Speaker Verification
(an overview)

Alexandros Xafopoulos
alexandr@zeus.csd.auth.gr
Phd student
Artificial Intelligence & Information Analysis laboratory
Informatics Dpt., Aristotle Univ. of Thessaloniki
Thessaloniki, GREECE

TICSP (Tampere International Center for Signal Processing) visitor, August 2001, TUT (Tampere Univ. of Technology), Tampere, Finland

Presentation Outline

- Framework
- Preprocessing -
- Features (Extraction, Noise Compensation-Channel Equalization, Selection)
- Matching - Modeling
- Decision Making -
- Performance Evaluation
- Experimental Results
- References

Framework

- Introduction
- Related Research Areas
- Generic Speaker Verification Process
- Speech Corpus Parameters
- Errors
- Applications

Introduction

- Motivation
  - Speech contains speaker specific characteristics

    Physiological: body parts (shape, size)
    - larynx (glottis - vocal cords)
    - pharynx, oral & nasal cavities (vocal tract)

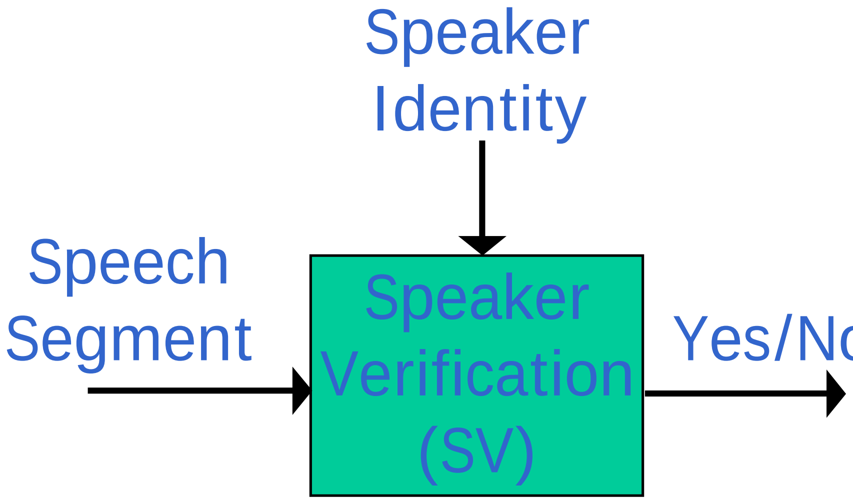
    Behavioral: way they are used
  - Voiceprint as a biometric (distinguishing trait)
  - Natural & economical way of identification

Introduction(2)

- Objective
  - Correct decision on a speaker's identity claim given a speech segment
- Definitions
  - Verification < Latin verus (true)
    - Claim: Speaker identity
    - Proof: Speech utterance
    - Binary decision to establish the truth
  - Client: speaker registered on the system
  - Impostor: speaker who claims a false identity
  - Model: set of parameters that represents a speaker or a group of speakers

Introduction(3)

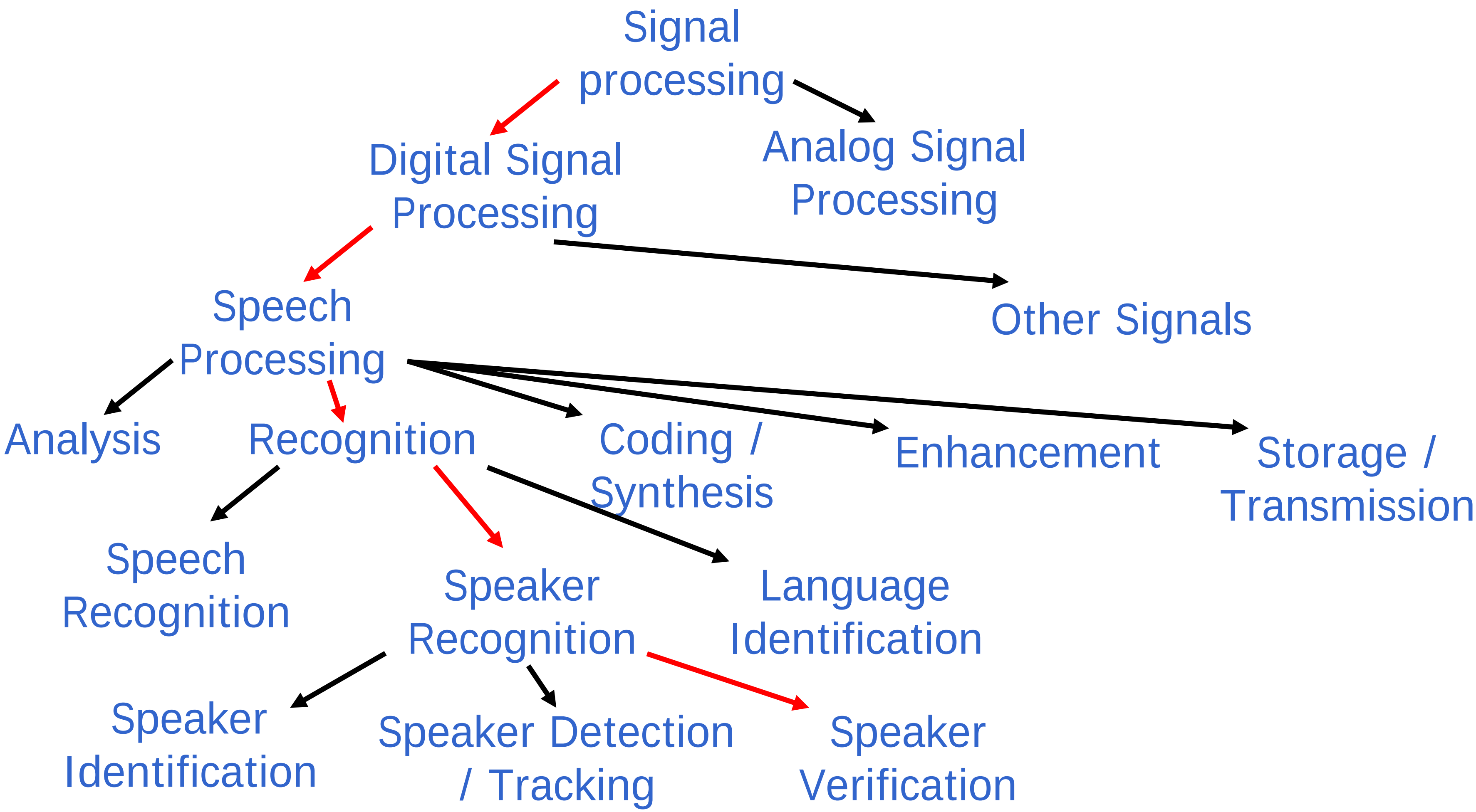
- Abstract schematic



- Example
  - Claimant: I am speaker A
  - SV system: Say: one two three
  - Claimant: One two three
  - SV system: You are not speaker A

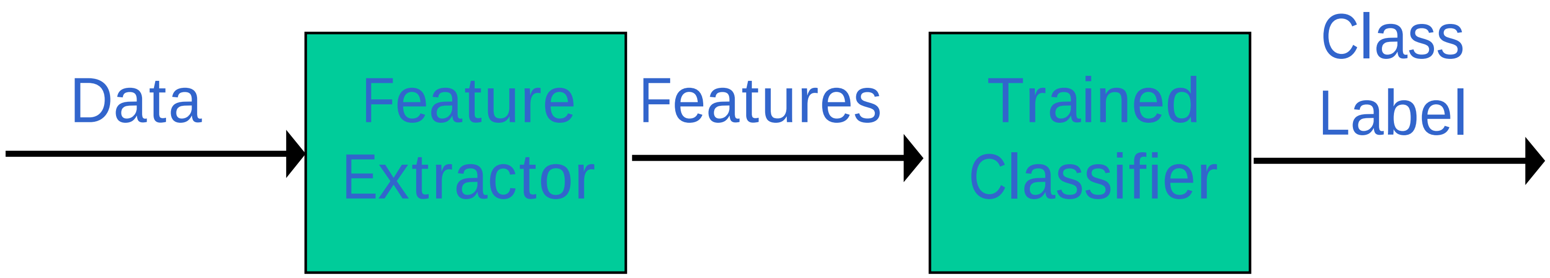
Related Research Areas

- Signal Processing



Related Research Areas(2)

- (Statistical) Pattern Recognition



Related Research Areas(3)

