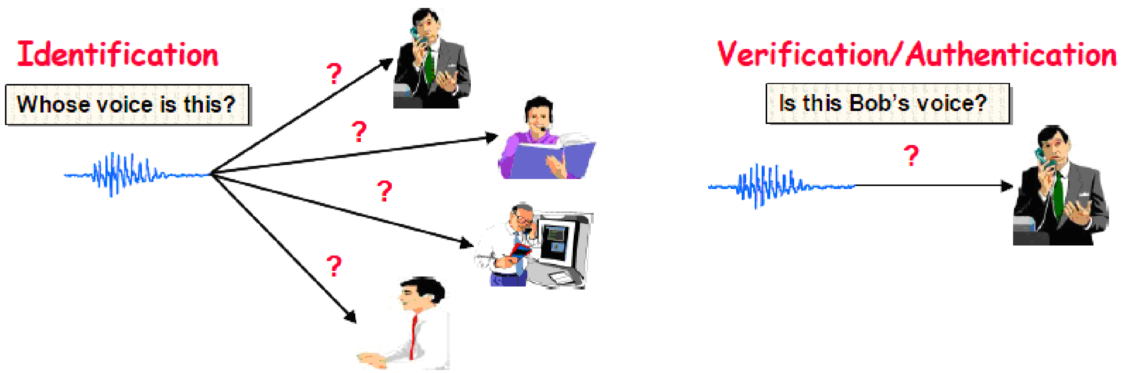
FEATURE EXTRACTION
MEL FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

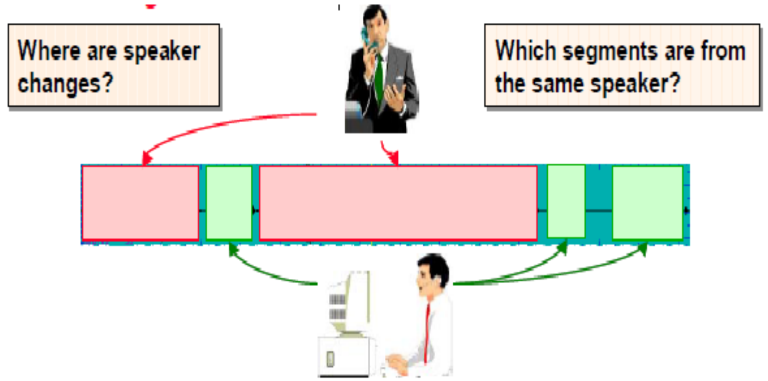
MUSTAFA YANKAYIŞ

CONTENTS

SPEAKER RECOGNITION

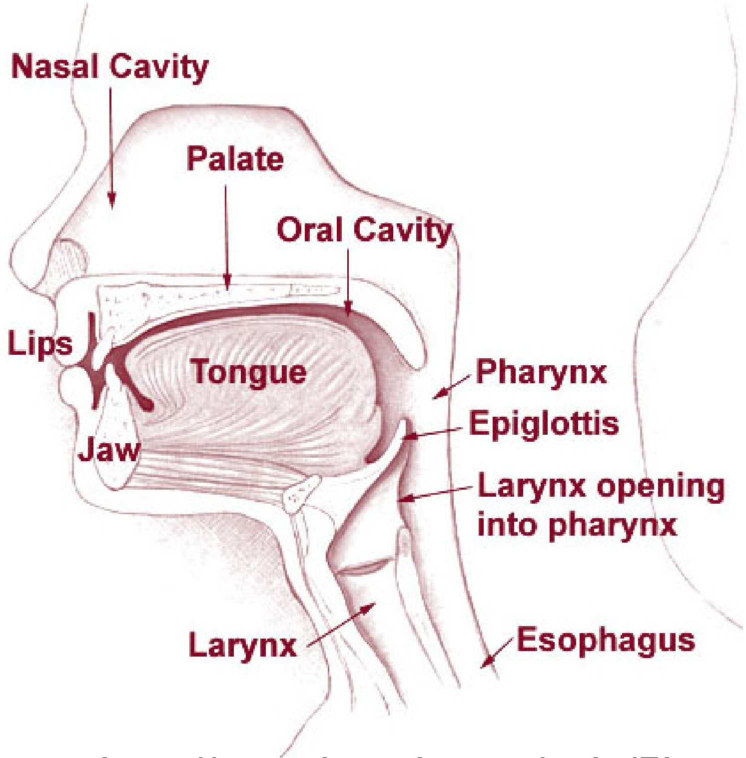
- Speaker recognition has two major tasks;
  - Speaker identification
  - Speaker verification
  - Speaker diarization
- Speaker recognition methods can be divided into;
  - Text – dependent
  - Text – independent





SPEECH PRODUCTION

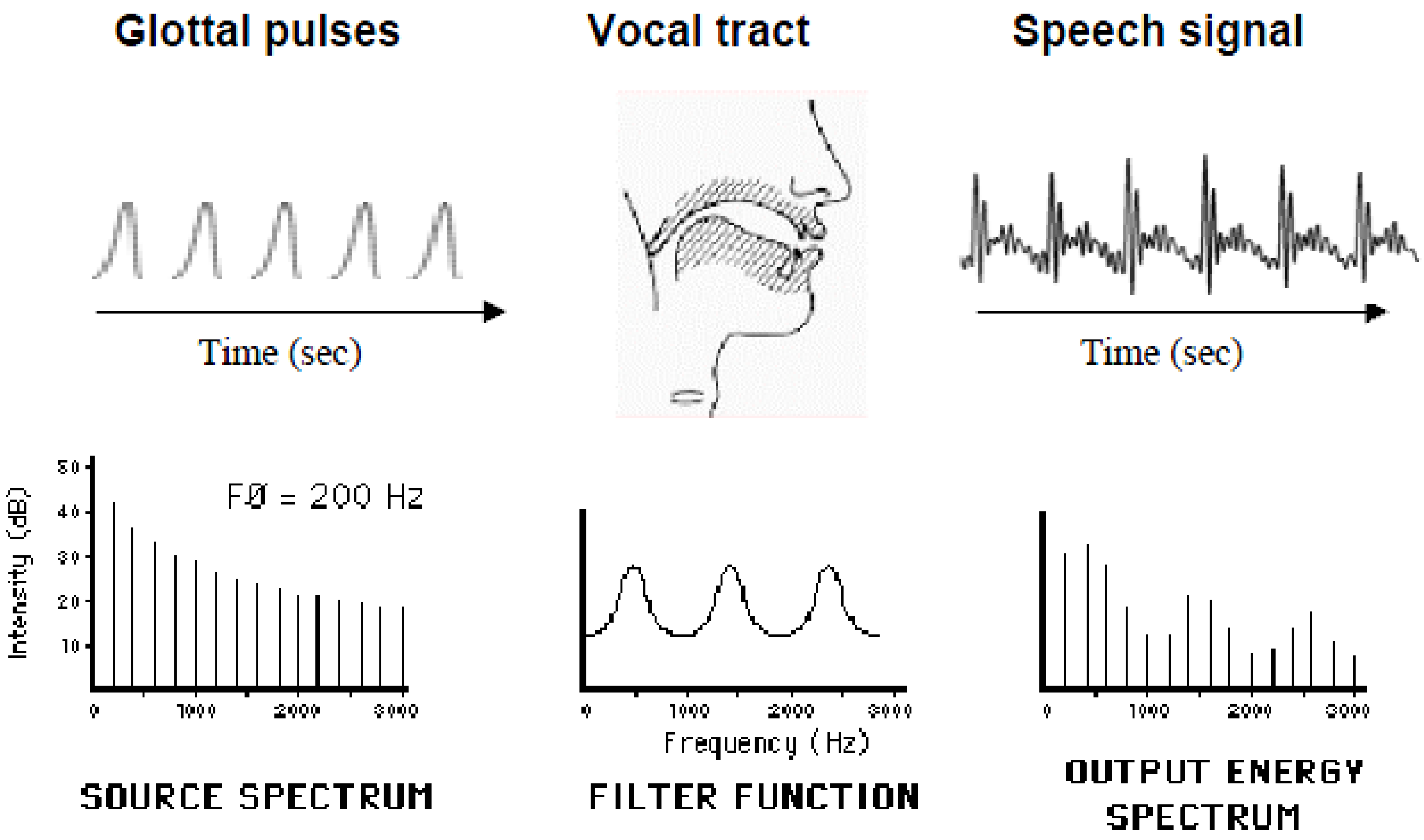
- The **vocal folds** are the sources for speech production in humans.
  - **Voiced:** The vocal folds vibrate
  - **Unvoiced:** The vocal folds do not vibrate.
- The **vocal tract** is the speech production organs above the vocal folds, which consist of the oral tract (tongue, pharynx, palate, lips, and jaw) and the nasal tract.



http://en.wikipedia.org/wiki/Pharynx

SPEECH PRODUCTION

- When **the glottal pulses signal** generated by **the vibration of the vocal folds** passes through **the vocal tract,** it is modified.
- **The vocal tract** works as **a filter,** and its frequency response depends on **the resonances** of the vocal tract.



FEATURE EXTRACTION

- Feature Extraction: characterization and recognition of **the speaker-specific information** contained in the speech signal.
- The feature extraction process transforms **the raw signal** into **feature vectors** in which **speaker-specific properties** are emphasized and **statistical redundancies** are **suppressed.**
- The signal during training and testing session can be greatly different due to many factors such as people voice **change with time**, **health condition** (e.g. the speaker has a cold), **speaking rate** and also **acoustical noise** and **variation recording environment** via microphone.

IDEAL FEATURES

- Stable over **time**
- Should occur **frequently** and **naturally** in speech
- Should not be susceptible to **mimicry**
- Easy to **measure** extracted speech features
- Shows little fluctuation from one **speaking environment** to another
- **Discriminate between speakers** while **being tolerant** of **intra speaker variabilities**(health, emotion, time…)
- Consistent against **the noise** by the transmission conditions.

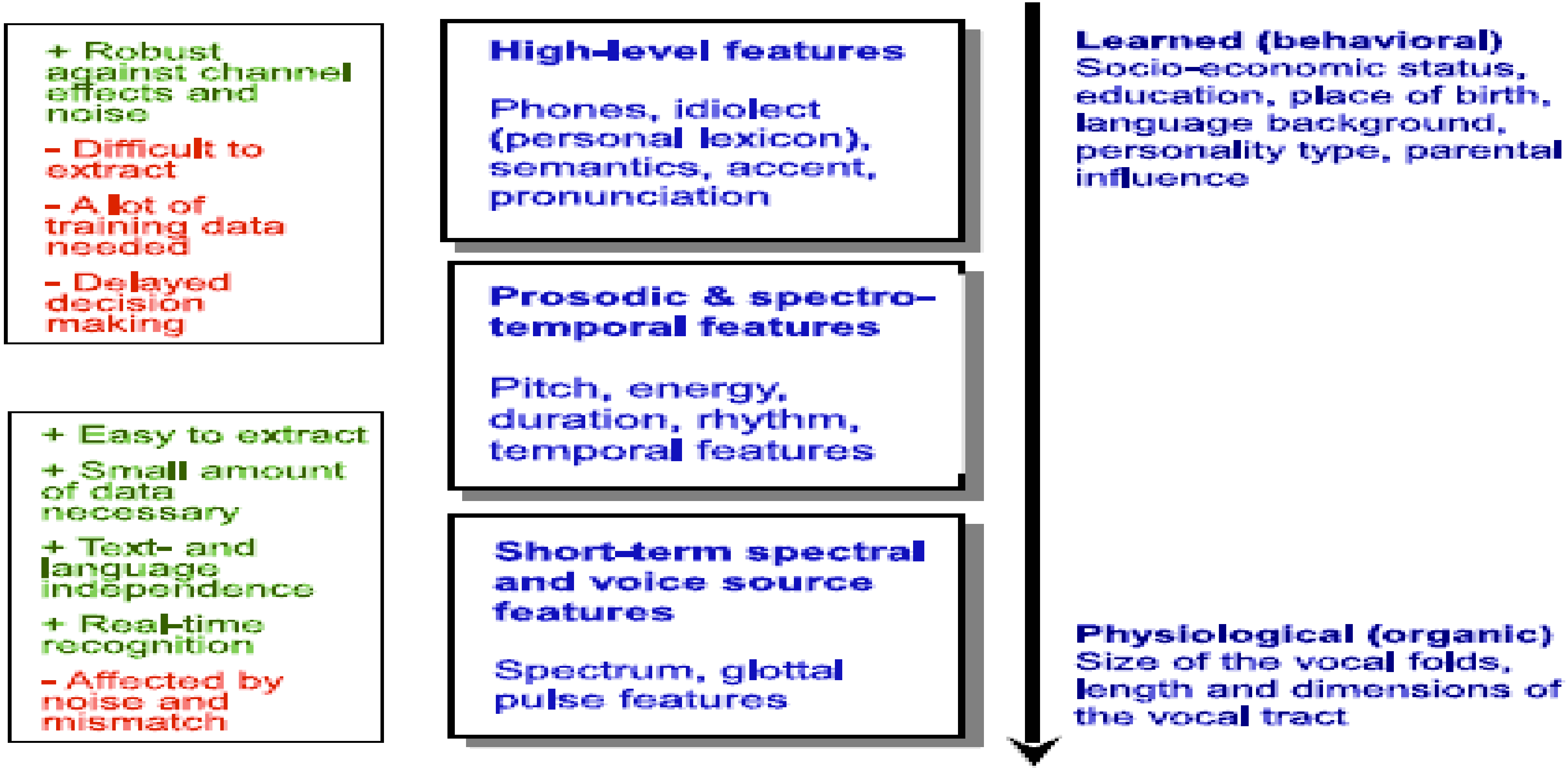
(Wolf (1972))

In practice, to obtain all the desired features simultaneously is very difficult (Reynolds, 1992).

