Speech Signal Processing and Speech Recognition

Bob Dunn

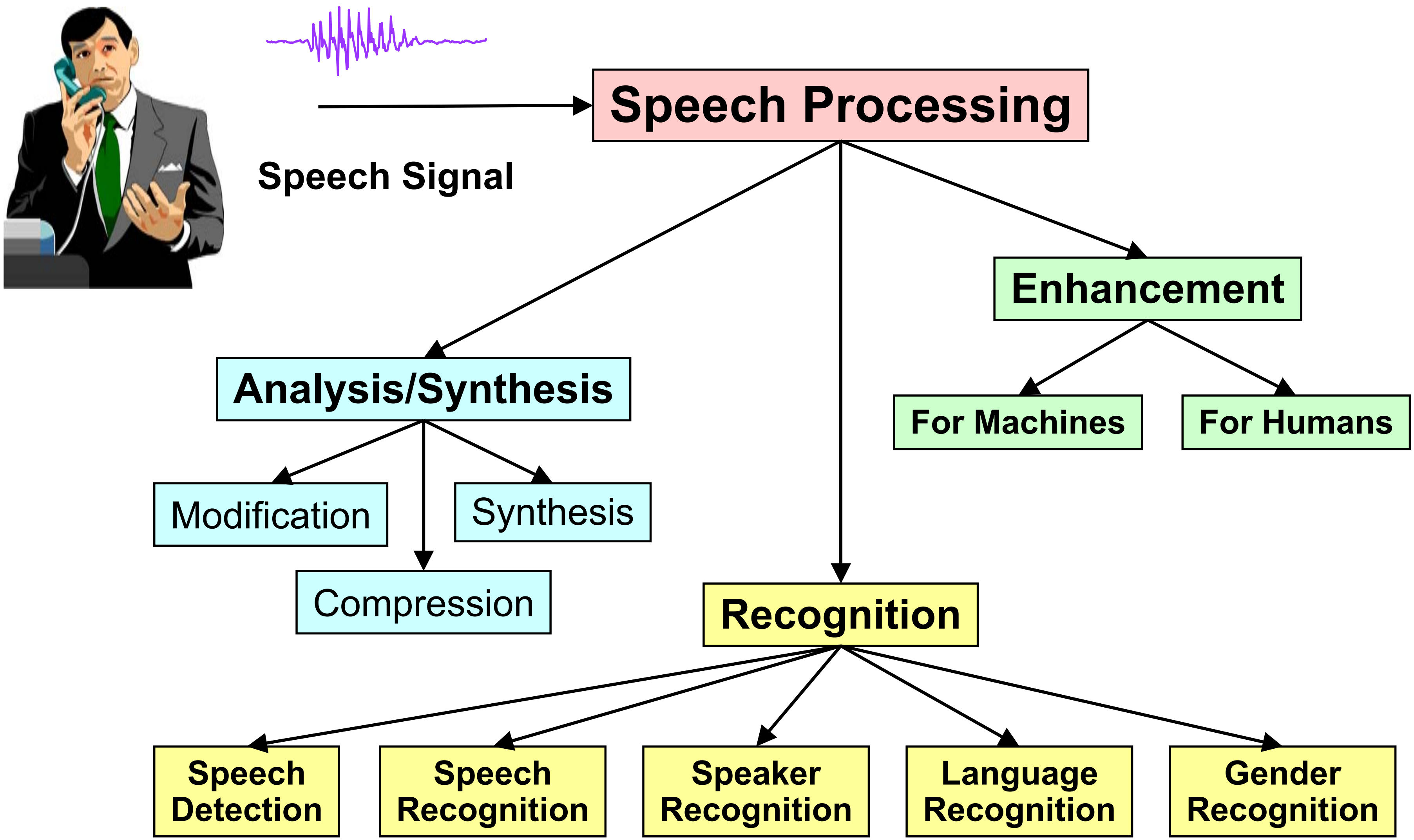
29 April 2003

This work was sponsored by the Department of Defense under Air Force contract F19628-00-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense.

Outline

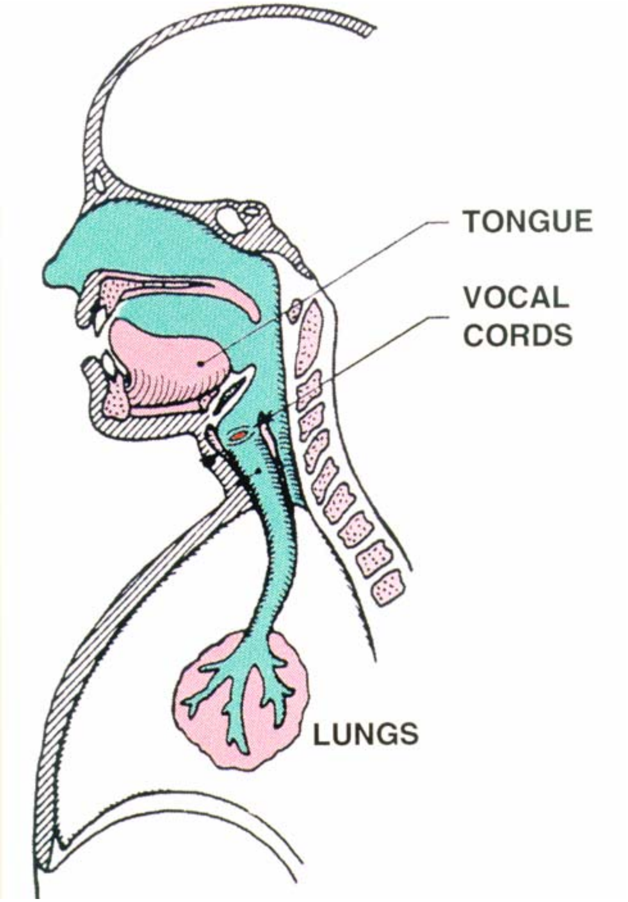
- Introduction
  - Speech Production and Modeling
- Sample Applications
  - Signal Processing
    - Modification, Enhancement
  - Recognition
    - Words, language, speaker
- Signal Processing for Recognition
  - Feature Extraction

Taxonomy & Function



Speech Production

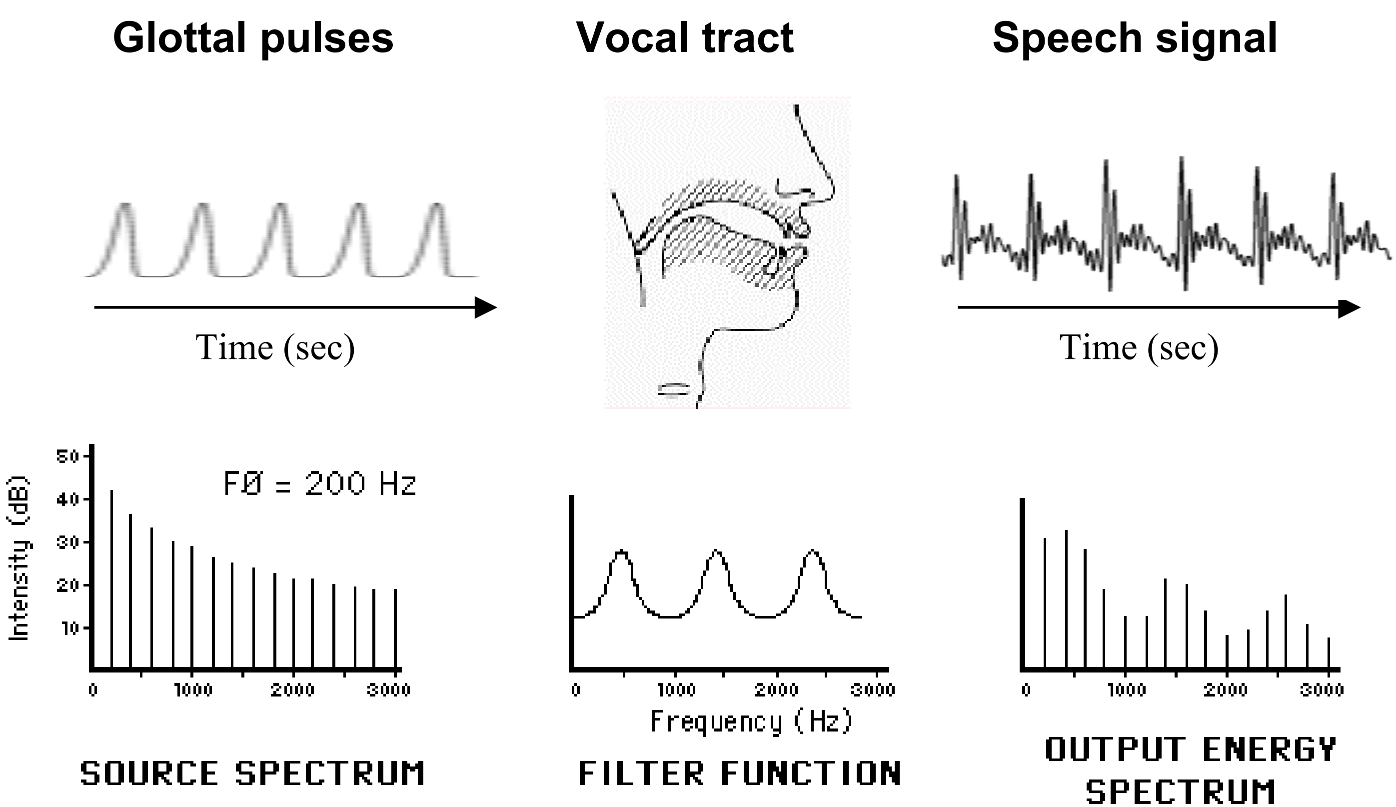
- **Sound production**
  - Air forced out of the lungs
  - Passes vocal cords
    - Voiced speech: vibration
    - Unvoiced speech: no vibration
  - Sound shaped by resonant vocal tract cavity