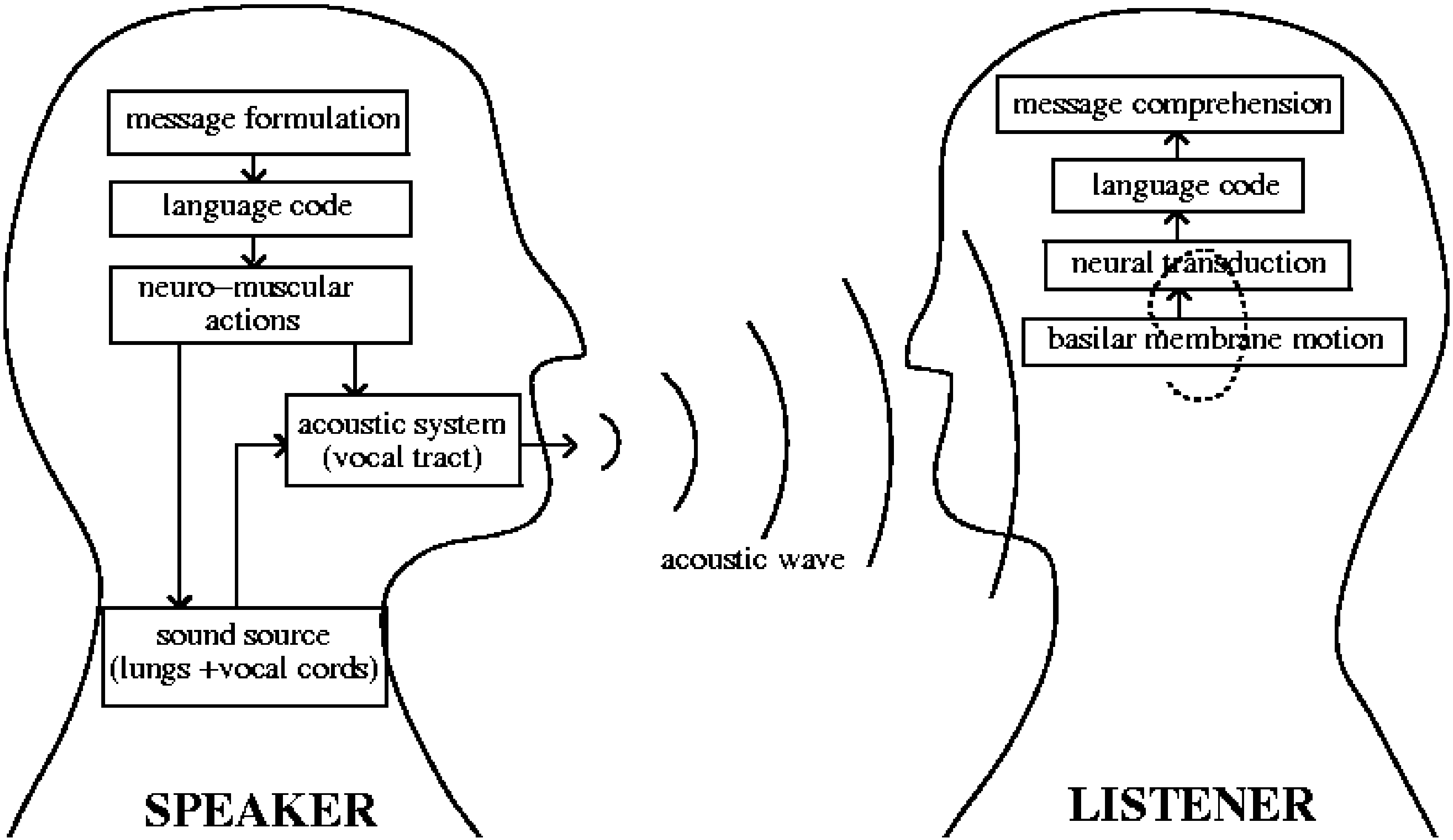
- Biometrics Technology
  - def: automatic recognition of a person based on his/her physiological or behavioral characteristics (biometrics)
  - desirable properties of biometrics [Jain_bk]
    - universality (found in every person)
    - uniqueness (different "value" for each person)
    - permanence (invariant with time)
    - collectability (quantitatively measurable)
    - performance (↗ accuracy vs. ↘ resources)
    - high acceptability (person’s willingness)
    - low circumvention (not easy to deceive)

Related Research Areas(4)

Framework

- Speech Science

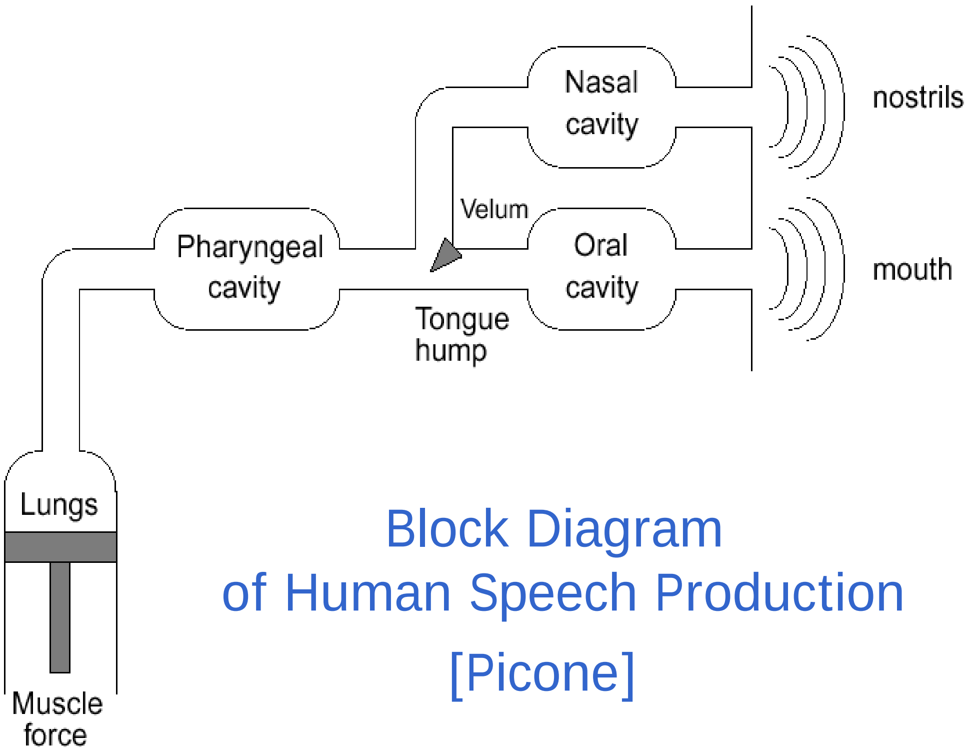
Communication by speech [Somervuo]



Related Research Areas(5)

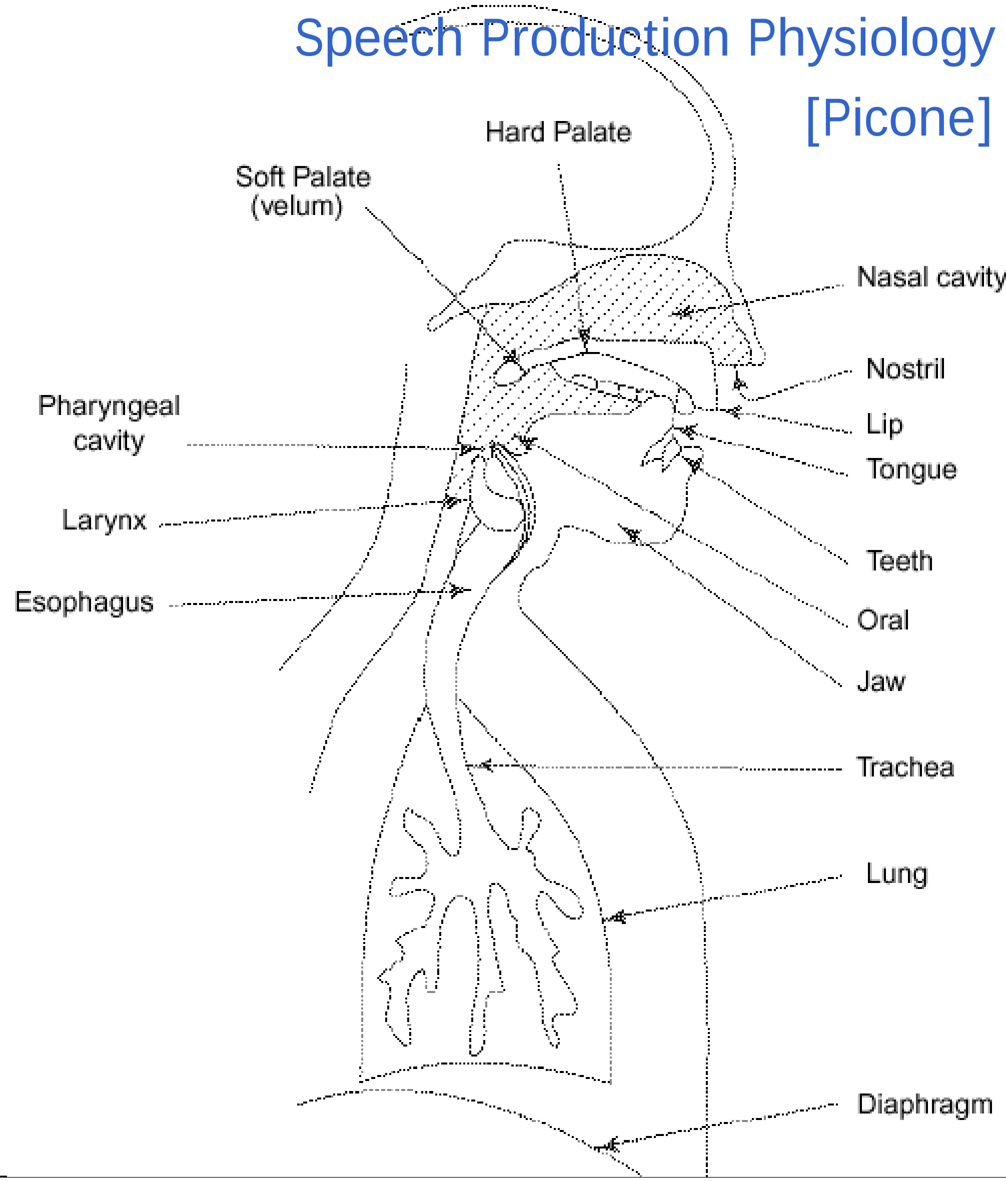
Framework

- Speech Science(2)



