Speech

Speech

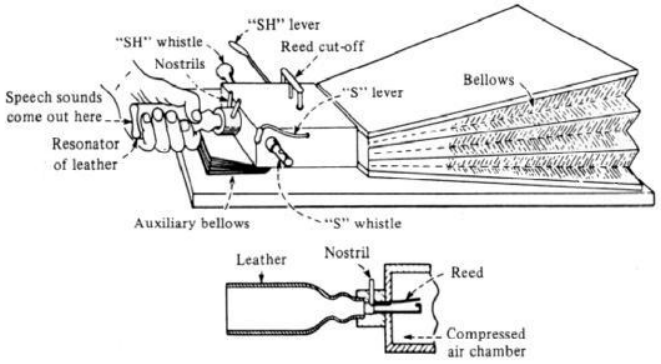
- Analysis
- Synthesis
- Recognition
- Data compression

Mechanical Synthesis

Wolfgang von Kempelen and Charles Wheatstone created a more sophisticated mechanical speech device...

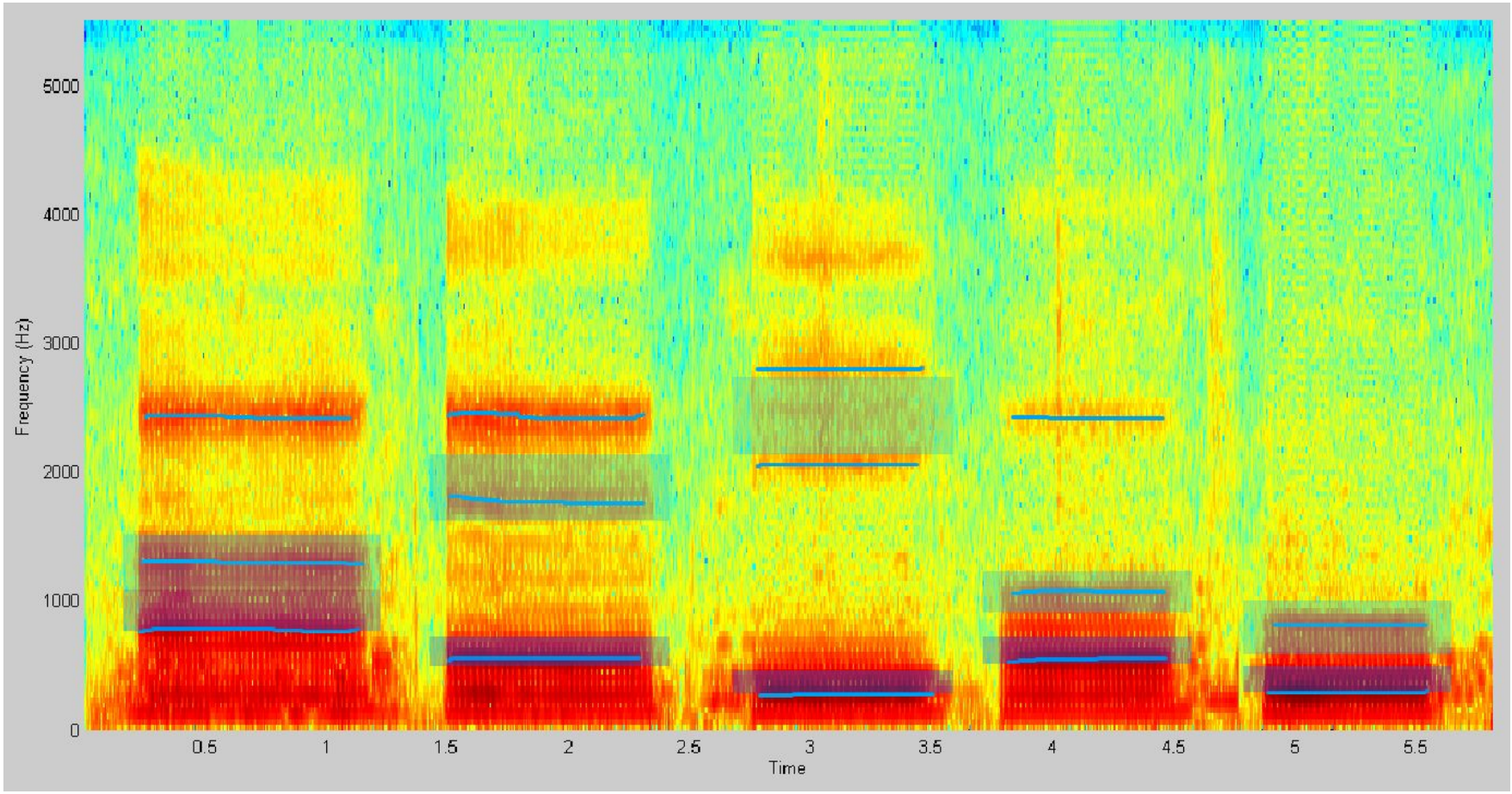
with independently manipulable source and filter mechanisms.

http://kempelen.info/vystava.html



Speech analysis -Vowel formants

Vowels (a,e,i, o, u) formants - example



Speech analysis
Vowel formants

Formants are defined as the spectral peaks of the sound spectrum |P(f)| of the voice.

In speech science and phonetics, formant is also used to mean an acoustic resonance of the human vocal tract.

Vowel first formant region:

u 200–400 Hz

o 400–600 Hz

a 800–1200 Hz

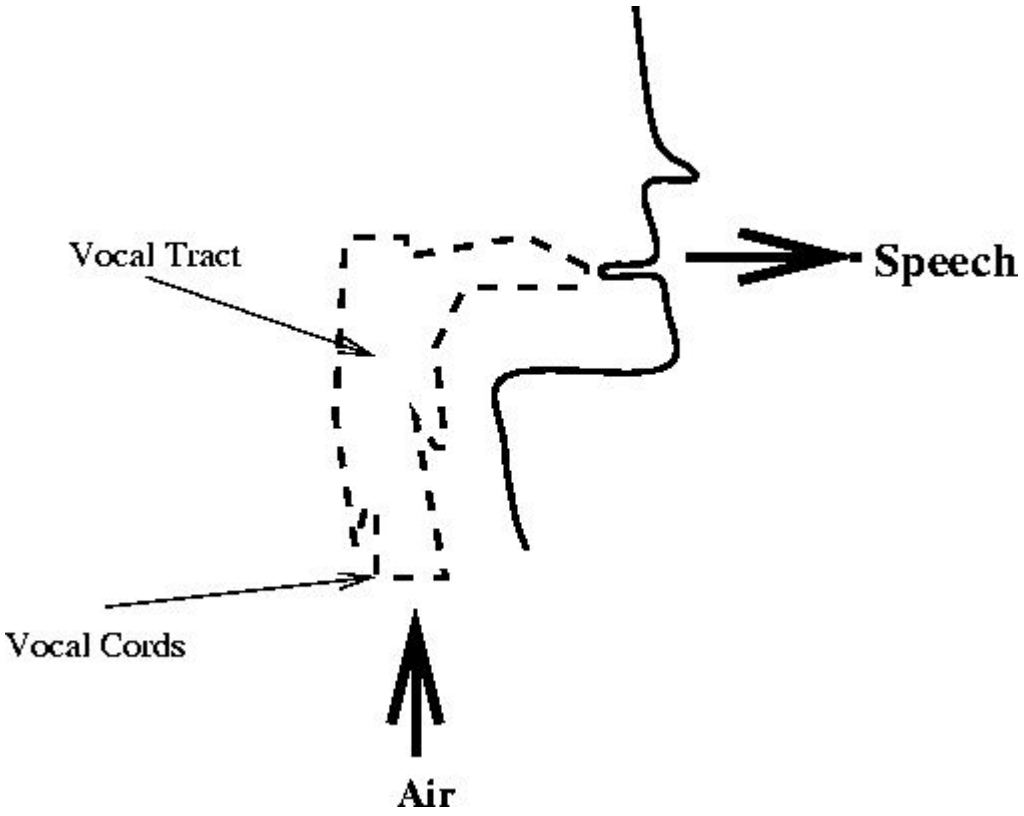
e 400–600 and 2200–2600 Hz

i 200–400 and 3000–3500 Hz

Model of Speech Production

Model of Speech Production

Physical Model of Speech Production

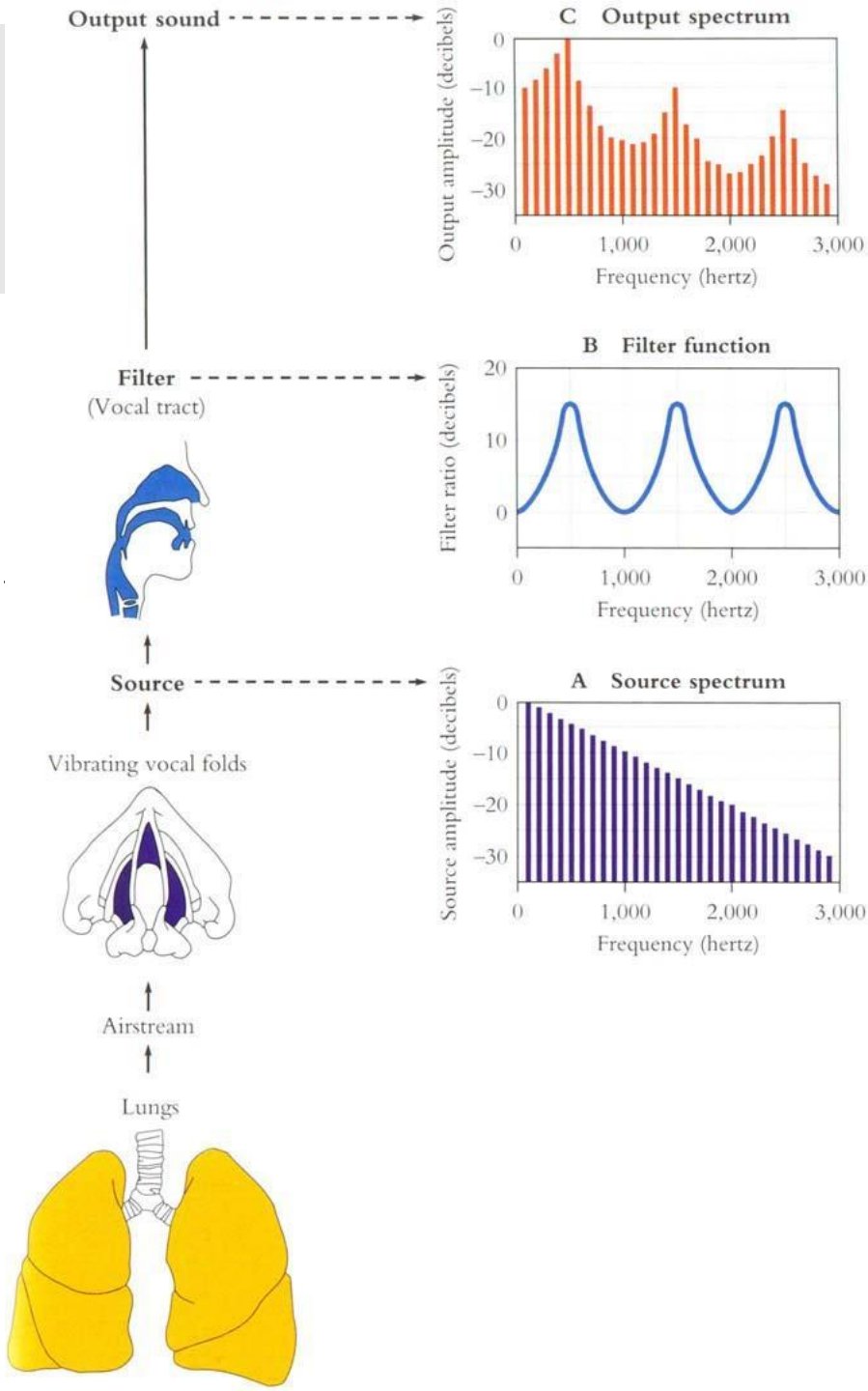


George Miller - Model of Speech Production

The **filter is given by**

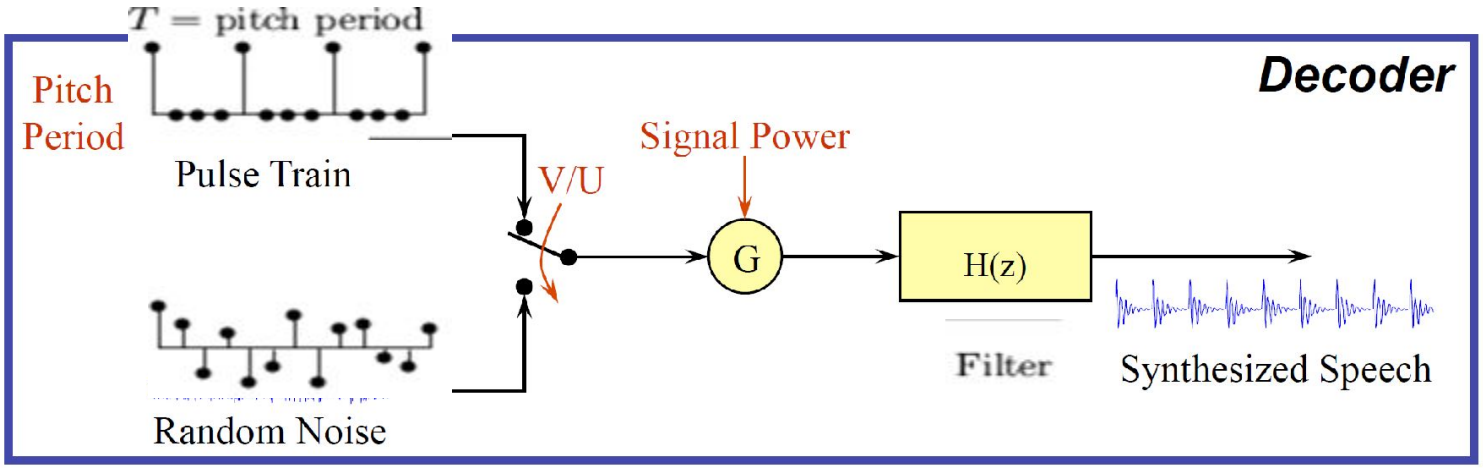
- the exact position of the articulators in the oral trac

-> So we want a way to separate these and use only the filter function



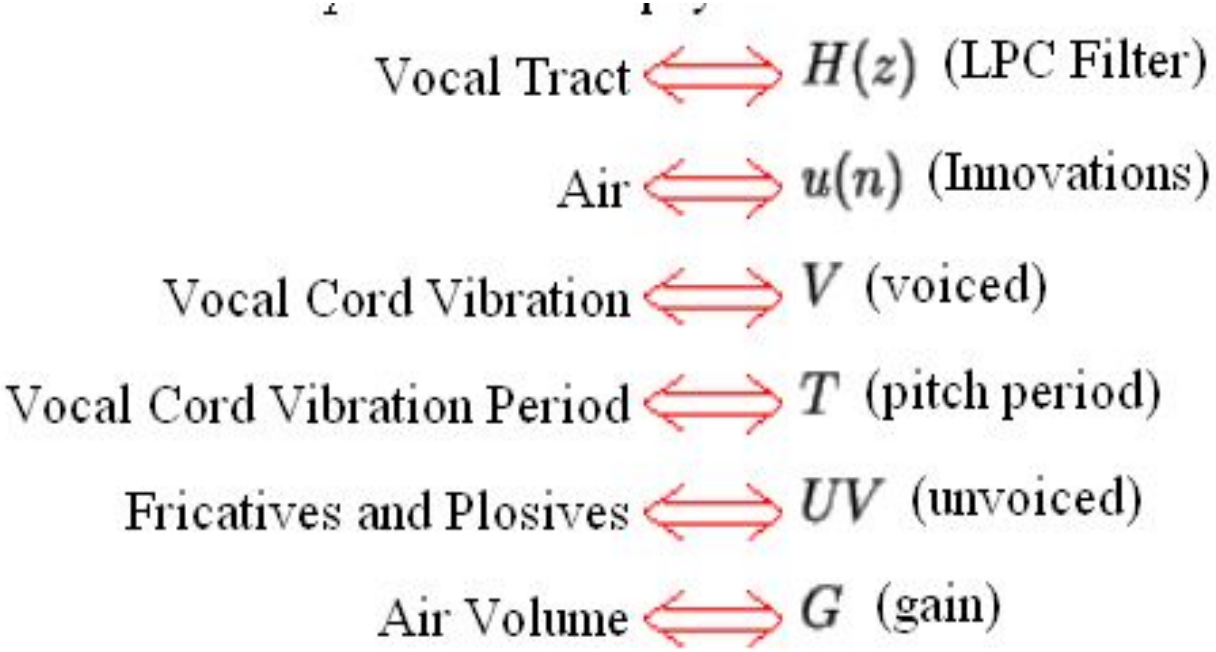
Model of Speech Production

Mathematical Model of Speech Production :