- **Source-filter model**
  - Source
    - Voiced speech: pulses
    - Unvoiced speech: stochastic
  - Filter
    - Time-varying resonant vocal tract cavity

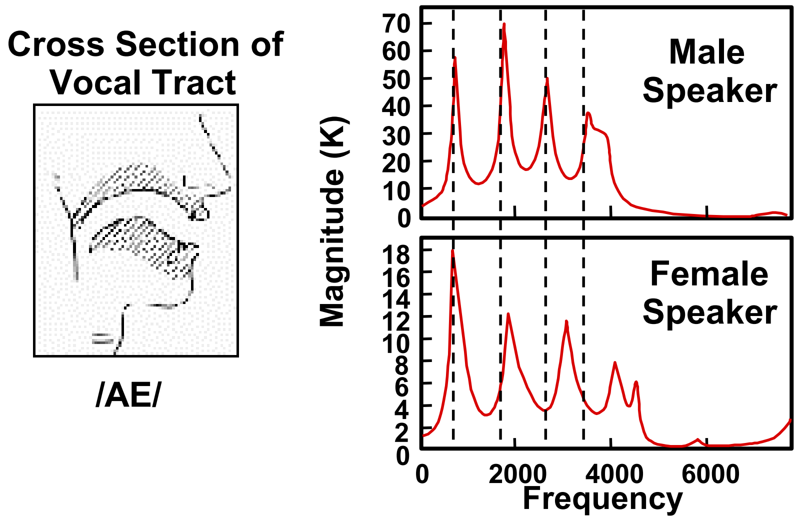
Speech Production

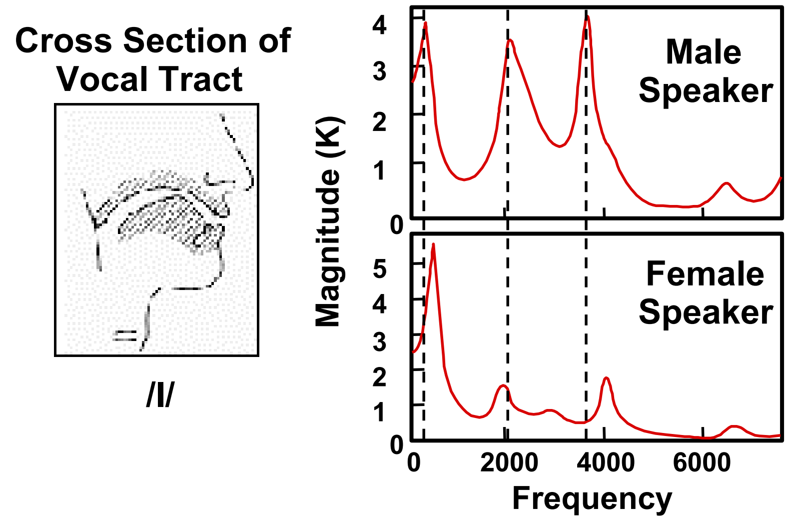
- **Speech production model: source-filter interaction**
  - Anatomical structure (vocal tract/glottis) conveyed in speech spectrum



Speech Characteristics

- **Different speakers will have different spectra for similar sounds**

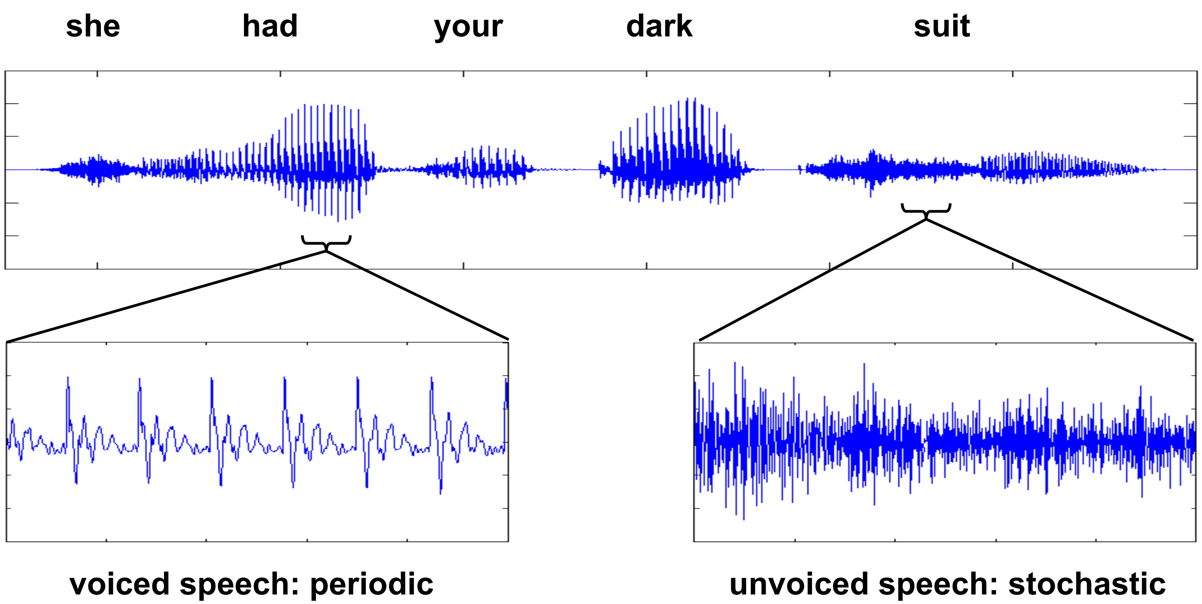




- **Differences are in location and magnitude of peaks in spectrum**
  - **Peaks are known as formants and represent resonances of vocal cavity**
- **The spectrum captures the format location and, to some extent, pitch without explicit formant or pitch tracking**

Speech Waveform

- **Speech waveform is quasi-stationary**
  - Appears stationary when analyzed over a short duration
- **Speech has two modes:**
  - Voiced speech - periodic
  - Unvoiced speech - stochastic



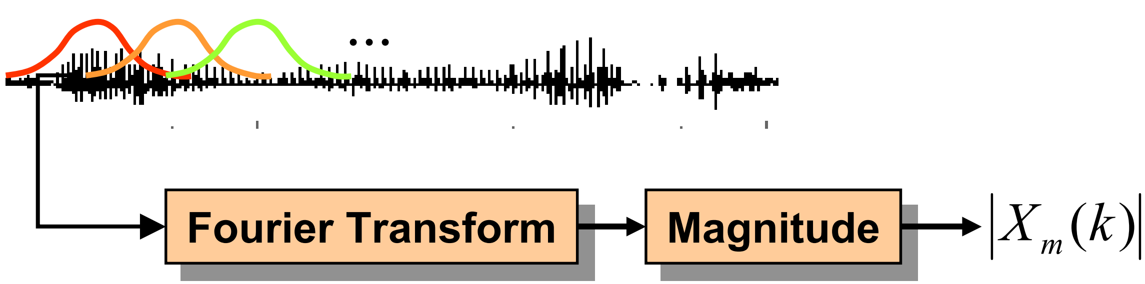
Frame-Based Analysis

- **The speech waveform is typically analyzed on a frame-by-frame basis**
  - **Windowed speech waveform on frame m:**

$$x_m(n) = w(n)s(n - mT)$$

  - **Discrete Fourier transform on frame m:**

$$X_m(k) = \sum_{n=0}^{N-1} x_m(n) e^{-j\frac{2\pi}{N}kn} \qquad 0 \le k \le N-1$$

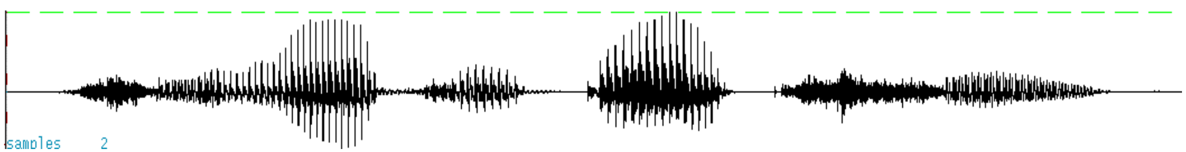


speech signal: $s(n)$
analysis window: $w(n)$
sample index: n
frame number: m
frame interval: T
frequency index: k
DFT length: N