Block Diagram
of Human Speech Production
[Picone]

Speech Production Physiology
[Picone]



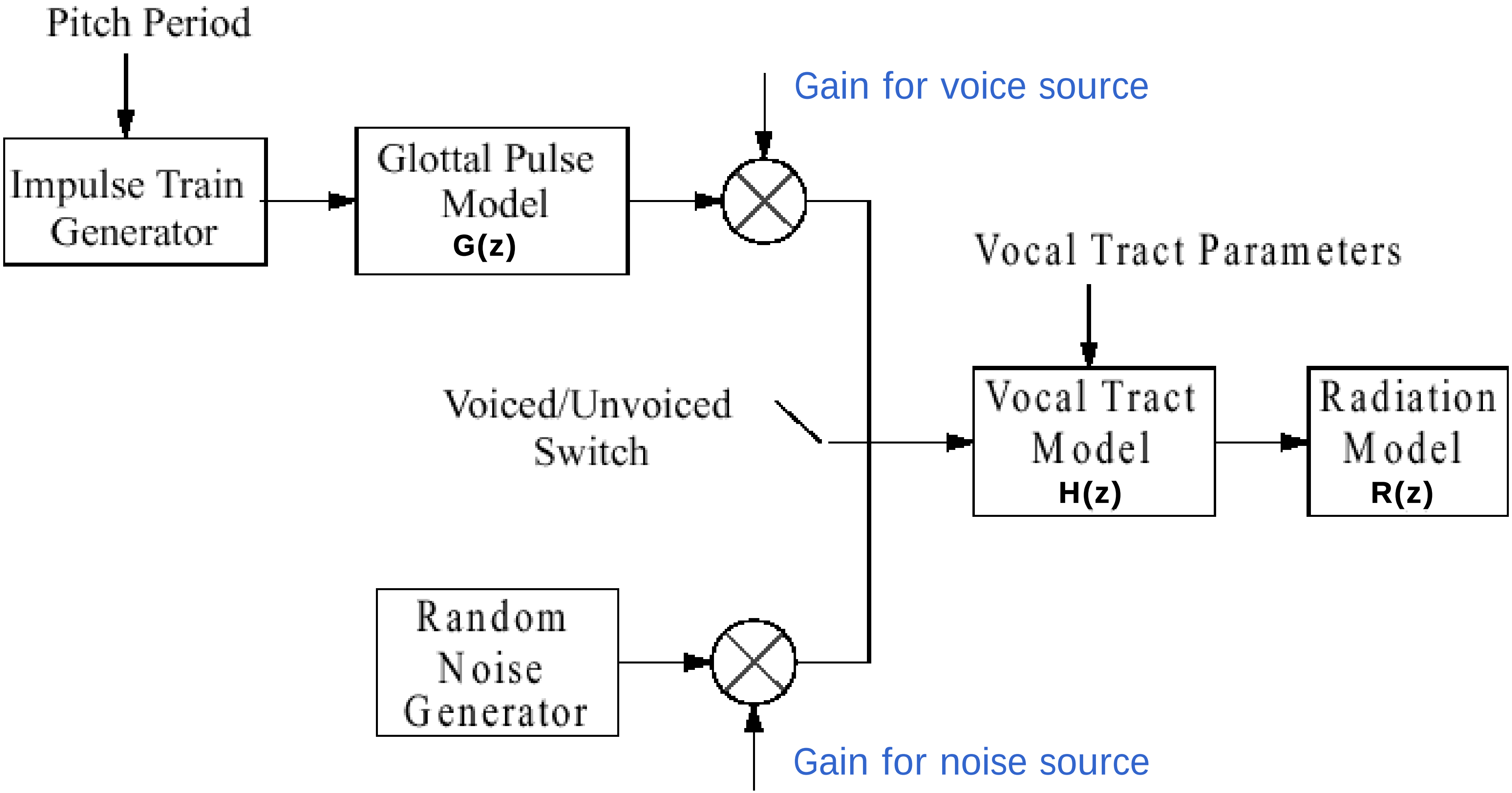
Related Research Areas(6)

Framework

- Speech Science(3)

[Morgan] (modified)

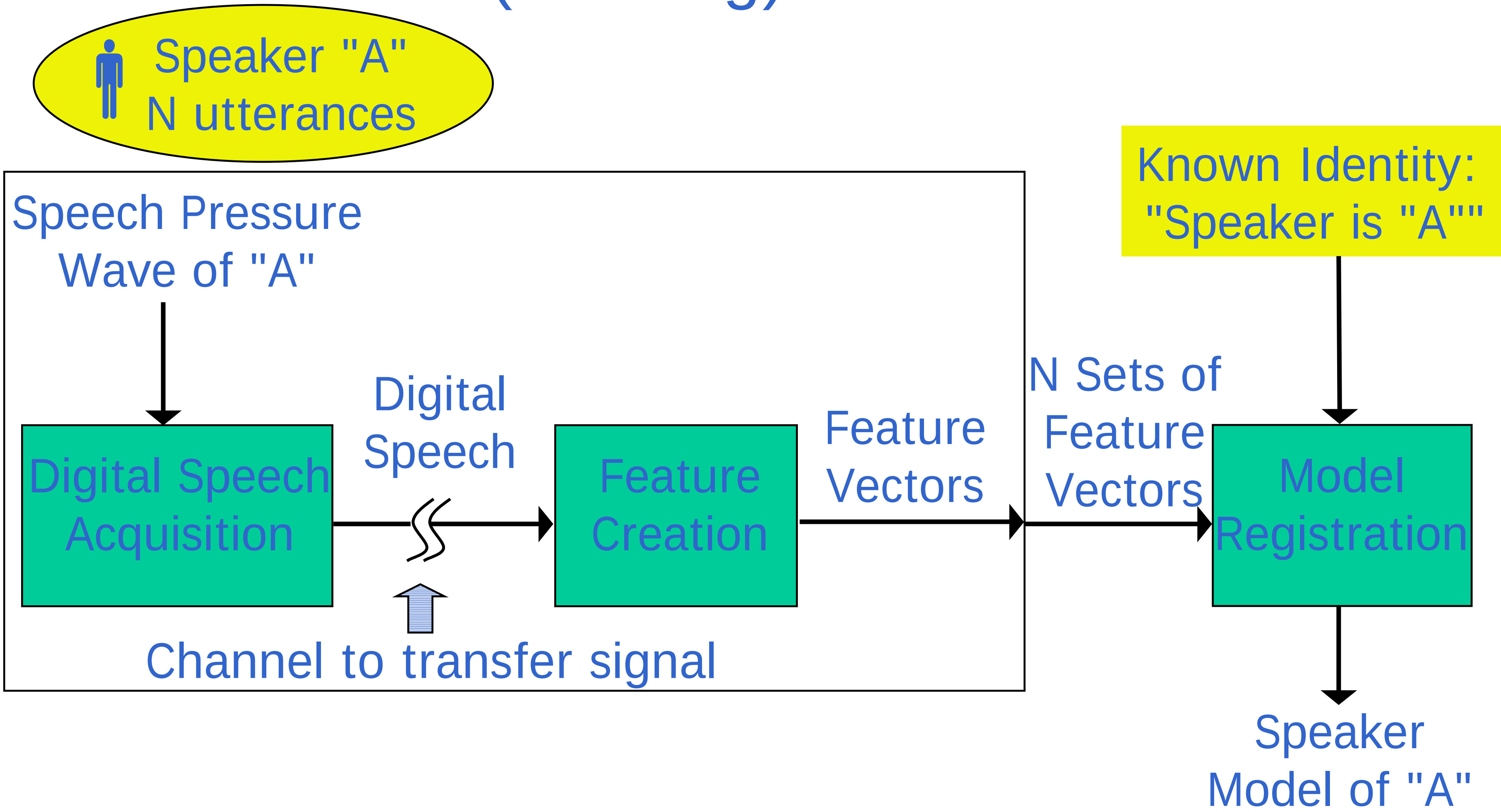
General Discrete-Time Model for Speech Production