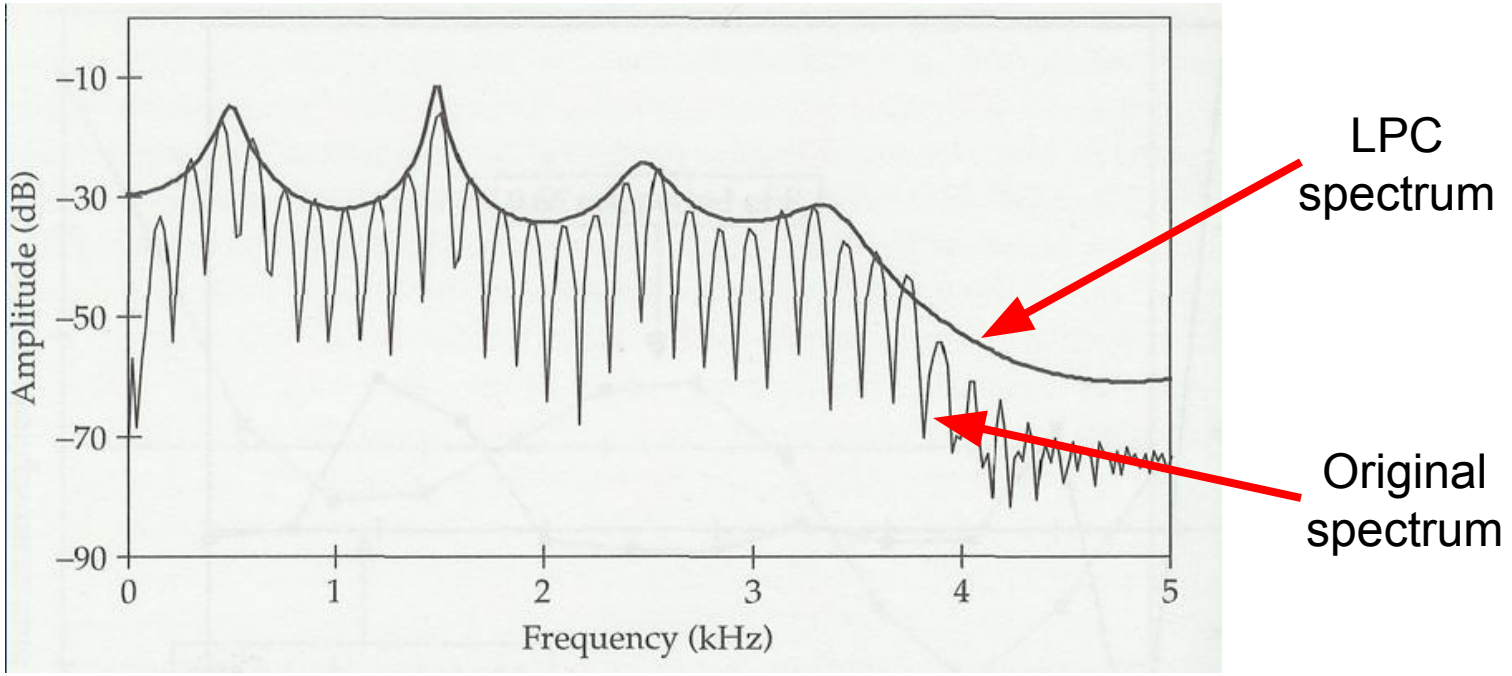
Model of Speech Production

The relationship between the physical and the mathematical models:



LPC model

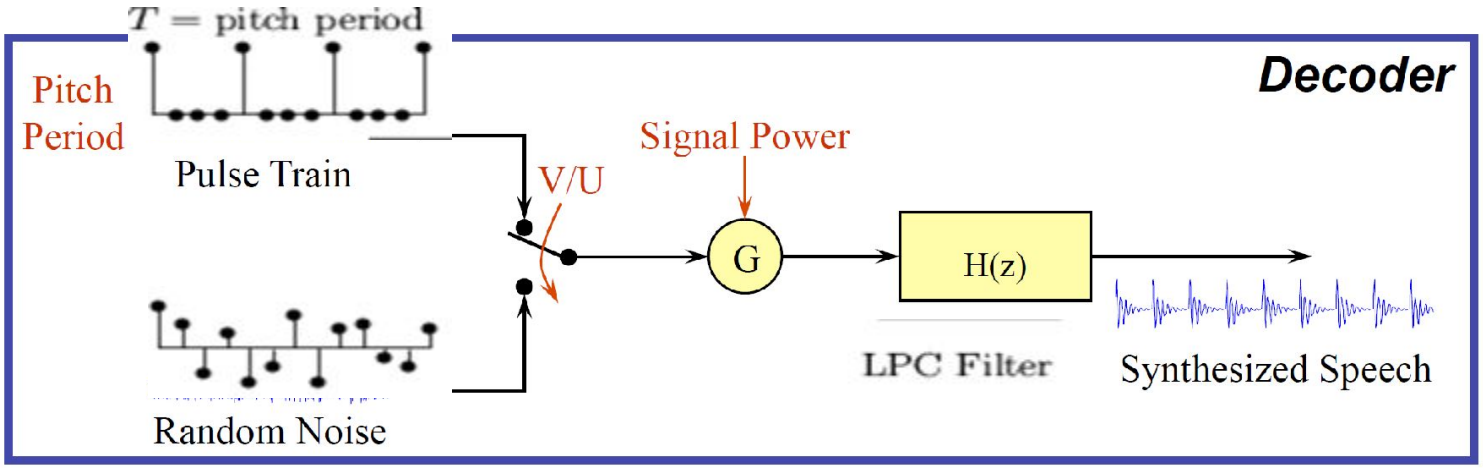
LPC Spectrum



Spectral representation of the LPC analysis.

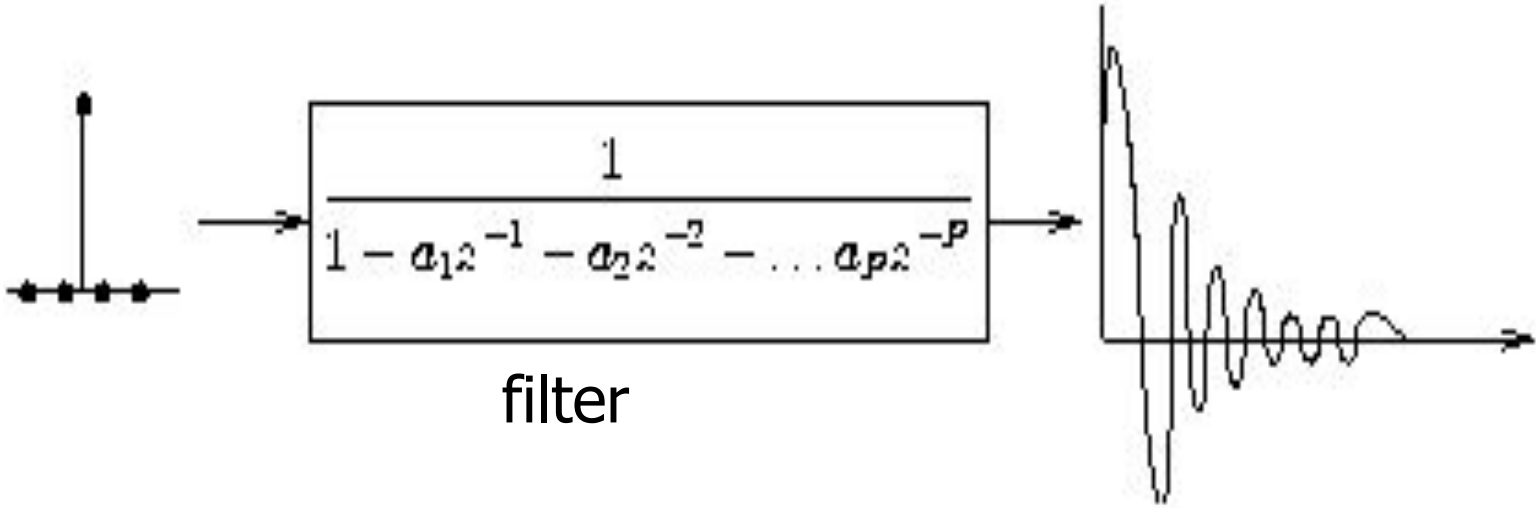
Model of Speech Production

Mathematical Model of Speech Production :

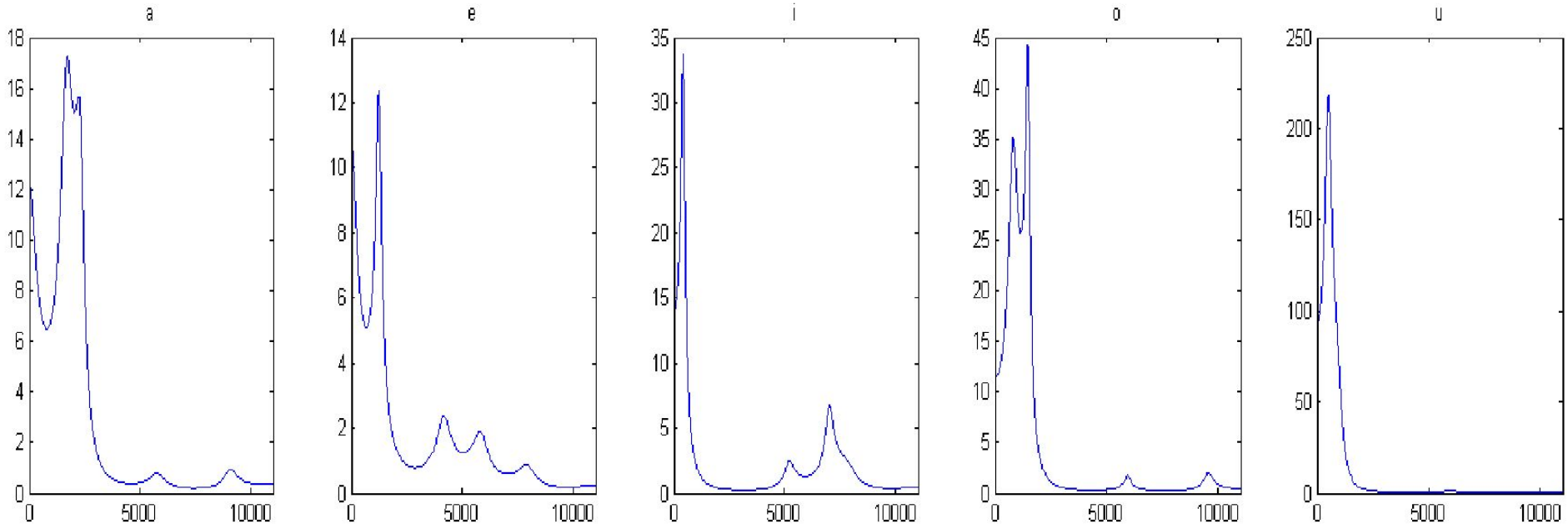


LPC – Linear predictive coding

LPC models - Linear prediction models the human vocal tract as a system that produces the speech signal



LPC Spectrum example – wovels a,e,i,o,u



Cepstrum

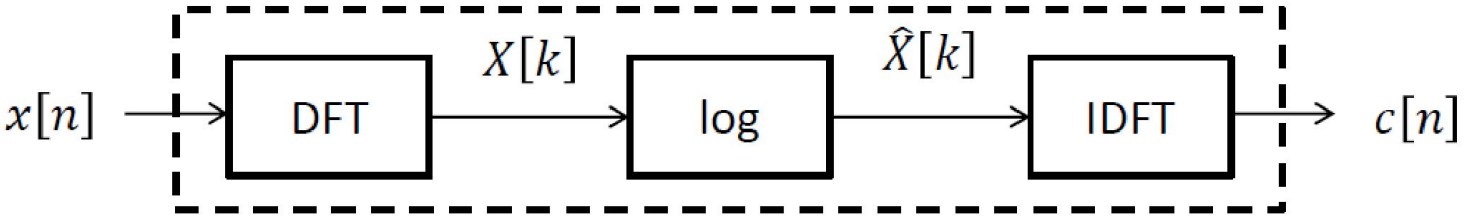
Cepstrum

The cepstrum is defined as the inverse DFT of the log magnitude of the DFT of a signal f(x)

$c(n) = F^{-1}(\log F(x(n)))$

where F is the DFT and F^{-1} is the IDFT

This original definition of Bogert, M. J. R. Healy, and J. W. Tukey, loosely framed in terms of spectrum analysis of analog signals, was motivated by the fact that the logarithm of the Fourier spectrum of a signal containing an echo has an additive periodic component depending only on the echo size and delay, and that further Fourier analysis of the log spectrum can aid in detecting the presence Speech Analysis of that echo.

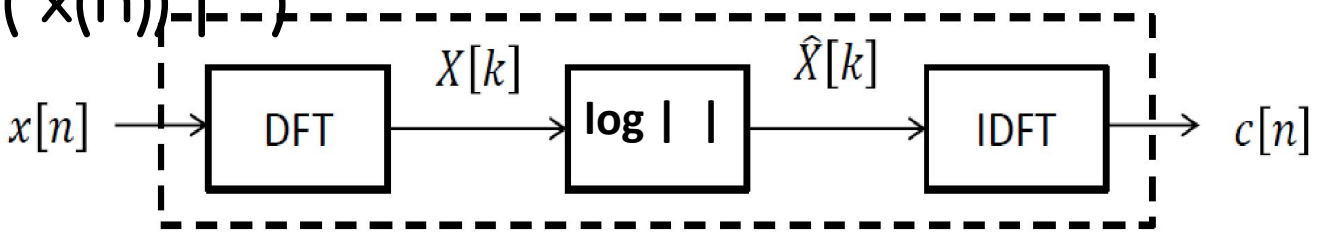


B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of times series for echos: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in Proceedings of the Symposium on Time Series Analysis, (M. Rosenblatt, ed.), New York: John Wiley and Sons, Inc., 1963

Definition of the real cepstrum ->*cepstrum*

The cepstrum was defined by Bogert, Healy, and Tukey to be the inverse Fourier transform of the log magnitude spectrum of a signal.

$$c(n) = F^{-1}(\log | F(x(n)) |)$$



where

F is the DFT and

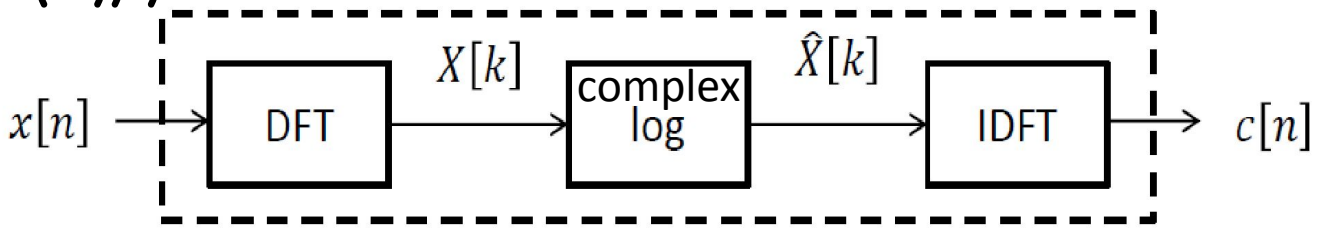
F^{-1} is the IDFT

B. P. Bogert, M. J. R. Healy, and J. W. Tukey, "The quefrency alanysis of times series for echos: Cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking," in Proceedings of the Symposium on Time Series Analysis, (M. Rosenblatt, ed.), New York: John Wiley and Sons, Inc., 1963

Definition of *complex cepstrum*

The *complex cepstrum* of a sequence *x* is calculated by finding the complex natural logarithm of the Fourier transform of *x* , then the inverse Fourier transform of the resulting sequence.

