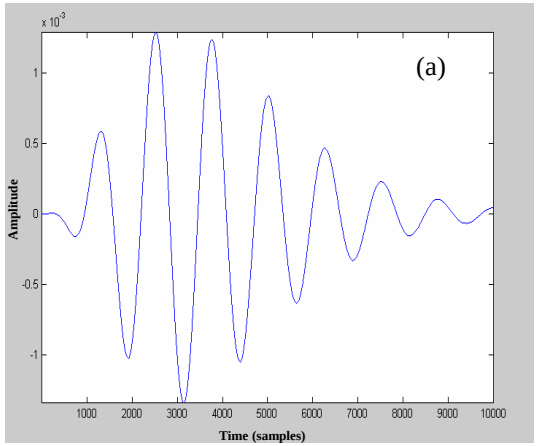


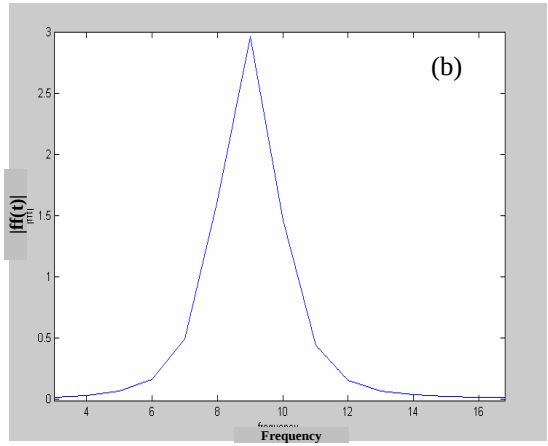
- Where:
- A is a high frequency control constant
- L is the cochlea length in (mm)
- a is the slop factor
- K is the low frequency control constant

Reverse Correlation (Revcor) technique

- Revcore technique states that, for a linear system, it is possible to extract the system parameters by operations on stochastic input and output signals (de-Boer and H. R. de Jongh 1978). The revcore function can be represented mathematically by the equation:

$$g(t) = t^{m} e^{-\alpha t} \cos(\beta t)$$

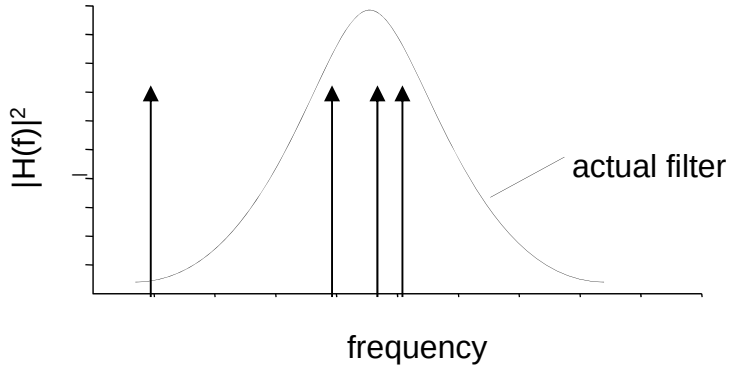




Critical band and equivalent rectangular bandwidth

- Critical band (CB) is the bandwidth of the human auditory filter at different characteristic frequencies positioned along the cochlea path. The bandwidth of the human auditory filter can be measured psycho-acoustically in masking experiments using a sine wave signal (single tone) and a broadband noise as a masker.

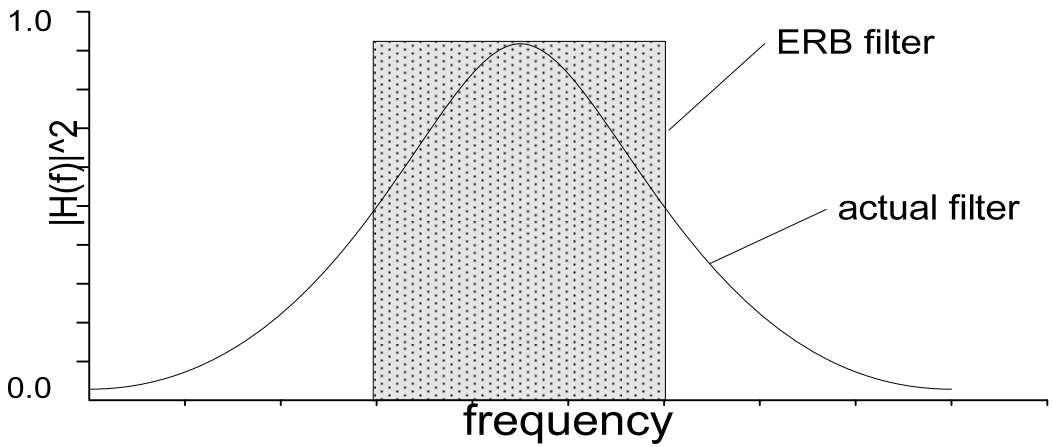
- Experiments show that sounds can be distinguished by ear only if they fall into different critical bands, and they practice the masking process on each other when they fall into the same critical band.



Equivalent rectangular bandwidth (ERB)

- The bandwidth of the actual auditory filter can be related to an equivalent rectangular bandwidth (ERB) filter that has a unit height and a bandwidth ERB. It passes the same power as the real filter does when subjected to a white noise input.

$$\mathrm{ERB} = \int_0^\infty |H(f)|^2 \, df$$



Formulae for the ERB

- Various formulas have been derived for the ERB values:

- Zwicker 1961

$$ERB_1 = 25 + 75(1 + 1.4f_c{}^2)^{0.69}$$

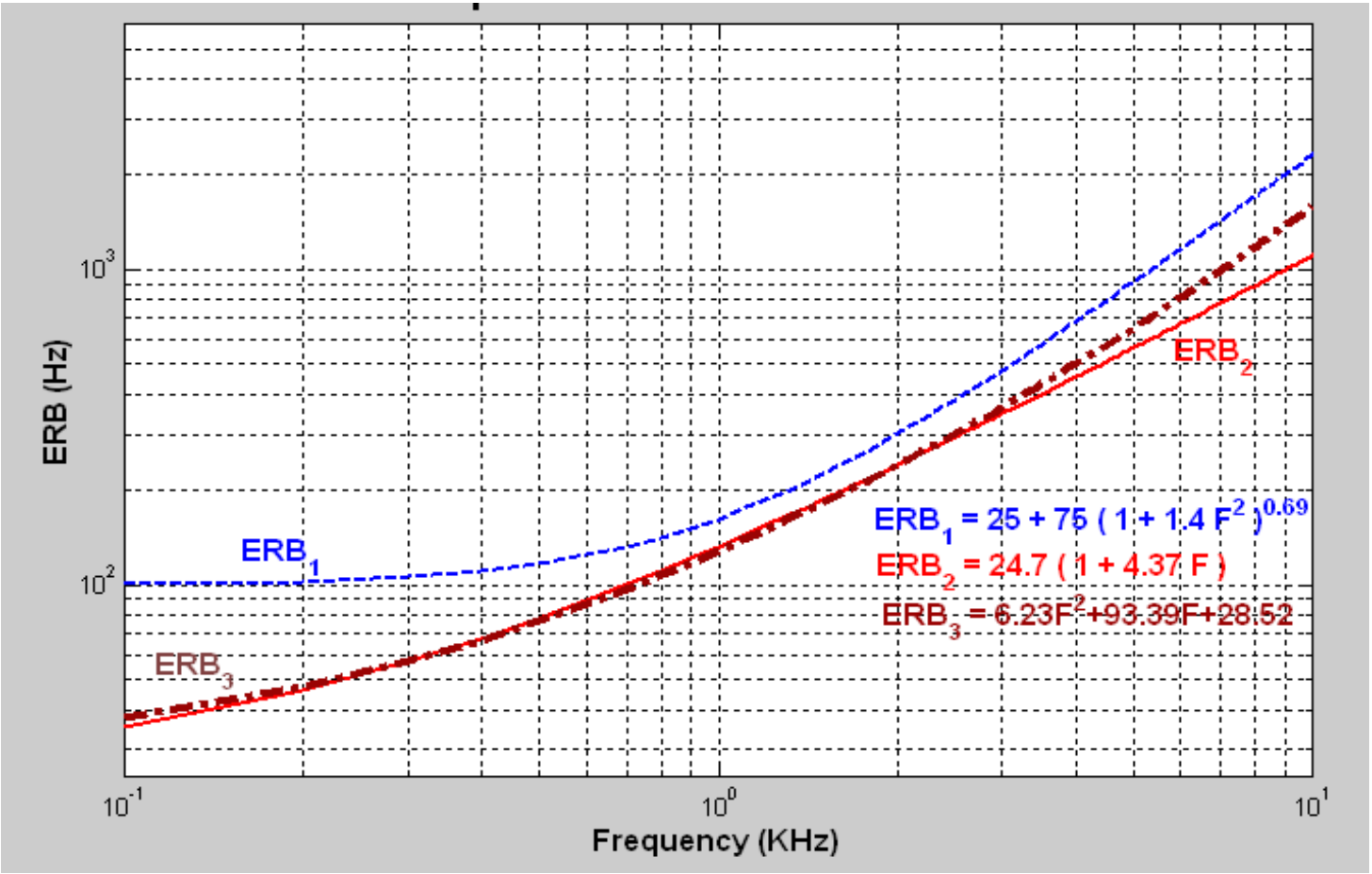
- Glasberg and Moore 1990

$$ERB_2 = 24.7(1 + 4.37f_c)$$

- Moore and Glasberg 1983

$$ERB_3 = 6.23f_c{}^2 + 93.39f_c + 28.52$$

Comparison of Different ERB Functions



General Formula for ERB

$$ERB = \left[\left(\frac{f_c}{Q} \right)^m + BW_{\min}{}^m \right]^{1/m}$$

Where f_c is the centre frequency, Q is the ear quality factor, which is the ratio between the centre frequency and its corresponding filter bandwidth, BW_{min} is the minimum bandwidth allowed, and m is the order.

Lyon recommended the following parameters (Slaney 1988):

Q = 8, BW_{min} = 125 Hz, and m = 2 to produce ERB_{Ly}

$$ERB_{Ly} = \sqrt{\left[\left(\frac{f_c}{8} \right)^2 + 125^2 \right]}$$

General Formula for ERB

Greenwood recommended:
Q = 7.24, BW_{min} = 22.85, m = 1 to form ERB_{Gr}

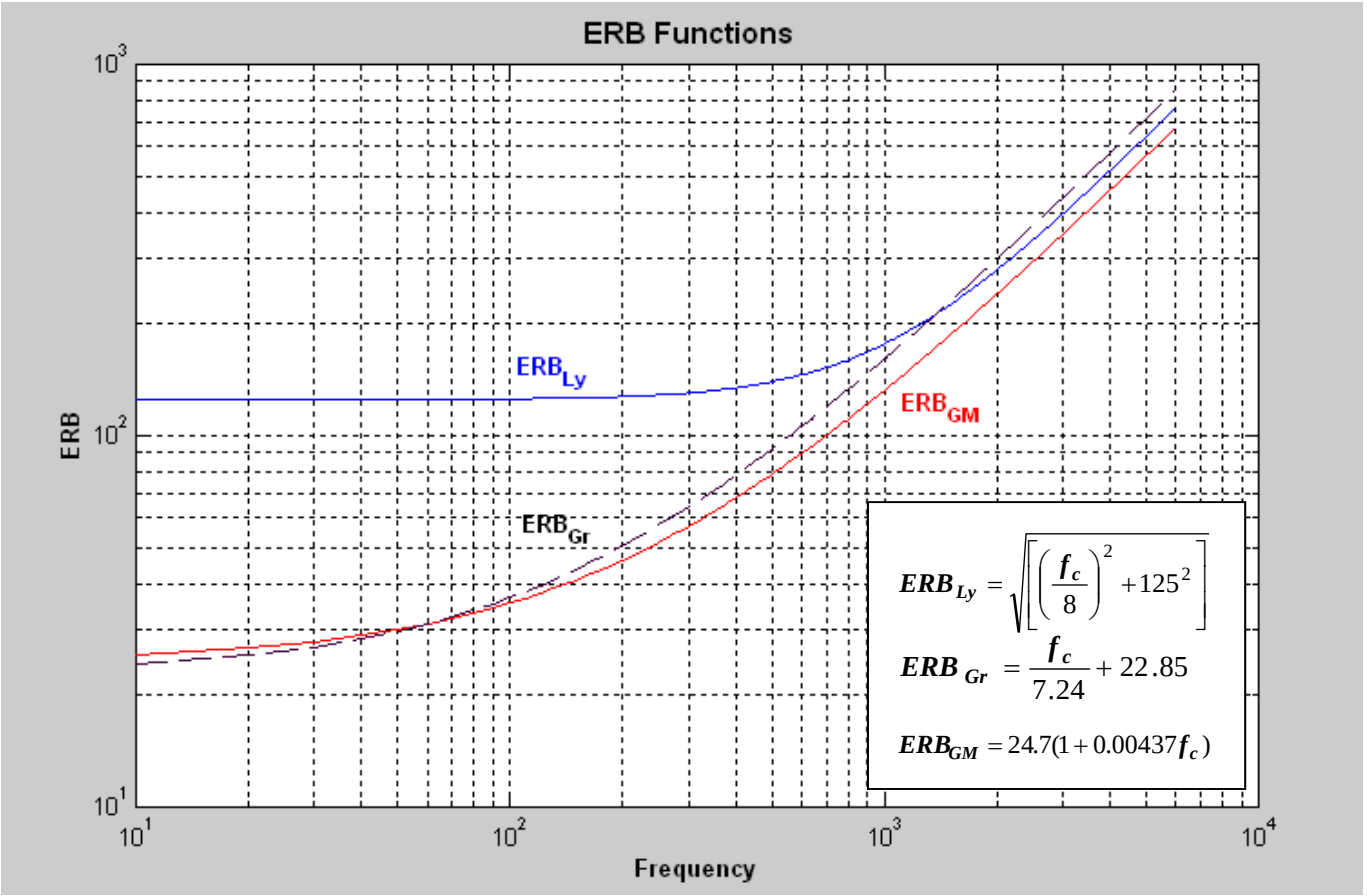
$$ERB_{Gr} = \frac{f_c}{7.24} + 22.85$$

Glassberg and Moore (Glasberg and Moore 1990) recommended
Q = 9.26, BW_{min} = 24.7, m = 1 to get ERB_{GM}

$$ERB_{GM} = \frac{f_c}{9.26} + 24.7 = 24.7(1 + 0.00437\, f_c)$$

ERB_{GM} is used in our approach as it approximates most of the other estimates.

Comparison Between Three ERB Definitions



Critical Band Number

- For a certain frequency, it represents the number of critical bands required until reaching that frequency.

  Let us consider the change in the critical-band number, z, as the frequency changes by df is given by:

$$dz = \frac{\Delta z}{\Delta f} df = \frac{1}{\Delta f / \Delta z} df = \frac{1}{ERB(f)} df$$