FEATURES

- Short-term spectral features
- Voice source features
- Spectral-temporal features
- Prosodic features
- High-level features

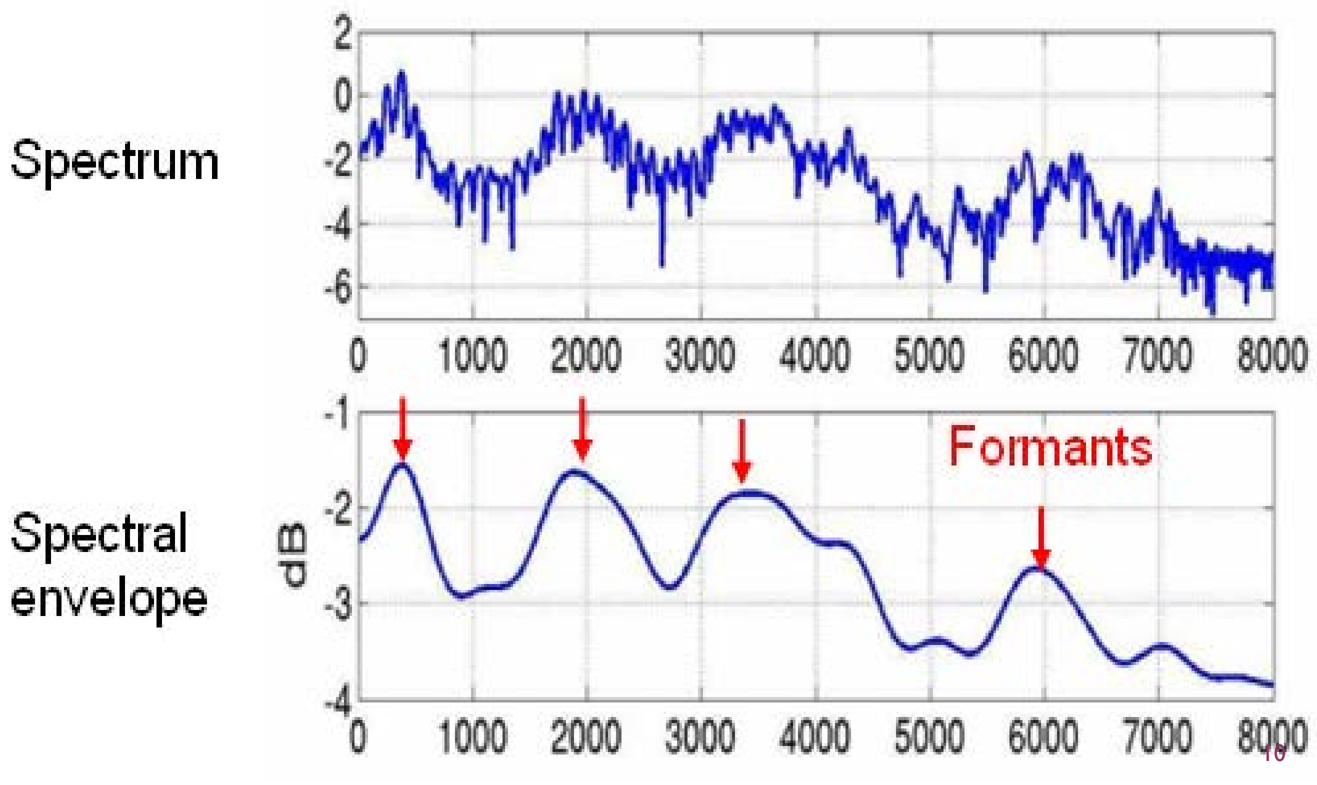
FEATURES



+ Robust against channel effects and noise
– Difficult to extract
– A lot of training data needed
– Delayed decision making
+ Easy to extract
+ Small amount of data necessary
+ Text- and language independence
+ Real-time recognition
– Affected by noise and mismatch
High-level features
Phones, idiolect (personal lexicon), semantics, accent, pronunciation
Prosodic & spectro-temporal features
Pitch, energy, duration, rhythm, temporal features
Short-term spectral and voice source features
Spectrum, glottal pulse features
Learned (behavioral)
Socio-economic status, education, place of birth, language background, personality type, parental influence
Physiological (organic)
Size of the vocal folds, length and dimensions of the vocal tract

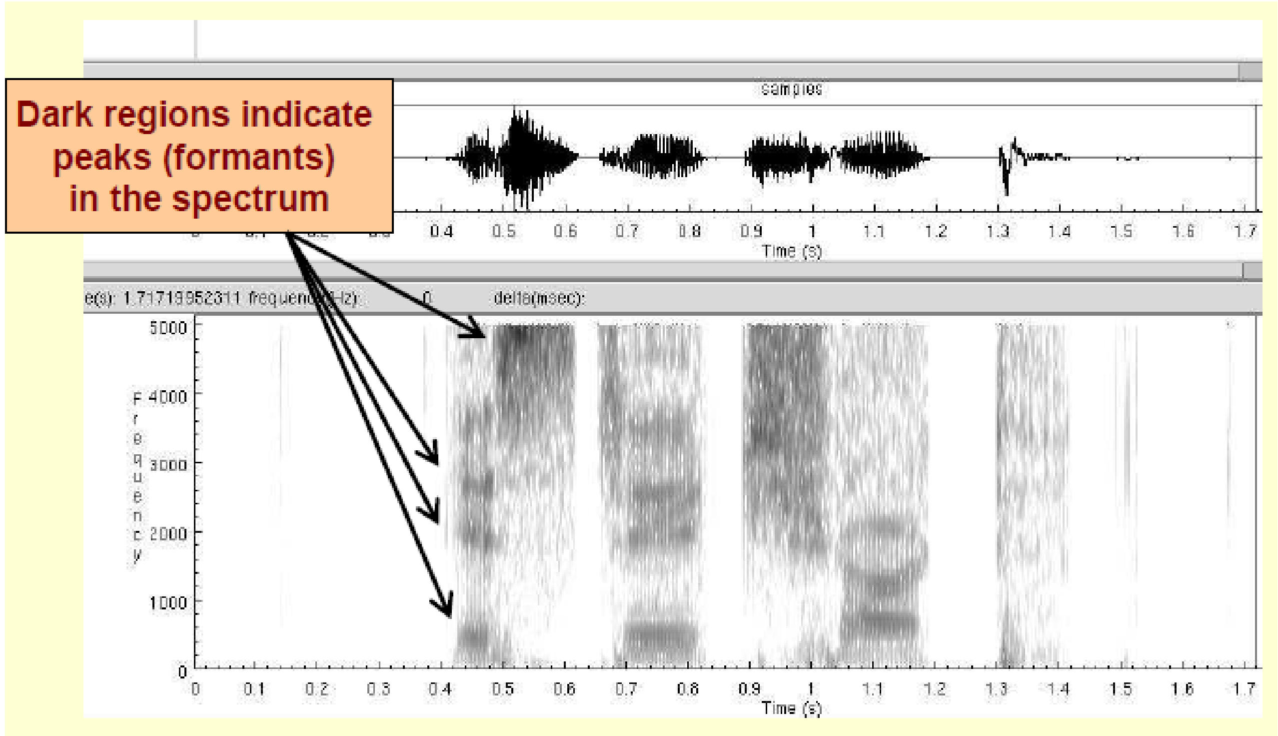
SHORT-TERM SPECTRAL FEATURES

- **Peaks** denote **dominant frequency components** in the speech signal
- **Vocal tract resonances**, also called **formants** are the **peaks** of the spectral envelope.
- The resonance frequencies (**formants**) are inversely proportional to the vocal tract length.
- **Formants** carry **the identity** of the sound

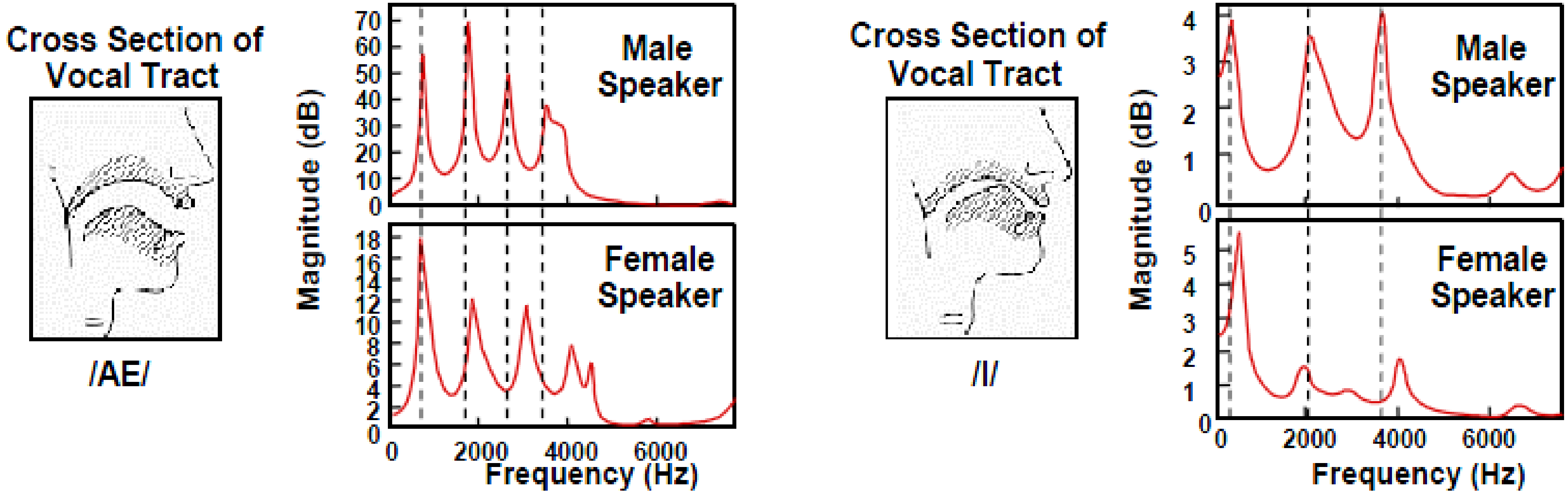


SHORT-TERM SPECTRAL FEATURES

- In speaker recognition, **the features** derived from **the vocal tract** characteristic **are most commonly used**. These features can be obtained from the **spectrogram** of the speech signal, thus are *categorized as* **Short-Term Spectral Features.**
- **Formants** are useful for evaluation of **text to speech** systems.
- **Spectrograms** of synthesized speech (TTS) should nearly **match** with natural sentences.



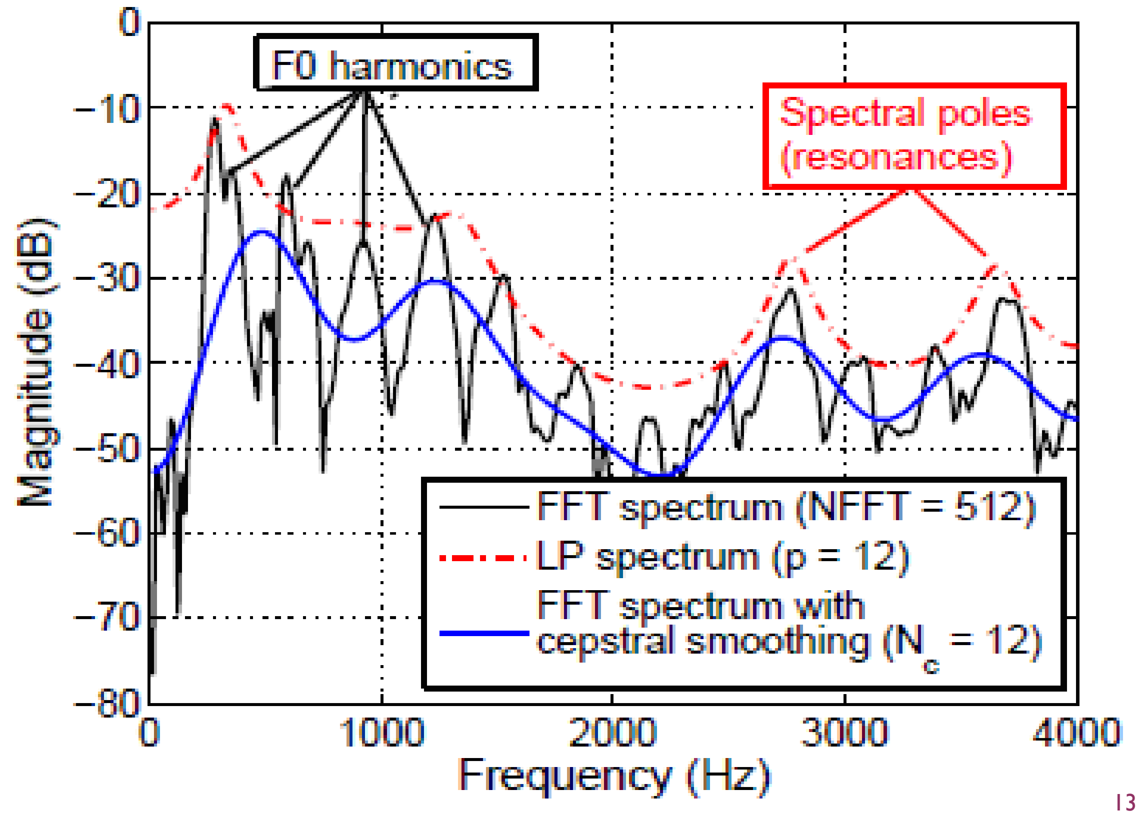
SHORT-TERM SPECTRAL FEATURES



- **Different speakers** will have **different spectra** for similar sounds
- Information of **the spectral envelope.** The speaker's vocal tract characteristics, the location and magnitude of **the peaks (formants)** in the spectrum.
- **Commonly used** for speaker recognition.
- Figure shows the spectral envelopes of two different speakers (one male, one female).