$\hat{c}(n) = F^{-1}(\log F(x(n)))$



where

F is the DFT and

F^{-1} is the IDFT

Complex logarithm $\log F(x(n)) = \log | F(x(n)) | + j*\arg(F(x(n)))$

Definition given by Oppenheim, Schafer and Stockham

RABINER, Lawrence R., et al. Introduction to digital speech processing. *Foundations and Trends® in Signal Processing*, 2007, 1. Jg., Nr. 1–2, S. 1-194.

Cepstral analysis as deconvolution

The speech signal can be modeled as the convolution of the source signal and filter

-> model of the vocal tract $\mathbf{x(n)= s(n) \otimes f(n)}$

Because these signals are convolved, they cannot be easily separated in the time domain –> we can however perform the separation as follows:

We apply the Fourier Transform, then we obtain:

$\mathbf{F(x(n))= F(x(n) * F(x(n)}$

Then we take the log of the magnitude of the Fourier Transform

$\mathbf{\log(F(x(n))= \log (F(x(n)) + \log(F(x(n))}$

which shows that source and filter are now just added together

We can now return to the time domain through the inverse FT

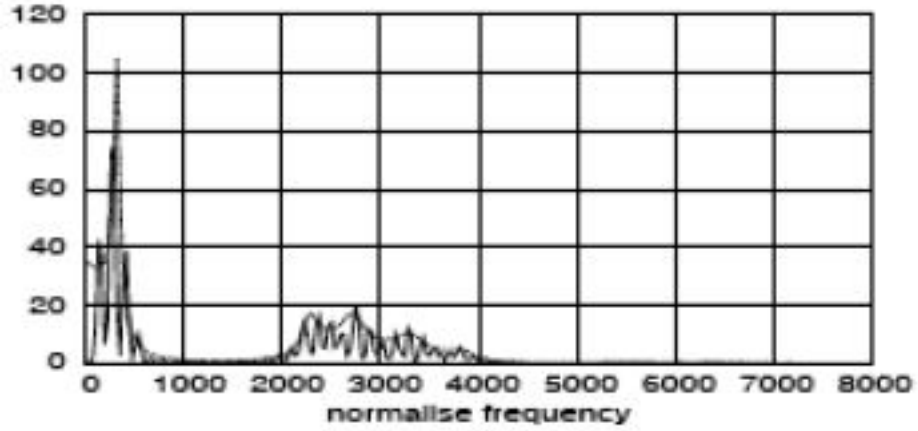
F^{-1} **(log(F(x(n)))=** F^{-1} **(log (F(x(n))) +** F^{-1} **(log(F(x(n)))**

Source-filter separation via the cepstrum

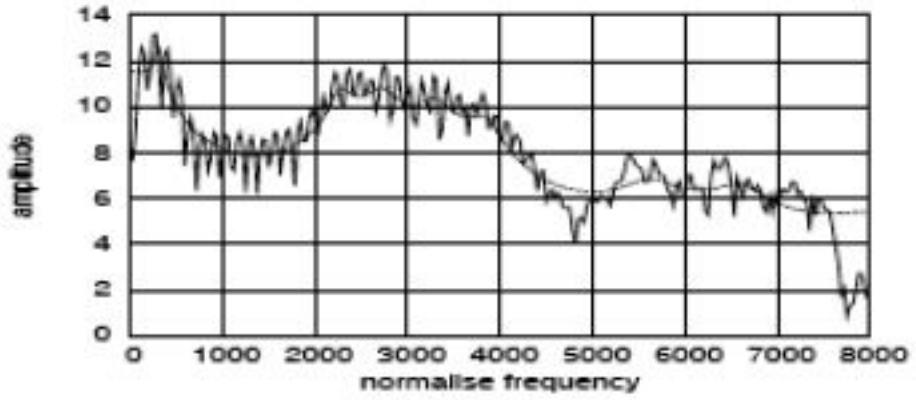
The cepstrum is useful because it separates source and filter

- If we are interested in the glottal excitation, we keep the high coefficients
- If we are interested in the vocal tract, we keep the low coefficients
- Truncating the cepstrum at different quefrency values allows us to preserve different amounts of spectral

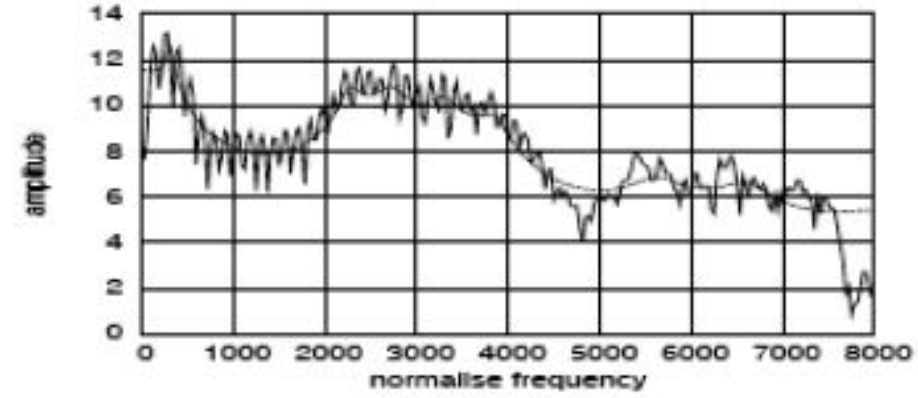
Cepstrum - example

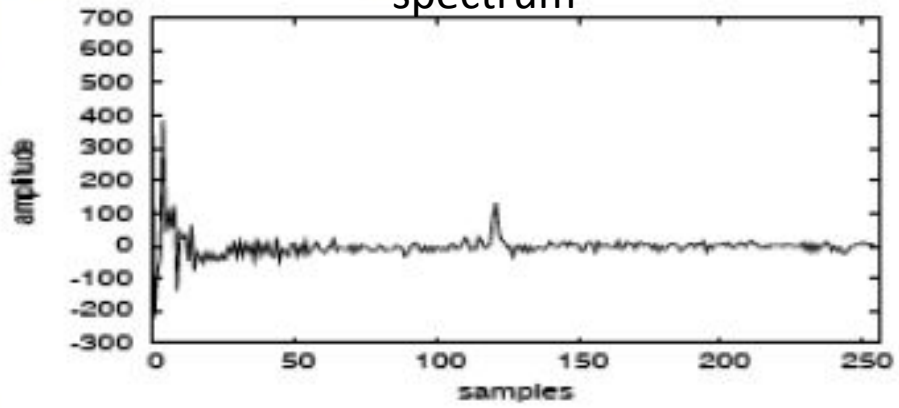


Spectrum

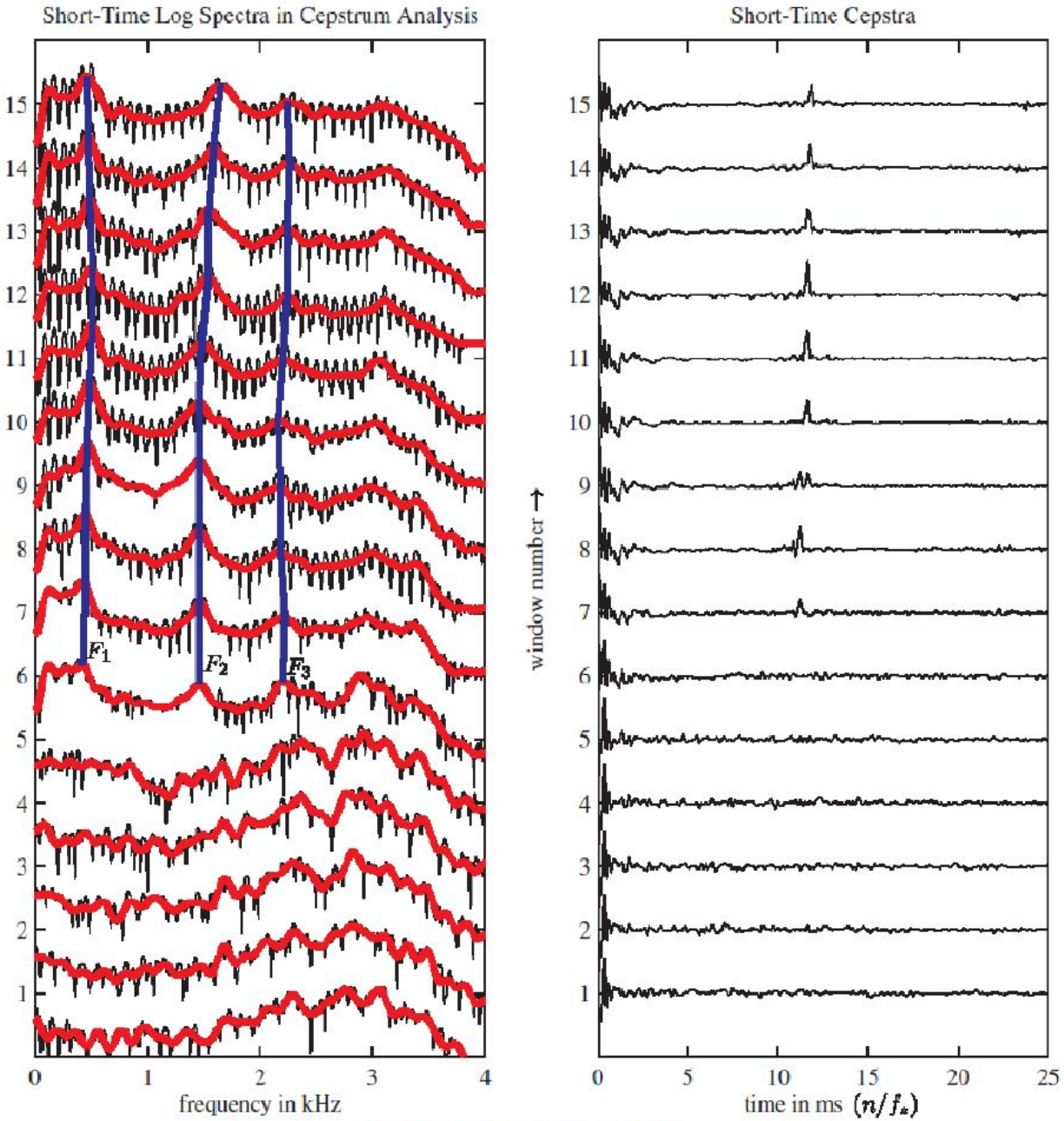


Log spectrum





Cepstrum analysis



[Rabiner & Schafer, 2007]

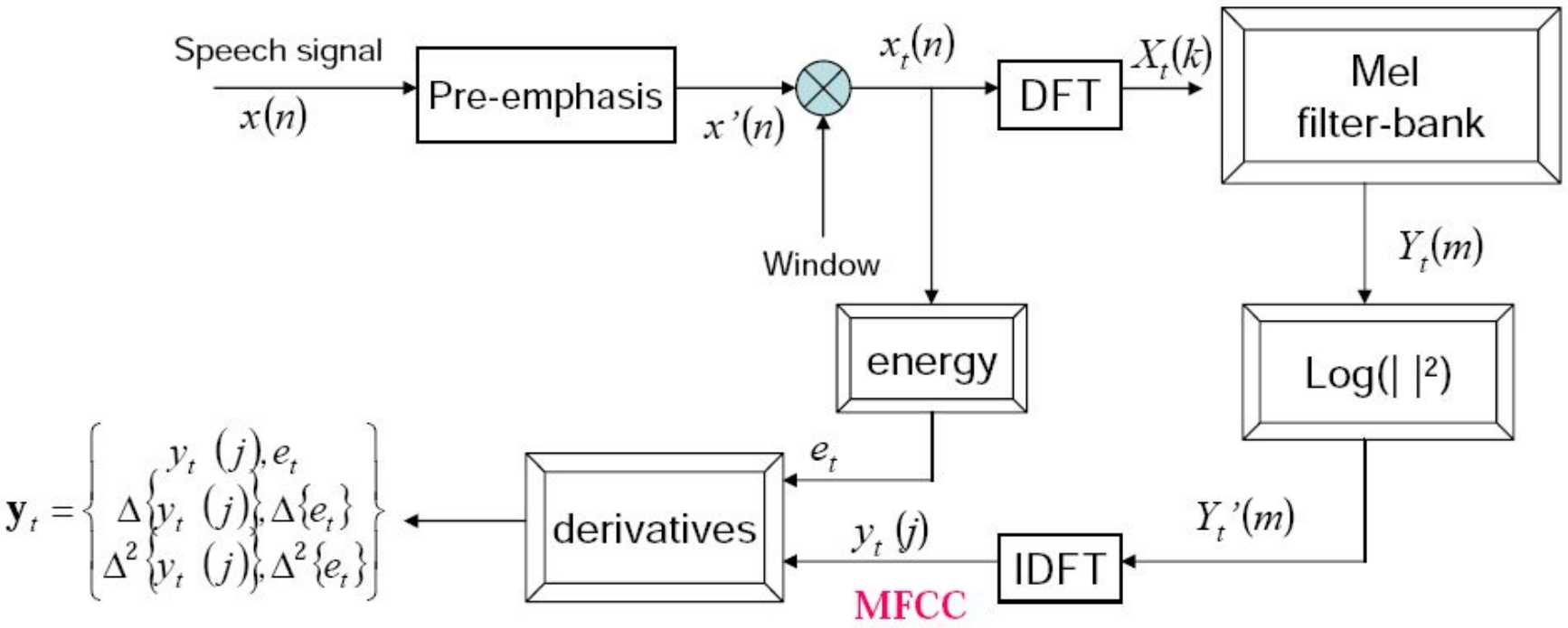
Mel-Frequency Cepstral Coefficient (MFCC)

Mel-Frequency Cepstral Coefficient(MFCC)