$$z = \int_0^{f_c} \frac{1}{ERB(f)} df$$

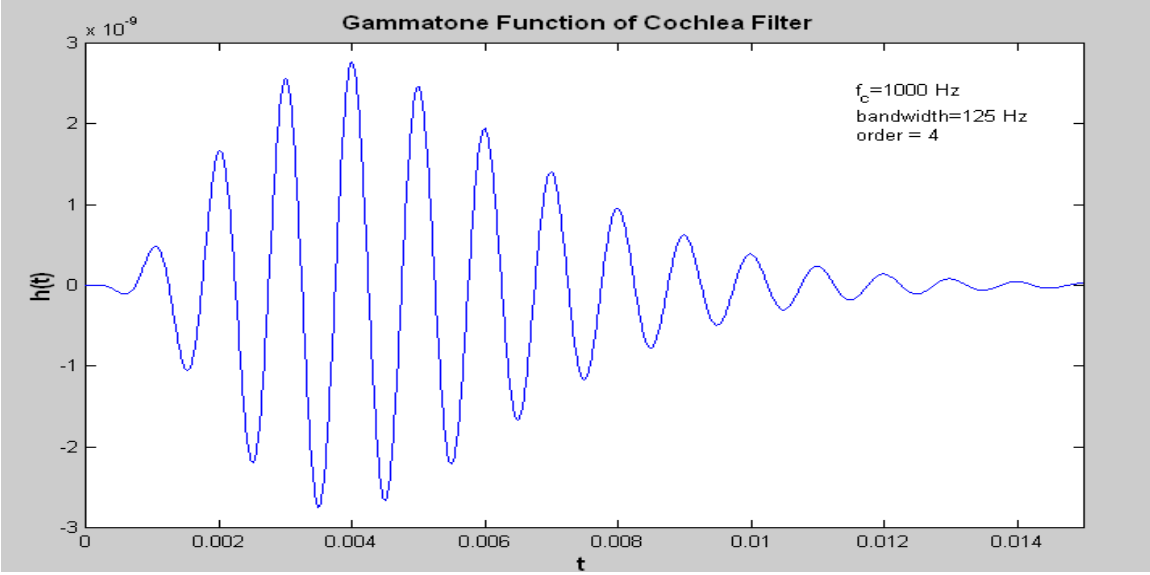
For $ERB(f) = 24.7(1 + 4.37 f)$

$$z = \int_0^{fc} \frac{1}{24.7(1 + 4.37 f)} df = 0.00926 \ln(4.37 f_c + 1)$$

Gammatone Filters

- The impulse response of these filters

$$h(t) = \gamma(n,b)t^{n-1}e^{-bt}\cos(wt+\phi)u(t)$$



ERB of Gammatone Filters

$$ERB = \int_0^{\infty} |H(f)|^2 \, df$$

$$|H(f)| = \frac{(n-1)!}{2} \cdot \frac{1}{\left[b^2 + 4\pi^2 (f - f_c)^2\right]^{n/2}}$$

$$ERB = 2\pi (n-1)! \frac{2^{-2(n-1)}}{\left[(n-1)!\right]^2} b$$

For n = 4, $ERB = 0.9817b$

$$b = 1.0186 \, ERB$$

Number of Channels and the Overlapping Spacing

$$z = \int_{f_L}^{f_H} \frac{1}{ERB(f)} df$$

$$ERB = \frac{f}{Q} + BW_{\min}$$ For m = 1,

$$z = \int_{f_L}^{f_H} \frac{Q}{f + BQ} df = Q \ln \frac{f_H + QB}{f_L + QB}$$ where B = BW_{min}

If the overlapping factor between the contiguous filters is ν then the number of channels, *N*, is related to z, as follows:

$$z = N.\nu$$

$$N = \frac{Q}{\nu}.\ln \frac{f_H + QB}{f_L + QB} = \frac{9.26}{\nu} \ln \frac{f_H + 228.7}{f_L + 228.7}$$ For Q = 9.26 and B = 24.7

Gammatone Filterbank

For a certain band f_L ➔ f_H with ν overlapping between filters

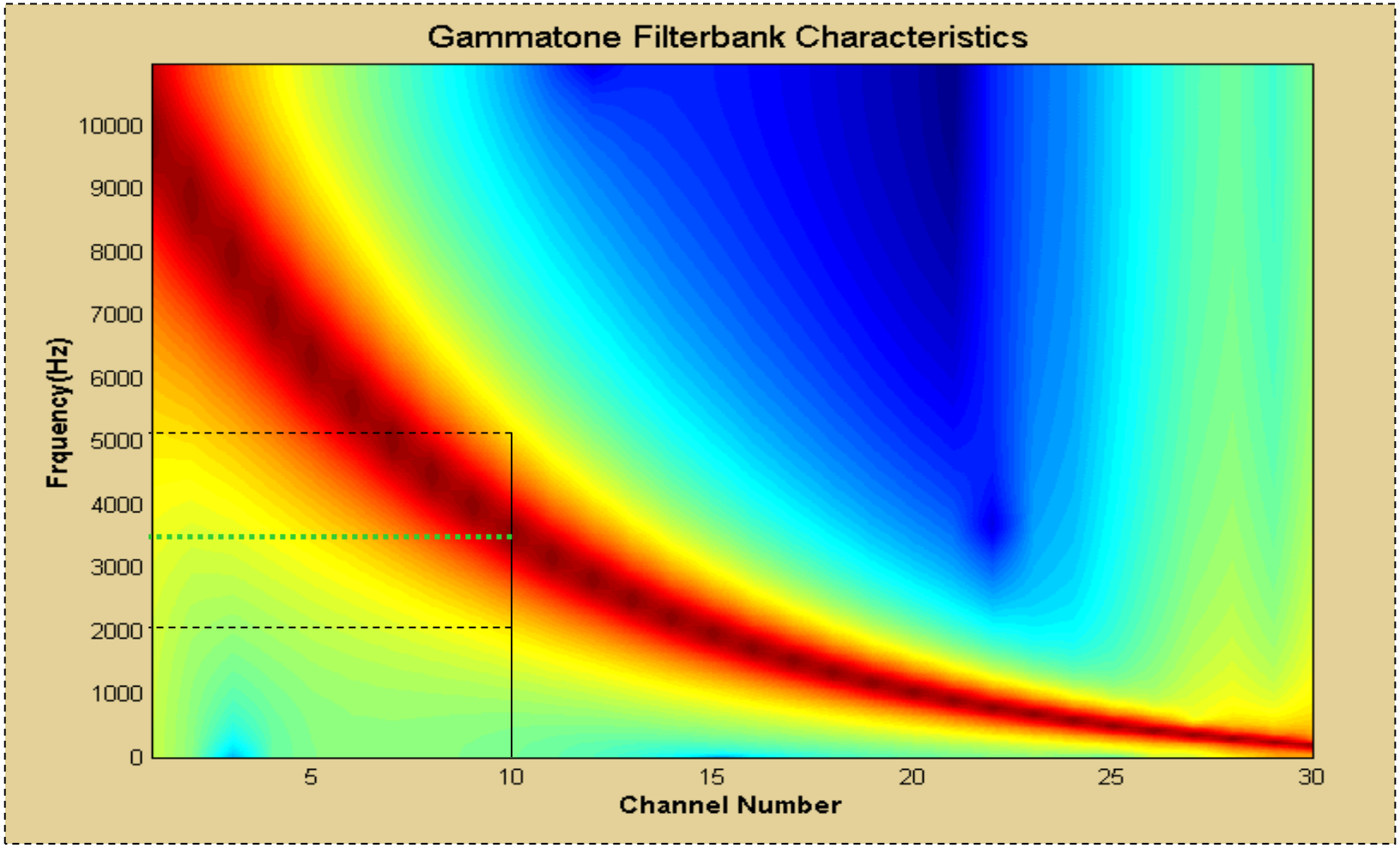
$$N = \frac{9.26}{\nu} \ln \frac{f_H + 228.7}{f_L + 228.7}$$