SHORT-TERM SPECTRAL FEATURES

- Linear Predictive Cepstral Coefficients(LPCC)
- Mel-Frequency Discrete Wavelet Coefficient(MFDWC)
- Mel-Frequency Cepstral Coefficients(MFCC)

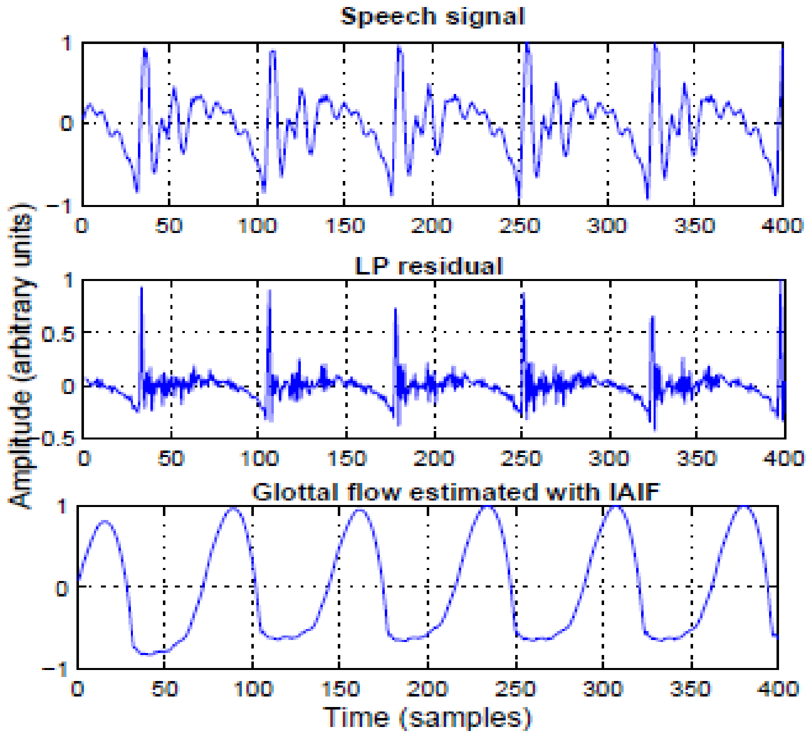


VOICE SOURCE FEATURES

- The features characterize the vocal folds oscillation are called **voice source features.**
- **The vibration** of **the vocal folds** depends on **the tension** exerted by the muscle, and **the mass** and **length** of **the vocal folds**.
- These characteristics vary between speakers, thus can be utilized for speaker recognition.
- Voice source features characterize the voice source (glottal pulses signal), such as **glottal pulse shape** and **fundamental frequency.**
- These features **cannot be directly measured** from **the speech signal**, because the voice source signal is **modified** when **passing though the vocal tract.**

VOICE SOURCE FEATURES

- **The voice source signal** is **extracted** from the speech signal by assuming **the voice source** and **the vocal tract** are **independent** of each other.
- Then **the vocal tract filter** can be first estimated using **the linear prediction model(LPC).**
- The voice source signal can be estimated by **inverse filtering** the speech signal. Here **S(z)** is **the speech signal, E(z)** is **the voice source signal**, and **H(z)** is **the response of the vocal tract filter.**



$$E(z) = S(z) \cdot \frac{1}{H(z)}$$

VOICE SOURCE FEATURES

- **The voice source features** depend on the source of the speech, namely **the pitch** generated by the vocal folds, so they are **less sensitive** to the content of speech than **short-term spectral features**, like **MFCCs features**.
- **The voice source features** are **not as discriminative as vocal tract features**, but **fusing** these two complementary features (short-term spectral features and voice source features) can **improve recognition accuracy**.
- **Wavelet Octave Coefficients of Residues (WOCOR)**

THE OTHER FEATURES

- **Spectral-temporal features**
  - **Formant transitions** and **energy modulations**.
- **Prosodic features**
  - **In linguistics,** prosody refers to **syllable stress**, **intonation patterns**, **speaking rate** and **rhythm of speech**.
- **High-level features:**
  - **Conversation-level features** of speakers, such as **speaker's characteristic vocabulary, the kind of words** the speakers tend to use in their conversations, called **idiolect**

FEATURE EXTRACTION METHODS

- Linear Predictive Coding (LPC)
- Mel-Frequency Discrete Wavelet Coefficient (MFDWC)
- Mel Frequency Cepstral Coefficients (MFCCs)

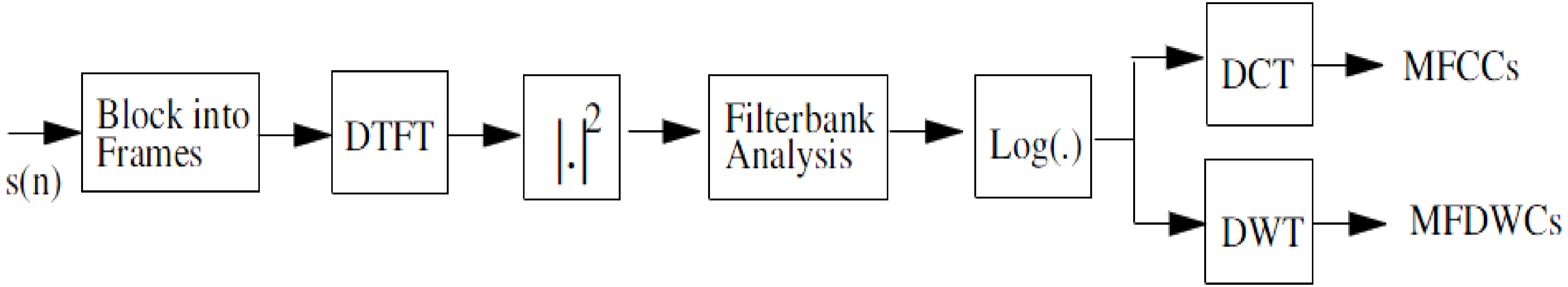
LINEAR PREDICTIVE CODING(LPC)

- **Linear prediction Coding (LPC)** is an alternative method for **spectral envelope estimation.**
- LPC is based on **the source-filter model** of speech production.
- The signal ***s*[*n*]** is predicted by a linear combination of its past values.

$$\tilde{s}[n] = \sum_{k=1}^{p} a_k s[n-k]$$

- However, unlike **MFCC**, the **LPCC** are **not based on perceptual frequency scale**, such as Mel-frequency scale.

MEL-FREQUENCY DISCRETE WAVELET COEFFICIENT (MFDWC)



Extraction of the MFCCs and MFDWCs

- Mel-Frequency Discrete Wavelet Coefficients are computed in **the similar way** as **the MFCC features.** The only difference is that a **Discrete Wavelet Transform (DWT)** is used to replace **the DCT** in the last step.
- **MFDWCs** were used in speaker verification, and it was shown that they give **better performance** than **the MFCCs in noisy environments**. An explanation for this improvement is **DWT** allows good localization both in time and frequency domain.

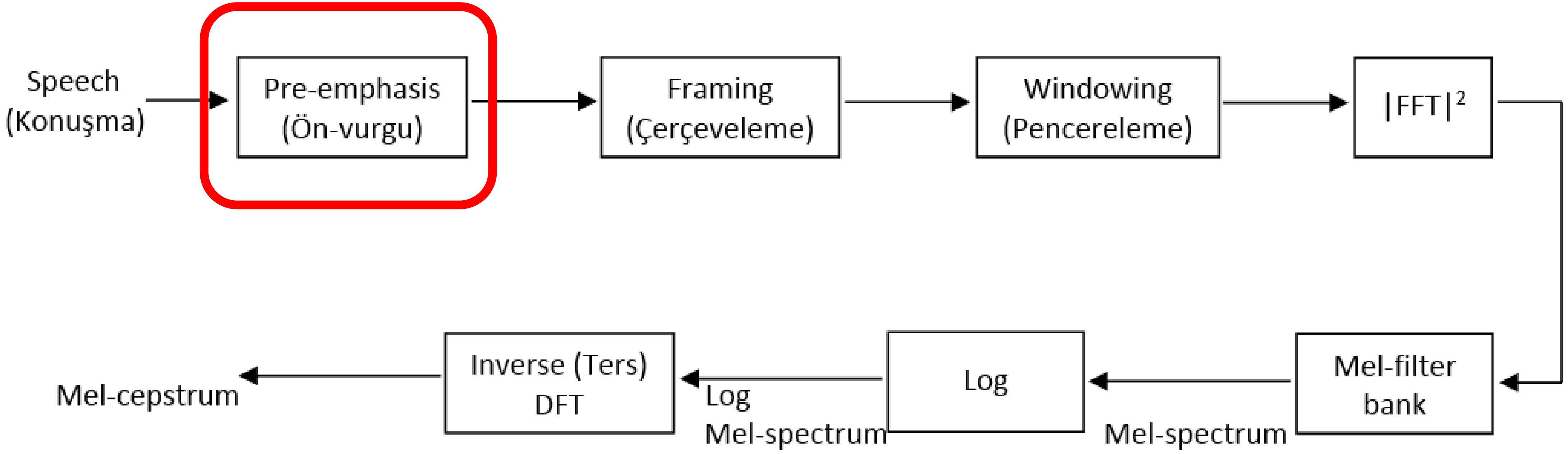
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)

- The Mel-Frequency Cepstral Coefficients (MFCC) features is **the most commonly used features** in speaker recognition.
- It combines the advantages of **the cepstrum analysis** with a **perceptual frequency scale based on critical bands.**

MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

- MFCC is based on human hearing perceptions which cannot perceive frequencies over 1Khz.
- In other words, in MFCC is based on known variation of the human ear's critical bandwidth with frequency.
- MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz.

MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



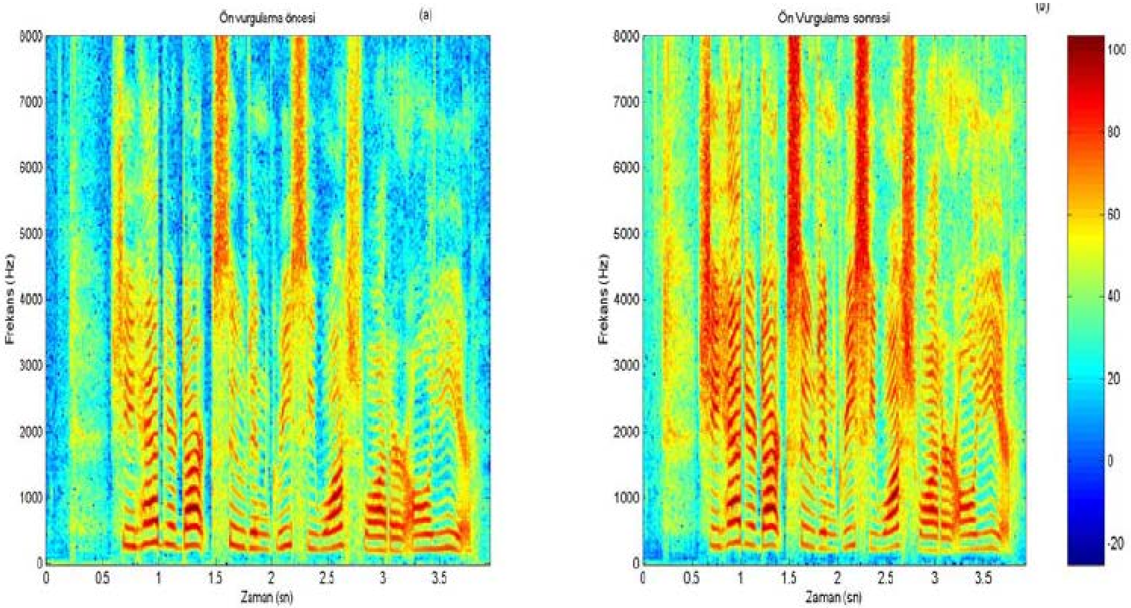
MFCC steps in speech analysis

MFCC – PRE-EMPHASIS

- Time-Frequency representation of a speech signal is referred to as **spectrogram.**
- This step processes the passing of signal through a filter which **emphasizes higher frequencies.** This process will **increase the energy of signal at higher frequency**

$$Y[n] = X[n] - 0.95\, X[n-1]$$

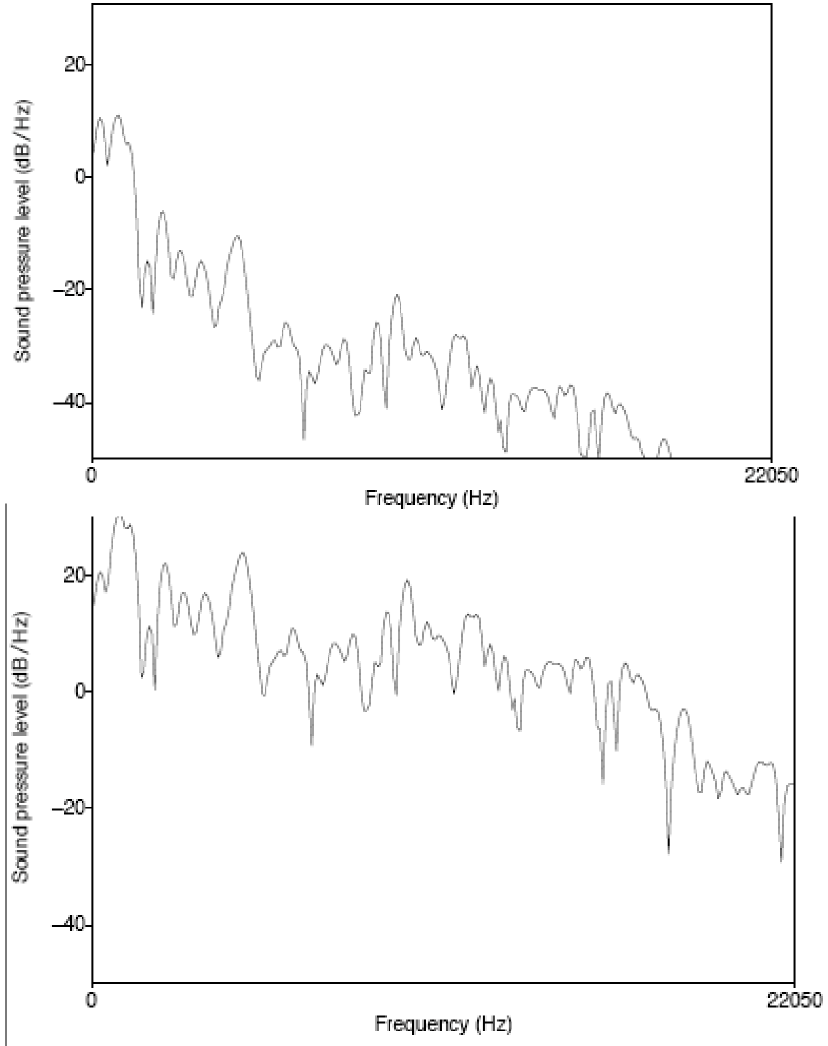
- **Pre-emphasis** is needed because **high frequency components** of the speech signal have **small amplitude** with respect to **low frequency components**.