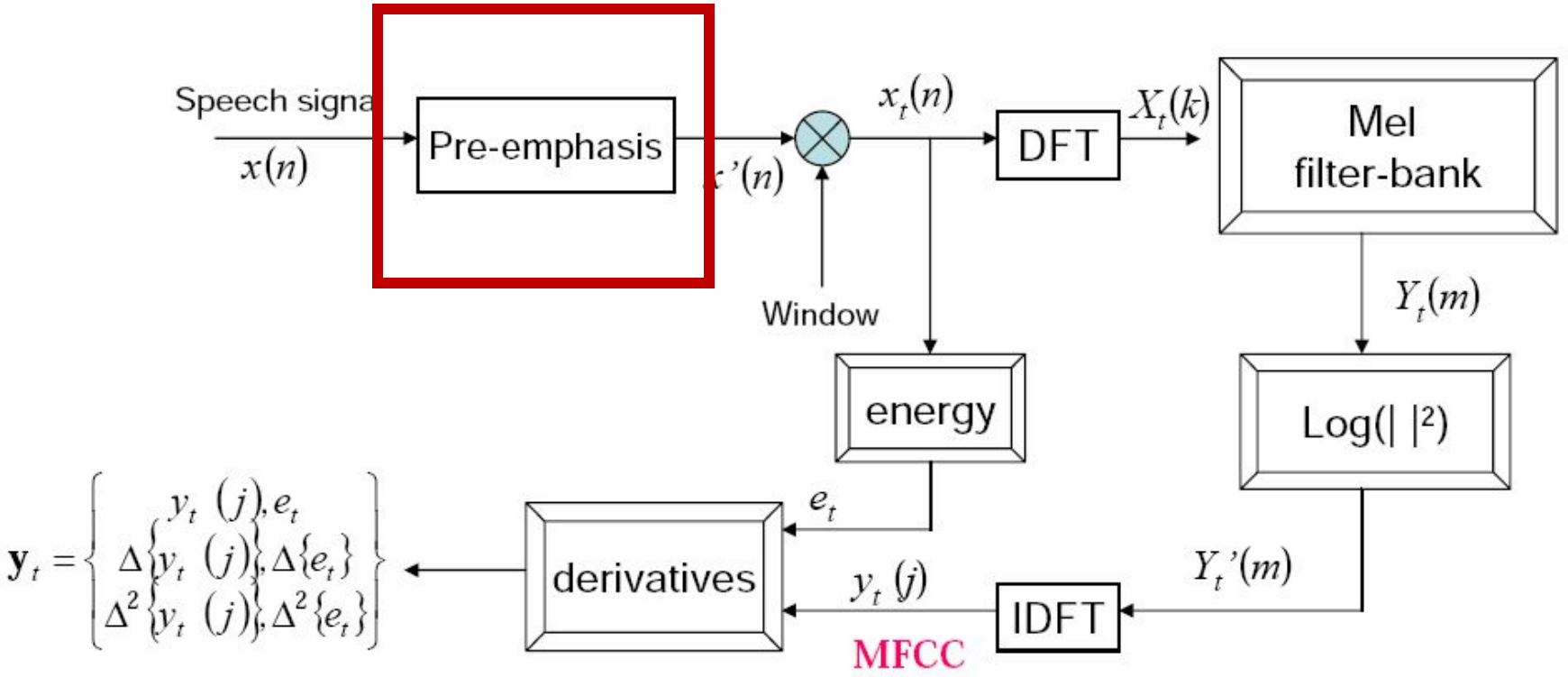
Calculation steps:

- Windowing in time domain
- Filtration using triangular bank of filters
- Cepstrum calculation
- Pre-emphasis boosting

MFCC - overview



Steps of MFCC calculation: Pre-Emphasis



Steps of MFCC calculation: Pre-Emphasis

Pre-emphasis: boosting the energy in the high frequencies

The spectrum for voiced segments has more energy at lower frequencies than higher frequencies.

- This is called **spectral tilt**
- Spectral tilt is caused by the nature of the glottal pulse

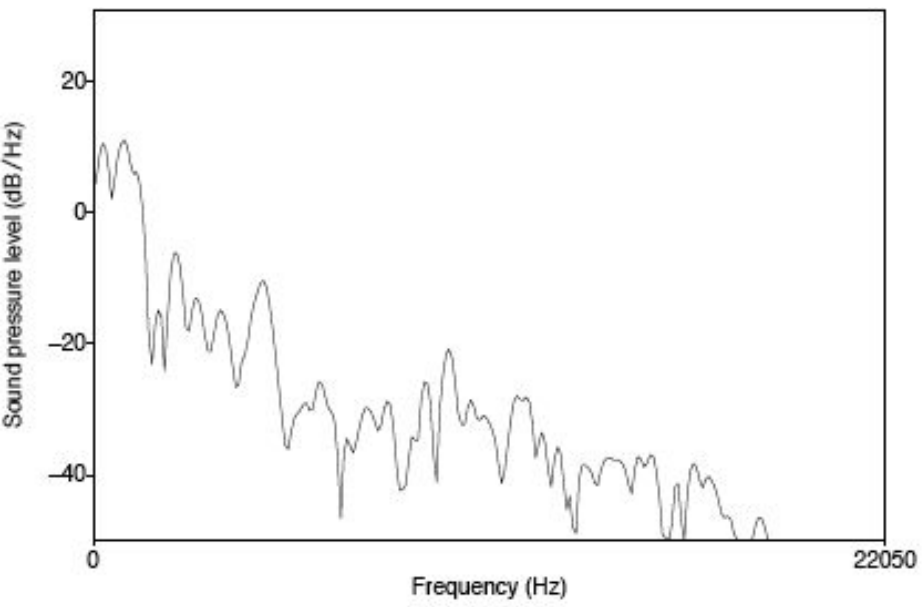
Boosting high-frequency energy gives more info to Acoustic Model

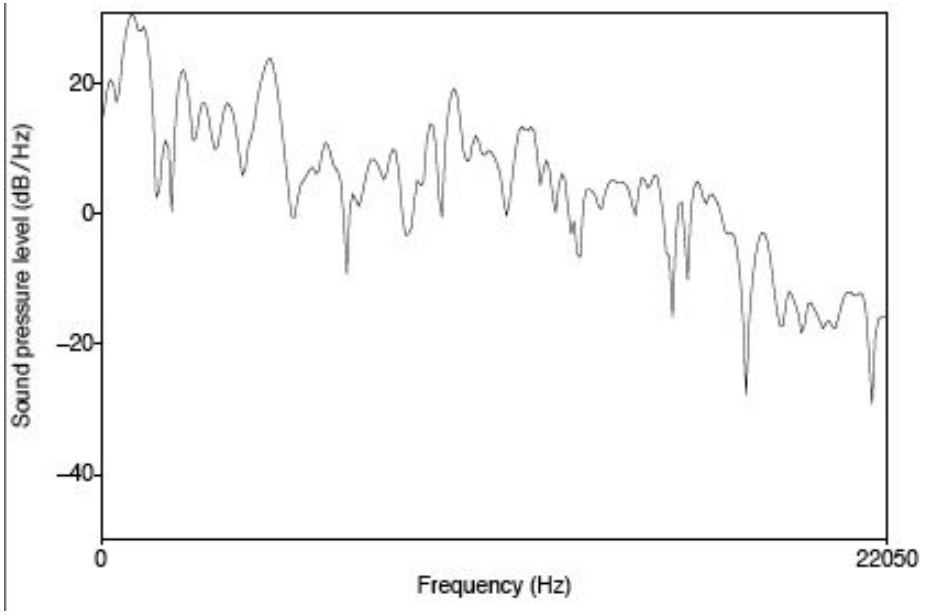
- Improves phone recognition performance

Example of pre-emphasis

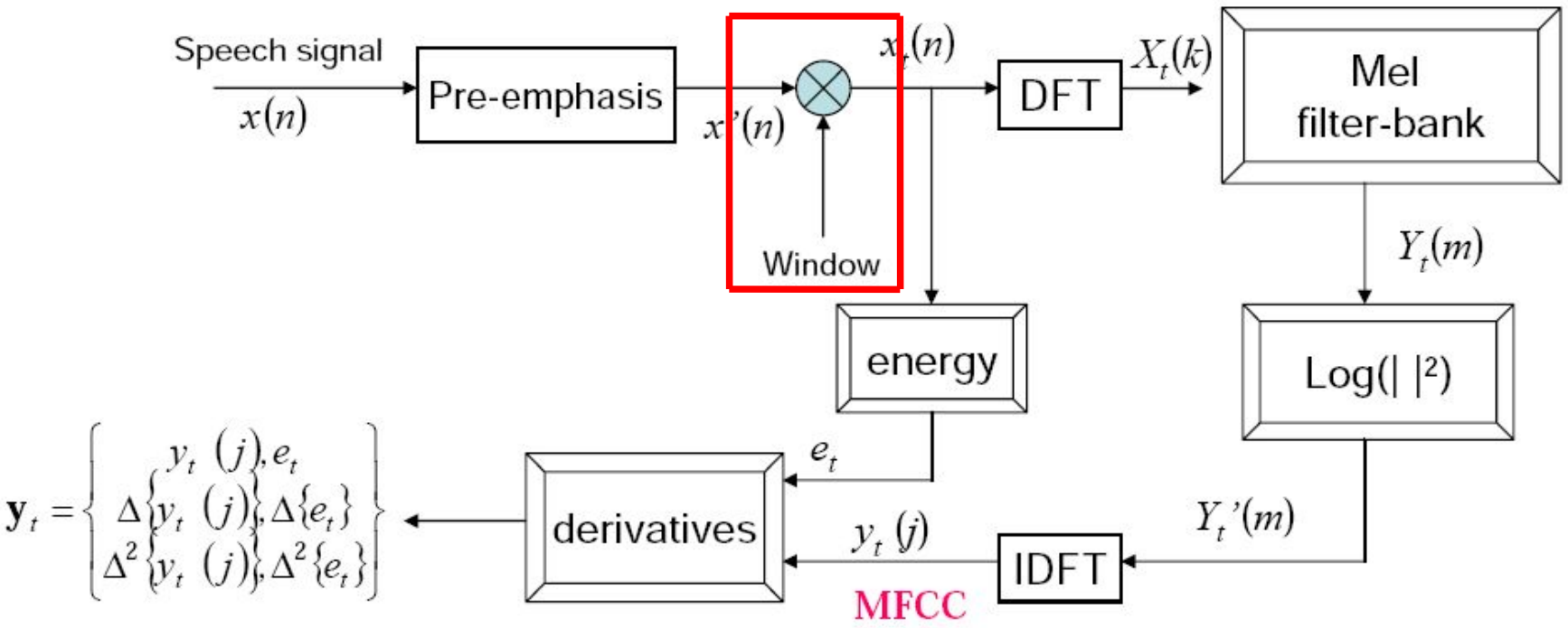
Before and after pre-emphasis

- Spectral slice from the vowel [aa]

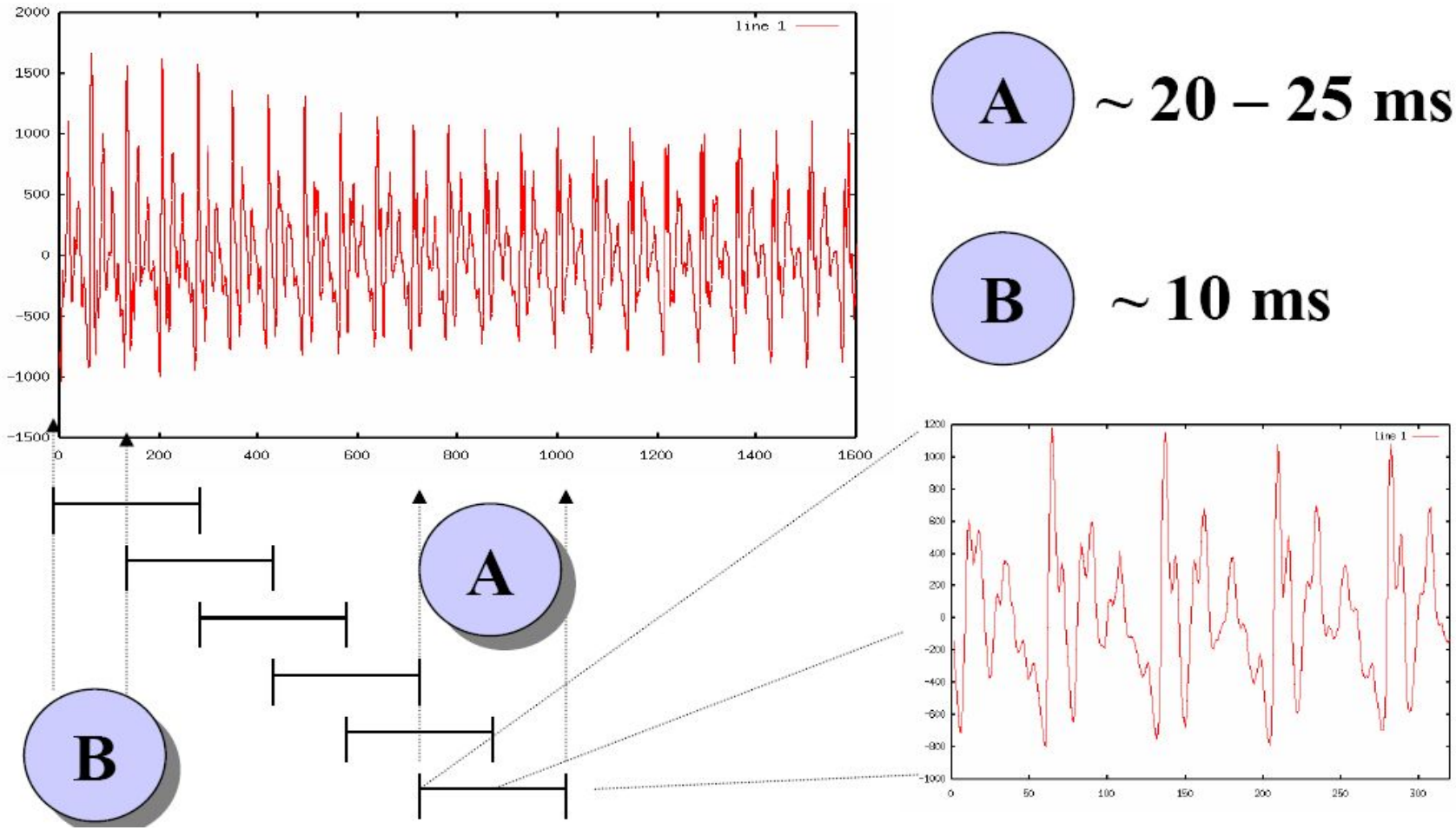




Steps of MFCC calculation: windowing



Steps of MFCC calculation: windowing



Slide from Bryan

Steps of MFCC calculation: windowing

Speech is not a stationary signal; we want information about a small enough region that the spectral information is a useful cue.