Spectogram

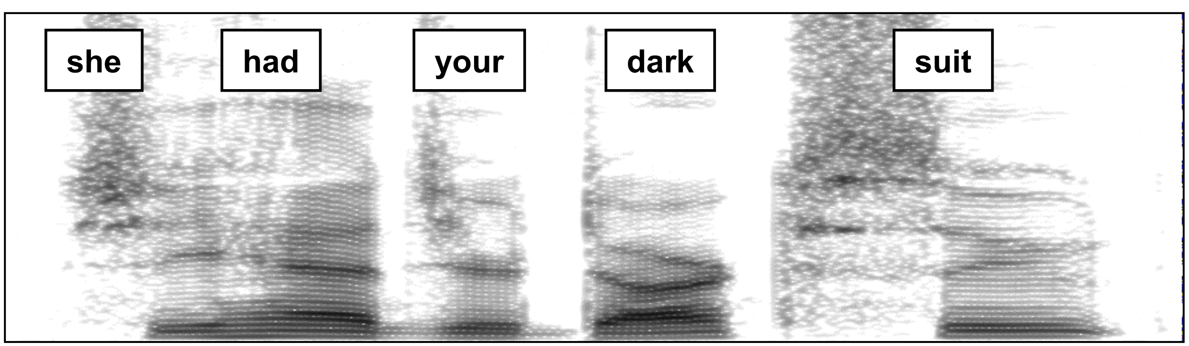
- **Speech is a continuous evolution of the vocal tract**
- **Spectogram displays the time and frequency evolution of the speech waveform**
  - **Formants: vocal tract resonances**
  - **Computed as** $\left|X_m(k)\right|$
- **Narrowband Spectogram:**
  - **Relatively long analysis window**
  - **High frequency resolution**

Speech harmonic are visible (horizontal striations)

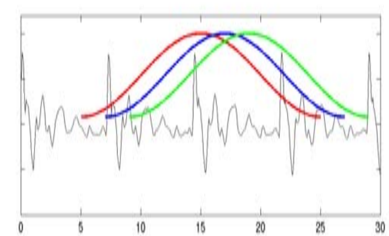


Narrowband Spectogram

Frequency



Frame Number (time)

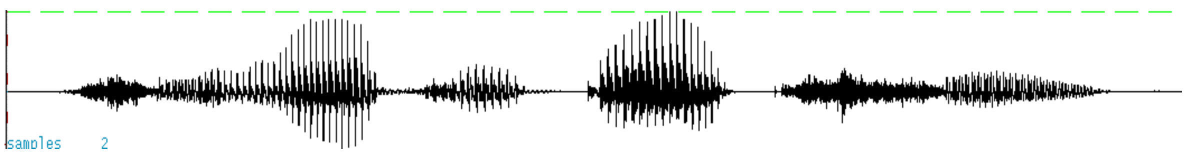


long analysis window

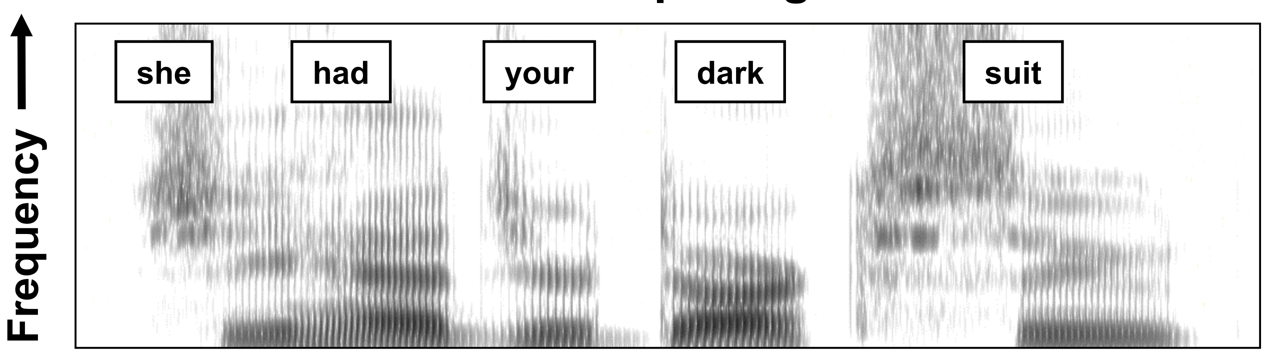
Spectogram

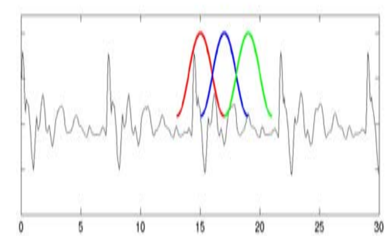
- Spectogram displays the time and frequency evolution of the speech waveform
  - Formants: vocal tract resonances
  - Computed as $\left|X_m(k)\right|$
- Wideband Spectogram:
  - Relatively short analysis window
  - High time resolution

    Pitch pulses are visible (vertical striations)



Wideband Spectogram





short analysis window

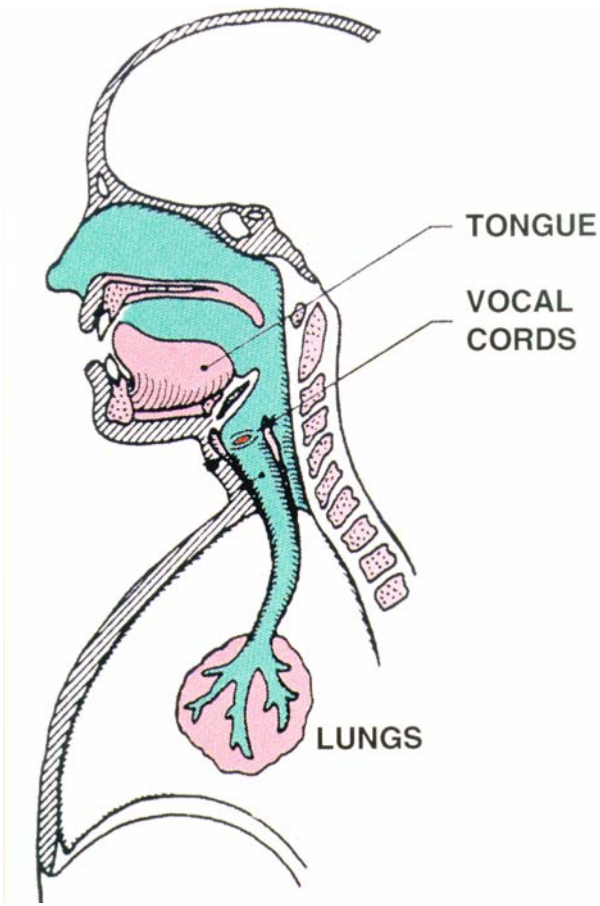
Outline

- **Introduction**
  - **Speech Production and Modeling**
- **Sample Applications**
  - **Signal Processing**
    - **Modification, Enhancement**
  - **Recognition**
    - **Words, language, speaker**
- **Signal Processing for Recognition**
  - **Feature Extraction**

Voice Modification

IEEE Signal Processing Society

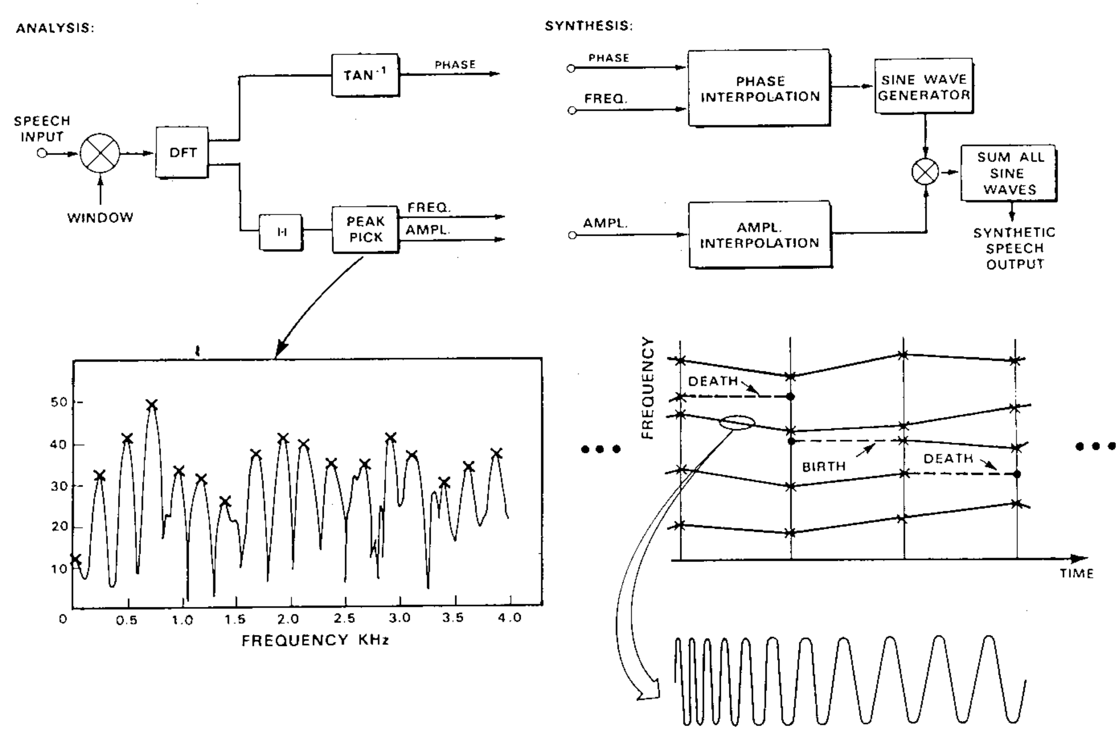
Speech Production



- **Features of Speech Production:**
  - vibration rate of vocal chords
  - vocal tract length and shape
- **Transformations:**
  - change vocal chord pitch
  - change vocal tract length by expanding/compressing spectrum
- **Example Applications:**
  - Online gaming / role playing
  - Voice disguise for anonymous TV interview
- **Example Transformations:**
  - Female Xform to Male
  - Male Xform to Female