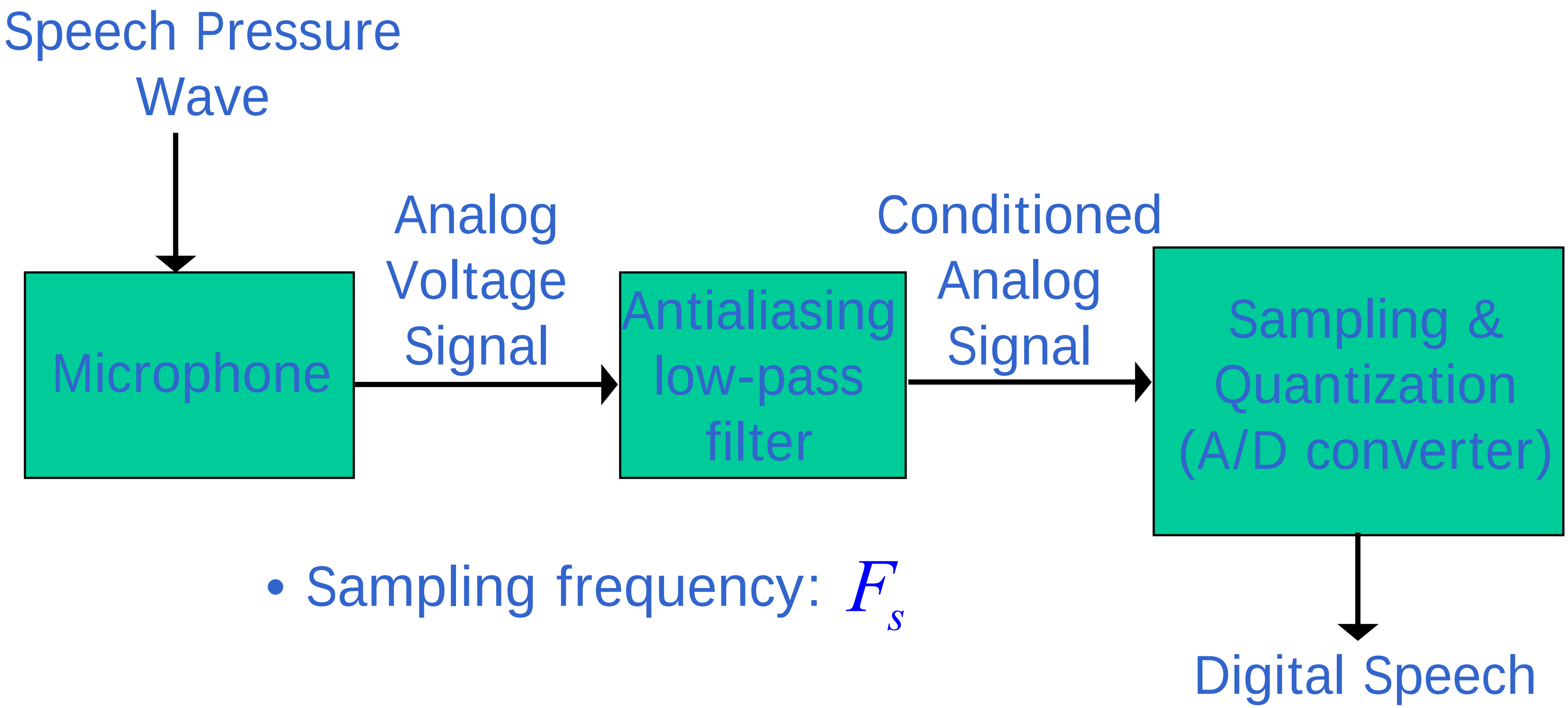
Generic Speaker Verification Process

Framework

- Enrollment (Training) module



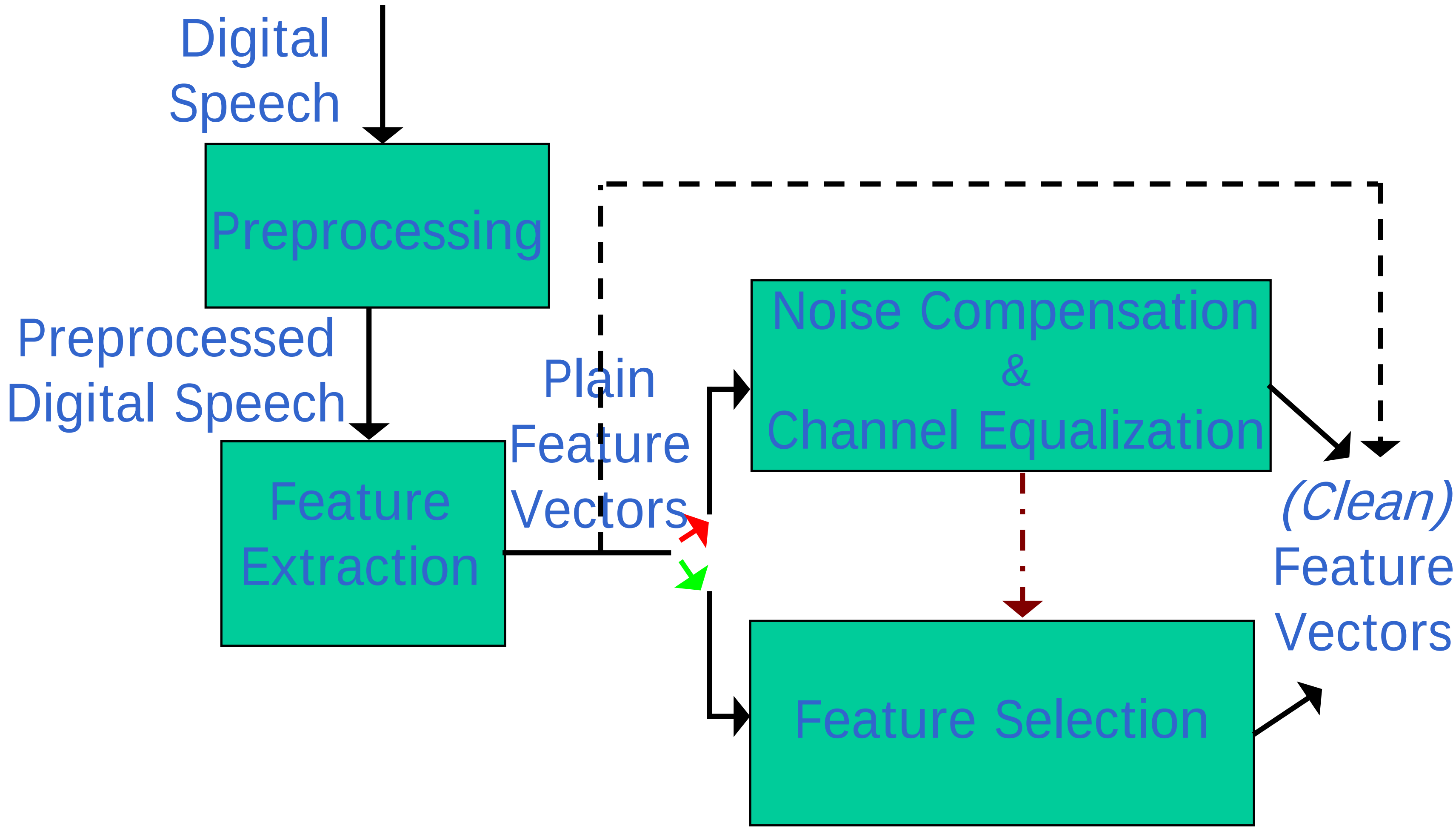
- Enrollment module(2)
  - Digital speech acquisition



- Sampling frequency: F_s

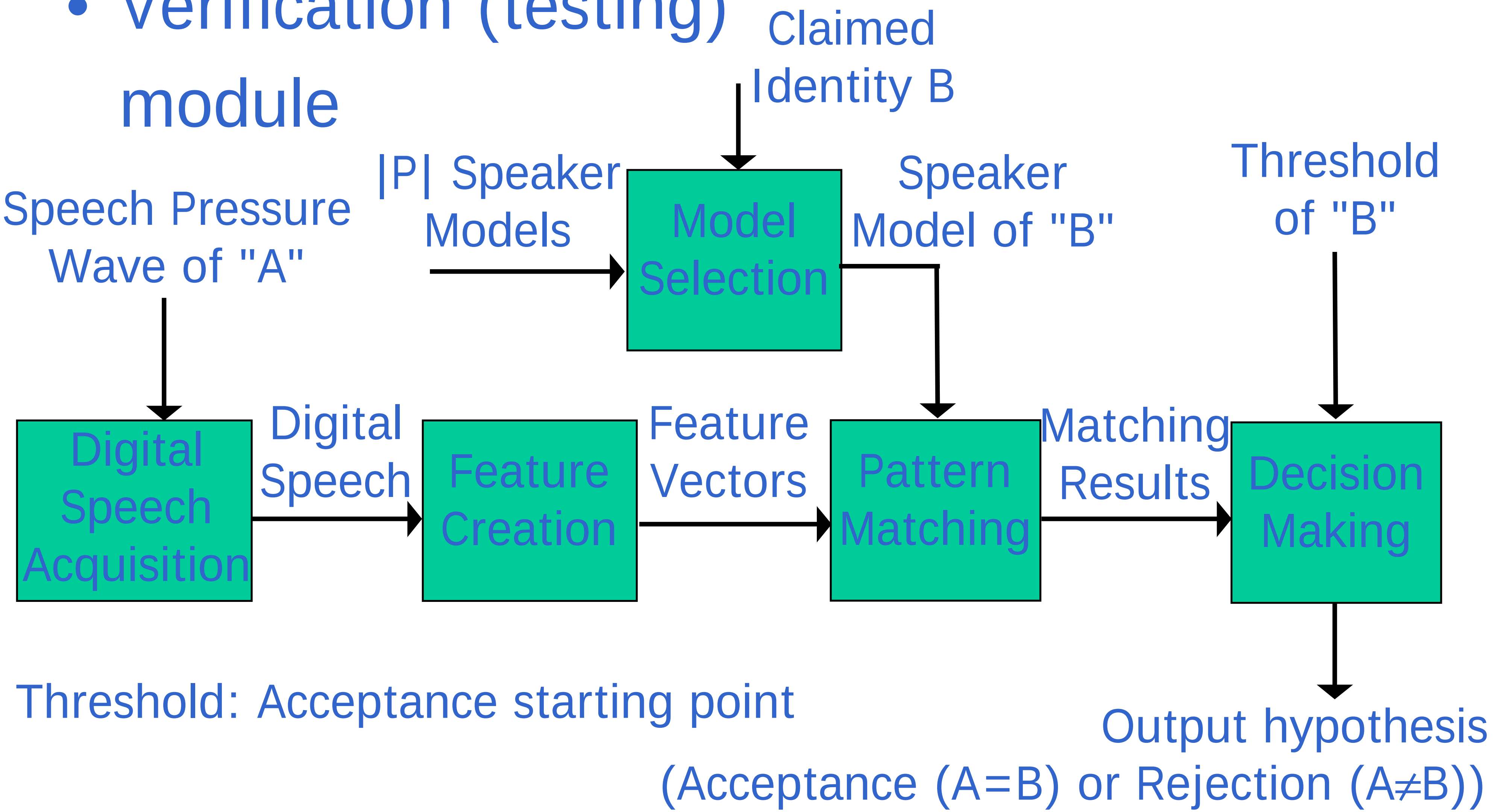
Generic SV Process(3)