Frames

- Frame size: typically, 10-25ms
- Frame shift: the length of time between successive frames, typically, 5-10ms

Rectangular window vs. Hamming window

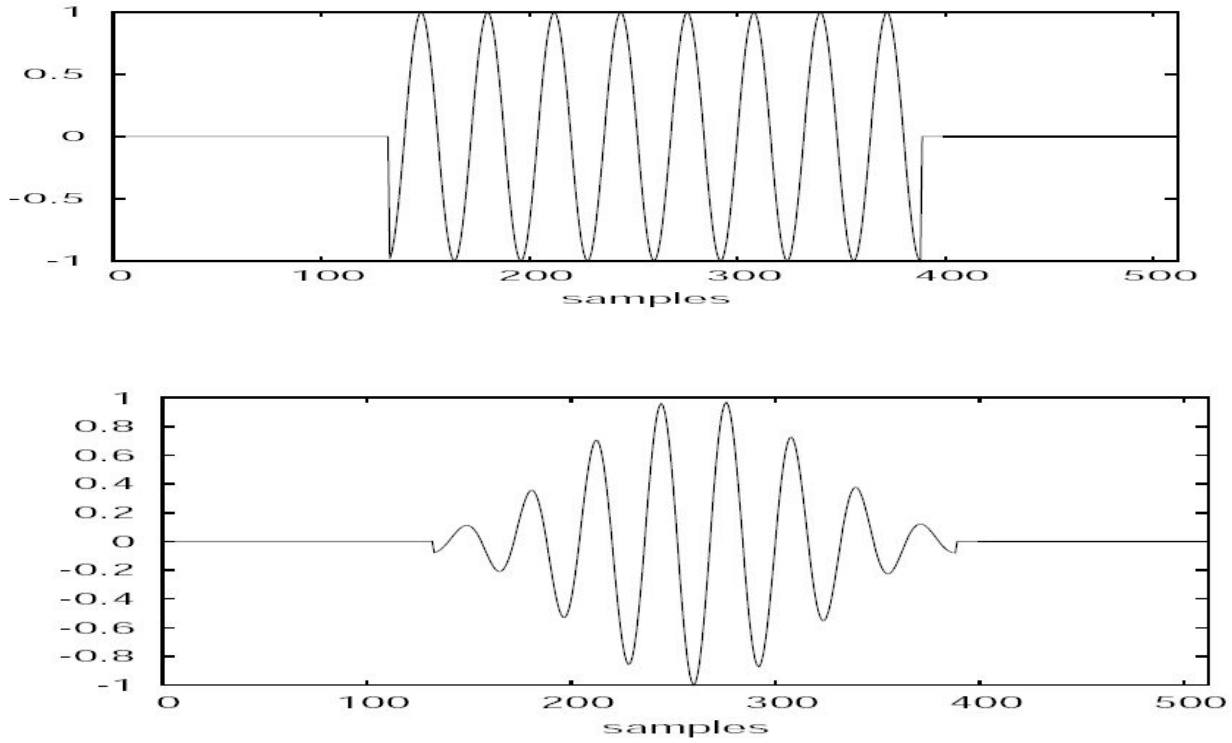
Rectangular window:

$$w[n] = \begin{cases} 1 & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

Hamming window

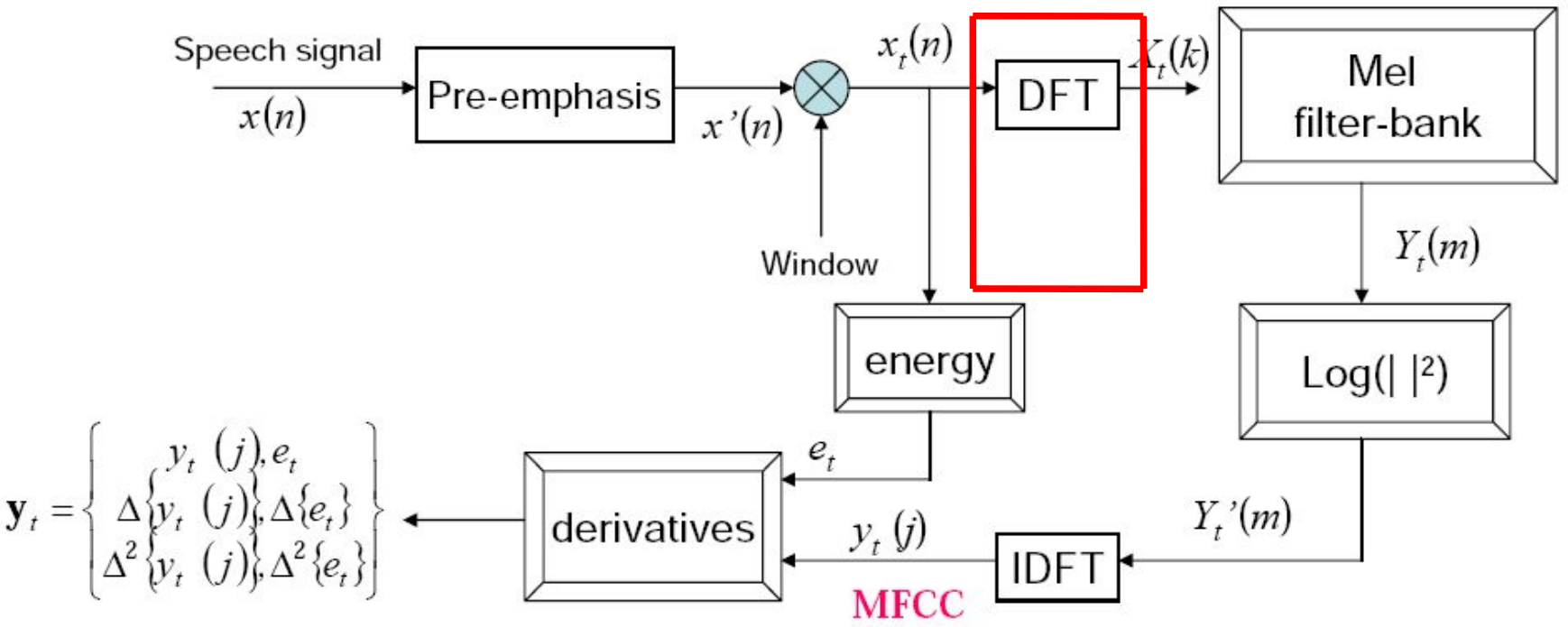
$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) & 0 \leq n \leq L-1 \\ 0 & \text{otherwise} \end{cases}$$

Steps of MFCC calculation: Windowing in time domain



(c) Hamming window

Steps of MFCC calculation: DFT

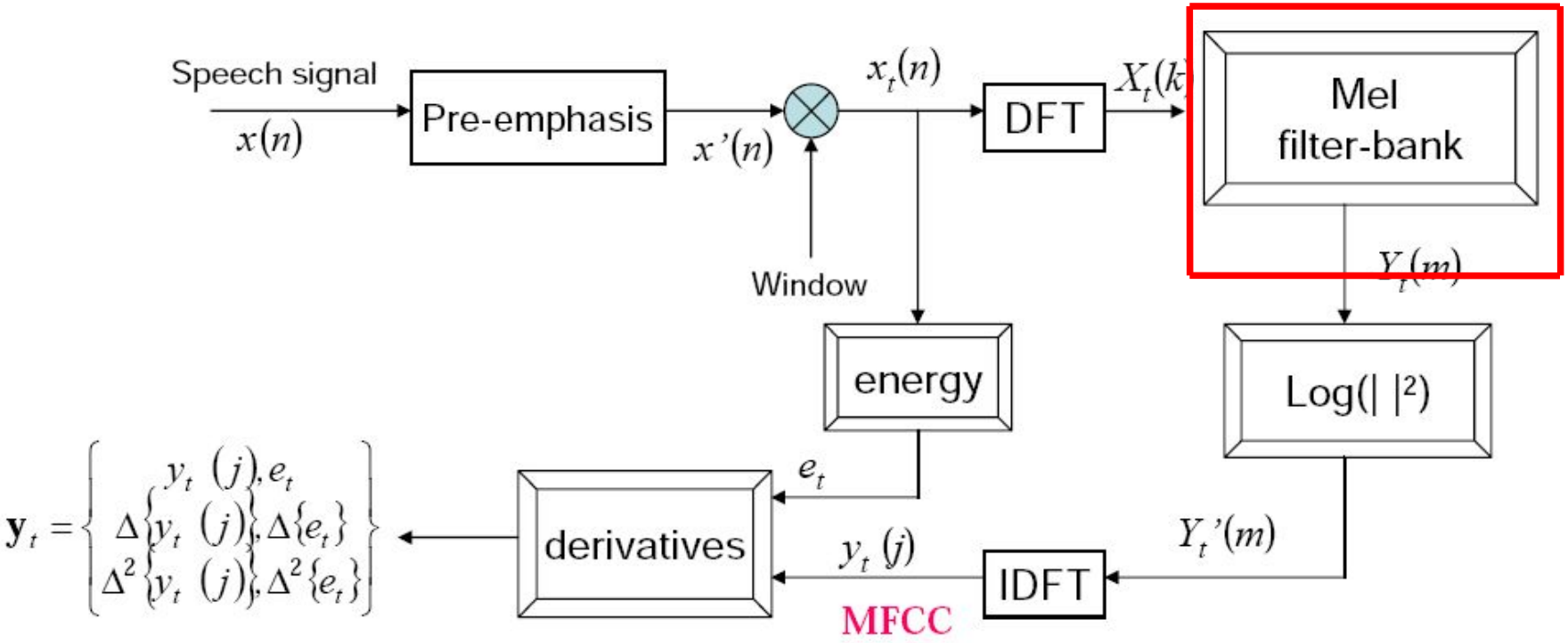


Discrete Fourier Transform

Standard algorithm for computing DFT:

- Fast Fourier Transform (FFT) with complexity N*log(N)
- Typical N=512 or 1024

Steps of MFCC calculation: Mel filter-bank



Steps of MFCC calculation: Mel filter-bank

Triangular bank of filters - Mel filter bank.

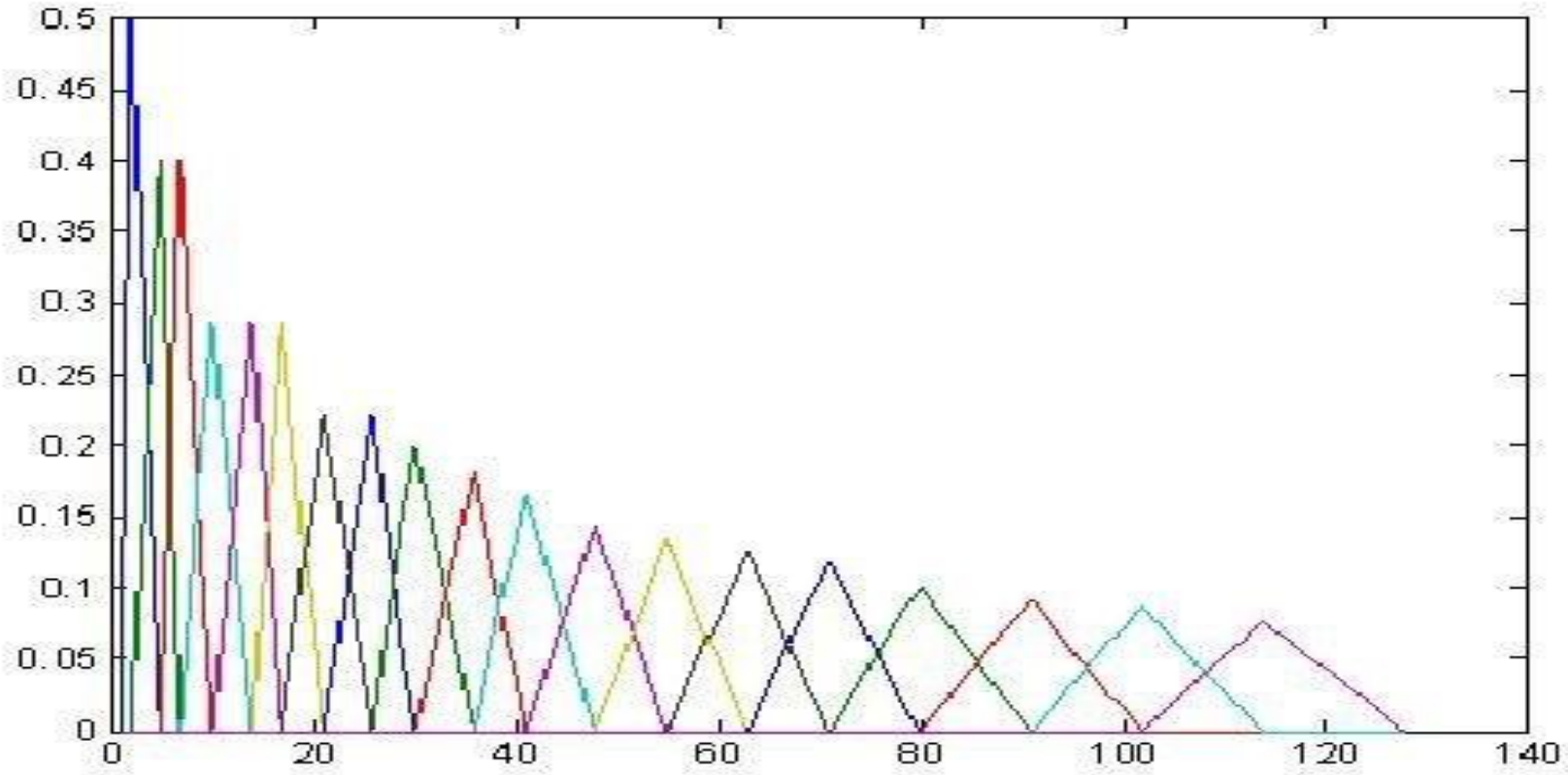
Each filter has a triangular shape, and unity response at its centre.

Its edges coincide with the adjacent filters' central frequencies.

The central frequencies are linearly spaced on the mel frequency scale, which results in an exponential interval between the filter centres onto the linear scale.