Sinusoidal Analysis/Synthesis

- **Analysis and synthesis based on a sinusoidal model**
  - Signal is sum of time varying sinusoids
  - Smooth frame concatenation via magnitude/phase interpolation
  - Accurate temporal structure via speech phase
- **Sinusoids can be stretched and compressed in time to alter articulation rate**

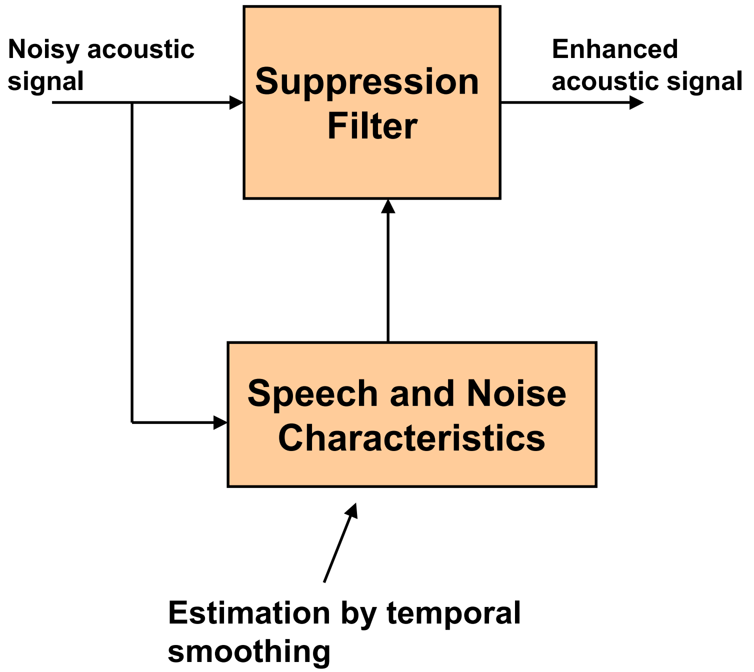


- **Example: Time scaling of female vocalist**
  - Original
  - Fast-slow articulation rate (Oscillates between 2:1 compression and 1:1.5 expansion) with unvoiced regions modified less than voiced regions

Wideband Interference Reduction

- **Single channel noise reduction**
  - No noise reference available
  - Existing systems have not improved intelligibility of enhanced speech
- **Spectral subtraction**
  - Speech spectrum = Noisy spectrum – Background spectrum
  - Background spectrum estimate required
- **Harmonic-driven systems**
  - Comb filtering of the noisy spectrum
  - Pitch estimate required
  - Unvoiced speech not enhanced
- **Wiener filter**
  - Filter gives best estimate of speech in a mean-squared error sense
  - Background and speech spectrum estimates required

Classic approaches require speech and/or noise measurements from the noisy acoustic signal



Narrowband Interference Reduction

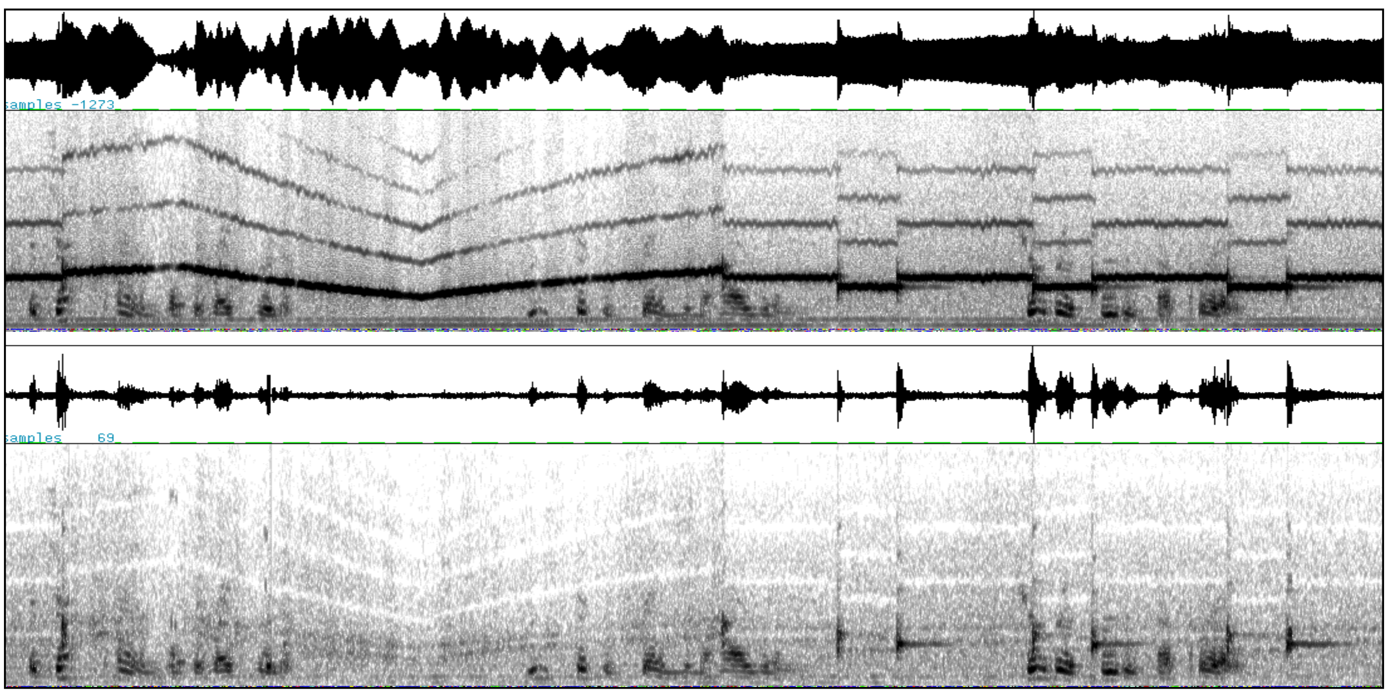
- Narrowband interference can be removed while leaving most of the speech intact
- Stationary tones can be easily removed with notch filters
- Time-varying tones can be tracked, estimated and removed
- Intelligibility may be improved