- Enrollment module(3): Feature creation



Generic SV Process(4)

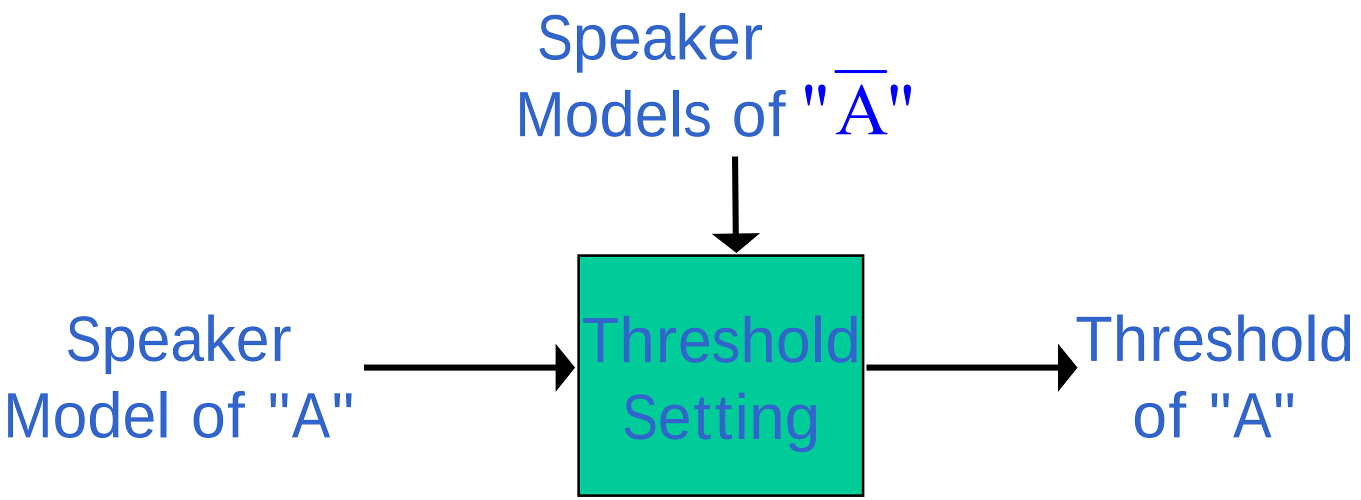
- Verification (testing) module



Threshold: Acceptance starting point

Generic SV Process(5)

- Threshold setting module



$\overline{A} : A_h$ (cohort model) or Ω (world model)

- Cohort model: competitive clients only
- World model: all the clients

Speech Corpus Parameters

- Text-dependency [Nedic]
  - Text dependent (or fixed phrase): verification (& enrollment) done on a fixed phrase (pass phrase), predetermined by the system
  - Text prompted: system list-selected/vocabulary-generated phrase prompted to the user
  - User customized: user list-selected/vocabulary-generated phrase
  - Text independent: user chosen unconstrained phrase
  - Language-dependency
- Vocabulary
  - Fixed or not
  - Size (|V|)

Speech Corpus Parameters(2)

- Population (Speakers)
  - Size (|P|)
  - Degree of similarity
    - gender, age, language, dialect, ...
- Speech Flow
  - Discrete Utterance (pauses betw. words)
  - Continuous
  - Spontaneous (natural)
- Training (system construction) - testing (evaluation) part
- Quantity (#sessions, #phrases, phrase duration)
- Quality of speech (Problems→)

Problems under real conditions

- Due to impostors
  - Mimicry by humans
  - Tape recorders & digital equipment for recording, editing & splicing sound
- Due to clients
  - Bad pronunciation
  - Extreme emotional states (e.g. anger)
  - Sickness / Allergies / Tiredness / Thirst
  - Aging

Problems under real conditions(2)

- Due to the input channel
  - Microphone / Communication channel / Digitizer quality
  - Channel mismatch (different channels for enrollment & verification request)
- Due to the environment
  - Environmental mismatch
  - Environmental noise
  - Poor room acoustics

Errors

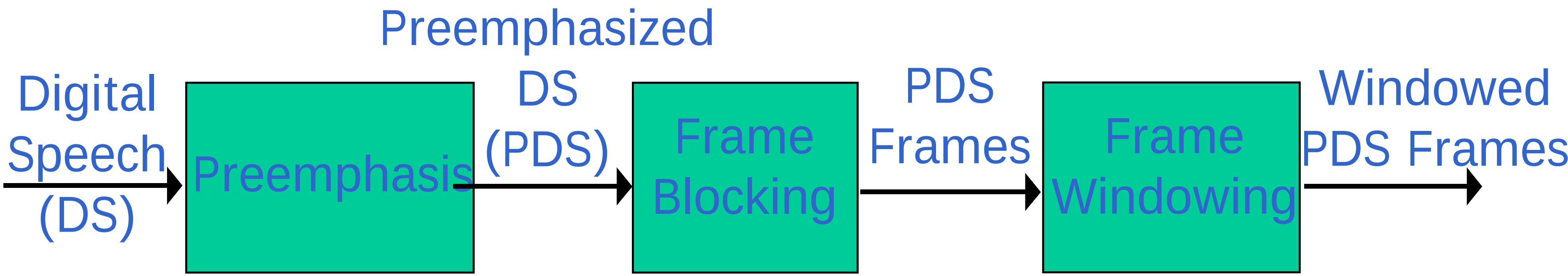
- False Rejection
  - A client request as himself/herself is rejected
  - High rate (rejected) client: goat [Koolwaaij]
  - Low rate (rejected) client: sheep
- False Acceptance
  - An impostor request as a client is accepted
  - High rate (victim) client: lamb
  - Low rate (victim) client: ram
  - High rate (accepted) impostor: wolf
  - Low rate (accepted) impostor: badger

Applications

- Access control to computers / databases / facilities
- Remote access to computer networks
- Electronic commerce
- Forensic
- Telephone banking [James]

Preprocessing

- Preemphasis
- Frame Blocking
- Frame Windowing
- Speech Activity Detection
- Signal Measures & Graphs



Preemphasis

- Preemphasis: Low order digital system to
  - spectrally flatten the signal (in favor of vocal tract parameters)
  - make it less susceptible to later finite precision effects
  - usually 1st order FIR filter:

$$s_{pe}(n) = s(n) - \alpha_{pe} s(n-1), \quad \alpha_{pe} \in [0.9, 1]$$

Frame Blocking

- Frame blocking (short-term(st) processing)
  - L successive overlapping (by M samples) frames

$$f(l;n) = s_{pe}(n + M(l-1)), \quad n = 0,\ldots,N-1, \quad l = 1,\ldots,L$$

  - window size/length: N samples = N/F_s sec (typically some msec)
  - frame rate/shift/period: M samples = M/F_s sec
  - Alternative: non-uniform frame rate

Frame Windowing

- Used to minimize the signal discontinuities at the beg. & end of each frame
  - Time (long window) vs. freq. (short) resolution

$$f_w(l;n) = f(l;n)w(n), \quad n = 0,\ldots,N-1$$

  - Window type:

Generalized Hanning: $$w_H(k) = w(k)\left[\alpha + (1-\alpha)\cos\left(\frac{2\pi}{N}k\right)\right] \qquad 0 < \alpha < 1$$

$$\alpha = 0.54, \qquad \textit{Hamming window}$$
$$\alpha = 0.50, \qquad \textit{Hanning window}$$

[Picone]

  - Modifications:

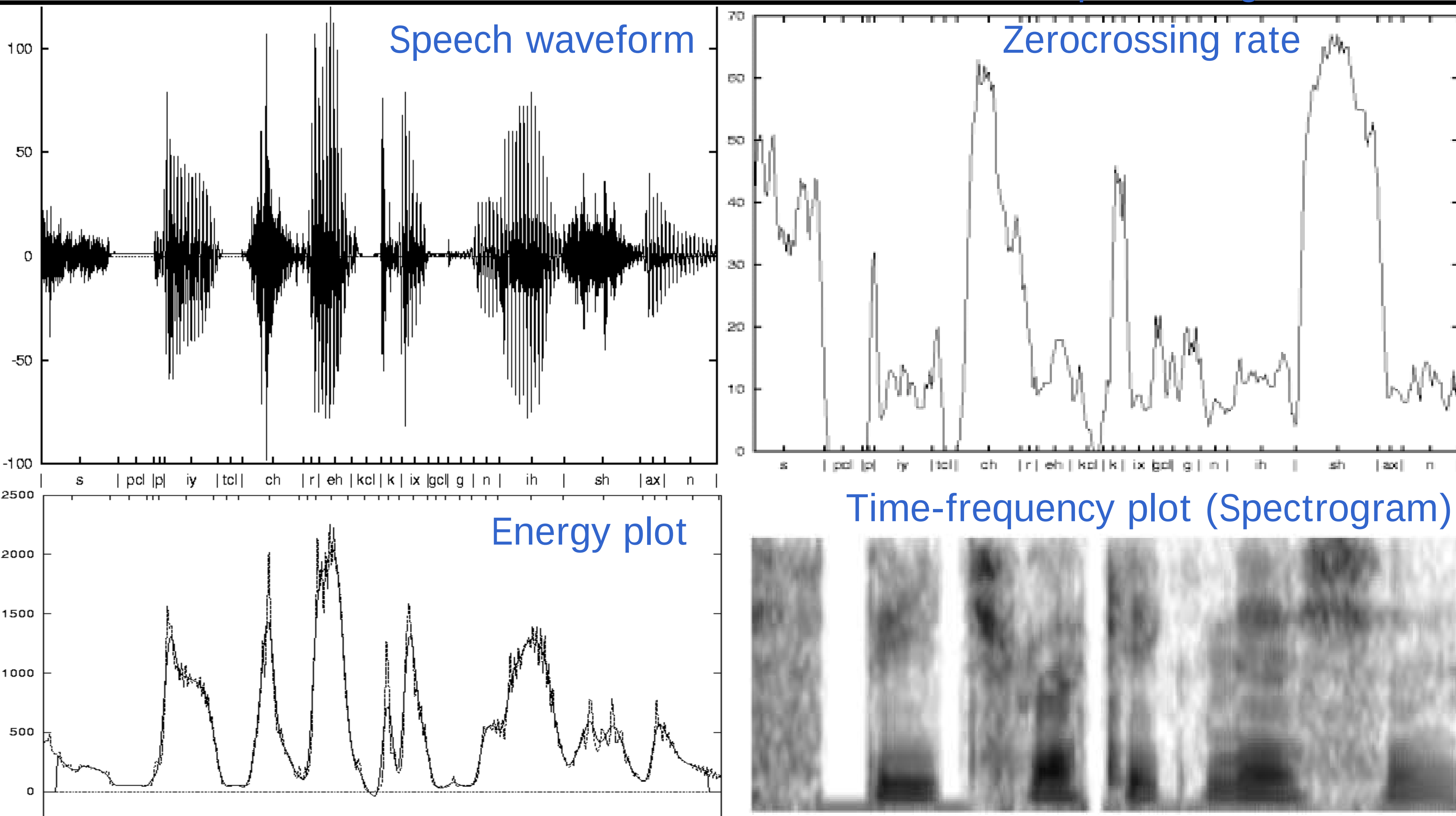
$$N \rightarrow N-1, \quad k \rightarrow n, \quad n = 0,\ldots,N-1$$

Speech Activity Detection

- Silence-speech detection
- Voiced-unvoiced discrimination
  - i.e. with or w/o fast vibration of the vocal cords
- Endpoint detection [Deller_bk]
- Word segmentation
- Applicable at several time points using several criteria-thresholds (energy, zero-crossing rate, feature-based, statistical)

Signal Measures & Graphs

Preprocessing



[Weingessel]

Features

- Feature Extraction
  - Features - General
  - Linear Prediction (LP)
  - Cepstrum (Complex - Real)
  - Mel Cepstrum
  - LP-derived Cepstrum
  - Other Cepstral Variants
  - Variants
  - Delta Cepstrum
  - Perceptual Linear Prediction (PLP) - Auditory Features

Features(2)

- Noise Compensation - Channel Equalization
  - Intra-frame Cepstral Processing
  - Inter-frame Cepstral Processing
  - Relative Spectral (RASTA) Processing
- Feature Selection
  - Principal Component Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Non Linear Discriminant Analysis (NLDA)

Features - General

- Mapping of each input speech interval (1 or more frames) to a multidimensional feature space (vector)
- Order N_{coef} : number of coefficients in each feature vector (dimensionality)
- Several kinds of coefficients proposed
- Ear performs spectral analysis→feature vectors usually consider local spectral energy estimates

Linear Prediction (LP)

- Speech sample as a linear combination of N_{LPC} previous samples (autoregressive (AR) model):

$$s(n) = \sum_{m=1}^{N_{LPC}} a_{LPC}(m) s(n-m) + Gu(n)$$

  - $a_{LPC}(m), \quad m = 1, \ldots, N_{LPC}$: LP coefficients (LPC)
  - $u(n)$: normalized excitation source
  - G : scale factor
  - $a_{LPC}(l;m), \quad m = 1, \ldots, N_{LPC}$: stLPC of frame l

Linear Prediction (LP)(2)

- Calculation of stLPC
  - Mean squared error minimization
  - Autocorrelation method
    - Levinson-Durbin (L-D) recursion
  - Covariance method
    - Cholesky (LU) decomposition

Initialization:

$$E_{LP}^{(0)} = R_n(0) \tag{26}$$

For $1 \le i \le N_{LP}$ {

$$k_{LP}(i-1) = -\frac{R_n(i) + \sum_{j=1}^{i-1} a_{LP}^{(i-1)}(j) R_n(i-j)}{E_{LP}^{(i-1)}} \tag{27}$$

$$a_{LP}^{(i)}(i) = k_{LP}(i-1) \tag{28}$$

For $1 \le j \le i-1$ {

$$a_{LP}^{(i)}(j) = a_{LP}^{(i-1)}(j) + k_{LP}(i-1)\, a_{LP}^{(i-1)}(i-j) \tag{29}$$

}

$$E_{LP}^{(i)} = (1 - k_{LP}(i-1)^2)\, E_{LP}^{(i-1)} \tag{30}$$

}

L-D recursion (*l* is implied, *R*: autocorrelation matrix) [Picone2]

Linear Prediction (LP)(3)

- LPC vectors
  - highly correlated
  - not orthonormal
- Distance: Itakura-Saito
  - Computationally expensive
- LPC processor [Rabiner_bk]

Cepstrum (Complex - Real)

- Special case of homomorphic signal proc. [Deller_bk]
- provides a method for separating the vocal tract info (system) from the glottal excitation
- Focuses on voiced segments
- Short-term complex cepstrum (stCC):

$$c_{CC}(l;m) = \mathrm{DFT}^{-1}\{\log_{10}(\mathrm{DFT}\{f_w(l;n)\})\}, \quad m = 1,\ldots,N_{CC}, \quad n = 0,\ldots,N-1$$

- Short-term real cepstrum (stRC):

$$c_{RC}(l;m) = \mathrm{DFT}^{-1}\{\log_{10}|\mathrm{DFT}\{f_w(l;n)\}|\}, \quad m = 1,\ldots,N_{RC}, \quad n = 0,\ldots,N-1$$

  – No phase information, usu. acceptable

Cepstrum (Complex - Real)(2)

- Distance of cepstrum based coefficients
  - Euclidean: vectors defined in an orthonormal space

$$D_{Eucl.}(l_1, l_2; RC) = \sum_{m=1}^{N_{RC}} \left(c_{RC}(l_2; m) - c_{RC}(l_1; m)\right)^2$$

  - Weighted Euclidean
    - weighted by the inverse of the corresponding covariance matrix element

Cepstrum (Complex - Real)(3)

- If the speech is considered as the output of the vocal tract system *v* having as input the glottal excitation *g*:

$$f_w(l;n) = g(l;n) * v(l;n), \quad n = 0,\dots,N-1$$

$$c_{RC}(l;m) = \mathrm{DFT}^{-1}\{\log_{10}|\mathrm{DFT}\{\mathrm{g}(l;n) * v(l;n)\}|\}, n = 0,\dots,N-1$$

$$= \mathrm{DFT}^{-1}\{\log_{10}|\mathrm{G}(l;k)| + \log_{10}|V(l;k)\}|\}, k = 0,\dots,N_{DFT}-1$$

$$= g_2(l;m) + v_2(l;m), \quad m = 0,\dots,N-1$$

- the 1^{st} coeffs represent the slowly varying vocal tract parameters & the remaining coeffs model the quickly varying excitation signal→selection of the 1^{st} N_{RC} coeffs excluding 0^{th}

Mel Cepstrum

- Mel
  - unit of measure of perceived frequency of a tone
  - non-linear correspondence to the physical freq. (like the human ear)