Steps of MFCC calculation: Mel filter-bank
Triangular bank of filters

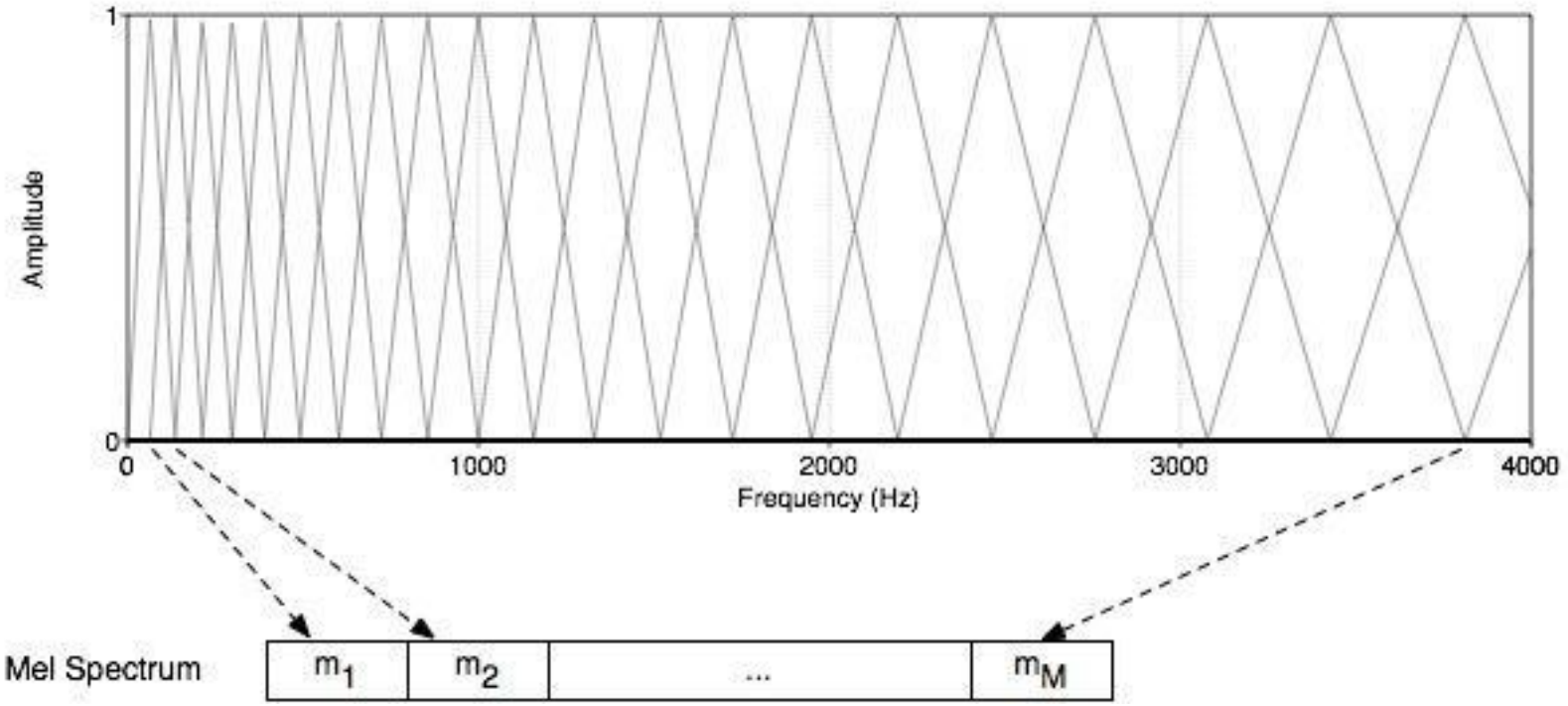
Triangular bank of filters



Steps of MFCC calculation
Mel Filter Bank Processing

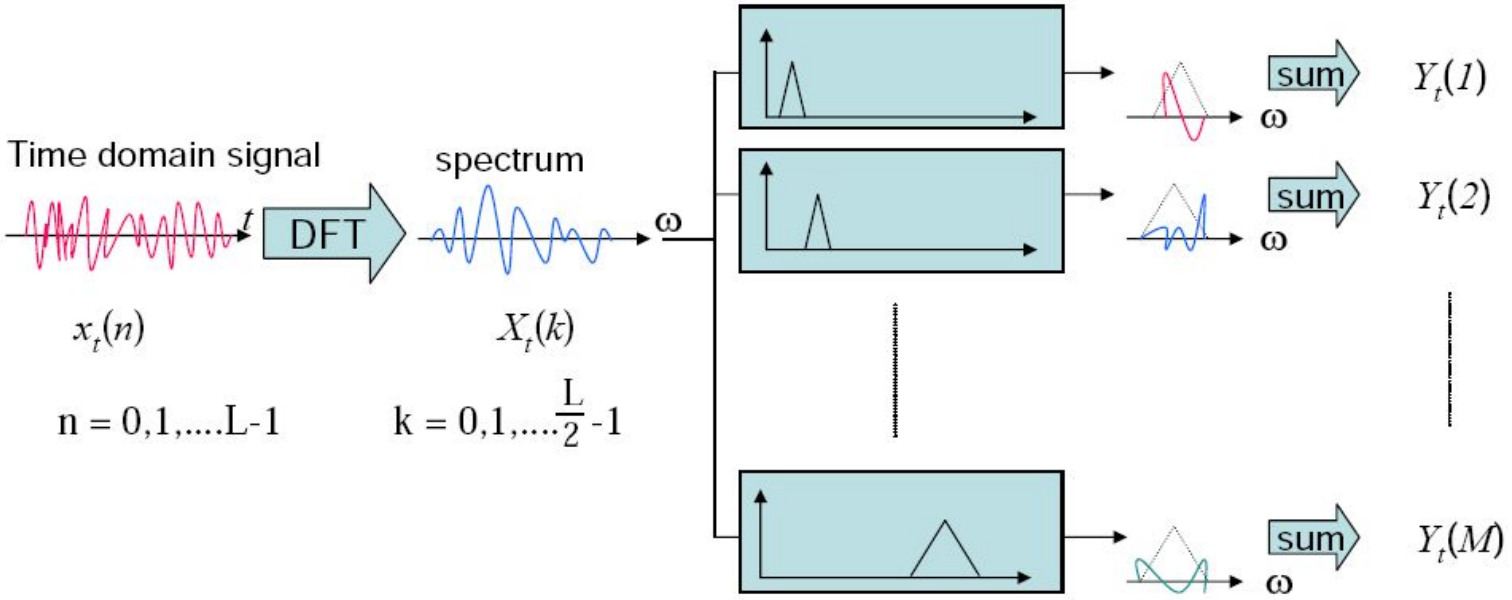
Mel Filter bank

- Uniformly spaced before 1 kHz
- logarithmic scale after 1 kHz

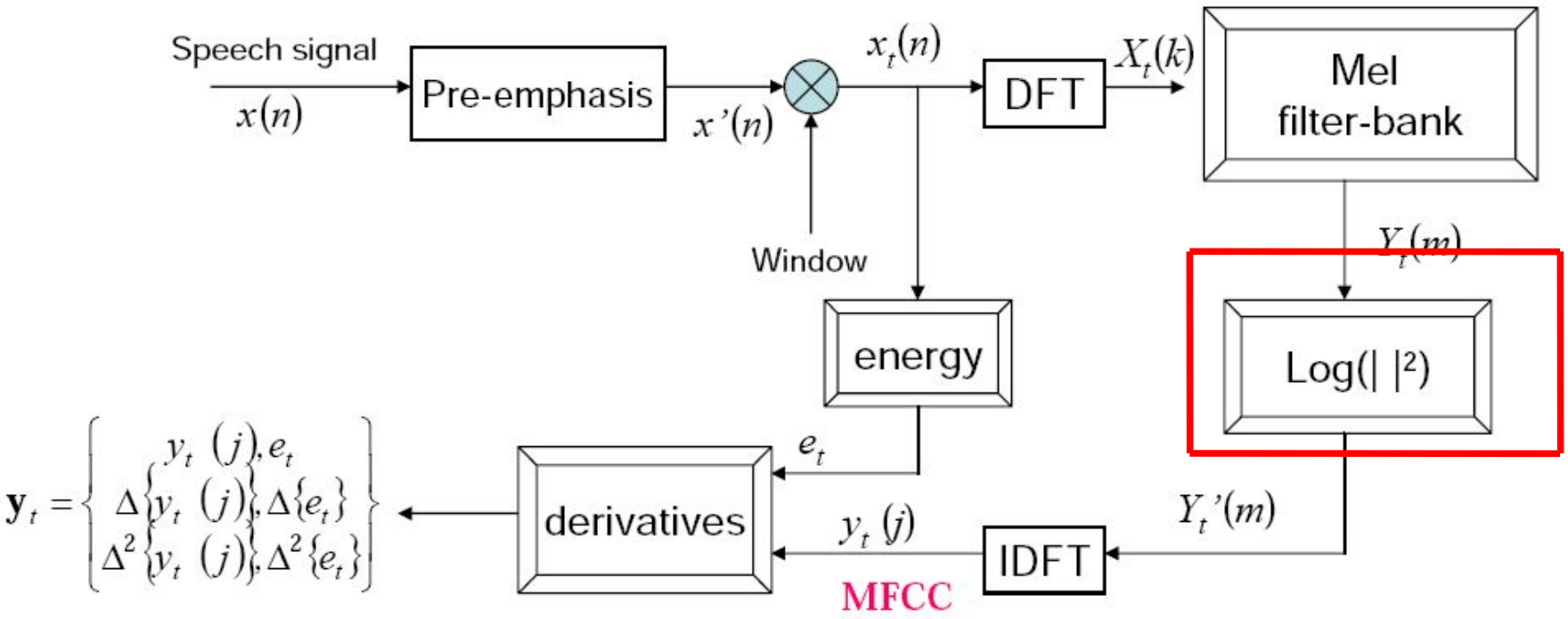


Steps of MFCC calculation
Mel Filter Bank Processing

- Apply the bank of filters according Mel scale to the spectrum
- Each filter output is the sum of its filtered spectral components

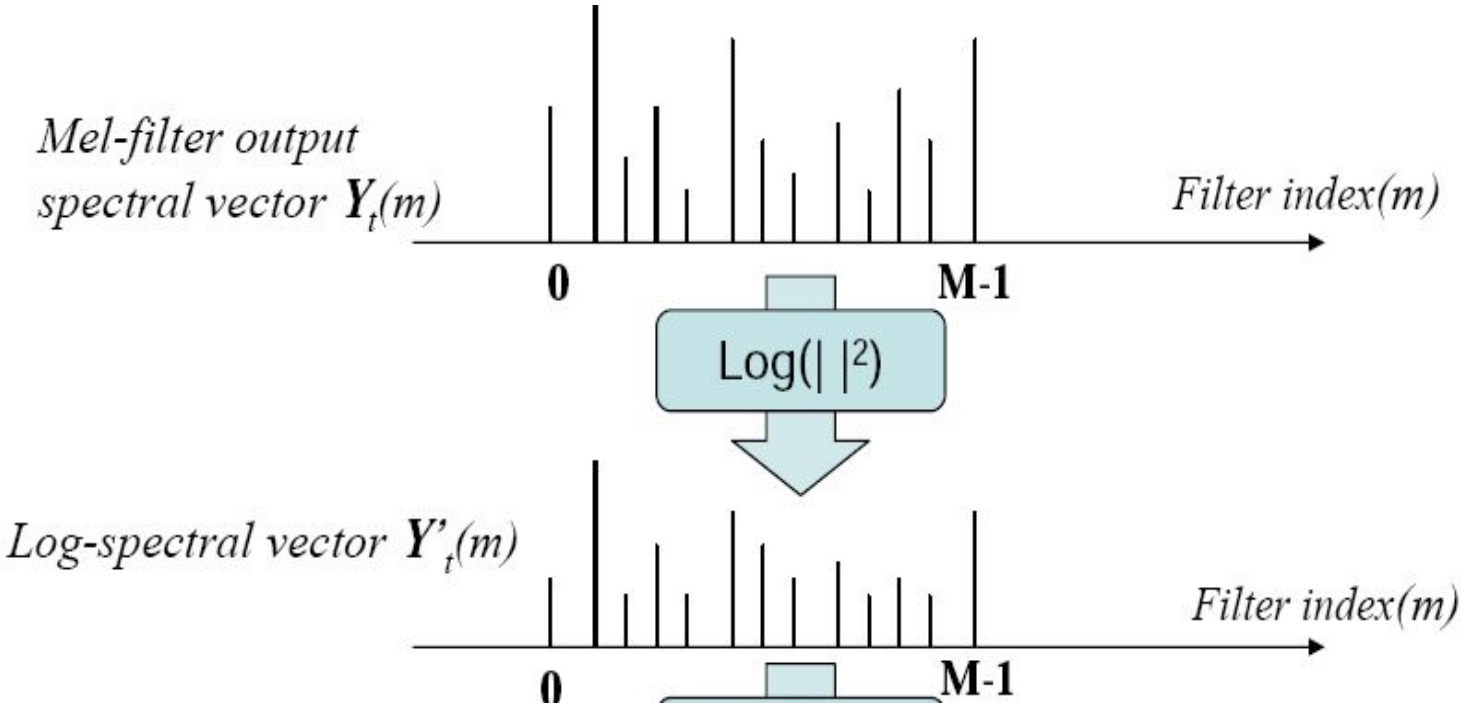


Steps of MFCC calculation: Log energy computation



Steps of MFCC calculation: Log energy computation

Compute the logarithm of the square magnitude of the output of Mel-filter bank



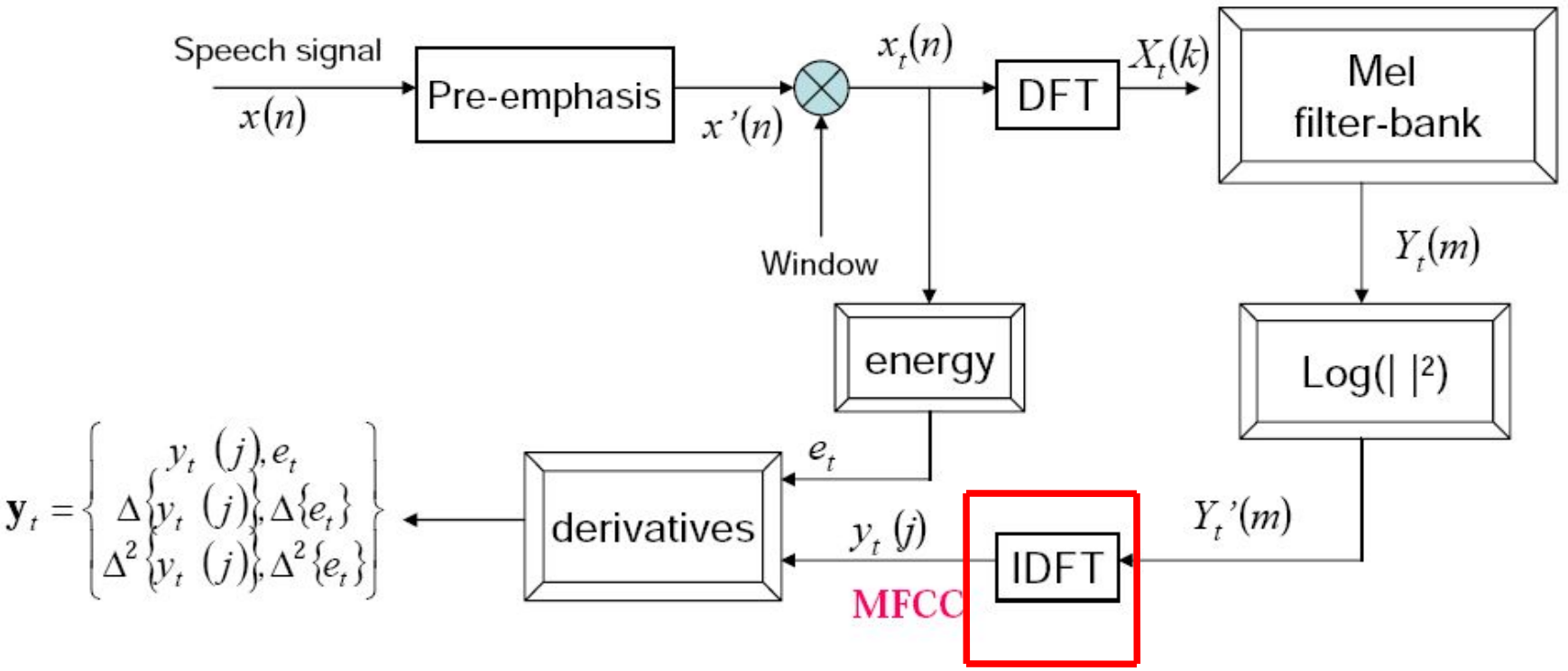
Steps of MFCC calculation: Log energy computation

Why log energy?

- Logarithm compresses dynamic range of values
  - Human response to signal level is logarithmic:

    Humans are less sensitive to slight differences in amplitude at high amplitudes than low amplitudes
- Makes frequency estimates less sensitive to slight variations in input
- Phase information not helpful in speech

Steps of MFCC calculation: IDFT



The Cepstrum –recap.