Original
Waveform

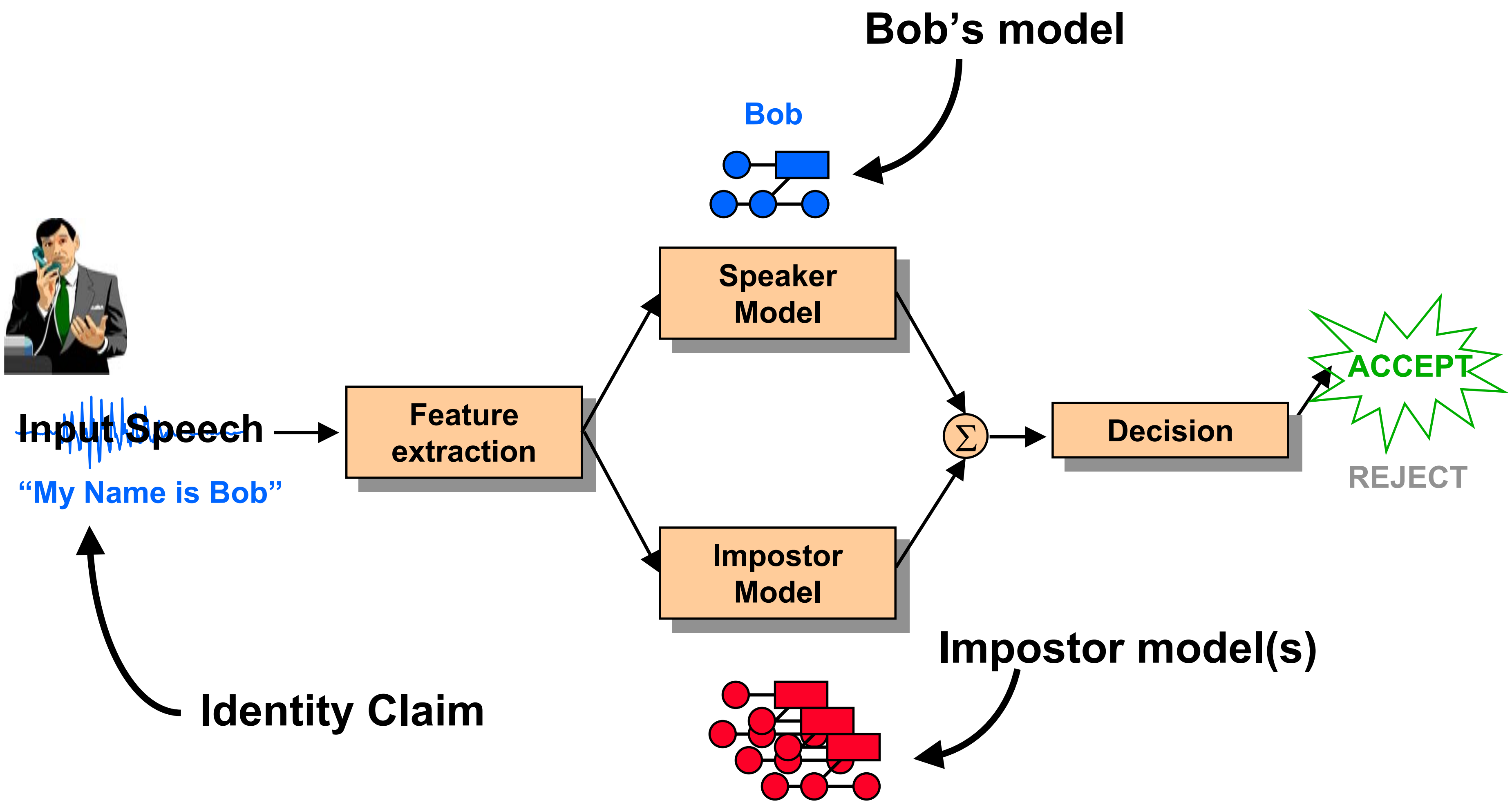
Spectogram

Processed
Waveform

Spectogram

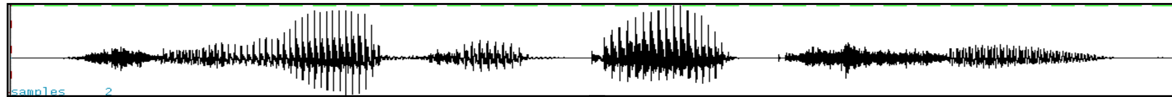


Speaker Verification System



Speech Recognition

acoustic signal



compute features → spectral-based features: mel cepstra

determine words → she had your dark suit

- **Find words that maximize $P(W \mid F)$:**

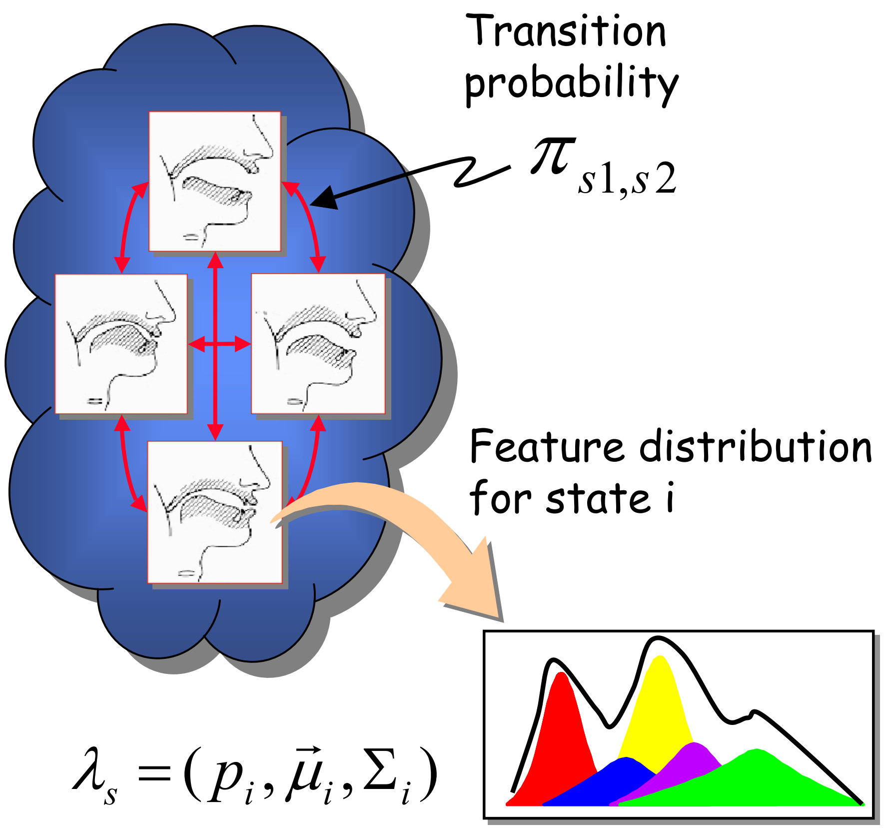
$$P(W \mid F) = \frac{P(F \mid W)P(W)}{P(F)} \qquad W = word, \quad F = features$$

  – $P(F \mid W)$: acoustic model (e.g. HMMs)
  – $P(W)$: language model (e.g. N-gram)

- **Maximizing $P(W \mid F)$ requires large scale searches**
  – Search algorithms are a critical system component
  – Pruning used

Acoustic Modeling

- Feature vectors generated from each speech state follow a Gaussian mixture distribution
- Transition between states based on modality of speech
  - Text-dependent case will have ordered states
  - Text-independent case will allow all transitions
- Model parameters
  - Transition probabilities
  - State mixture parameters
- Parameters are estimated from training speech using Expectation Maximization (EM) algorithm

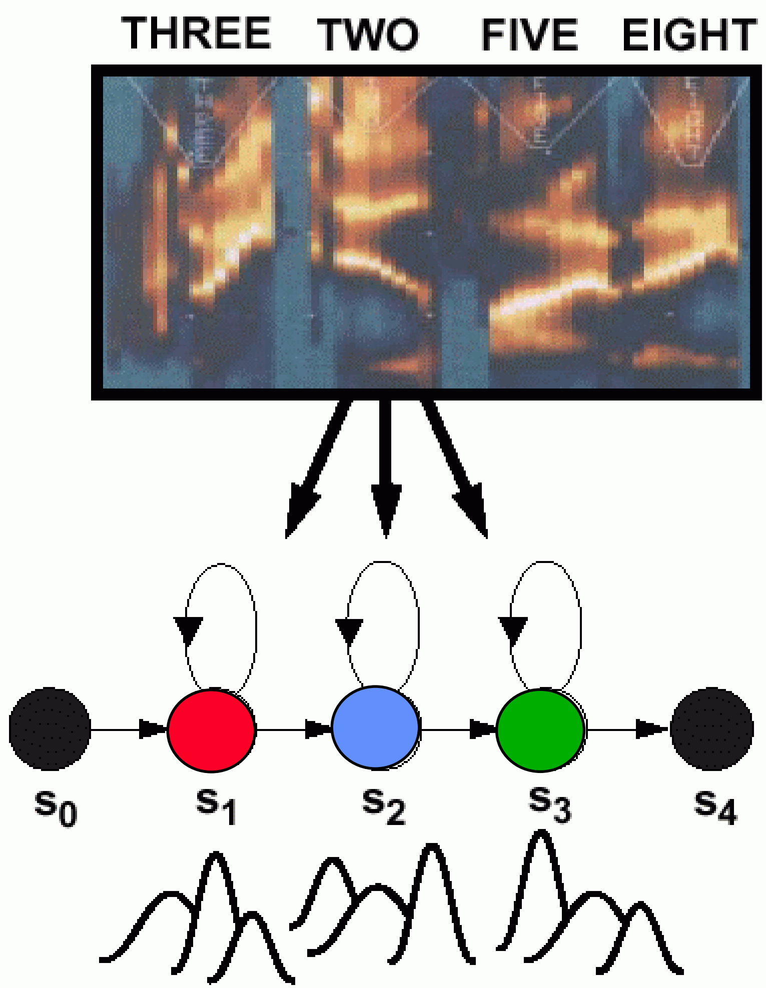


$$\lambda_s = (p_i, \vec{\mu}_i, \Sigma_i)$$

$$p(\vec{x} \mid \lambda_s) = \sum_{i=1}^{M} p_i b_i(\vec{x})$$

Hidden Markov Models

- **HMMs encode the temporal evolution of the features (spectrum)**
- **HMMs represent underlying statistical variations in the speech state (e.g., phoneme) and temporal changes of speech between the states.**
- **This provides a statistical model of a sound is produced**
- **Designer needs to set**
  - **Topology (# states and allowed transitions)**
  - **Number of mixtures**



Outline

- **Introduction**
  - **Speech Production and Modeling**
- **Sample Applications**
  - **Signal Processing**
    - **Modification, Enhancement**
  - **Recognition**
    - **Words, language, speaker**
- **Signal Processing for Recognition**
  - **Feature Extraction**

Speaker Recognition

- **Humans use several levels of perceptual cues for speaker recognition**

High-level cues (learned traits)

Difficult to automatically extract

Hierarchy of Perceptual Cues

Semantics, diction, pronunciations, idiosyncrasies	Socio-economic status, education, place of birth
Prosodics, rhythm, speed intonation, volume modulation	Personality type, parental influence
Acoustic aspect of speech, nasal, deep, breathy, rough	Anatomical structure of vocal apparatus

Low-level cues (physical traits)

Easy to automatically extract

- **There are no exclusive speaker identity cues**
- **Low-level acoustic cues most applicable for automatic systems**