$$f_{mel} = 2595 \log_{10}\left(1 + \frac{f_{Hz}}{700}\right)$$

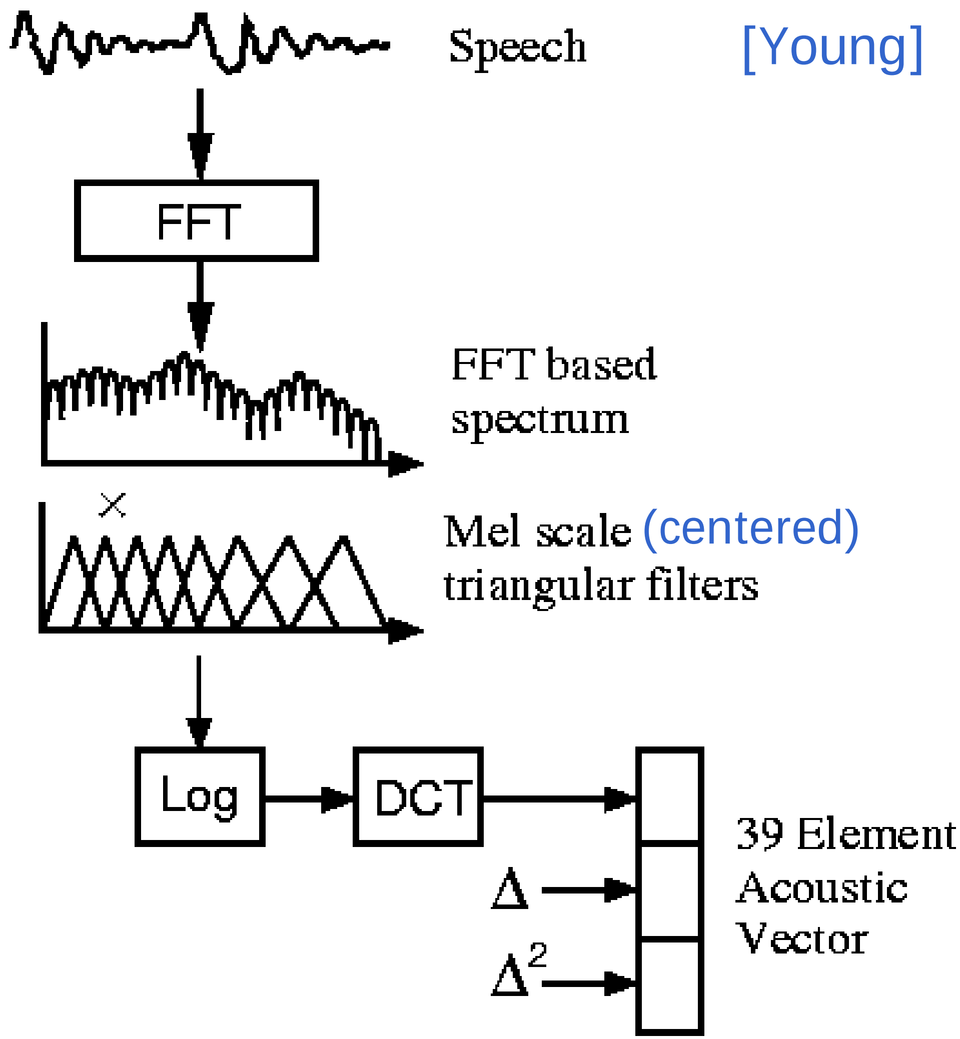
  - mel freq. cepstral coefficients (MFCCs):

$$c_{MFCC}(l;m), \quad m = 1,\ldots,N_{MFCC}$$
$$N_{FFT(mel)}, N_{filters(mel)}$$

  - generalized case [Vergin]

Mel-cepstral feature generation (frame *l*)

[Young]



LP derived Cepstrum

- LP Cepstral Coefficients (LPCCs):

$$m = 1,\ldots,N_{LPC}:$$

$$c_{LPCC}(l;m) = a_{LPC}(l;m) + \sum_{k=1}^{m-1} \frac{k}{m} c_{LPCC}(l;k)\, a_{LPC}(l;m-k)$$

$$m = N_{LPC}+1,\ldots,N_{LPCC}:$$

$$c_{LPCC}(l;m) = \sum_{k=m-N_{LPC}}^{m-1} \frac{k}{m} c_{LPCC}(l;k)\, a_{LPC}(l;m-k)$$

Proven to be equivalent to CC but faster computed

Other Cepstral Variants

- Linear Freq. Cepstral Coefficients (LFCCs)
  - Like MFCCs but:

    filters are uniformly spaced on the Hz scale
- Mel-warped LPCCs (MLPCCs) [Kuitert]
  - CC not directly derived from LPC
  - 1st compute the log magnitude spectrum of LPC
  - then warp the freq. axis to correspond to the mel axis

Variants

- Discrete Wavelet Transform (DWT) instead of FFT [Krishnan]
- Application of other type than triangular filters
- Application of the logarithm before the triangular filters

Delta Cepstrum

- [Milner]:

$$\Delta c_{GCC}(l;m) = \frac{\sum_{k=-K}^{K} k c_{GCC}(l+k;m)}{\sum_{k=-K}^{K} k^2}, \quad m = 1,\ldots,N_{GCC}$$

- Higher order:

$$\Delta c_{GCC} -> \Delta\Delta c_{GCC} \; \& \; c_{GCC} -> \Delta c_{GCC}, \ldots$$

- Inclusion of temporal information

PLP - Auditory Features

- Perceptual Linear Prediction (PLP) [Hermansky]
  - Spectral scale: non-linear Bark scale

$$f_{Bark} = 13\,\mathrm{atan}\left(\frac{0.76 f_{Hz}}{1000}\right) + 3.5\,\mathrm{atan}\left(\frac{f_{Hz}^2}{7500^2}\right)$$

  - Spectral features smoothed within freq. bands
- Auditory Features [Kumar]
  - Imitates signal proc. performed by the ear
  - cochlear modeling

Intra-frame Cepstral Processing

Noise Compensation - Channel Equalization

[Mammone]

- Liftering - weighting
  - low order coeffs: sensitive to overall spectral slope
  - high order: sensitive to noise
  - →tapered window (bandpass liftering)

$$w(m) = 1 + \frac{N_{GCC}}{2} \sin\left(\frac{\pi m}{N_{GCC}}\right), \quad m = 1, \ldots, N_{GCC}$$

$$c_{w-GCC}(l;m) = w(m) c_{GCC}(l;m), \quad m = 1, \ldots, N_{GCC}$$

- Adaptive Component Weighting (ACW)
  - motivation: all frames don't have same distortion

Inter-frame Cepstral Processing

- Cepstral Mean Subtraction (CMS)
  - mean (over a num of frames) subtraction (tackles training-testing discrepancy)

$$c_{CMS-GCC}(l;m) = c_{GCC}(l;m) - \mathrm{avg}_k(c_{GCC}(k;m)), \quad m = 1,\ldots,N_{GCC}$$

  - lowpass filtering
  - eliminates communication channel spectral shaping
- Pole Filtered CMS (PFCMS): cepstrum poles modification

RASTA Processing

- Relative Spectral Filtering (RASTA) [Hermansky]
  - bandpass filtering in the log-spectral domain
  - suppresses spectral components that change more slowly or quickly than in typical speech
  - RASTA-PLP
    - Microphone (type, position) robustness

Feature Selection Introduction

- Goal
  - find a transformation to a relatively low-dimensional feature space that preserves the information pertinent to the application while enabling meaningful comparisons to be performed using measures of similarity
- Processing of features
  - Principal Component Analysis (PCA) (or Karhunen Loève Expansion-KLE)
    - seeks a lower dimensional representation that accounts for variance of the features
    - not necessarily optimum for class discrimination
  - Linear Discriminant Analysis (LDA) [Jin]
  - Non Linear Discriminant Analysis (NLDA) (using MLP) [Konig]

Matching - Modeling

- Matching - Modeling Introduction
- Template Matching Methods
  - DTW (Dynamic Time Warping)
  - VQ (Vector Quantization)
  - LVQ (Learning Vector Quantization)
- Statistical Measures
  - AHS (Arithmetic-Harmonic-Sphericity)
- Generative Models
  - HMMs (Hidden Markov Models)
  - GMMs (Gaussian Mixture Models)

Matching – Modeling(2)

- Neural Networks (NNs)
  - Feed-forward NNs
  - SOMs (Self Organizing Maps)
  - RNNs (Recurrent NNs)
- NNs & Combined Methods
  - Neural Tree Networks (NTNs)
  - DTW-SOM
- Support Vector Machines (SVMs)
- Sub-band Processing Introduction

Matching - Modeling Introduction

- Modeling: creation of (speaker) models
- Model: Can be considered as the output of a proper proc. of a speaker's set of feature vectors
- Matching: computation of a match score betw. the input feature vectors & some speaker model
- Methods [Wassner]
  - Template Matching
    - deterministic
    - score: distance betw. a test speaker (feature vectors of an) utterance & a reference speaker model
    - better score: min distance

Matching-Modeling Introduction(2)