Intuition:

- Separating the **source** and **filter**
- Speech waveform is created by
  - A glottal source waveform
  - Passes through a vocal tract which because of its shape has a particular filtering characteristic

Articulatory facts:

- The vocal cord vibrations create harmonics
- The mouth is an amplifier

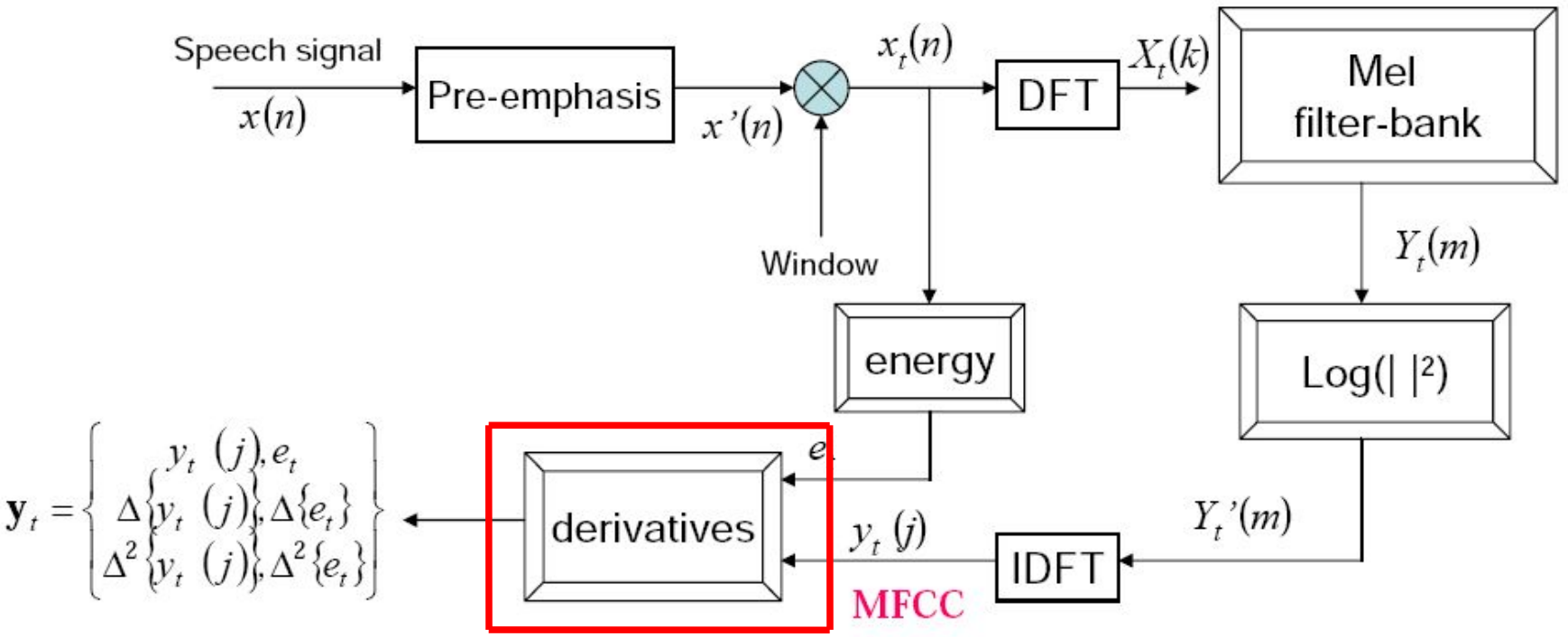
MFCC – Cepstrum

MFCCs cepstrum employ the DCT instead of the IDFT

Since the log power spectrum is real and symmetric instead of the IDFT can be used Inverse Discrete Cosine Transform (IDCT)

In general we'll just use the first 12 cepstral coefficients...

Steps of MFCC calculation: derivatives



Typical MFCC features

Window size: 25ms

Window shift: 10ms

Pre-emphasis coefficient: 0.97

MFCC:

- 12 MFCC (mel frequency cepstral coefficients)
- 1 energy feature
- 12 delta MFCC features
- 12 double-delta MFCC features
- 1 delta energy feature
- 1 double-delta energy feature

Total 39-dimensional features

Why is MFCC so popular?

Efficient to compute

Incorporates a perceptual Mel frequency scale

Separates the source and filter

IDFT(DCT) decorrelates the features

- Improves diagonal assumption in HMM modeling

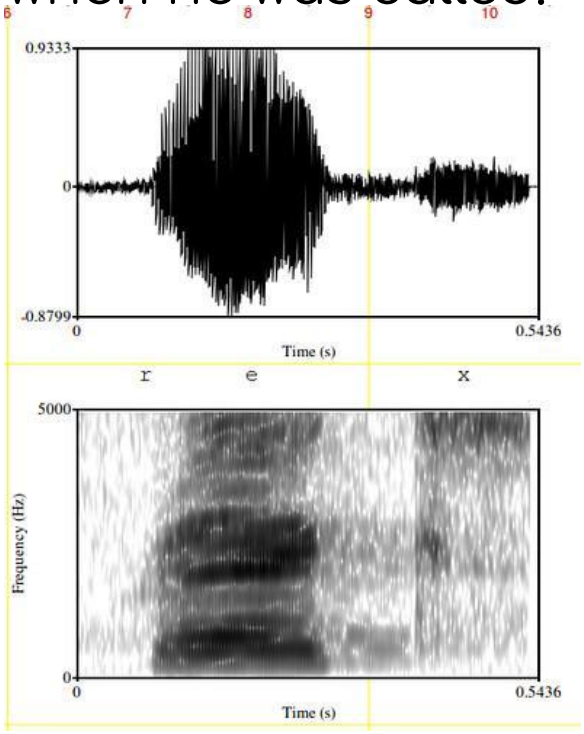
Speech recognition

„Radio Rex“ from 1920s - the first speech recognition machine

The first machine that recognized speech was probably a commercial toy named “Radio Rex” which was sold in the 1920’s.

Rex was a celluloid dog that moved if acoustic energy by 500 Hz had been detected.

Since 500 Hz is roughly the first formant of the vowel [eh] in “Rex”, the dog seemed to come when he was called.

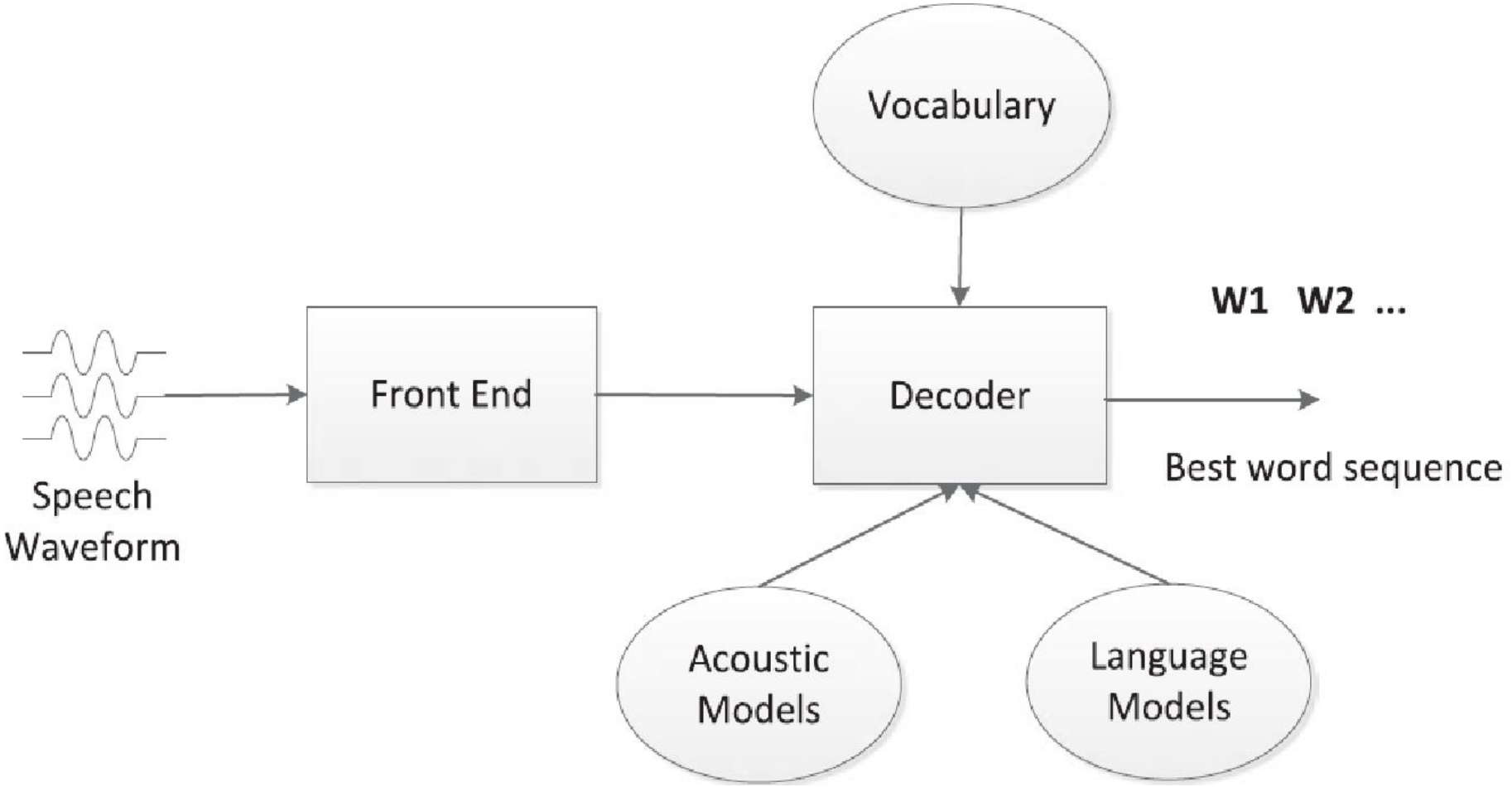


The speech signal and spectrogram of the word "Rex"

Speech recognition

- Isolated word versus continuous speech
- Speaker dependent versus speaker independent systems
- Small versus vocabulary systems

Speech recognition system architecture



Speech recognition

Used methods are based on:

traditional approach - classical way : feature generation + classification

- Mel Frequency Cepstral Coefficients (MFCC) features + Hidden Markov Model (HMM) for classification

- Dynamic Time Warping

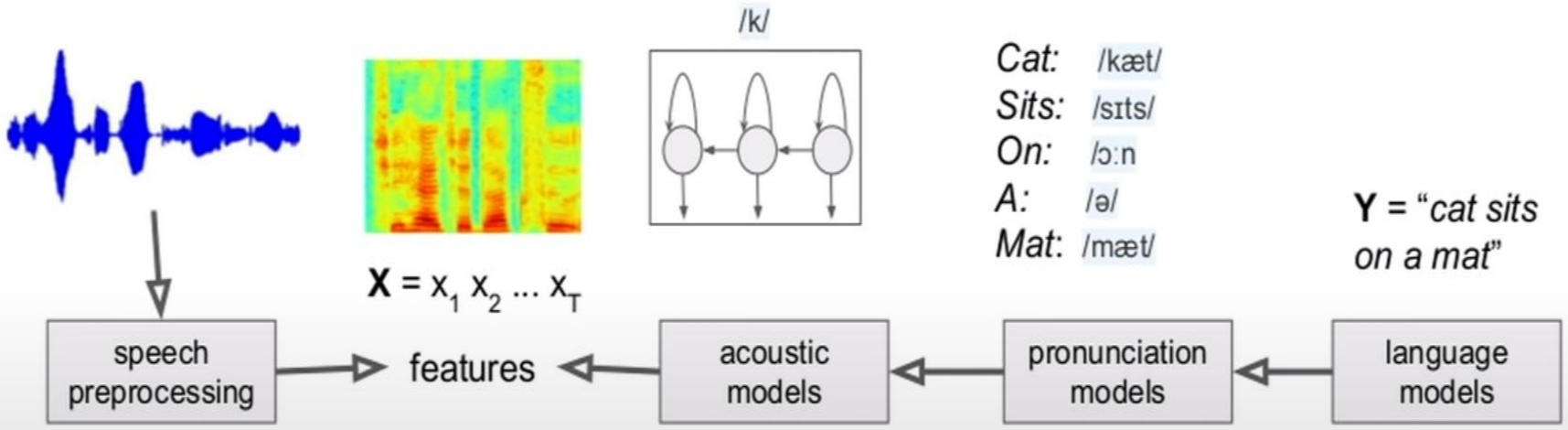
novel approach :

- Methods of Deep Neural Networks

Speech recognition - *traditional approach*

Speech Recognition -- the classical way

- Building a statistical model of speech starting from text sequences $\mathbf{Y} = y_1 y_2 \ldots y_L$ to audio features $\mathbf{X} = x_1 x_2 \ldots x_T$



source: http://stanfordonline.stanford.edu/

Features

- MFCCs are the most common representation in traditional audio signal processing
- log-mel spectrograms are the dominant feature in deep learning (novel approach), followed by raw waveforms or complex spectrograms.

Raw waveforms:

- avoid hand-designed features,
- allow to better exploit the improved modeling capability of deep learning models,
- higher computational costs and data requirements,

Log-mel spectrograms:

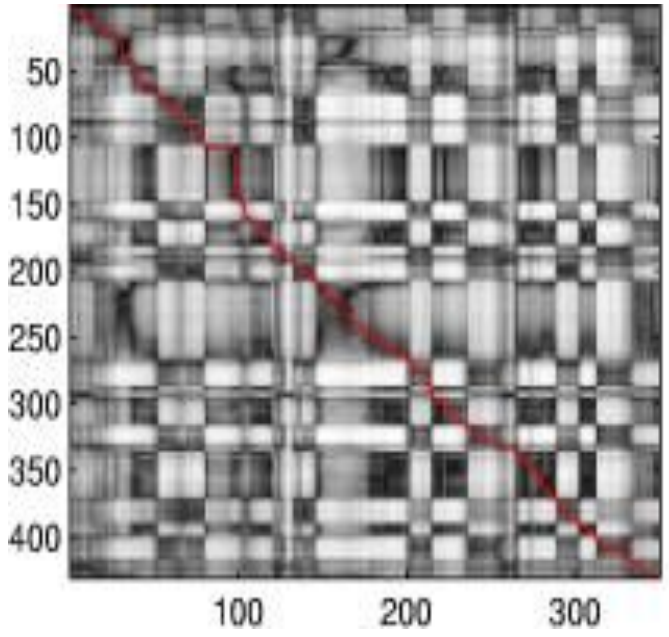
- for analysis tasks, such as ASR, MIR, or environmental sound recognition, log-mel spectrograms provide a more compact representation,
- less data and training to achieve results

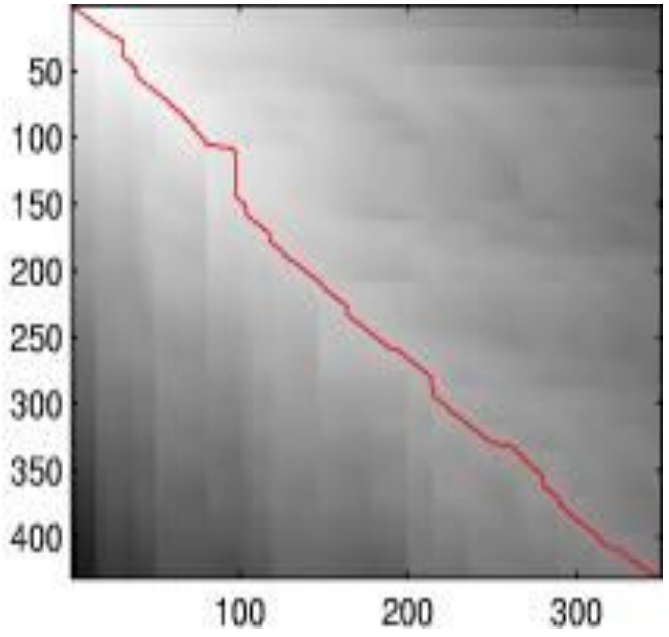
Classification

- Hidden Markov Model (HMM): dominant traditional approach
- Deep Neural Networks : CNNs, RNNs and CRNNs can model temporal sequences, and solve sequence classification, sequence labelling and sequence transduction tasks.
  - CNNs (Convolutional Neural Networks)
    - have a fixed receptive field, which limits the temporal context taken into account for a prediction
    - makes it very easy to widen or narrow the context used.
  - RNNs (Recurrent Neural Networks)
    - can base their predictions on an unlimited temporal context
    - require processing the input sequentially, making them slower to train and evaluate on modern hardware than CNNs.
  - CRNNs (Convolutional Recurrent Neural Networks)
    - offer a compromise in between, inheriting both CNNs and RNNs advantages and disadvantages.