Features for Speaker Recognition

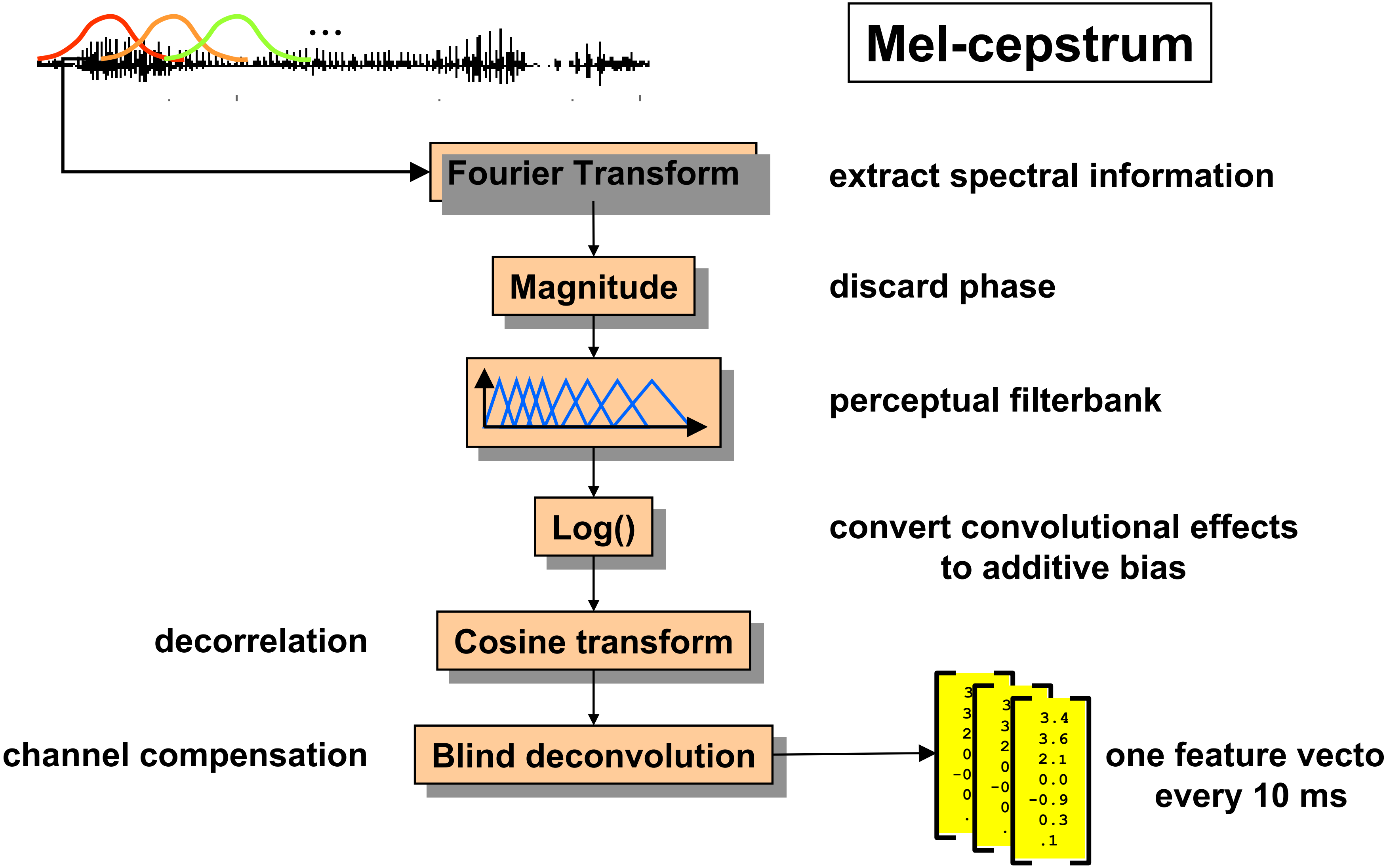
- Desirable attributes of features for an automatic system (Wolf '72)

Practical	• Occur naturally and frequently in speech • Easily measurable
Robust	• Not change over time or be affected by speaker's health • Not be affected by reasonable background noise nor depend on specific transmission characteristics
Secure	• Not be subject to mimicry

- No feature has all these attributes
- Features derived from spectrum of speech have proven to be the most effective in automatic systems
  - These features are also most effective for speech recognition

Feature Extraction

IEEE Signal Processing Society



Features for Speaker Recognition

- Primary feature used in speaker recognition systems are cepstral feature vectors
- Perceptually-based filter-bank
  - Mimics cochlear filters in the ear
  - Removes pitch information
  - Reduces number of features
  - Bandwidth constrained to remove out-of-channel noise
- Log() function turns linear convolutional effects into additive biases
  - Easy to remove using blind-deconvolution techniques
- Cosine transform helps decorrelate elements in feature vector
  - Less burden on model and empirically better performance
- Cepstral mean subtraction (CMS)
  - Blind deconvolution removes convolutional channel effects
- 1st order delta features appended

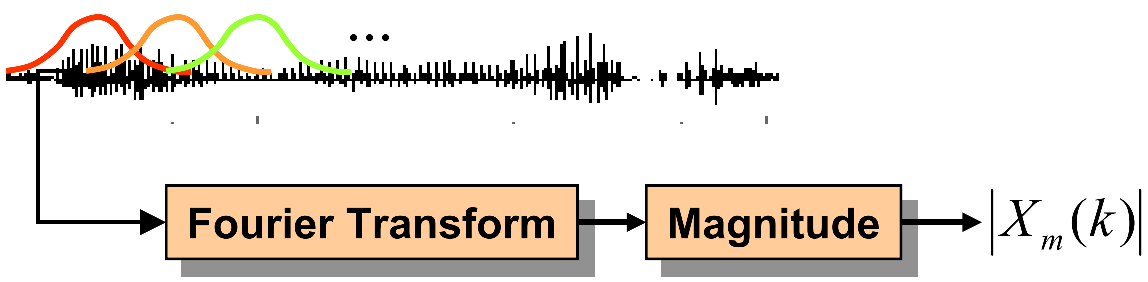
Frame-Based Analysis

- **The speech waveform is typically analyzed on a frame-by-frame basis**
  - **Windowed speech waveform on frame m:**

$$x_m(n) = w(n)s(n - mT)$$

  - **Fourier transform on frame m:**

$$X_m(k) = \sum_{n=0}^{N-1} x_m(n) e^{-j\frac{2\pi}{N}kn} \qquad 0 \le k \le N-1$$



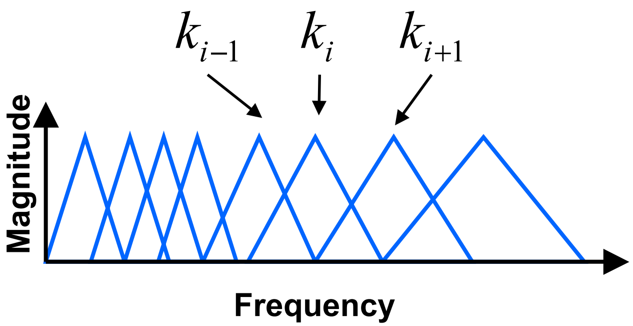
speech signal: $s(n)$
analysis window: $w(n)$
sample index: n
frame number: m
frame interval: T
frequency index: k
DFT length: N

Simulated Perceptual Filterbank

- Perceptually based filters created using Mel frequency scale
- Mel-scale center frequencies are approximately spaced:
  - linearly below 1000 Hz
  - logarithmically above 1000 Hz

$$F_{mel}(i) = \begin{cases} 100\,(i+1) & i < 10 \\ 1.1 F_{mel}(i-1) & i \geq 10 \end{cases}$$

- At each center frequency, a triangular filter extend from the previous to the following center frequency



Triangular Filter Definition

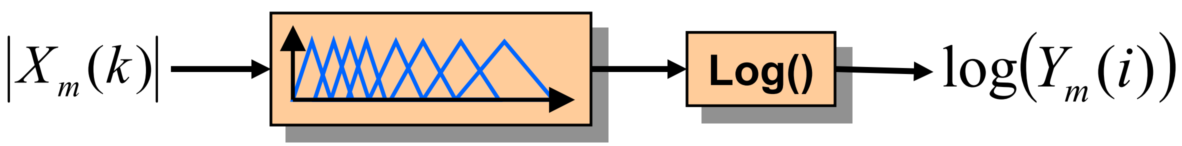
$$F(i,k) = \begin{cases} 0 & k \leq k_{i-1} \\ (k - k_{i-1})/(k_i - k_{i-1}) & k_{i-1} < k < k_i \\ 1 & k = k_i \\ (k - k_i)/(k_{i+1} - k_i) & k_i < k < k_{i+1} \\ 0 & k \geq k_{i+1} \end{cases}$$

Apply Filterbank

- **Apply simulated filterbank to DFT magnitude:**

$$Y_m(i) = \sum_{k=0}^{N-1} F(i,k) \left| X_m(k) \right|$$

- **Apply log to filterbank output**



Discrete Cosine Transform

- **Cosine transform helps decorrelate elements in feature vector**
  - **less burden on model and empirically better performance**
  - **definition:**

$$c(k) = \alpha(k) \sum_{n=0}^{N-1} x(n) \cos\left[\frac{\pi(2n+1)k}{2N}\right] \quad 0 \le k \le N-1$$