High frequnecy components in **figure b** are more explicit than those in **figure a**

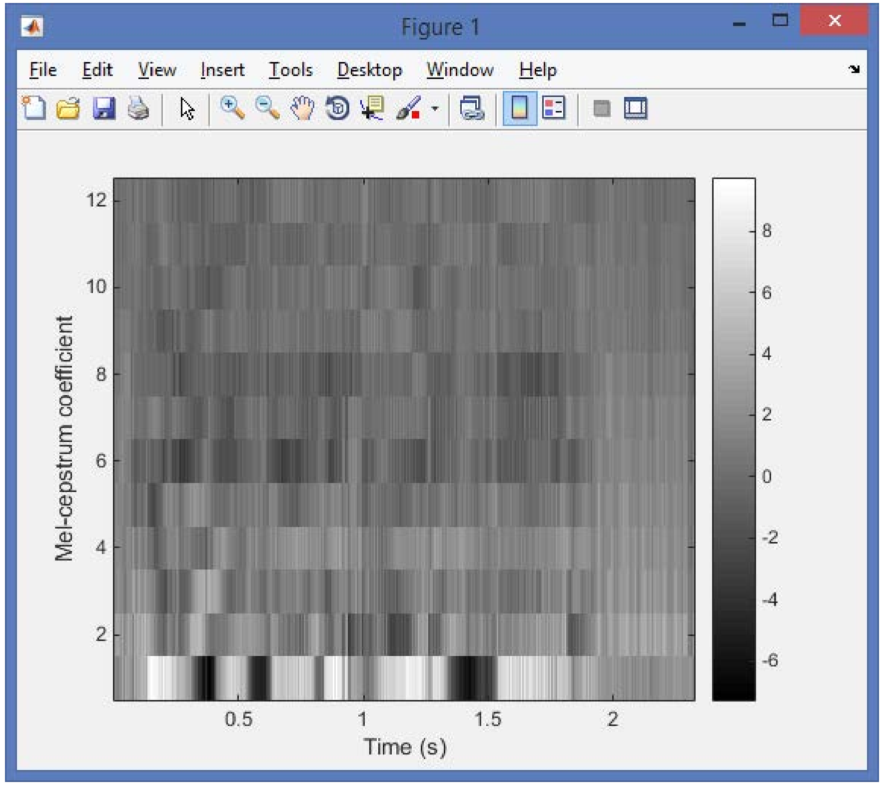
MFCC – PRE-EMPHASIS

- Pre-emphasis: boosting the energy in the high frequencies
- Q: Why do this?
- A: 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**

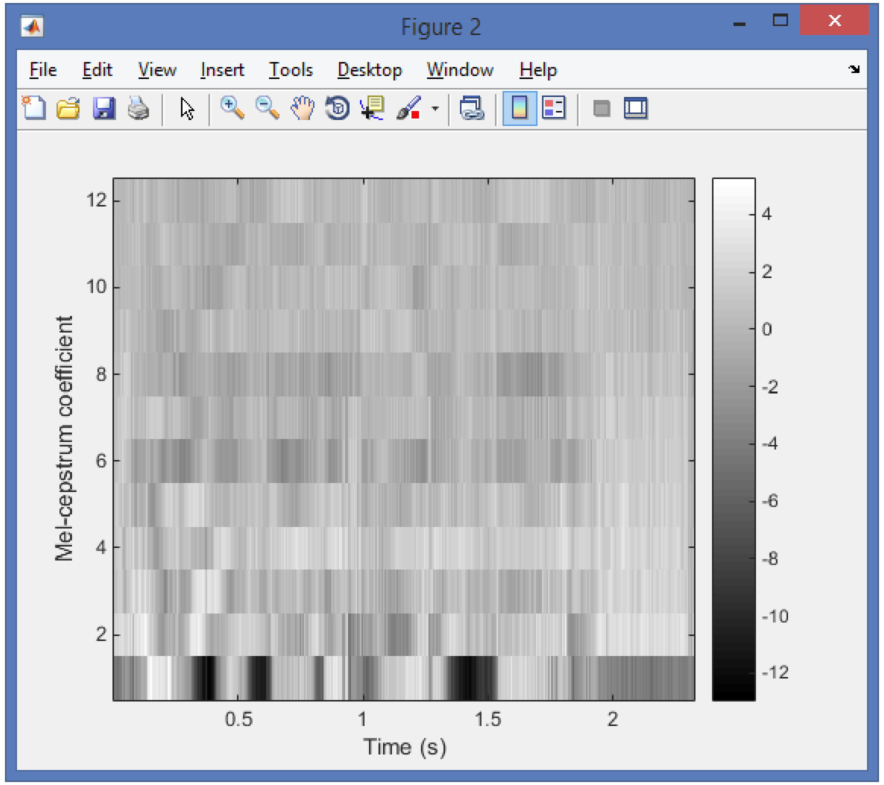


MFCC – PRE-EMPHASIS

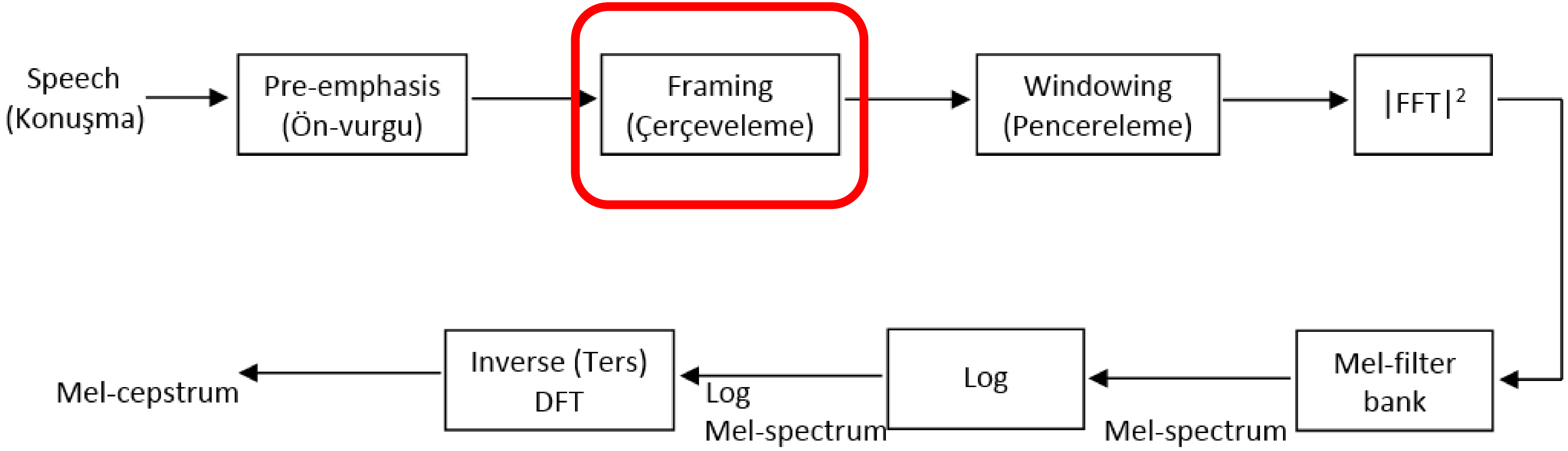
WITHOUT PRE-EMPHASIZING



WITH PRE-EMPHASIZING



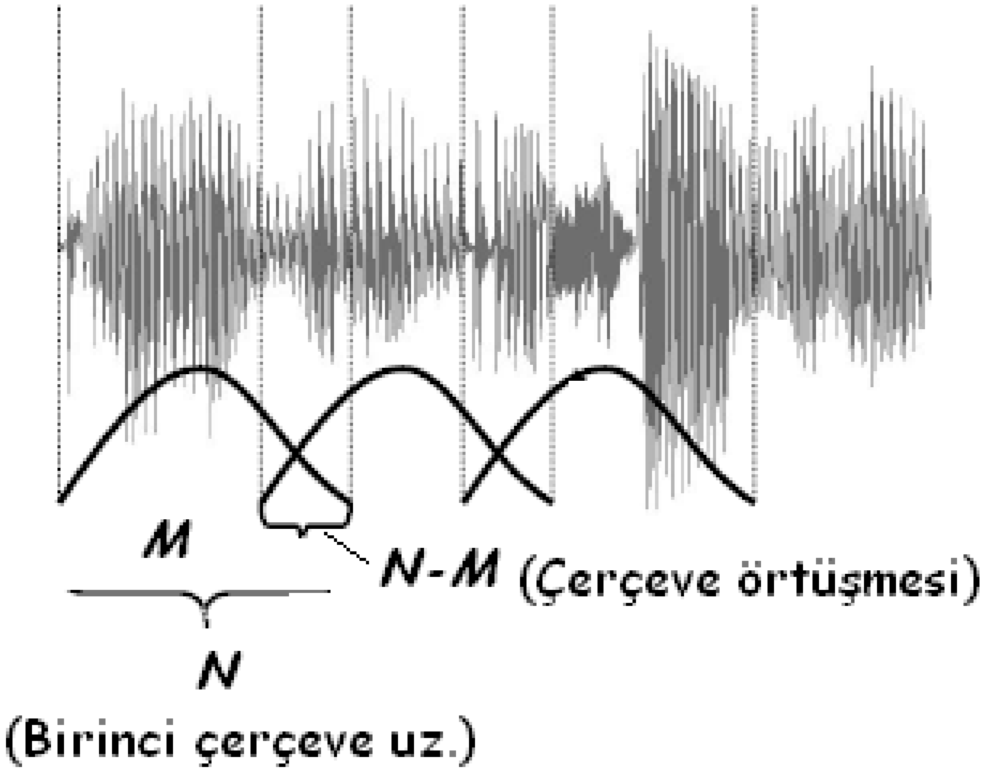
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

MFCC - FRAMING

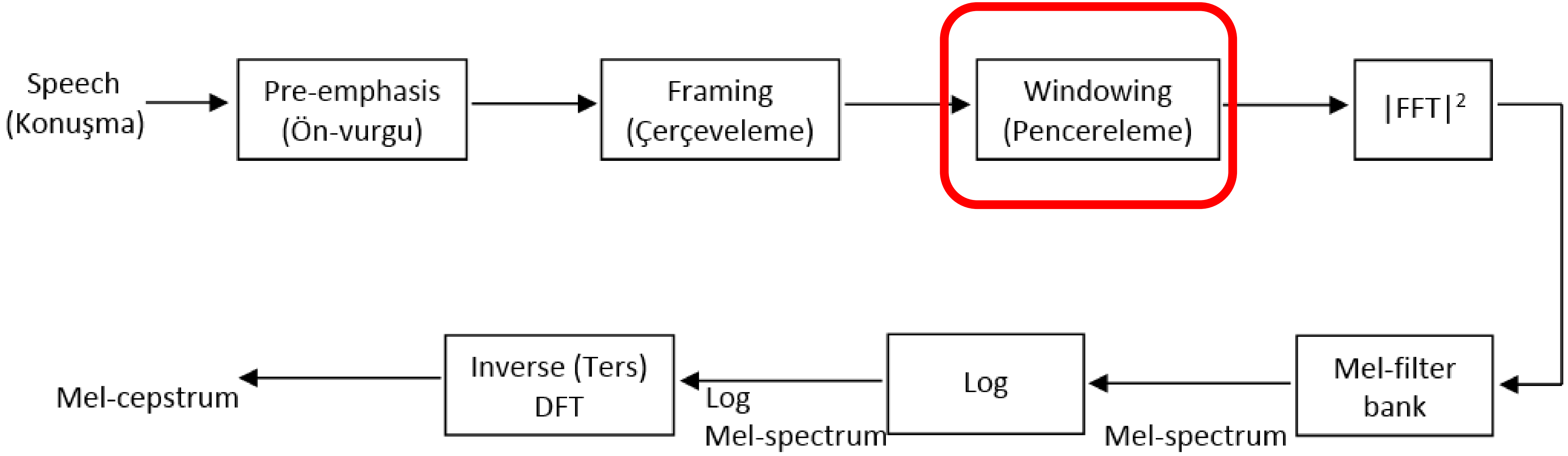
- The width of **the frames** is generally about **30ms** with an **overlap** of about **20ms** (10ms shift).
- Each frame contains **N sample points** of the speech signal.
- Overlap rate of frames, between %30 and % 75 of the length of the frames. (Kinnunen, 2003).
- **M=100 ve N=256** (Lindasalwa Muda, 2010)



MFCC - FRAMING

- This is why we frame the signal into 20-40ms frames. If the frame is much shorter we don't have enough samples to get a reliable spectral estimate, if it is longer the signal changes too much throughout the frame.
- It is assumed that although the speech signal is **non-stationary,** but is **stationary** for a short duration of time.