For $1 \le n \le N$

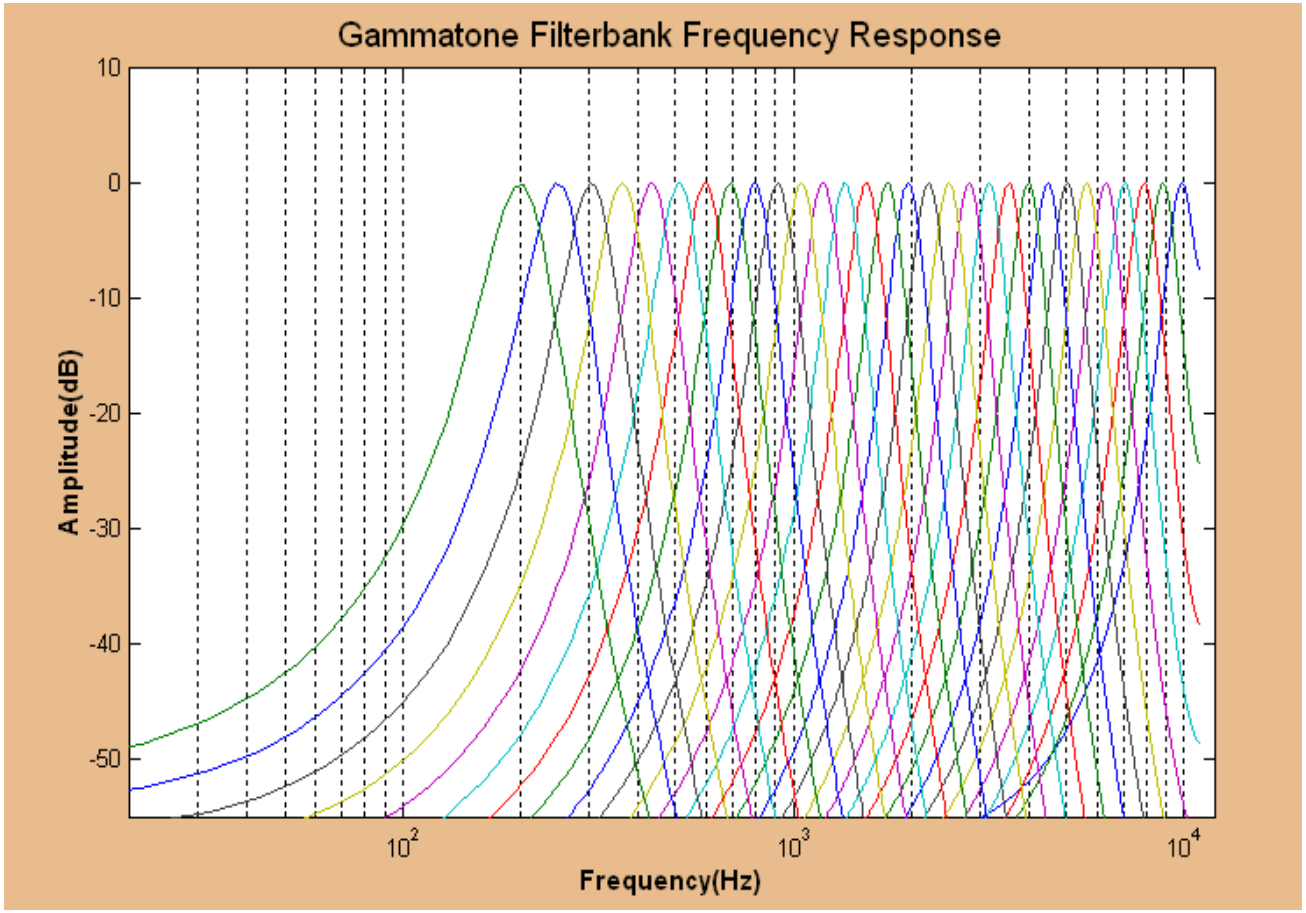
$$f_c(n) = -228.7 + (f_H + 228.7) e^{-\frac{\nu n}{9.26}}$$

$$ERB(n) = 24.7\{1 + 4.37 f_c(n)\}$$

Characteristics of the GTF

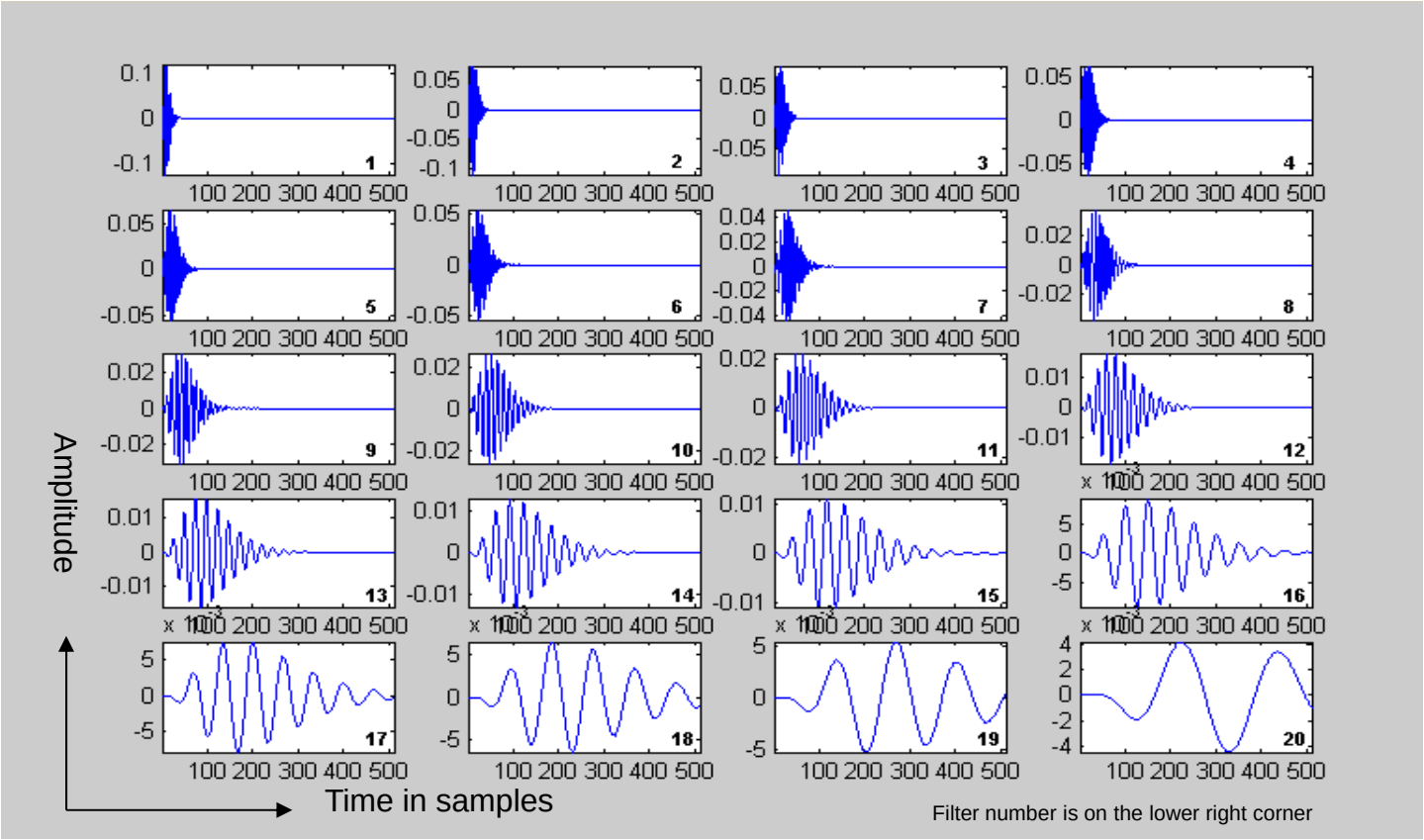


Gammatone Filterbank



Frequency response of a 30-channel filterbank, covering 200-11025 Hz band

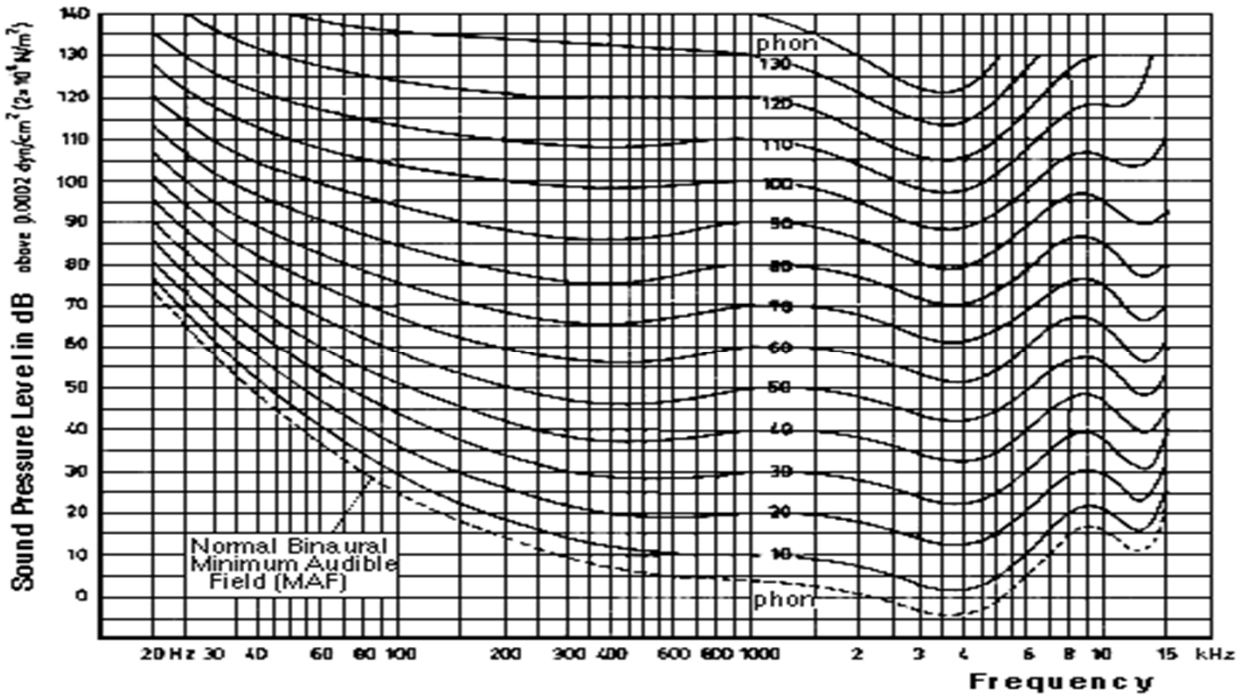
Gammatone Filterbank



Impulse responses of a 20-filters Gammatone filterbank.