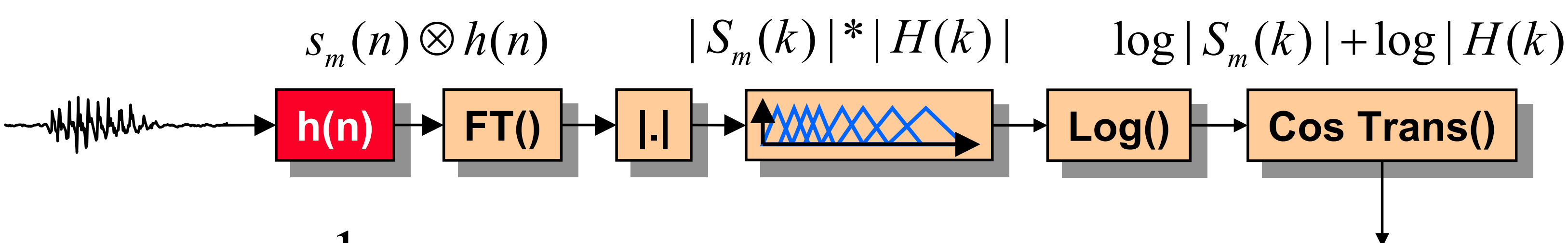
$$\alpha(0) = \sqrt{1/N}, \quad \alpha(k) = \sqrt{2/N} \quad 1 \le k \le N-1$$

$$\log\big(Y_m(i)\big) \longrightarrow \boxed{\textbf{Cosine transform}} \longrightarrow c_m(q)$$

$$c_m(q) = \alpha(k) \sum_{i=0}^{I-1} \log\big(Y_m(i)\big) \cos\left[\frac{\pi(2i+1)q}{2I}\right]$$

Channel Compensation

- **Blind deconvolution is used to help remove convolutional channel effects**
  - **cepstral mean subtraction (CMS) is applied to the cepstral vectors**



$$c_m(q) = s_m(q) + h(q)$$

$$\bar{c}(q) = \frac{1}{M}\sum_m c_m(q) = \bar{s}(q) + h(q)$$

$$c_m(q) - \bar{c}(q) = s_m(q) - \bar{s}(q)$$

- **Some speaker information is lost, but generally CMS is highly beneficial to performance**

Cepstral Mean Subtraction

- **The cepstral mean is computed for each quefrency, q, as:**

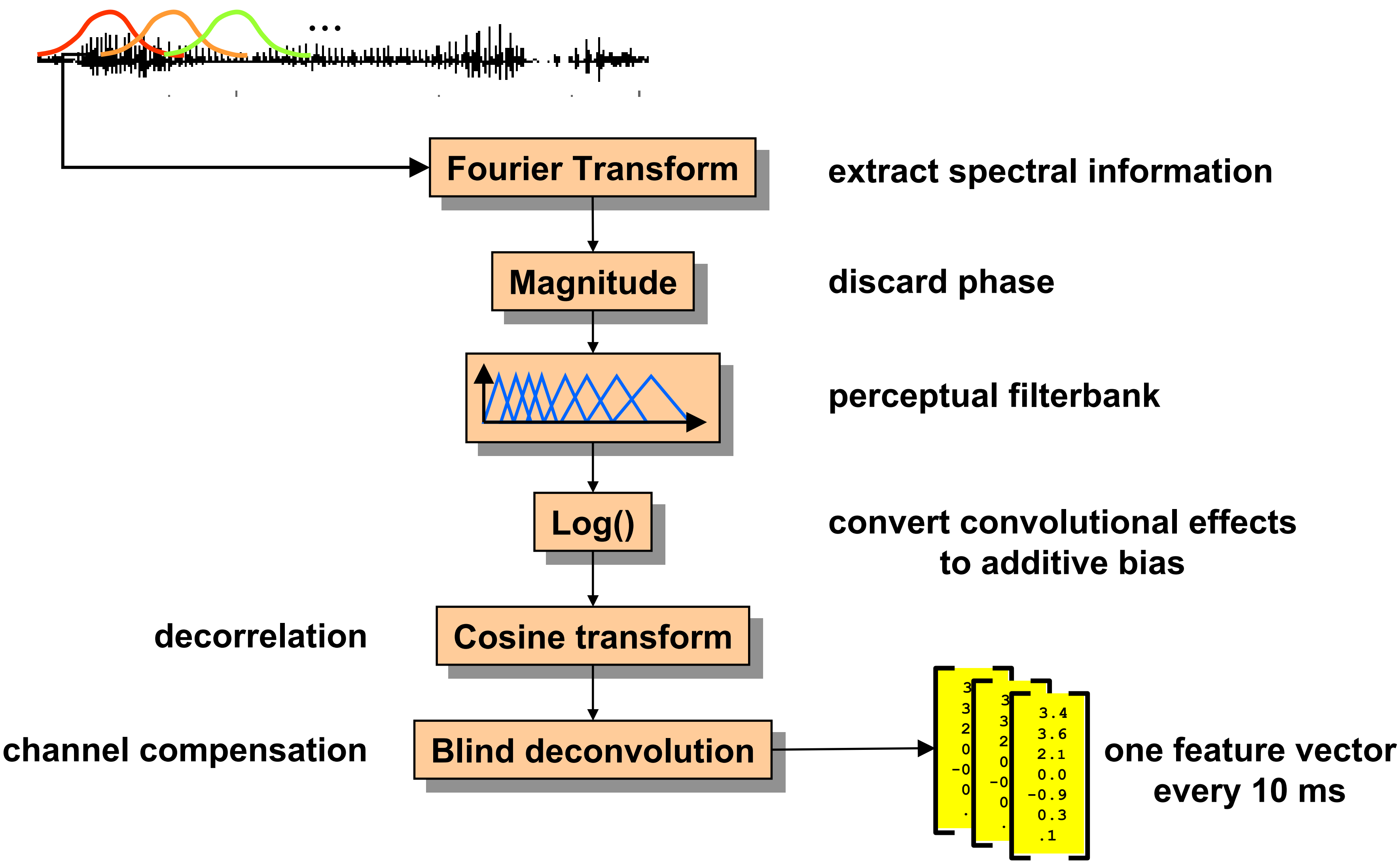
$$\bar{c}(q) = \frac{1}{M}\sum_{m=0}^{M-1} c_m(q)$$

- **Cepstral mean subtraction is non-causal**
  - **A short term mean can be used when causal system is required**

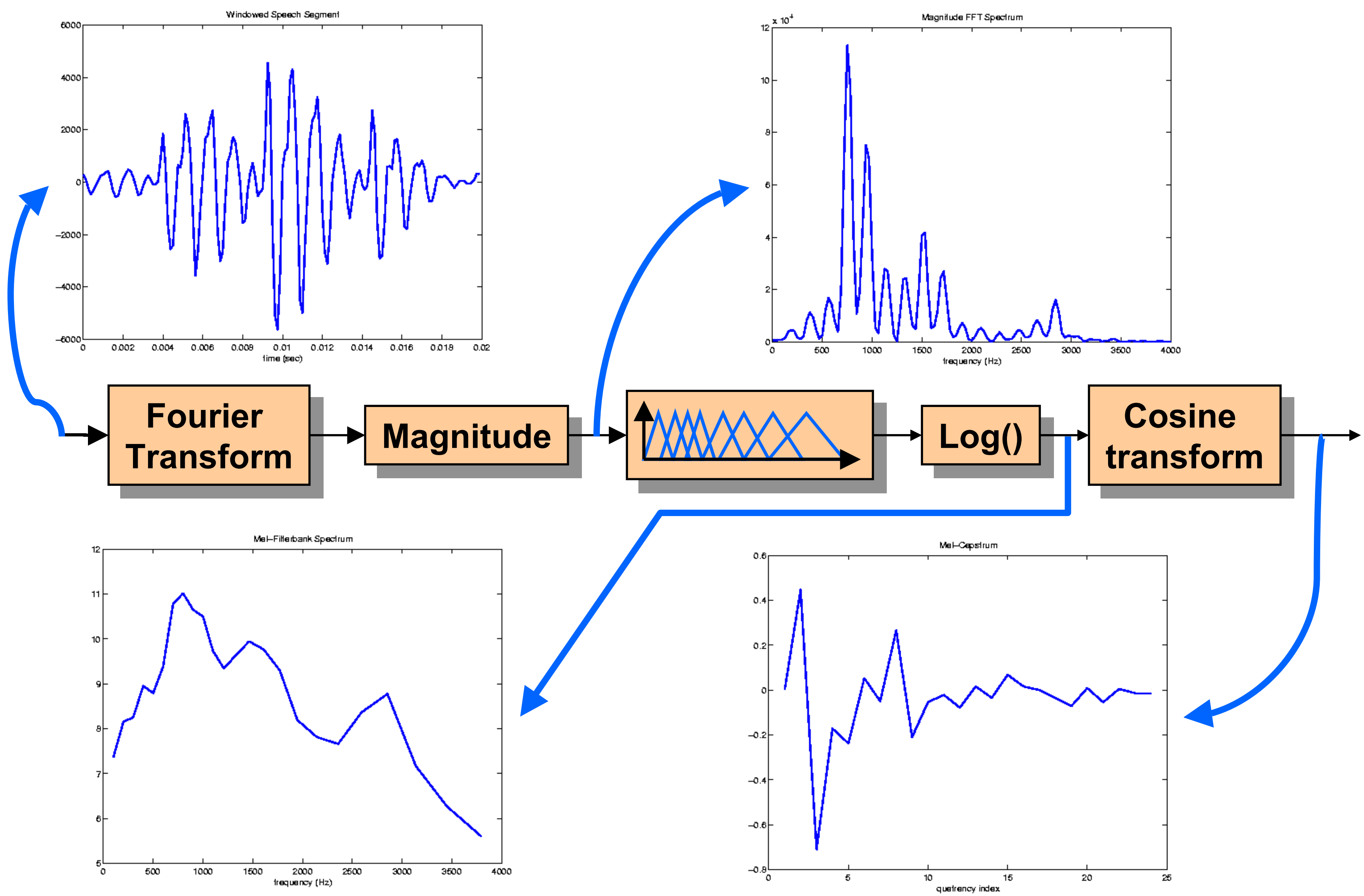
$$c_m(q) \longrightarrow \boxed{\textbf{Blind deconvolution}} \longrightarrow f_m(q)$$