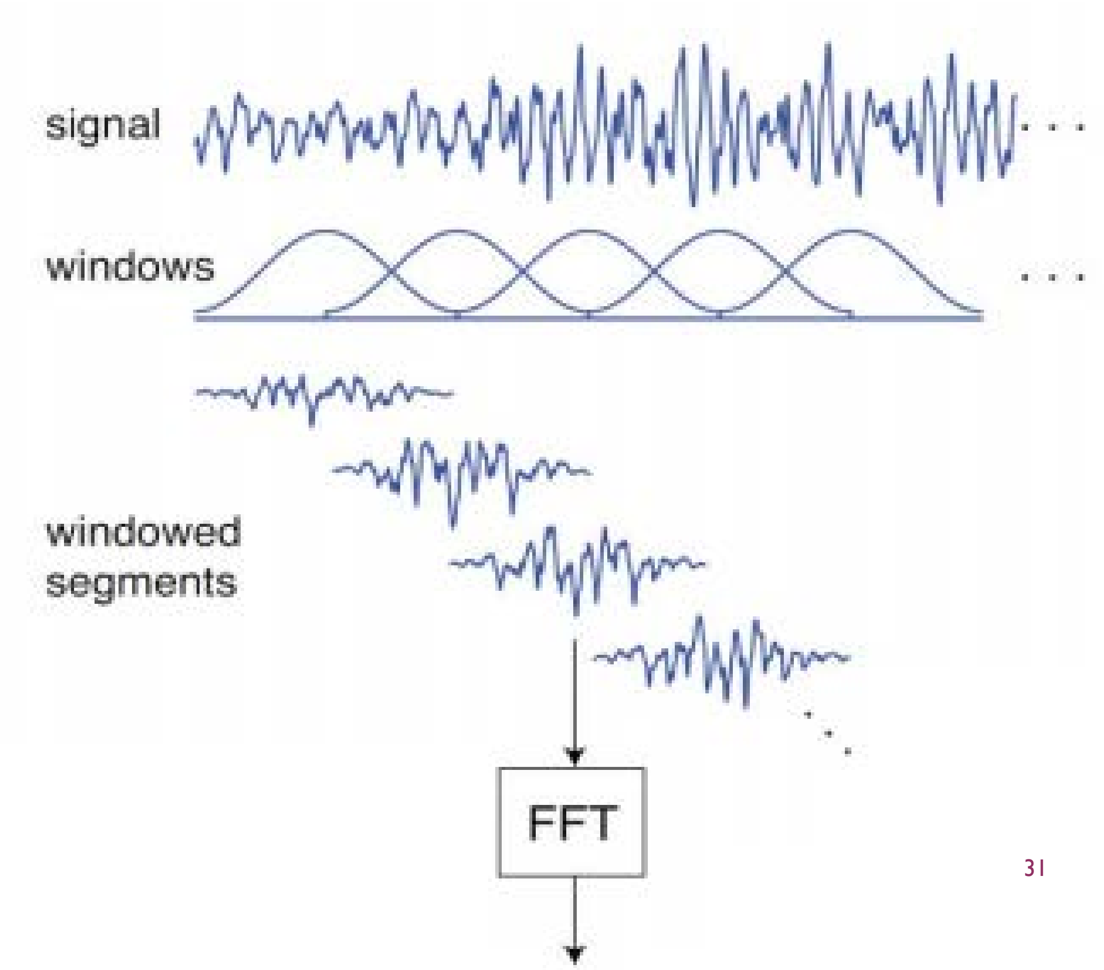
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

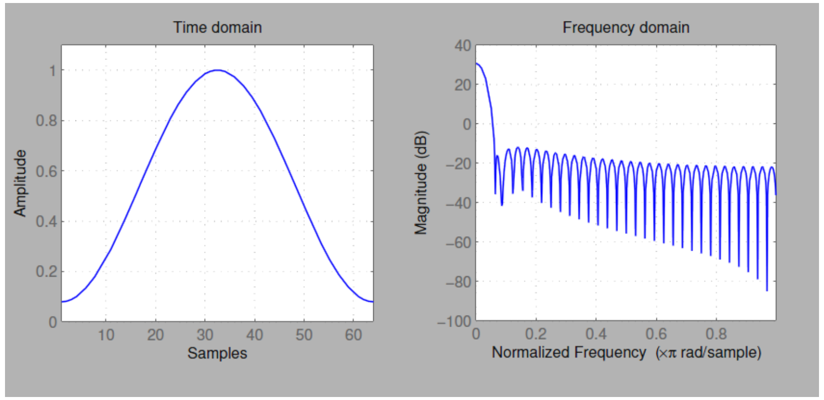
MFCC - WINDOWING

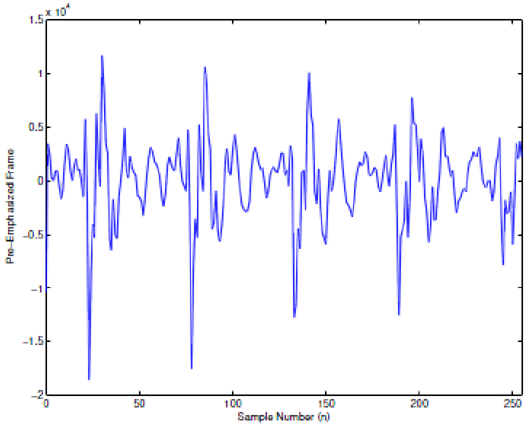
- **The window function** is used **to smooth the signal** for the computation of the **DFT**.
- The DFT computation makes an assumption that the input signal repeats over and over. **The discontinuity** in the frame is prevented (Rabiner ve Juang, 1993).
- If there is a discontinuity between the first point and the last point of the signal, artifacts occur in the DFT spectrum.
- By multiplying a window function to smoothly attenuate both ends of the signal towards zero, this unwanted artifacts can be avoided.



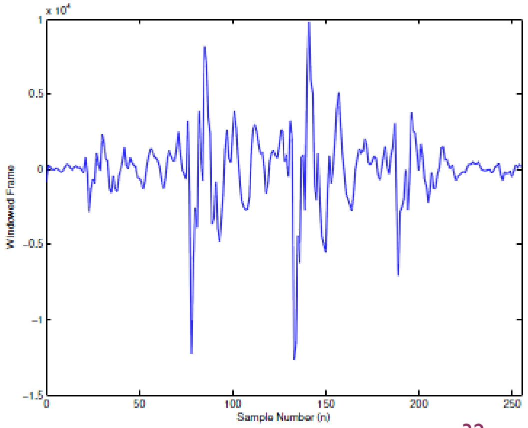
MFCC - WINDOWING

- The objective is to reduce **the spectral effects.**
- **Windowing functions** commonly used : Hamming, Hanning, Blackman, Gauss, rectangular, and triangular…
- **The hamming window** is usually used in speech signal spectral analysis, because its spectrum falls off rather quickly so the resulting frequency resolution is better, which is suitable for detecting formants.





Pre-Emphasized Frame of audio $N = 256$



Windowed Frame $N = 256$

MFCC – WINDOWING – THE HAMMING WINDOW

If the window is defined as W (n), $0 \le n \le N-1$ where

N = number of samples in each frame

Y[n] = Output signal

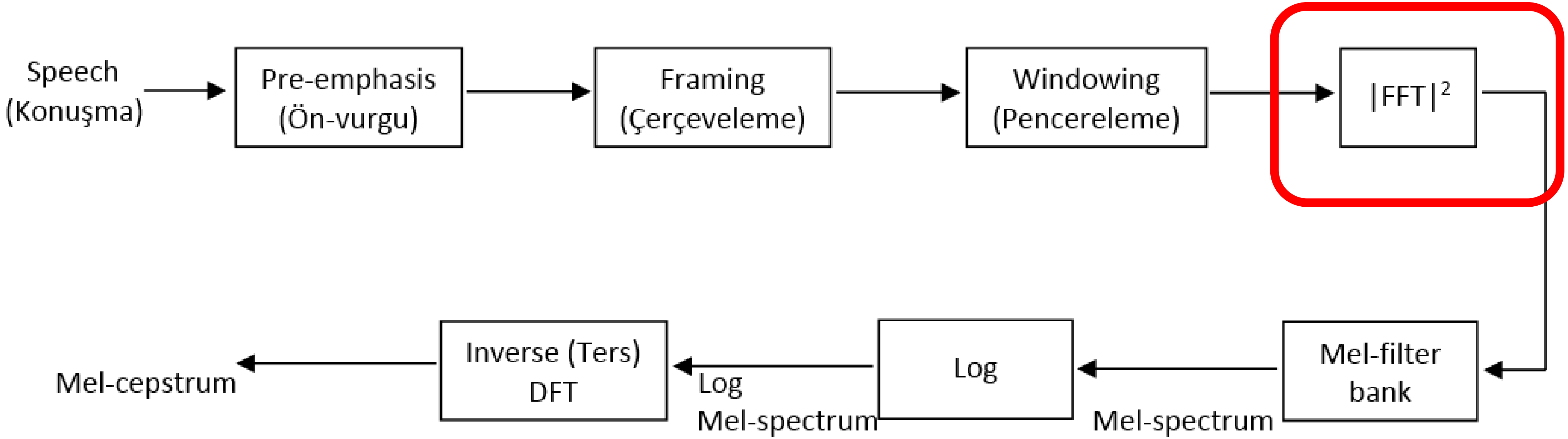
X (n) = input signal

W (n) = Hamming window, then the result of windowing signal is shown below:

$$Y(n) = X(n) \times W(n)$$

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

MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

MFCC – FAST FOURIER TRANSFORM (FFT)

- To **convert** each frame of N samples **from time domain** into **frequency domain.**
- The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain.
- This statement supports the equation below:

$$Y(w) = FFT\left[h(t) * X(t)\right] = H(w) * X(w)$$

- If **X (w), H (w) and Y (w)** are the Fourier Transform of **X (t), H (t) and Y (t)** respectively.

MFCC – FAST FOURIER TRANSFORM (FFT)

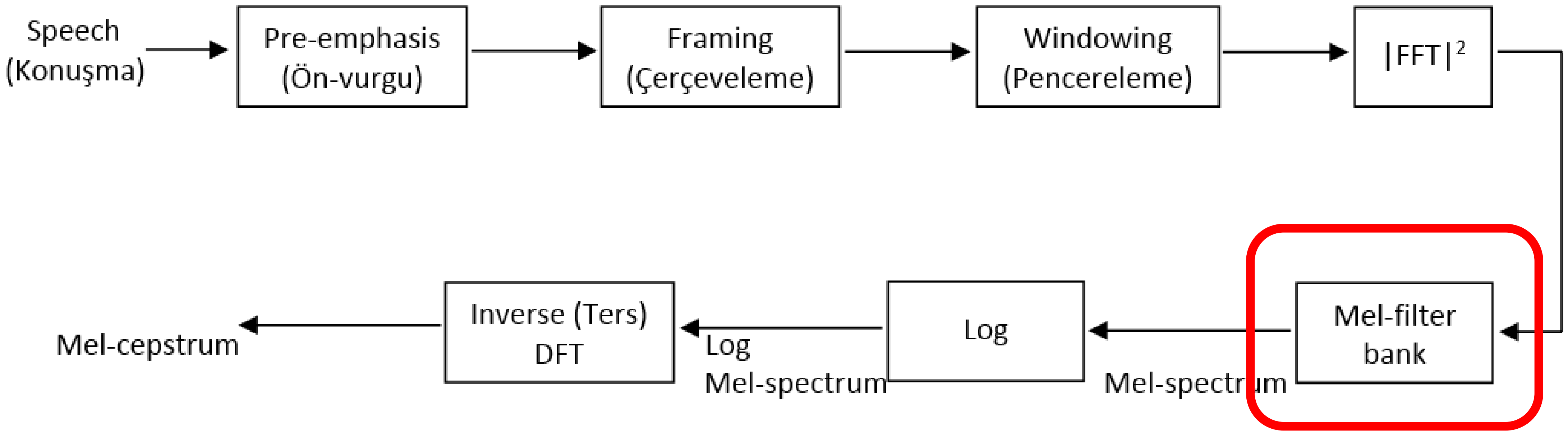
- Input:
  - Windowed signal x[n]…x[m]
- Output:
  - For each of N discrete frequency bands
  - **A complex number X[k]** representing **magnitude** and **phase** of that frequency component in the original signal
- **Discrete Fourier Transform (DFT)**

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\frac{\pi}{N}kn}$$

The amplitude spectrum of the signal passed through the window is calculated by FFT.

- **FFT size can be 512, 1024 or 2048**

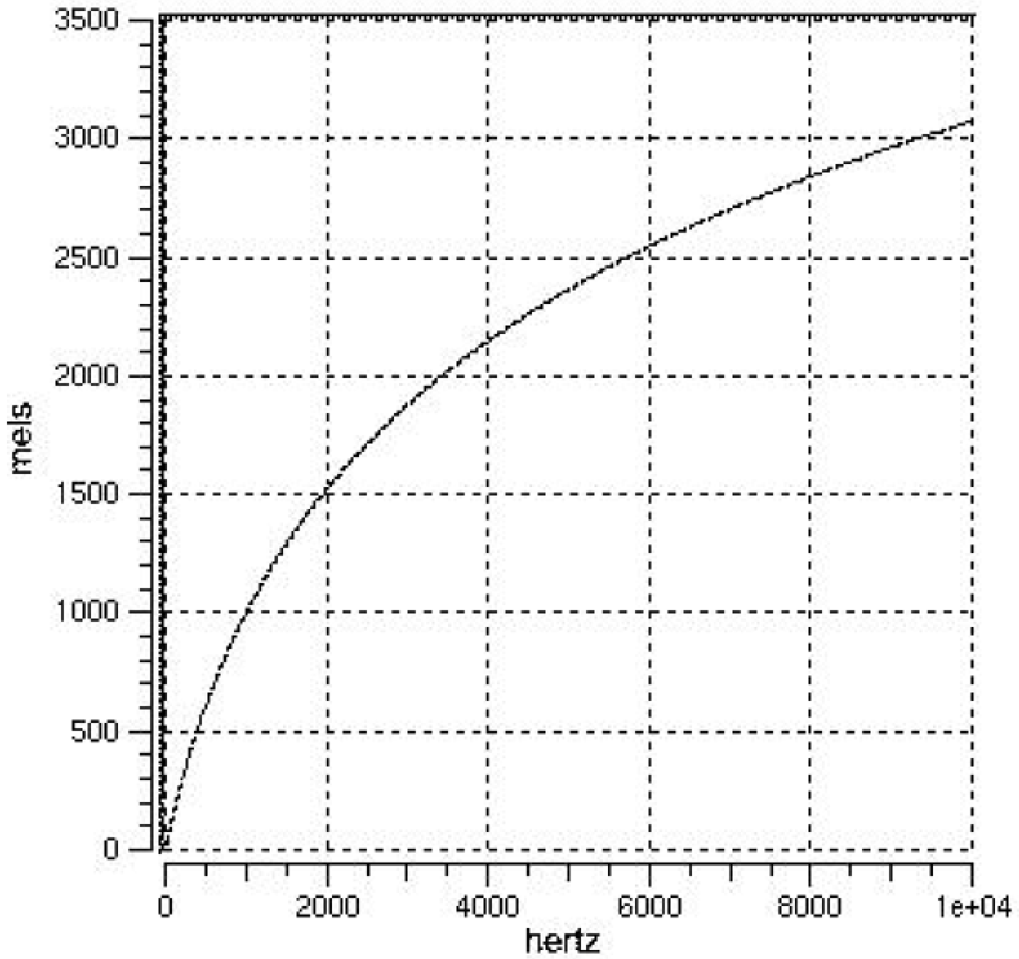
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

MFCC - MEL-FILTER BANK PROCESSING

- Human hearing is not equally sensitive to all frequency bands
- **Less sensitive** at higher frequencies, **roughly > 1000 Hz**
- I.e. human perception of frequency is non-linear.