Equal Loudness Contours

This graph shows that the ear is not equally sensitive to all frequencies

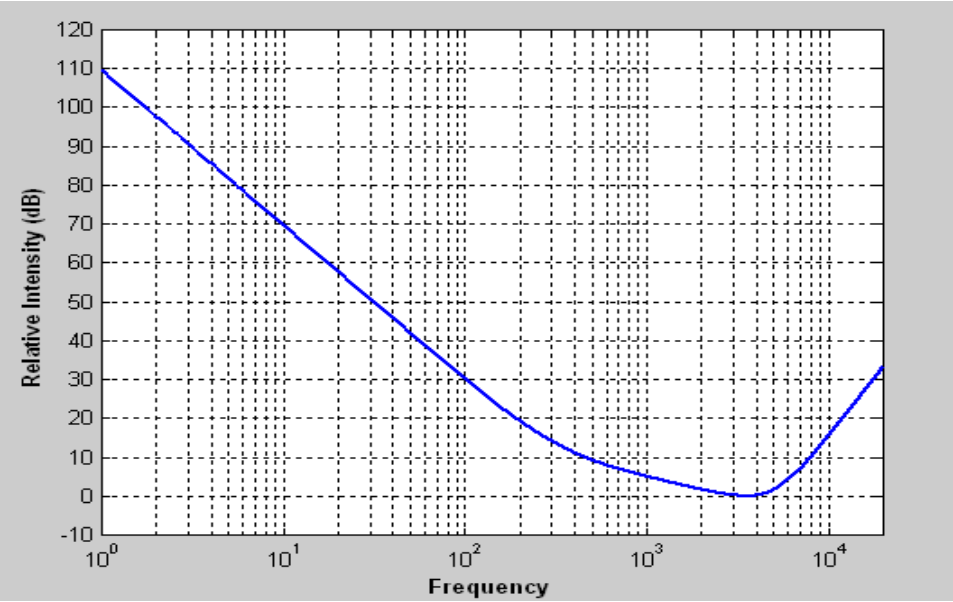


ISO recommendation R226 of equal loudness contours for pure tones and normal threshold of hearing for persons aged 18-25 years.

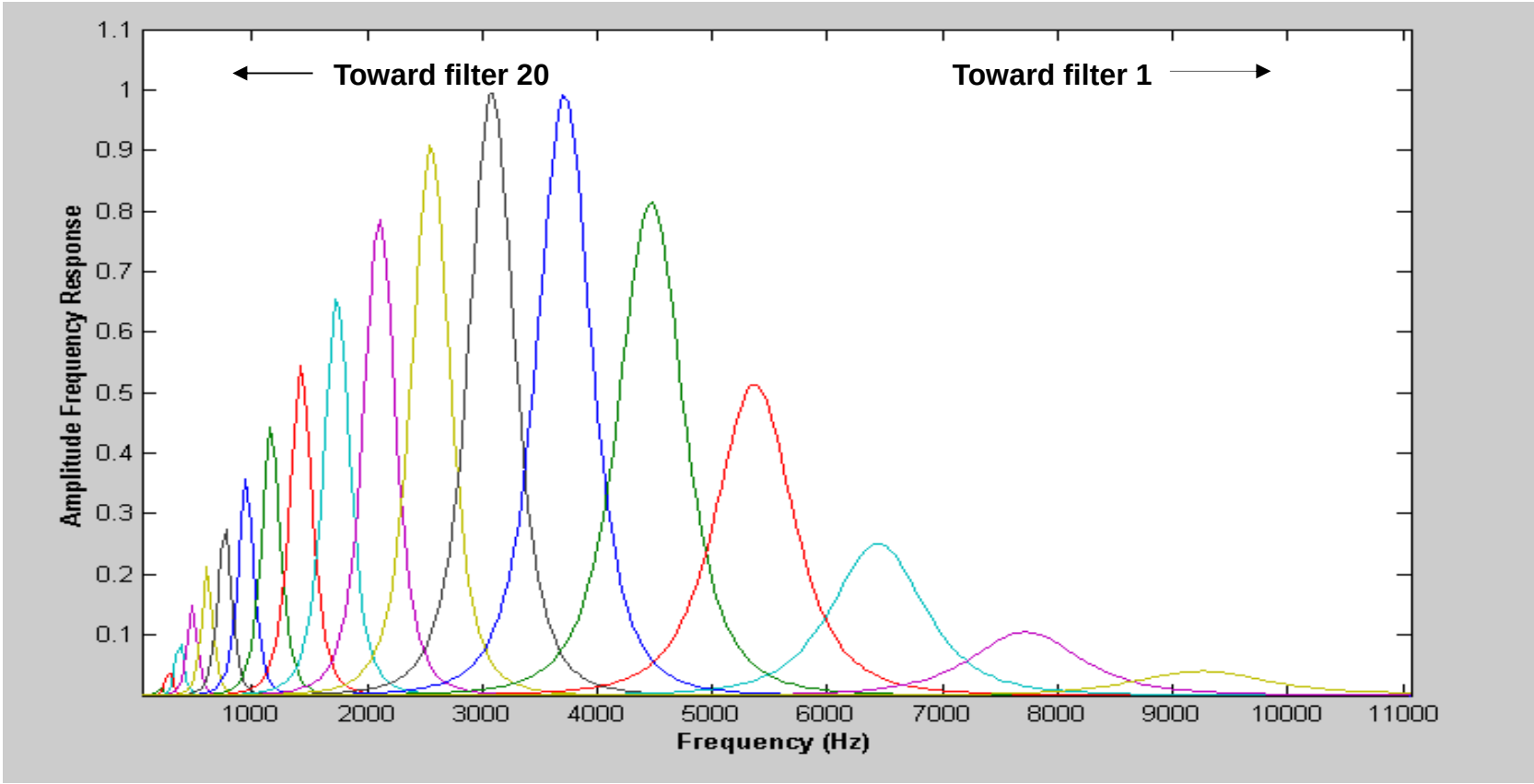
Equal Loudness Preemphasis Filter

The non-uniformity of the loudness sensing can be compensated for by a filter with the following transfer function

$$E(\omega) = \frac{\omega^4(\omega^2 + 56.8\times10^6)}{(\omega^2 + 6.3\times10^6)^2.(\omega^2 + 0.38\times10^9).(\omega^6 + 9.58\times10^{26})}$$