Dynamic Time Warping

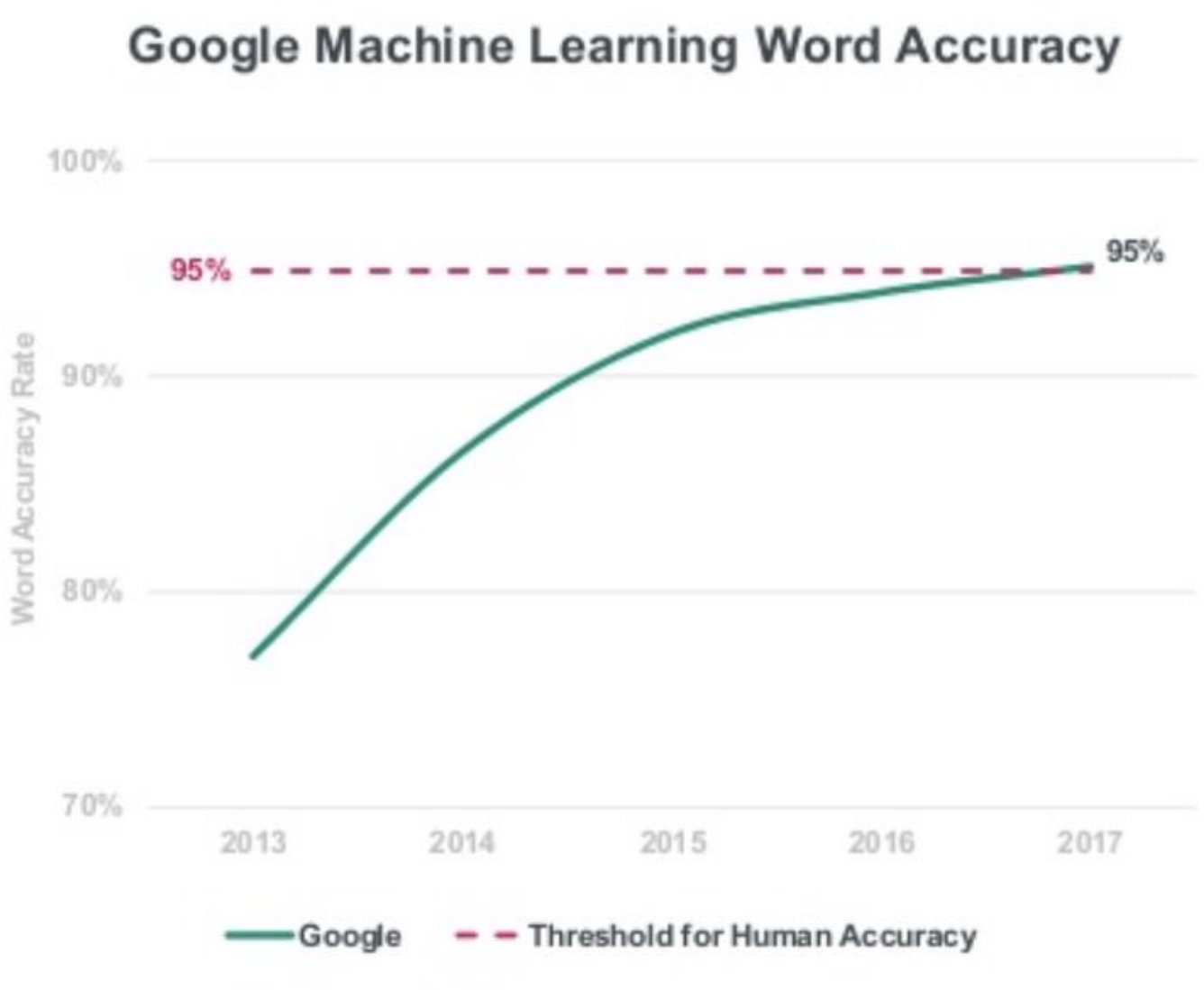
- Create a similarity matrix for the two utterances
- Use dynamic programming to find the lowest cost path



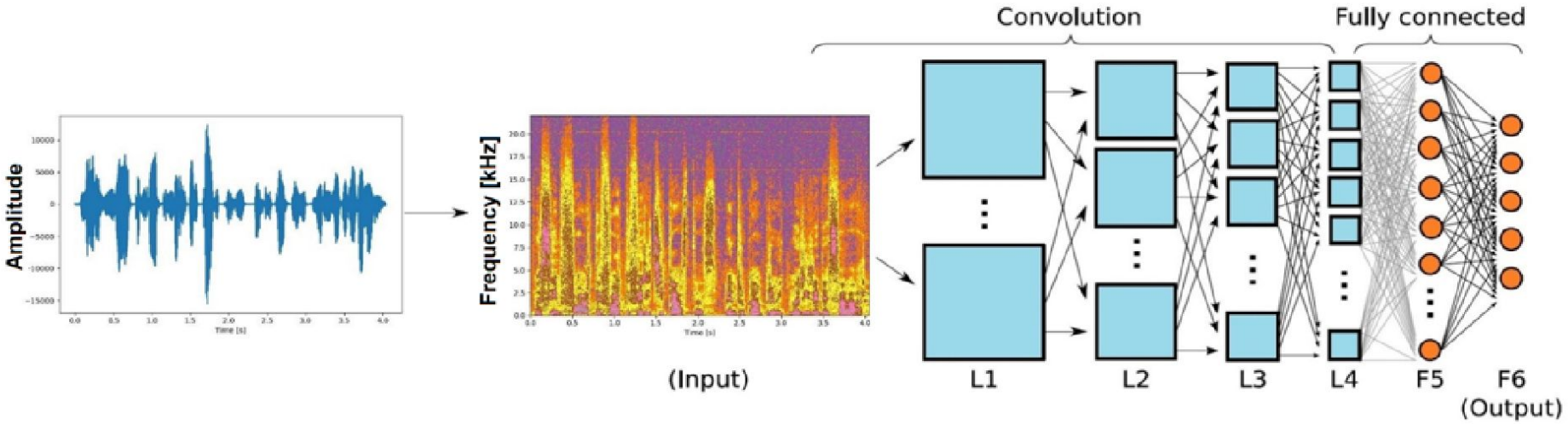


Speech recognition using deep learning

Voice Recognition Accuracy



Speech recognition using CNN - example



Convolutional Neural Network

Speech recognition using RNN - example

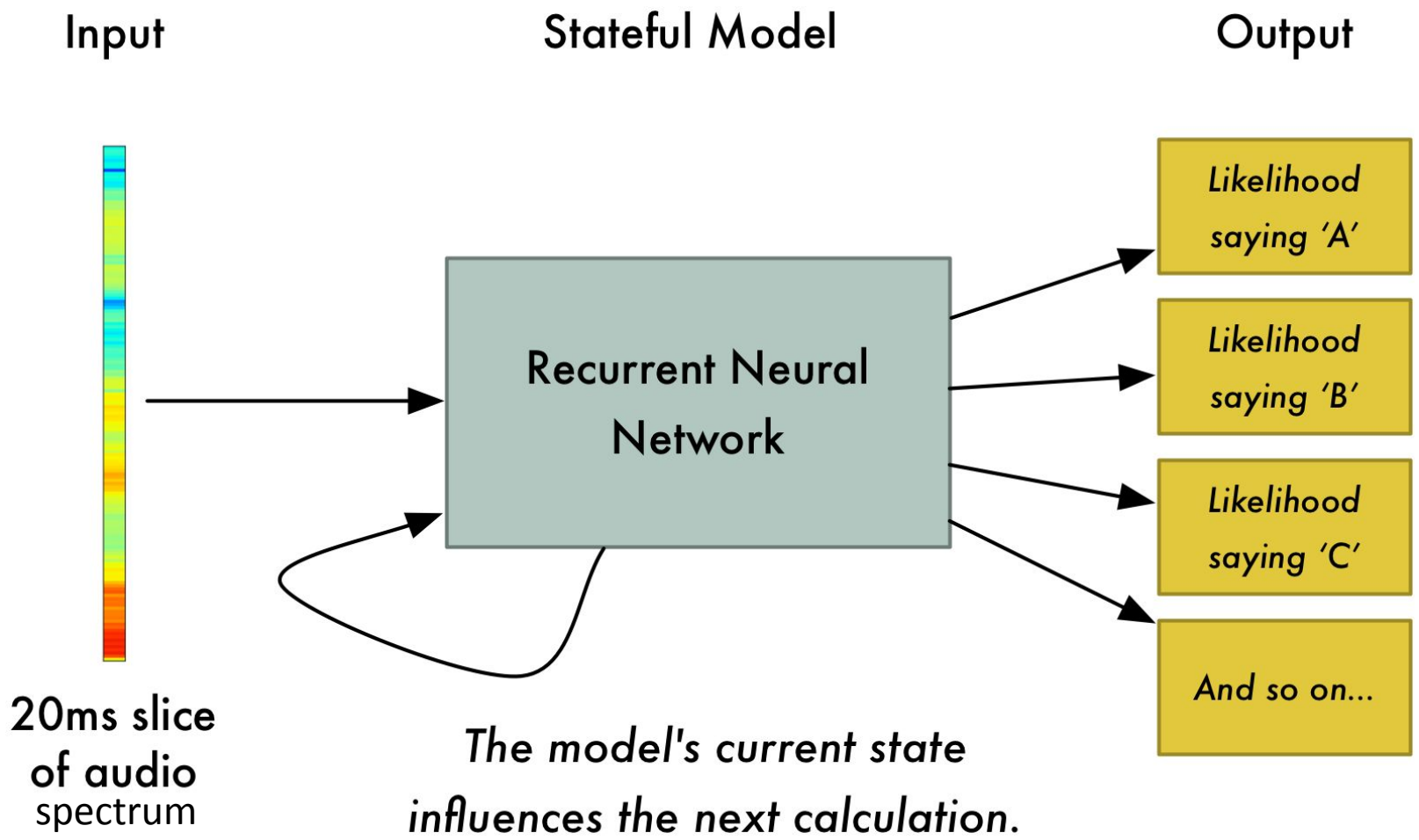


Image from:https://medium.com/@ageitgey/machine-learning-is-fun-part-6-how-to-do-speech-recognition-with-deep-learning-28293c162f7a

Speech recognition using LSTM, Reinforcement learning and more...

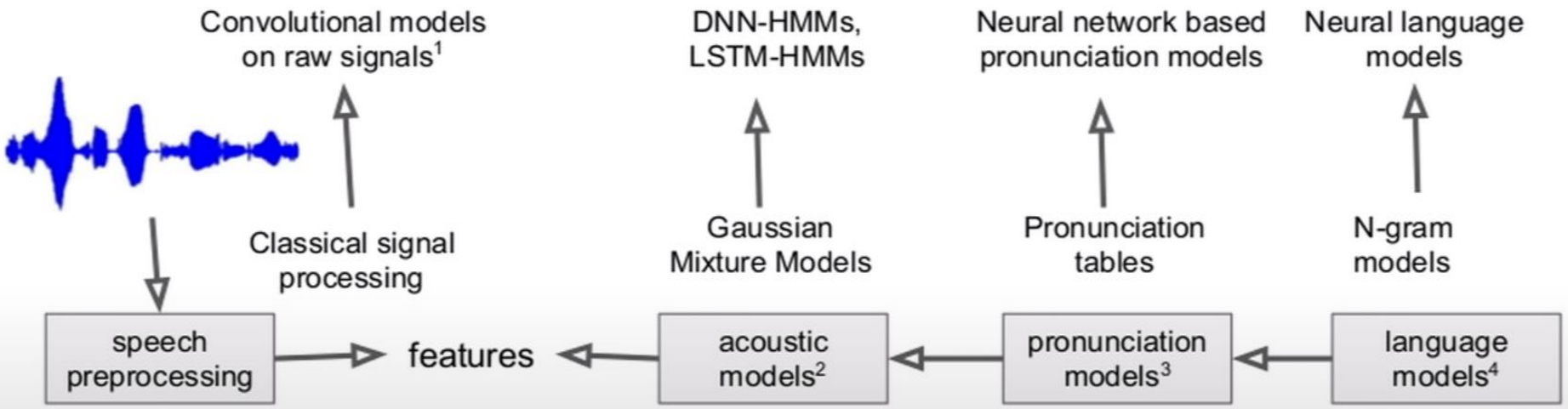
Speech recognition using deep larning:

LSTM Long Short-Term Memory
Reinforcement learning

Speech Recognition - DNN model

Speech Recognition -- the neural network invasion

- Each of the components seems to be better off with a neural network



1. Jaitly, Navdeep, and Geoffrey Hinton. "Learning a better representation of speech soundwaves using restricted boltzmann machines." *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011.
2. Hinton, Geoffrey, et al. "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups." *IEEE Signal Processing Magazine* 29.6 (2012): 82-97.
3. Rao, Kanishka, et al. "Grapheme-to-phoneme conversion using long short-term memory recurrent neural networks." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.
4. Mikolov, Tomas, et al. "Recurrent neural network based language model." *Interspeech*. Vol. 2. 2010.

source: http://stanfordonline.stanford.edu/

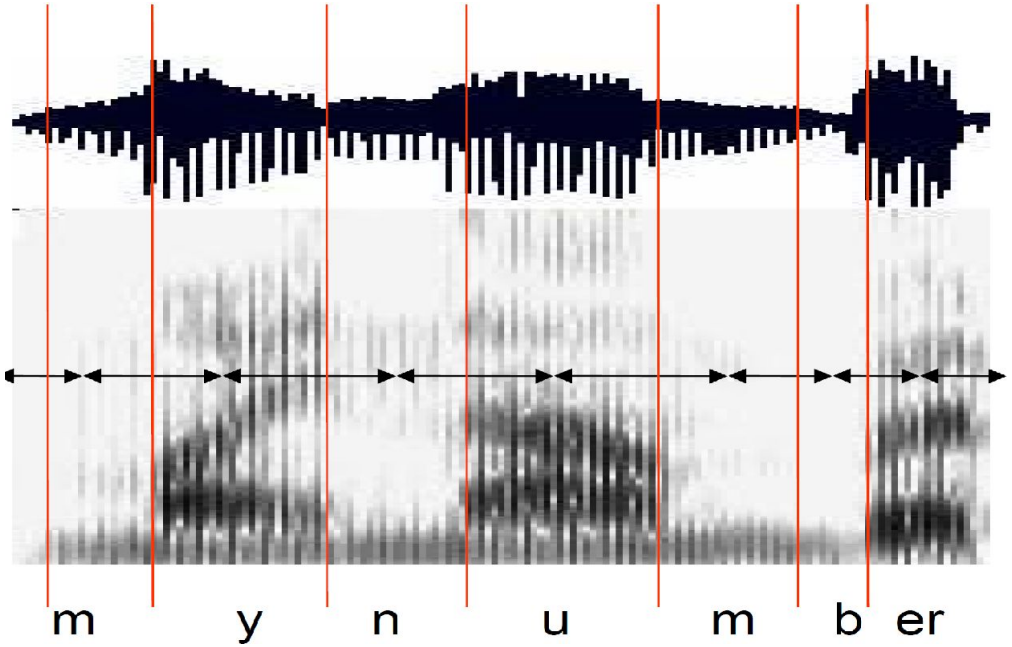
Speech synthesis

Speech Synthesis - Unit selection

Diphone / triphone-based speech units

In phonetics, a diphone is an adjacent pair of phones.

It is usually used to refer a recording of the transition between two phones.



Speech generation with Neural Networks

- WaveNet, a deep generative model of raw audio waveforms
- Talking Machines
- WaveNets are able to generate speech which mimics any human voice and which sounds more natural than the best existing Text-to-Speech systems, reducing the gap with human performance by over 50%.

Readings

https://www.inf.ed.ac.uk/teaching/courses/asr/lectures-2023.html