MFCC - MEL-FILTER BANK PROCESSING

- **Mel** (melody) is a unit of pitch. Mel-frequency scale is approximately **linear** up to the frequency of **1KHz** and then becomes close to **logarithmic** for the higher frequencies.
- Human ear acts **as filters** that concentrate on only certain frequency components. **Band-pass filters.**
- These filters are **non-uniformly** spaced on the frequency scale, with **more filters** in the **low frequency regions** and **less filters** in the **high frequency regions.**

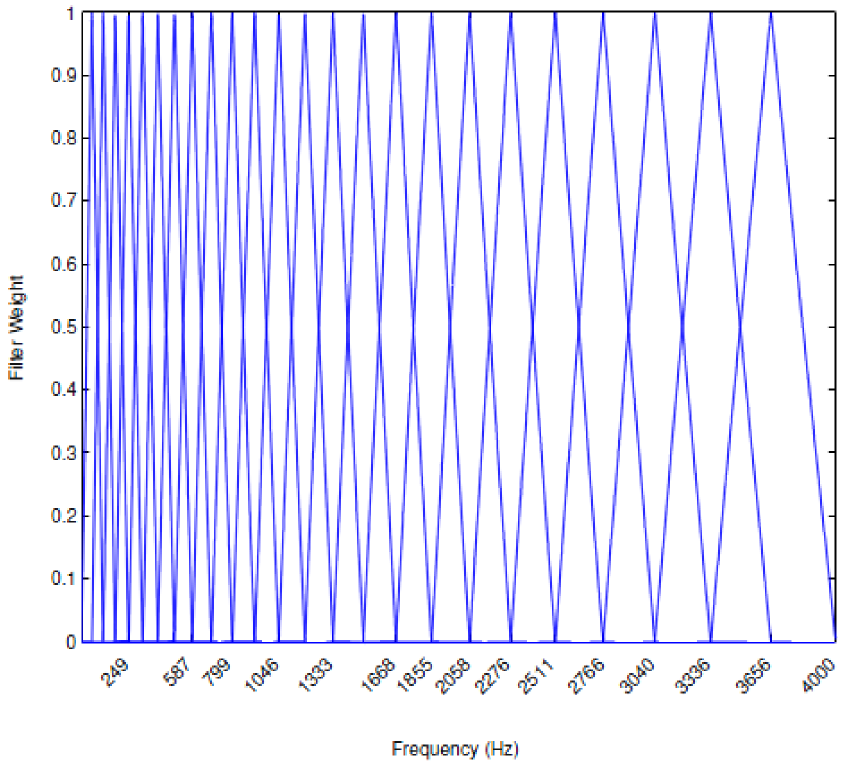
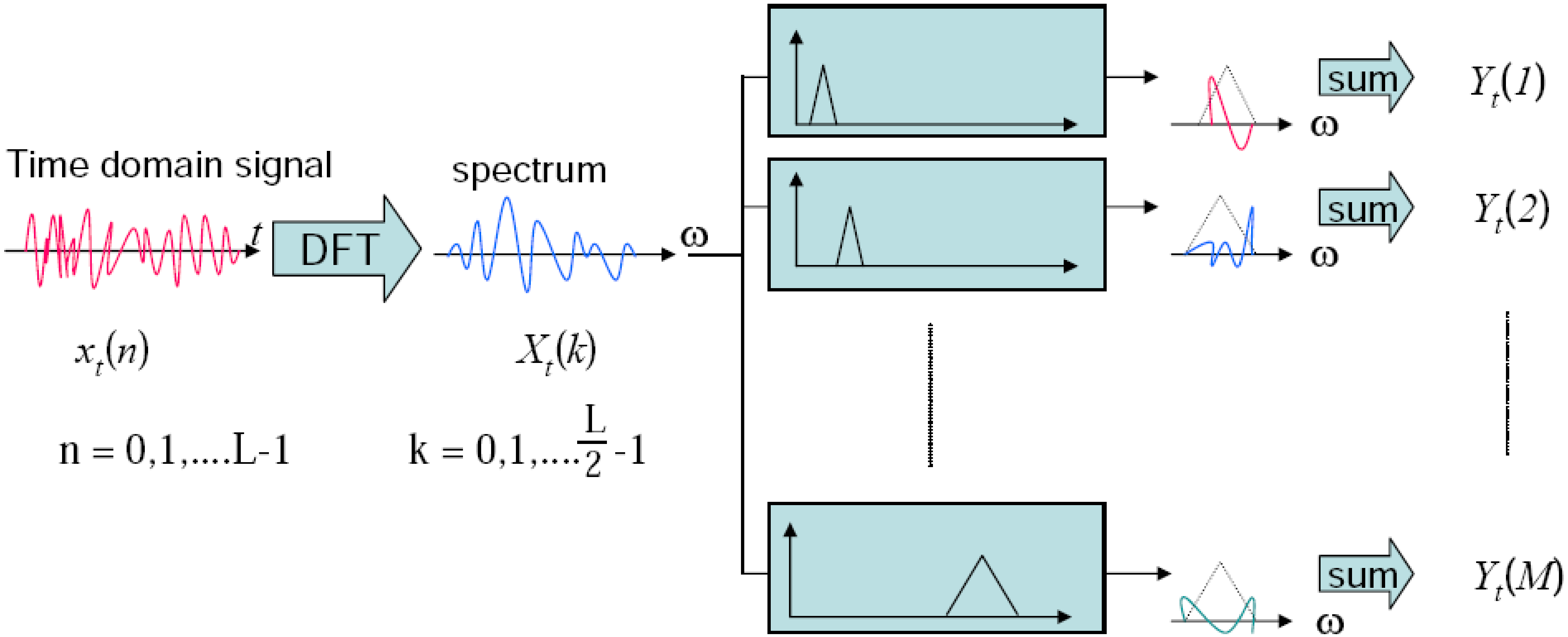


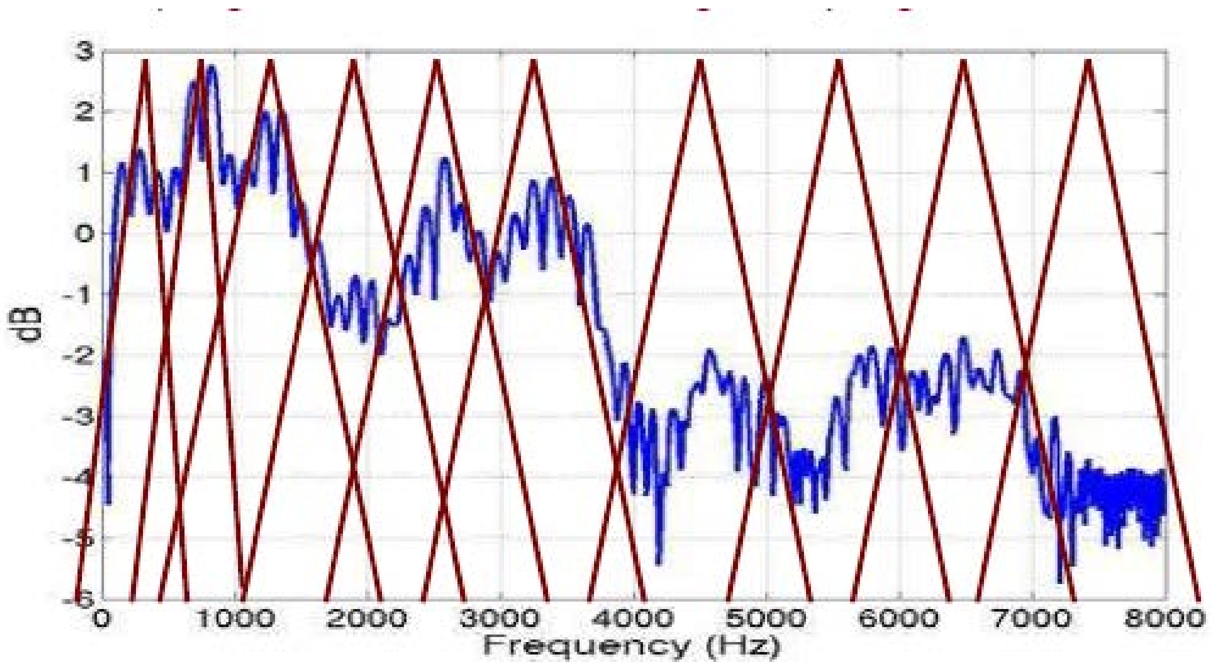
Figure shows the plot of pitch (Mel) versus frequency.

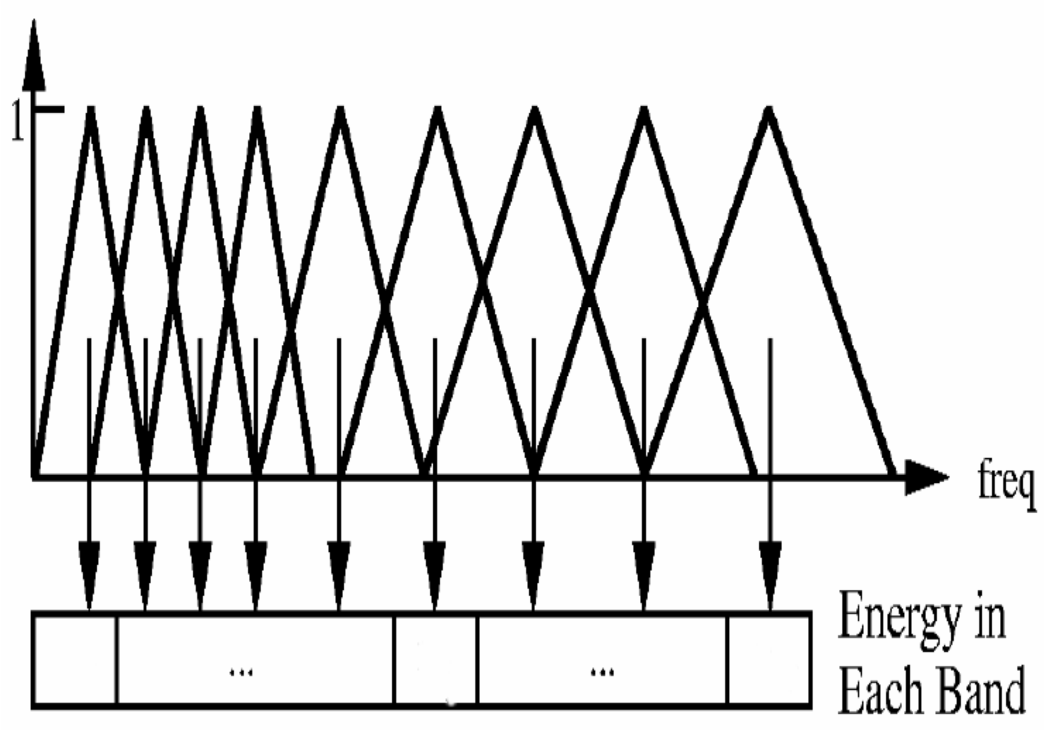
MFCC - MEL-FILTER BANK PROCESSING

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



MFCC - MEL-FILTER BANK PROCESSING





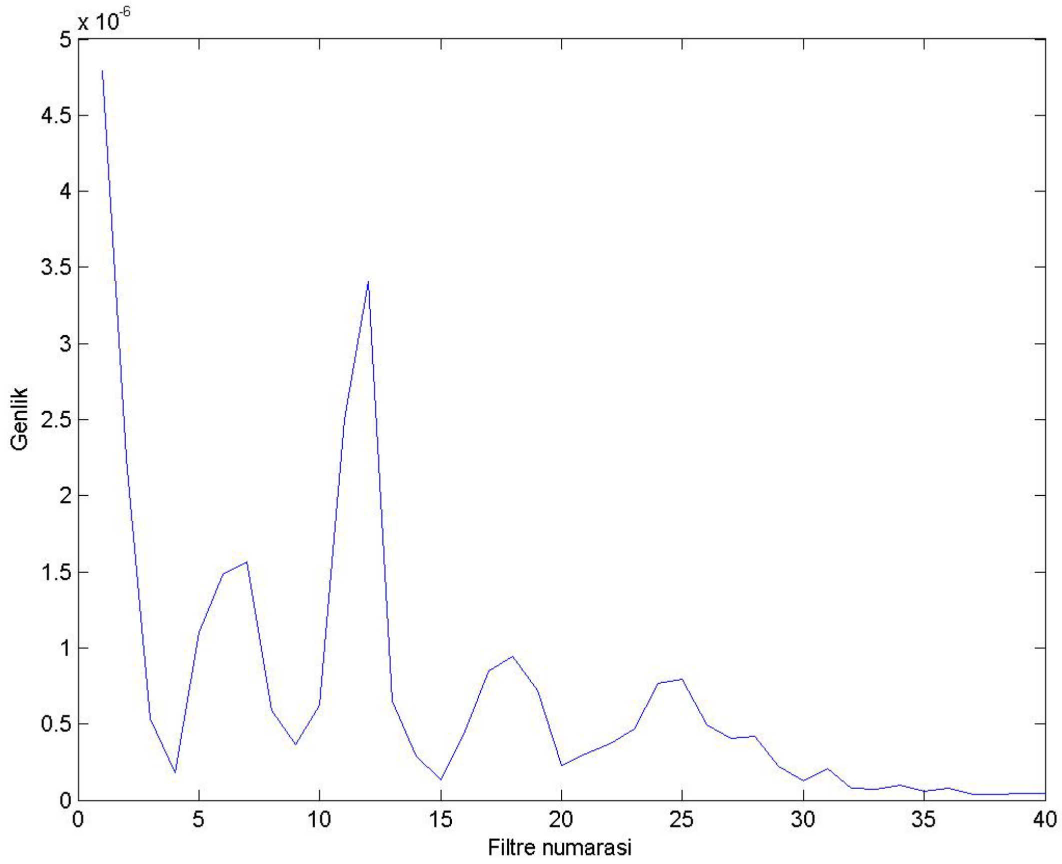
- Where ‘f’ denotes the real frequency, and mel(*f*) denotes the perceived frequency.