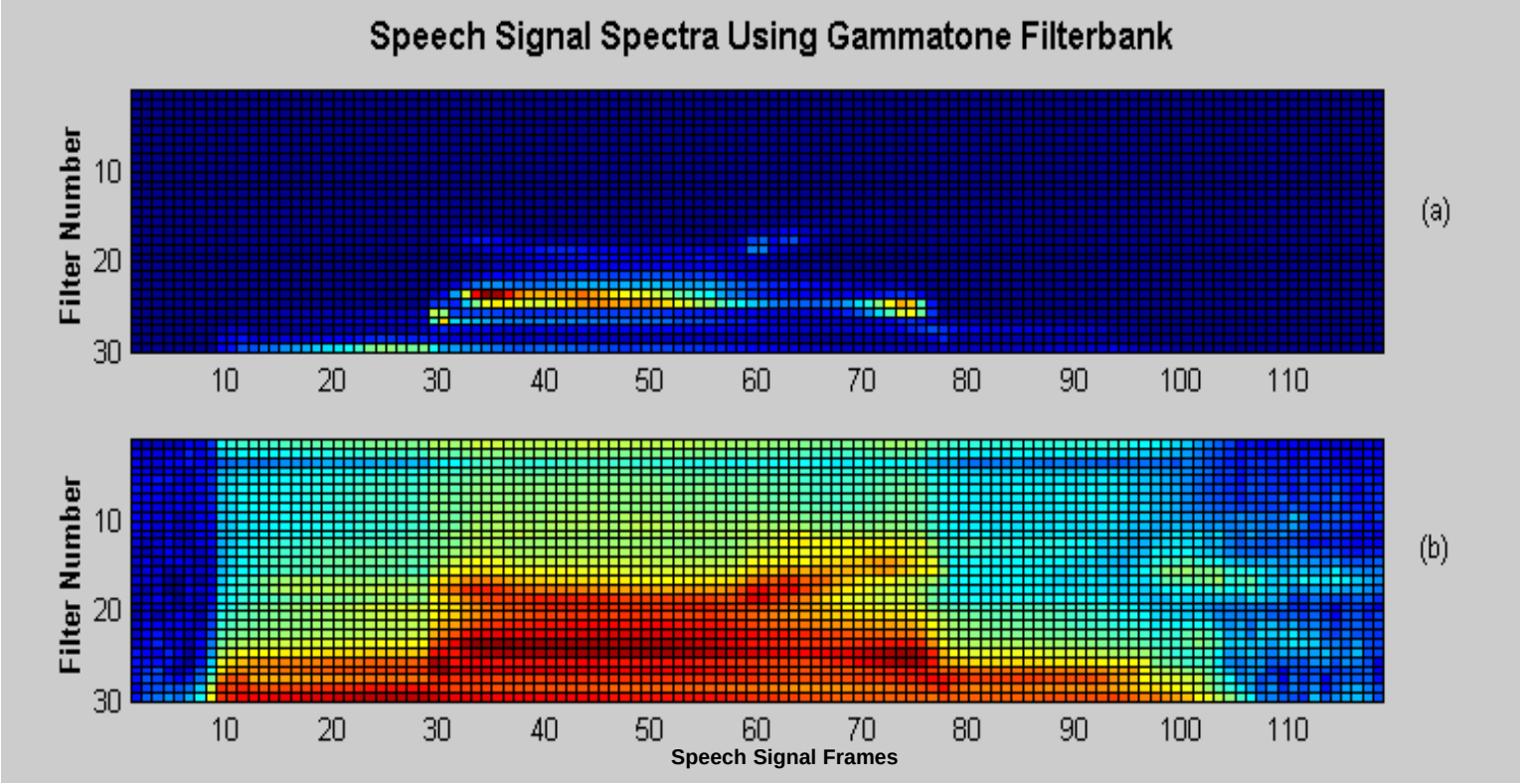


Gammatone Filterbank



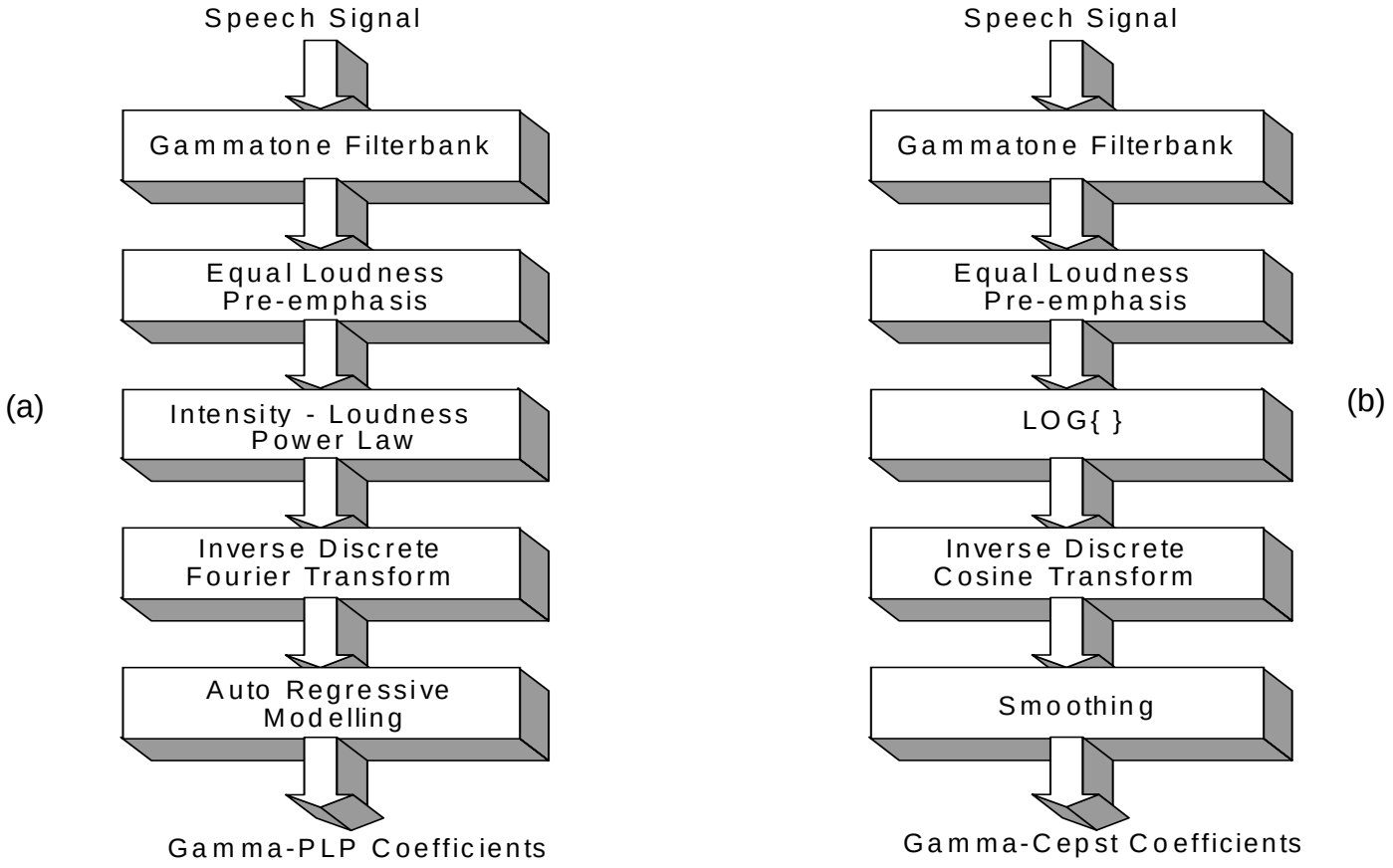
Amplitude frequency responses of a 20-filters Gammatone filterbank after subjecting the filters to the equal loudness pre-emphasis filter.

Speech Signal GTF Frequency Analysis



Speech signal analysis of a spoken digit “9” using 30 Gammatone filters. (a) Spectra of the speech signal, (b) Log spectra of the speech signal.

Feature Extraction Paradigms



Block diagrams of two feature extraction paradigms.