$$f_m(q) = c_m(q) - \bar{c}(q)$$

Feature Extraction



Example Feature Extraction

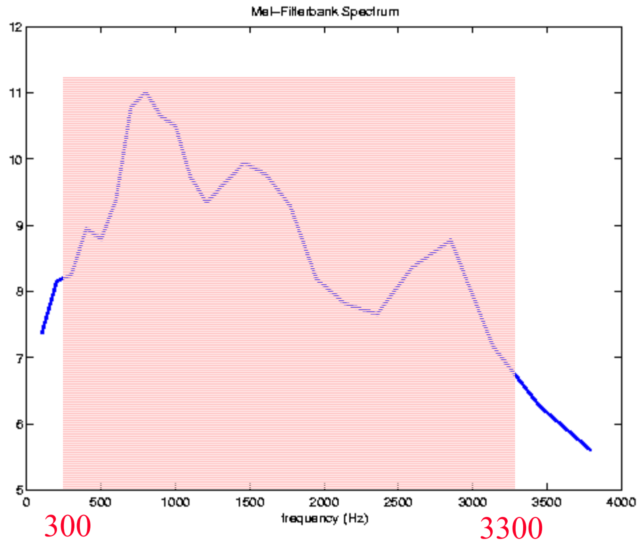


Additional Processing

- Additional processing steps for speaker recognition features
- To help capture some temporal information about the spectra, delta cepstra are often computed and appended to the cepstra feature vector
  - 1st order linear fit used over a 5 frame (50 ms) span

$$\frac{\partial c_m(t)}{\partial t} \approx \Delta c_m(q) = \frac{\sum_{k=-K}^{K} k c_m(q+k)}{\sum_{k=-K}^{K} k^2}$$

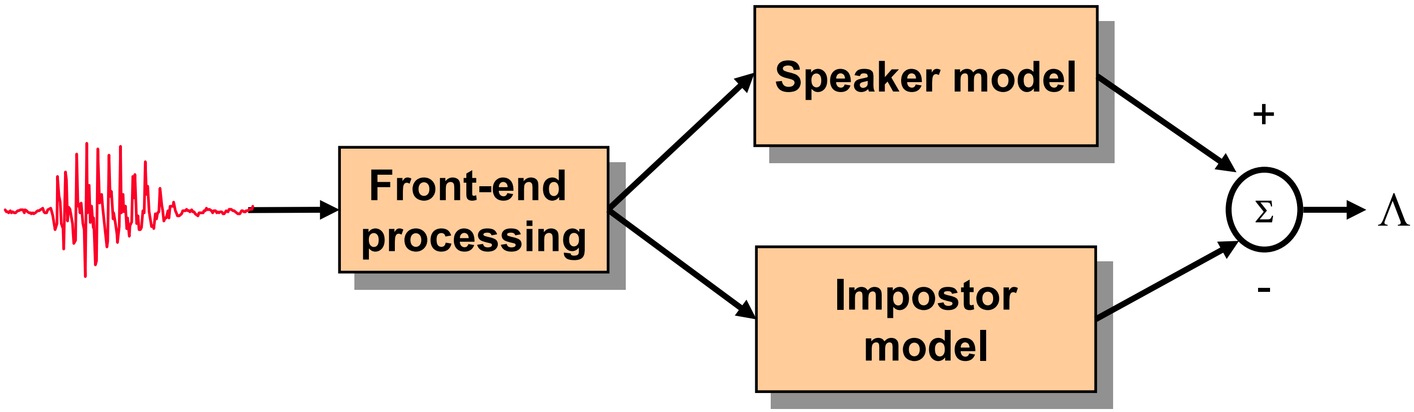
- For telephone speech processing, only voice pass-band frequency region is used
  - Use only output of filters in range 300-3300 Hz



Speaker Verification

- **Usually the log-likelihood ratio is used**

$$LLR = \Lambda = \log p(S \mid H1) - \log p(S \mid H0)$$



$\Lambda > \theta$ Accept

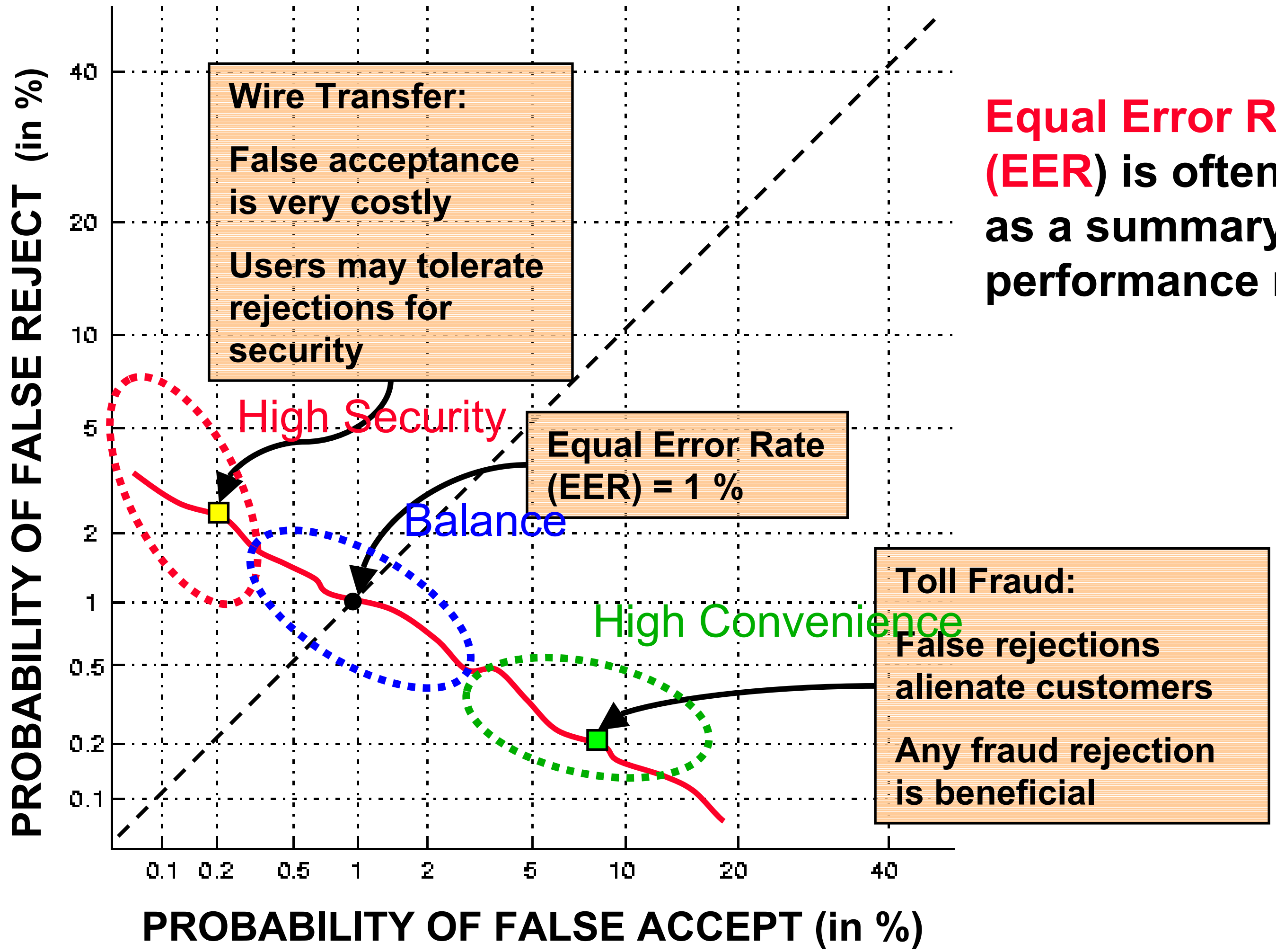
$\Lambda < \theta$ Reject

- **The H1 likelihood is computed using the claimed speaker model**

- **Requires an alternative or impostor model for H0 likelihood**

Speaker Verification Errors

Application operating point depends on relative costs of the two errors



Equal Error Rate (EER) is often quoted as a summary performance measure

Range of Verification Performance