$$F(Mel) = [2595 * \log 10[1 + f]700]$$

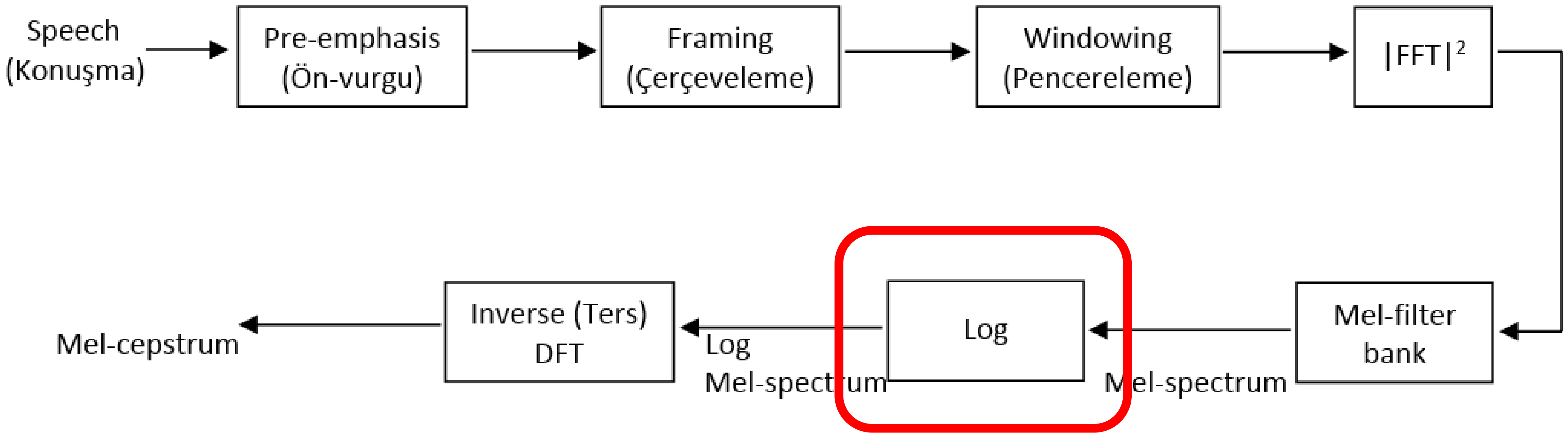
MFCC - MEL-FILTER BANK PROCESSING

Typically P=24 to P=30 filters in the Mel bank, but in Slaney's obtaining MFCC method (1998) used;

- Speech signal (1*512)
- 40 Mel filters (40*512)
- Output is (1*40)



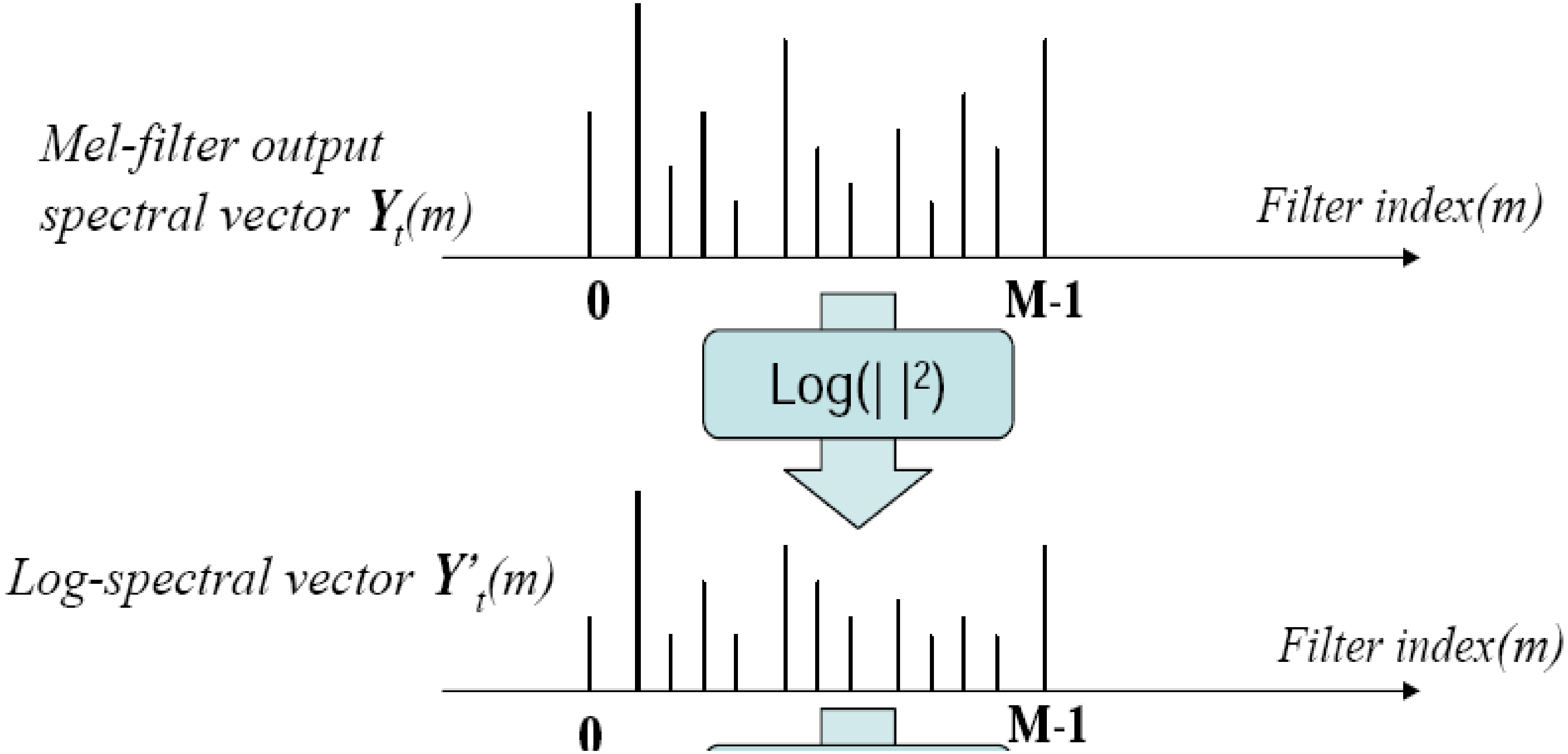
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

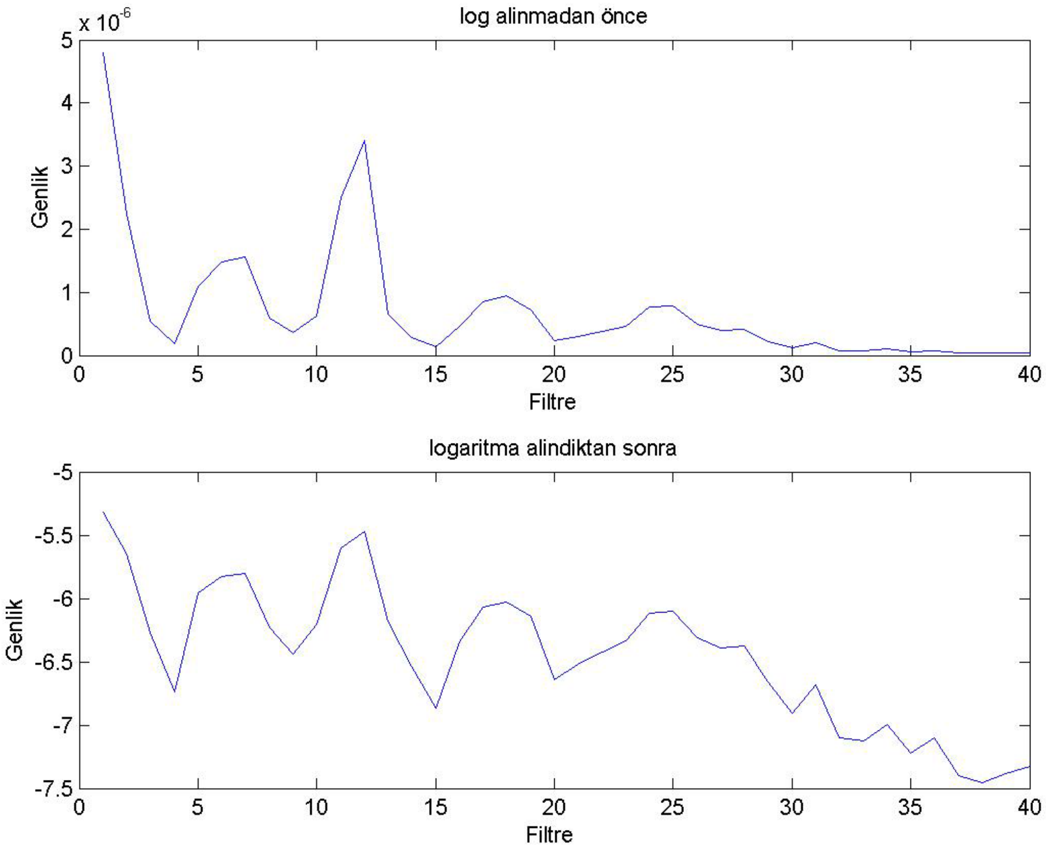
MFCC - LOG ENERGY COMPUTATION

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

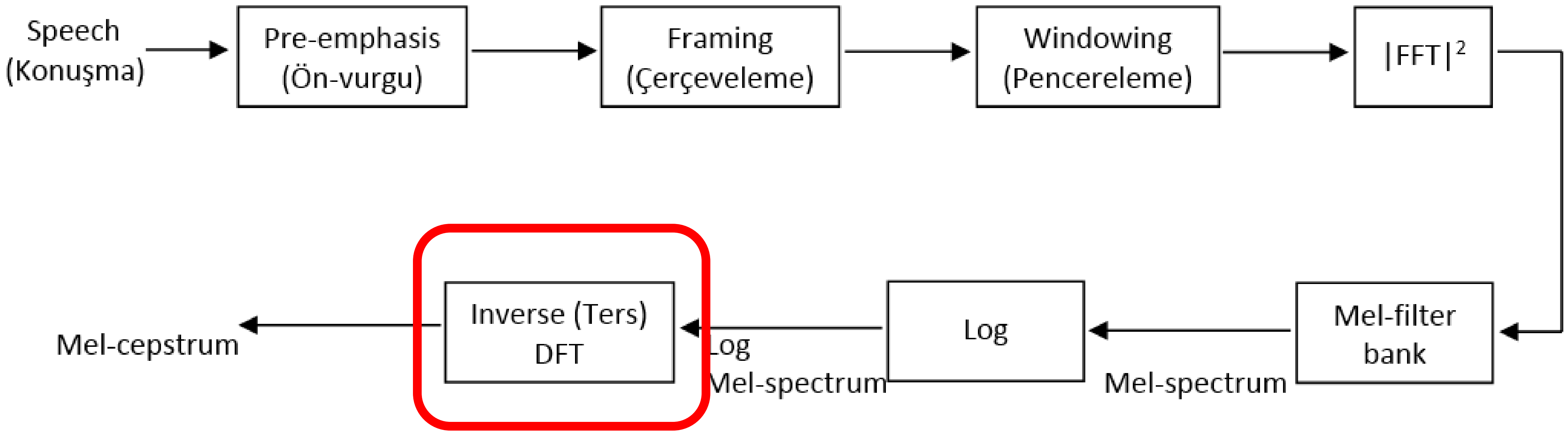


MFCC - LOG ENERGY COMPUTATION

- Why log energy?
- Logarithm **compresses dynamic range** of values
  - Human response to signal level is logarithmic
  - Humans **less sensitive** to **slight differences** in amplitude **at high amplitudes** than **low amplitudes**
- Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike)
- Phase information not helpful in speech