Feature Evaluation Based on F-Ratio

- F-ratio is a measure of the feature effectiveness. It is the ratio of the between class variance (B) to the within class variance (W). For the i^{th} feature in the j^{th} class of K classes:

$$F_i = \frac{B_i}{W_i}$$

$$B_i = \frac{1}{K}\sum_{j=1}^{K}(\mu_{ij} - \mu_i)^2$$

$$W_i = \frac{1}{K}\sum_{j=1}^{K} W_{ij}$$

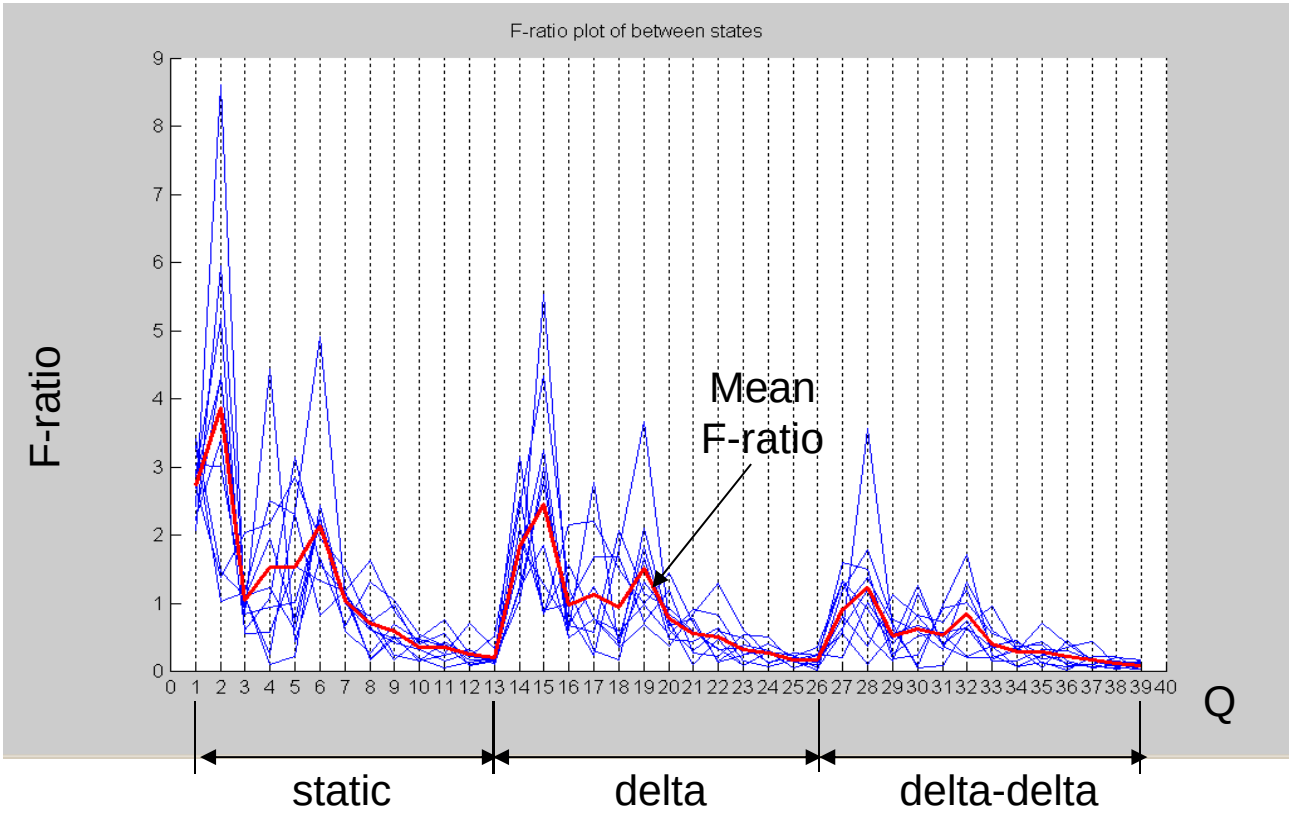
F-Ratio Based on HMM

- HMM satisfies the F-ratio conditions
  - Features have Gaussian distribution.
  - Diagonal covariance implies uncorrelated features

For K states in each model and for H models we have:

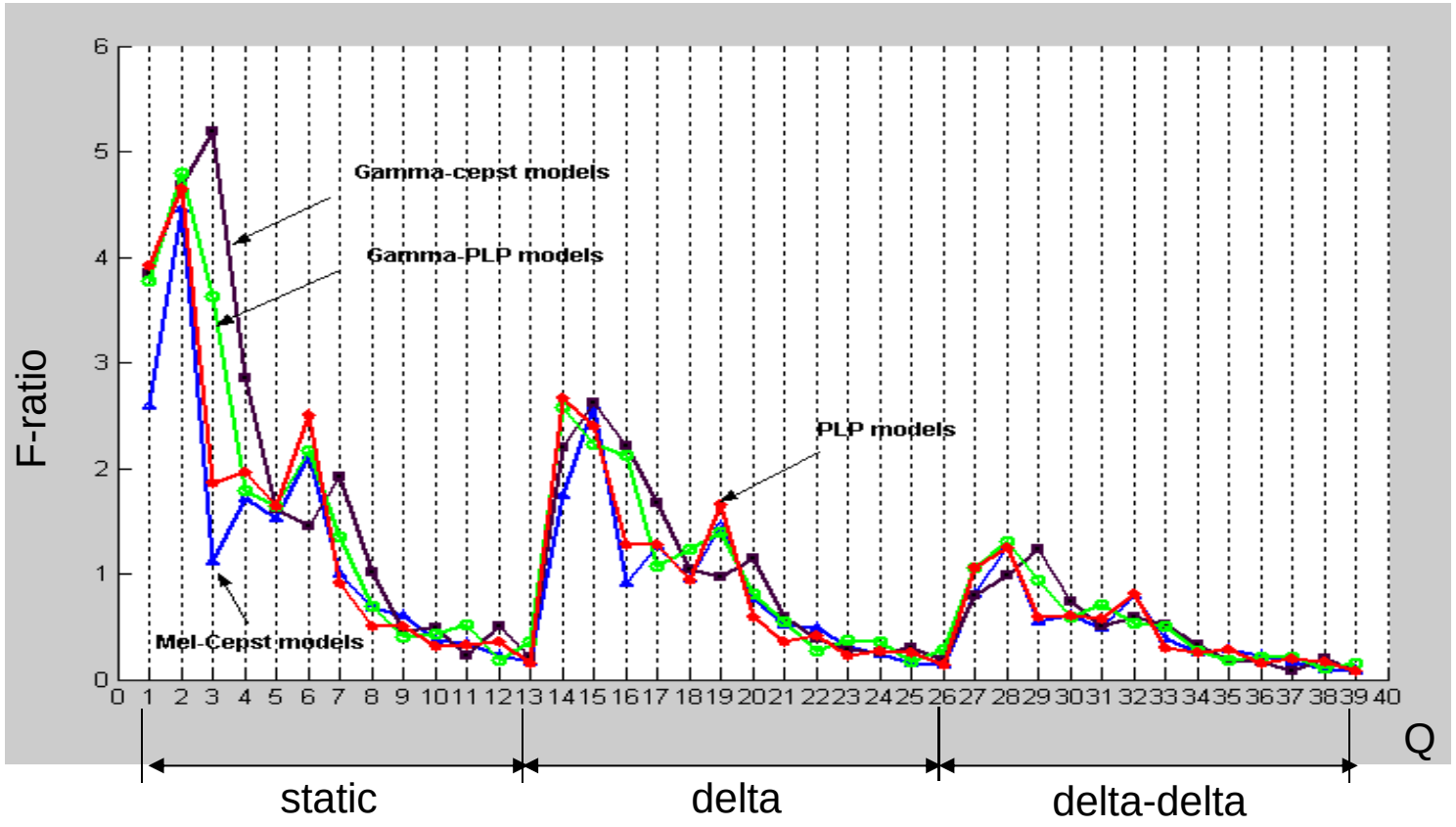
$$F^{ave} = \frac{1}{H}\sum_{i=1}^{H} F_i$$