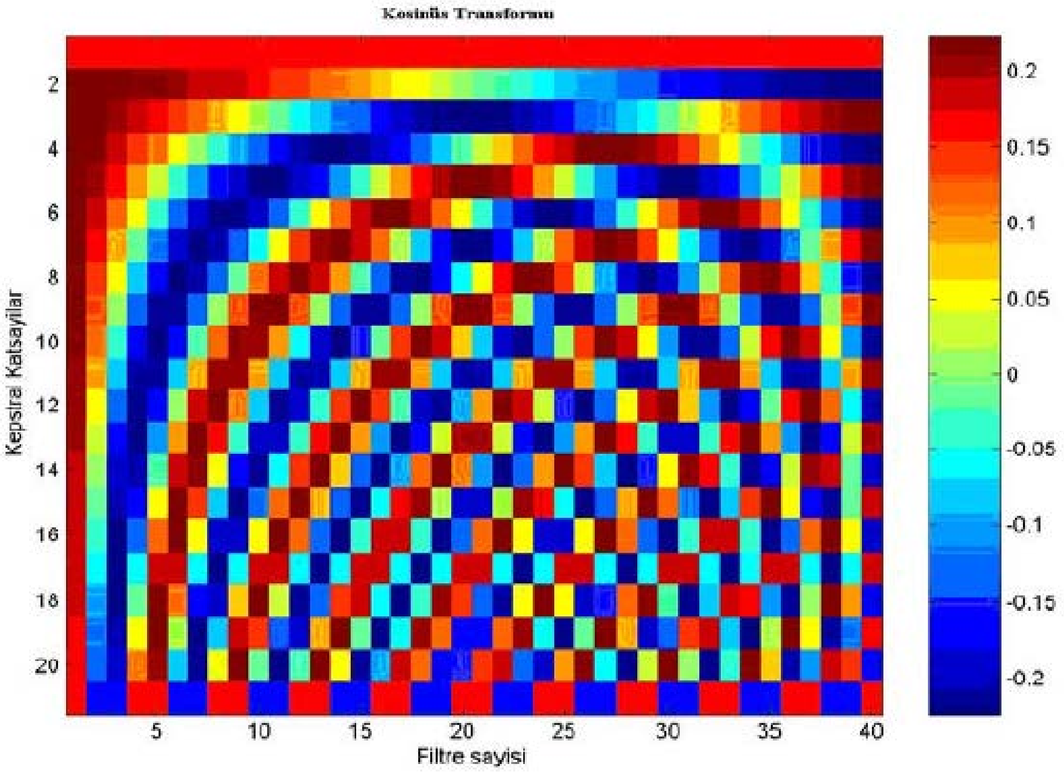
MEL-FREQUENCY CEPSTRAL COEFFICIENTS(MFCC)



MFCC steps in speech analysis

MFCC – DISCRETE COSINE TRANSFORM

- This is the process to **convert the log Mel spectrum** into **time domain** using **Discrete Cosine Transform (DCT).**
- **The result** of the conversion is called **Mel Frequency Cepstrum Coefficient.**
- The set of coefficient is called **acoustic vectors.**
- Therefore, each **input utterance** is **transformed** into **a sequence of acoustic vector.**



Şekil 12: Ayrık kosinüs dönüşümü

MFCC – DISCRETE COSINE TRANSFORM

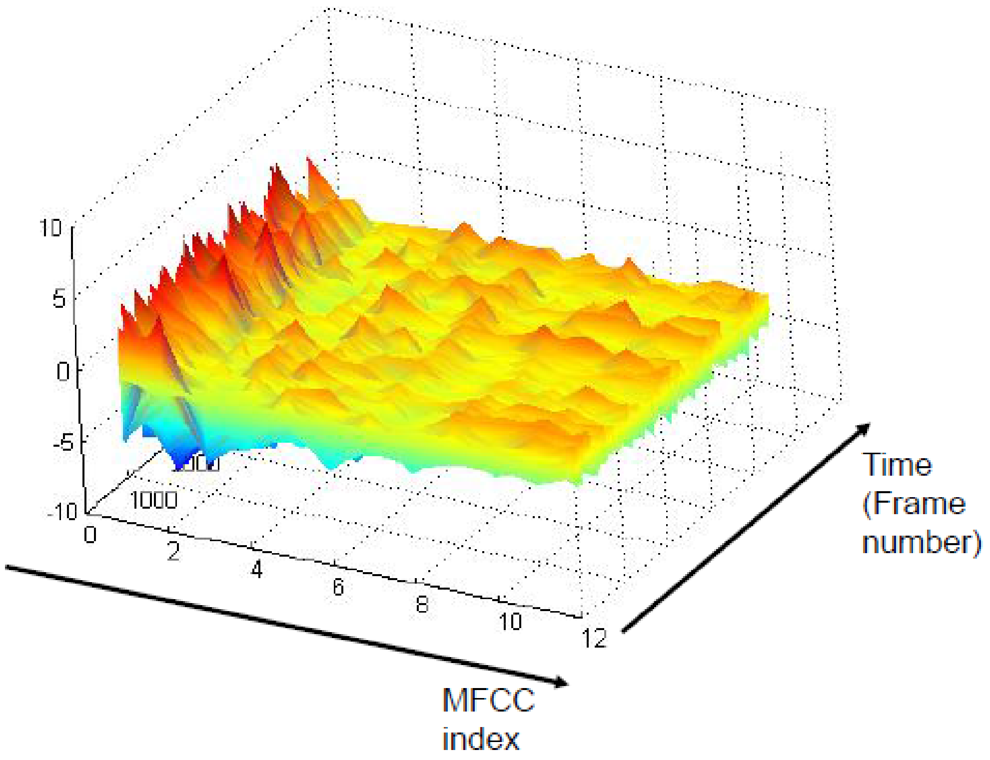
- The cepstrum requires **Fourier analysis**
- But we're going **from frequency** space back to **time**
- So we actually apply **inverse DFT**

$$y_t[k] = \sum_{m=1}^{M} \log(|Y_t(m)|) \cos(k(m-0.5)\frac{\pi}{M}), \;\; \text{k=0,...,J}$$

- Details for signal processing gurus: Since the log power spectrum is real and symmetric, inverse DFT reduces to a Discrete Cosine Transform (DCT)

HOW DO THE MFCCS LOOK LIKE?

- During the feature extraction stage a database of **"voiceprints"** is created in order to be used as a reference in the feature matching stage.
- A voiceprint represents the most **basic**, yet **unique**, features of the speech command in the **frequency domain**.
- A **voiceprint** is merely **a matrix of numbers** in which each number represents **the energy or average power** that is heard in a particular frequency band during a specific interval.



TEMPLATE MATCHING

- The input utterance is converted to a set of feature vectors

Hello → h | he | e | el | l | lo | o

- Time alignment may need to be done

h | he | e | el | l | lo | o

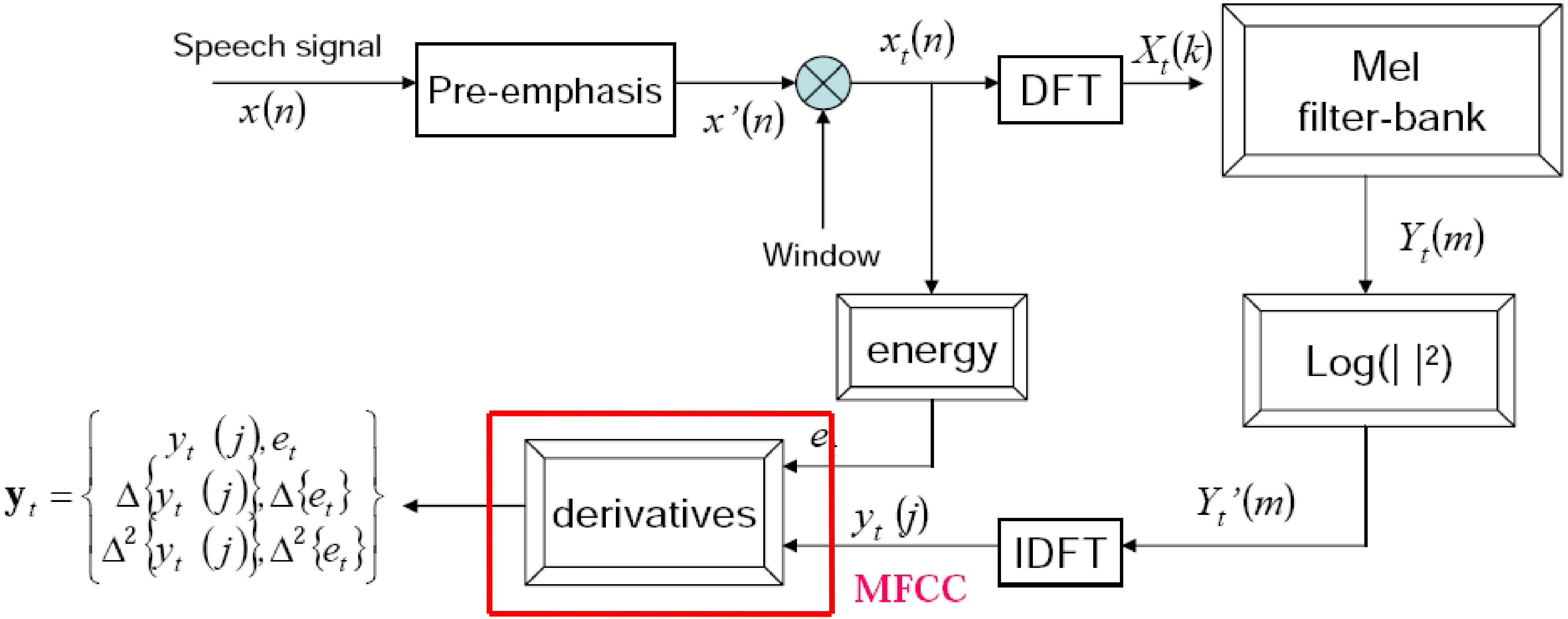
h | he | e | el | l | lo | o

- Calculate similarity between each captured vector with the registered speaker template or model

h ⊗ h = .90 similarity

he ⊗ he = .60 similarity, .75 overall

MFCC-ADDITIONAL FEATURES



MFCC-ADDITIONAL FEATURES

- **Dynamic cepstral coefficients**
  - The cepstral coefficients do not capture energy
  - So we add an **energy feature**

$$Energy = \sum_{t=t_1}^{t_2} x^2[t]$$

- Speech signal is not constant
  - **slope of formants,**
  - change **from stop burst to release**
- So in addition to the cepstral features
- Need to model changes in the cepstral features over time.
  - “delta features”
  - “double delta” (acceleration) features

TYPICAL MFCC FEATURES

- Window size: 25ms
- Window shift: 10ms
- FFT size 512, 1024 or 2048
- Pre-emphasis coefficient: 0.97
- P=24 to P=30 filters in the Mel bank
- 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

MFCC – AREAS OF USAGE

- **•Speech classification**
- –Automatic speech recognition
- –Speaker identification
- –Language identification
- –Emotion recognition
- **•Music information retrieval**
- –Musical instrument recognition
- –Music genre identification
- –Singer identification

- **•Speech synthesis, coding, conversion**
- –Statistical parametric speech synthesis
- –Speaker conversion
- –Speech coding
- **•Others**
- –Speech pathology classification
- –Identification of cell phone models
- •etc …

MFCC – SAMPLE IMPLEMENTATIONS

- •Voicebox (Matlab)
- http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
-
- •Rastamat (Matlab)
- http://labrosa.ee.columbia.edu/matlab/rastamat/
-
- •Bob toolkit
- http://idiap.github.io/bob/
-
- •Hidden Markov model (HTK) toolkit
- http://htk.eng.cam.ac.uk/

RESULTS

Çerçeve süresi değişimi tanıma başarımını değiştirmemektedir. (TIMIT)

Hamming pencereleme fonksiyonu kullanılması (NTIMIT)

Tablo I. Çerçeveleme sürelerinin konuşmacı tanımaya etkisi (%)

Veritabanları	Çerçeveleme süreleri (msn.)			
	30	25	20	15
TIMIT	99.4	99.4	99.4	99.4
NTIMIT	67.9	67.9	69.9	68.1

Tablo II. Pencereleme fonksiyonlarına bağlı olarak konuşmacı tanıma oranları (%)

Pencereleme fonk.	Veritabanları	
	TIMIT	NTIMIT
Hamming	99.4	69.9
Hanning	99.4	69.3
Blackman	99.7	68.4
Gauss	99.7	67.6
Dikdörtgen	99.1	64.9
Üçgen	99.4	67.6

Mel Frekansı Kepstrum Katsayılarındaki Değişimlerin Konuşmacı Tanımaya Etkisi, *Ömer ESKİDERE*, *Figen ERTAŞ*

RESULTS

Ön vurgulamanın uygulanmadığı veya güç spektrumu alındıktan sonra uygulanması durumlarında en iyi konuşmacı tanıma başarımı elde edilmektedir. (TIMIT)

İşarete çerçevelemeden önce önvurgulama uygulanması (NTIMIT)

Tablo IV. Önvurgulamanın konuşmacı tanıma üzerine etkisi (%)

Önvurgulama uygulama şekilleri	Veritabanları	
	TIMIT	NTIMIT
Çerçevelemeden önce	99.4	70.2
Çerçevelemeden sonra	99.4	60.1
Pencerelemeden sonra	99.4	69.1
Güç spektrumu alındıktan sonra	99.7	67.3
Önvurgulama yok	99.7	69.9

Slaney'in (1998) önerdiği Mel ölçek kullanılması (TIMIT ve NTIMIT)

Tablo V. İki farklı Mel ölçek için konuşmacı tanıma oranları (%)

Veritabanları	Mel Ölçek	
	Davis ve Mermelstein (1980)	Slaney (1998)
TIMIT	99.4	99.7
NTIMIT	67.9	70.2

Mel Frekansı Kepstrum Katsayılarındaki Değişimlerin Konuşmacı Tanımaya Etkisi, *Ömer ESKİDERE, Figen ERTAŞ*

REFERENCES

- Time-Frequency Analysis for Voiceprint (Speaker) Recognition, Wen-Wen Chang
- Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques, Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi
- Mel Frekansı Kepstrum Katsayılarındaki Değişimlerin Konuşmacı Tanımaya Etkisi, ***Ömer ESKİDERE***, ***Figen ERTAŞ***
- Feature extraction of speech with Mel frequency cepstral coefficients (MFCCs)
- **http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/**