F-Ratio Characteristics



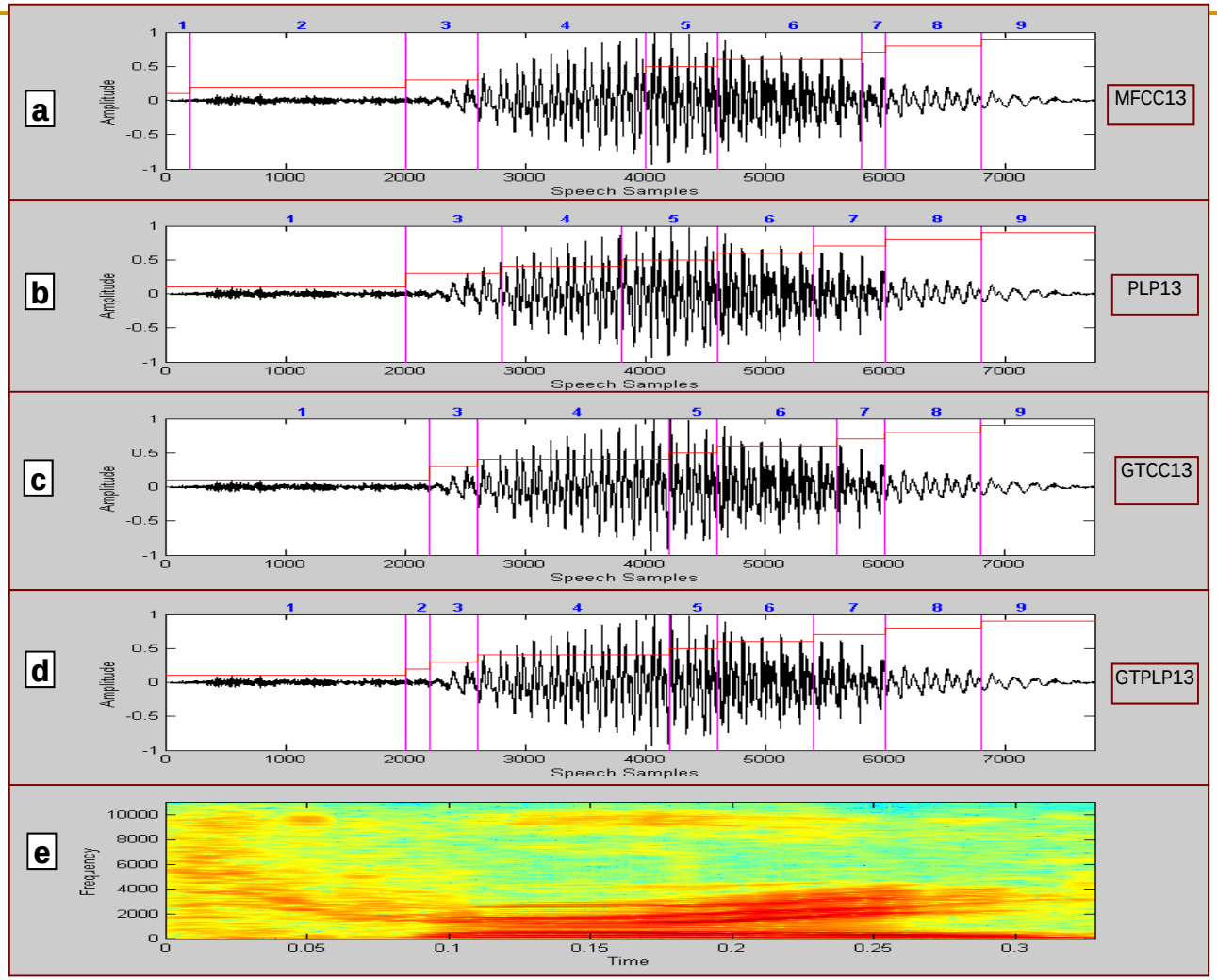
F-ratio of the between states procedure. The thick red line indicates the mean of the between states F-ratio.

Performance Evaluation



Classification properties based on F-ratio calculations of different feature extraction paradigms.

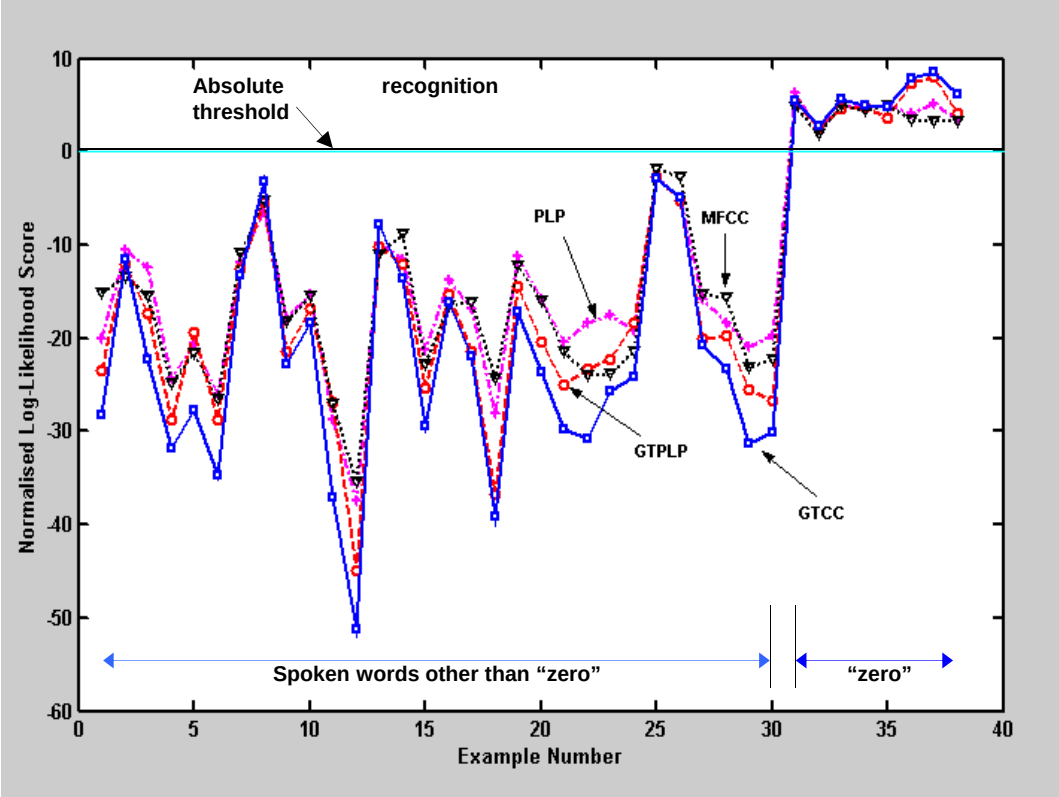
Feature	Rank	F-ratio MFCC	Rank	F-ratio GTCC	Rank	F-ratio GTPLP	Rank	F-ratio PLP
1	2	4.46	3	5.19	2	4.8	2	4.65
2	1	2.59	2	4.68	1	3.78	1	3.92
3	15	2.59	1	3.84	3	3.62	14	2.67
4	6	2.12	4	2.86	14	2.58	6	2.5
5	14	1.75	15	2.63	15	2.22	15	2.4
6	4	1.72	16	2.21	6	2.17	4	1.96
7	5	1.53	14	2.2	16	2.12	3	1.87
8	19	1.45	7	1.92	4	1.79	19	1.66
9	17	1.27	17	1.68	5	1.65	5	1.64
10	28	1.26	5	1.61	19	1.4	17	1.28
11	3	1.11	6	1.45	7	1.34	16	1.28
32	12	0.23	11	0.23	34	0.26	25	0.25
33	36	0.22	13	0.21	37	0.21	23	0.23
34	13	0.17	38	0.19	36	0.21	37	0.2
35	37	0.16	35	0.19	35	0.18	38	0.16
36	25	0.15	26	0.18	12	0.18	13	0.15
37	26	0.14	36	0.17	25	0.17	36	0.15
38	38	0.1	37	0.08	39	0.15	26	0.14
39	39	0.08	39	0.07	38	0.12	39	0.08
Mean F-ratio		**0.8839**		**1.1471**		**1.0753**		**0.9862**



Shows the states of the word three as detected by its four static features based CDHMMs.

(a) Model MFCC13 is constructed from 13 static mel scale coefficients.
(b) Model PLP13 is constructed from 13 static perceptual linear prediction coefficients.
(c) Model GTCC13 is constructed from 13 static Gammatone cepstral coefficients.
(d) Model GTPLP13 is constructed from 13 static Gammatone PLP coefficients.
(e) The spectrogram of the input signal to envisage the frequency content of each state.

Classification Performance



	Margin
MFCC	21.51
PLP	21.78
GTPLP	25.14
GTCC	28.95

Recognition Rate Performance

	DATASET-I	DATASET-II
Mel-cespt	100	95.2
PLP	100	96.1
Gamma-PLP	100	97.8
Gamma-cepst	100	98.9

DATASET-I : 10 digits
DATASET-II : 31 words
S/N ratio = 20 dB

Conclusions

- Efficient auditory motivated technique is introduced.
- It is mainly based on Gammatone filterbank (GTF).
- GTF composed of non uniform bandpass filters imitating the frequency resolution of the cochlea.
- Two paradigms: Gamma-cepst and Gamma-PLP are investigated.
- Classification performance based on the F-ratio figure of merit has been investigated as it is a strong cue to the recognition performance.
- Gamma-cepst feature set outperforms the other feature sets.

Thank You