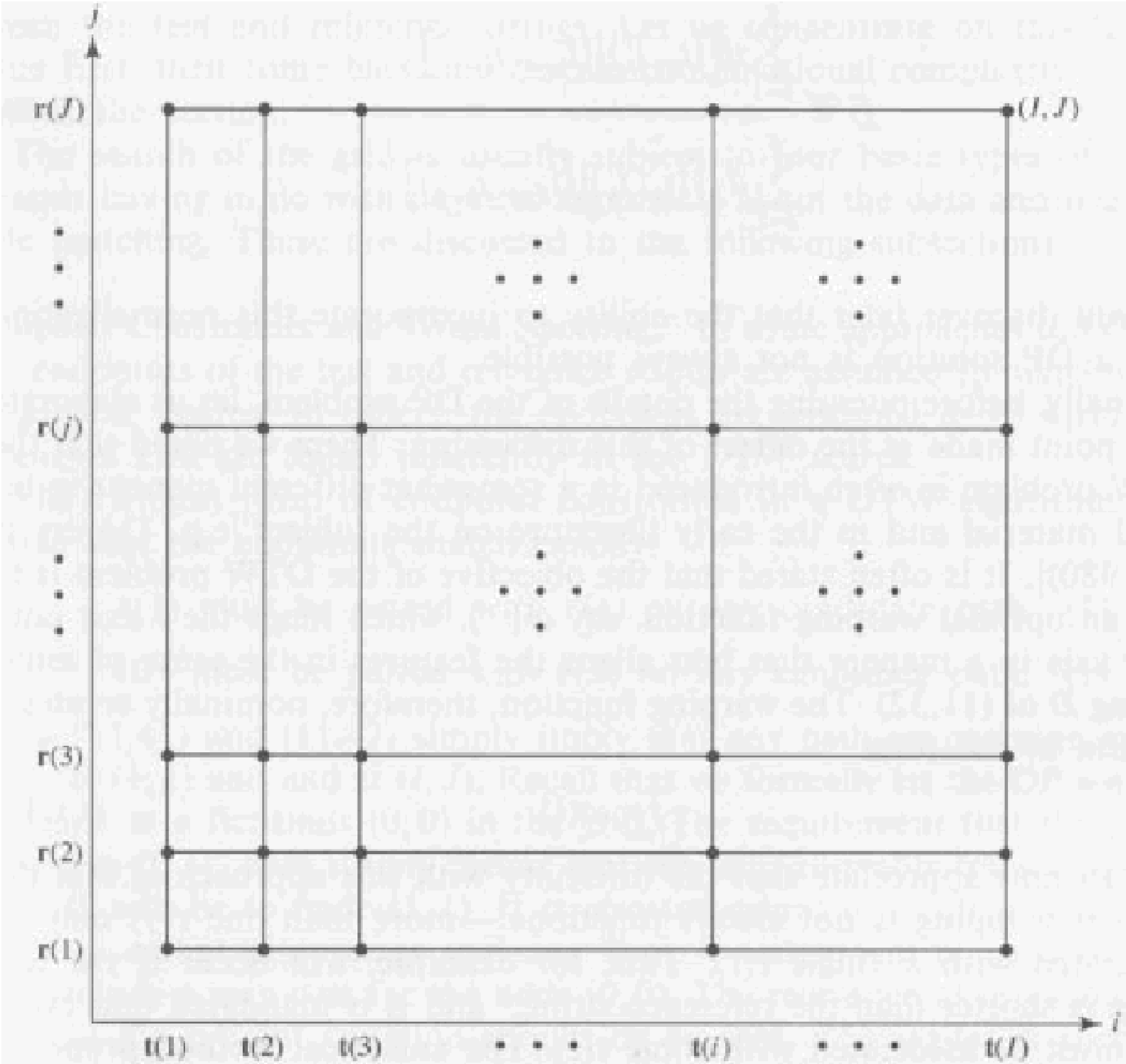
- Methods(2)
  - Stochastic Approach
    - probabilistic matching
    - score: prob. of generation of a speech utterance by the claimed speaker $P(U \mid S_c)$
    - better score: max probability
    - Parametric speaker model: specific pdf is assumed & its appropriate parameters (e.g. mean vector, covariance matrix) can be estimated using the Maximum Likelihood Estimation (MLE) e.g. multivariate normal model

Template Matching Methods

- Dynamic Time Warping (DTW)
  - dynamic comparison betw. a test & a reference (model) matrix (set of feature vectors)
  - computes a distance betw. the test & ref. patterns
  - allows time alignment at different costs
  - uses Dynamic Programming (DP)
  - text dependent cases

Template Matching Methods(2)

- Dynamic Time Warping (DTW)(2)



The DP grid with test (**t**) & reference (**r**) feature vectors at respective frame indices [Picone]

Template Matching Methods(3)

- Dynamic Time Warping (DTW)(3)
  - distances-costs on the DP grid (*i*,*j* frame indices, *k* step index)
    - Node $d_N(i_k, j_k)$

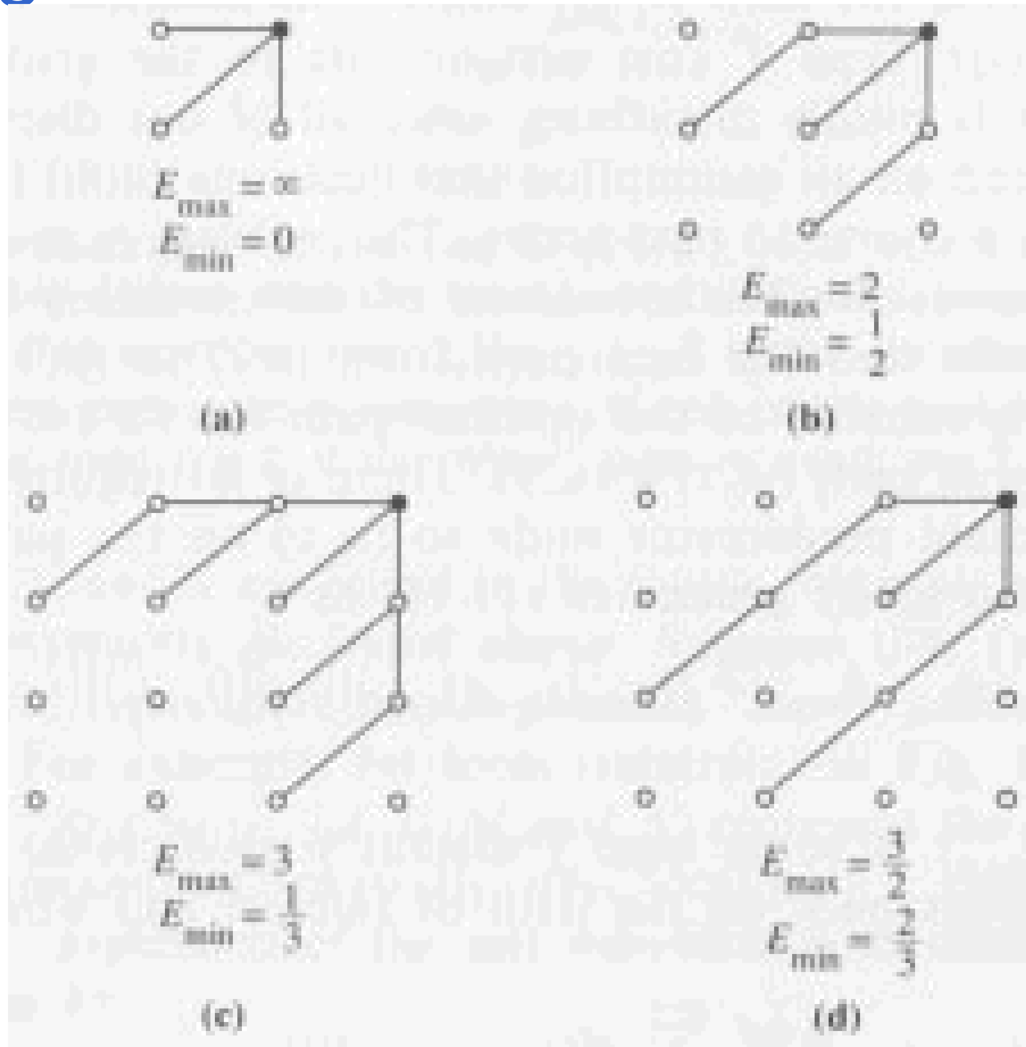
      e.g. $$D_{Eucl.}(i_k, j_k; LPCC) = \sum_{m=1}^{N_{LPCC}} \left(c_{LPCC}^{test}(j_k; m) - c_{LPCC}^{ref}(i_k; m)\right)^2$$
    - Transition $d_T[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]$ e.g. $[i_k - i_{k-1}] + [j_k - j_{k-1}]$
    - Both $d_B[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]$ (Type 4)
      - e.g. $d_N(i_k, j_k) \times d_T[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]$
    - Global $$D = \sum_{k=1}^{K} d_B[(i_k, j_k) \mid (i_{k-1}, j_{k-1})]$$
      - *K*: number of transitions

Template Matching Methods(4)

- Dynamic Time Warping (DTW)(4)
  - DTW search constraints
    - Endpoint Constraints (bottom left(S) - top right(E) corners)
      - endpoint relaxation: $\Delta iS, \Delta jS, \Delta iE, \Delta jE$ max points allowed in each direction
    - Monotonicity (going up & right) $i_{k-1} \le i_k \wedge j_{k-1} \le j_k$
    - Global Path Constraints (global movement area)
      - permissible slope or
      - permissible window $\left| j_k - i_k \right| \le W$

Template Matching Methods(5)

- Dynamic Time Warping (DTW)(5)
  - DTW search constraints(2)
    - Local Path Constraints (local movement area)



Sakoe & Shiba local constraints on DTW path search [Picone]

Template Matching Methods(6)

- Dynamic Time Warping (DTW)(6)
  - The minimum cost final endpoint provides the distance betw. a test & a reference phrase
  - Training-Modeling [Deller_bk]
    - Casual: Unaltered feature strings form models
    - Averaging feature strings of utterances
    - The stochastic techniques possess superior training methods

Template Matching Methods(7)

- Vector Quantization (VQ)
  - Uses intra-vector dependencies to break-up a (feature) vector space in cells (unsupervised)
  - follows Linde-Buzo-Gray (LBG) algorithm
  - speaker model: codebook
  - codebook: set of prototype vectors (codevectors)
  - codevector: vector computed from "similar" single (feature) vectors (e.g. representing a phoneme) (phoneme: basic speech unit)
  - handles text independent cases
  - goal: data structure "discovery" by finding how the data is clustered

Template Matching Methods(8)

- Learning Vector Quantization (LVQ)
  - Predefined classes, labeled data
  - defines the class borders according to the nearest neighbor rule
  - supervised version of VQ
    - quantization of feature vectors by codevectors based on a distance
    - (gradual) update of codevectors
  - set of variants (e.g. LVQ1,2,3)
  - goal: to determine a set of prototypes that best represent each class.

Statistical Measures

- Second Order Statistical Measures (SOSM) [Bimbot]
  - E.g. Arithmetic-Harmonic-Sphericity (AHS)
    - speaker model: covariance matrix of feature vectors
    - Distance=min(=0) iff all eigenvalues of test & reference covariance matrices are equal

Generative Models

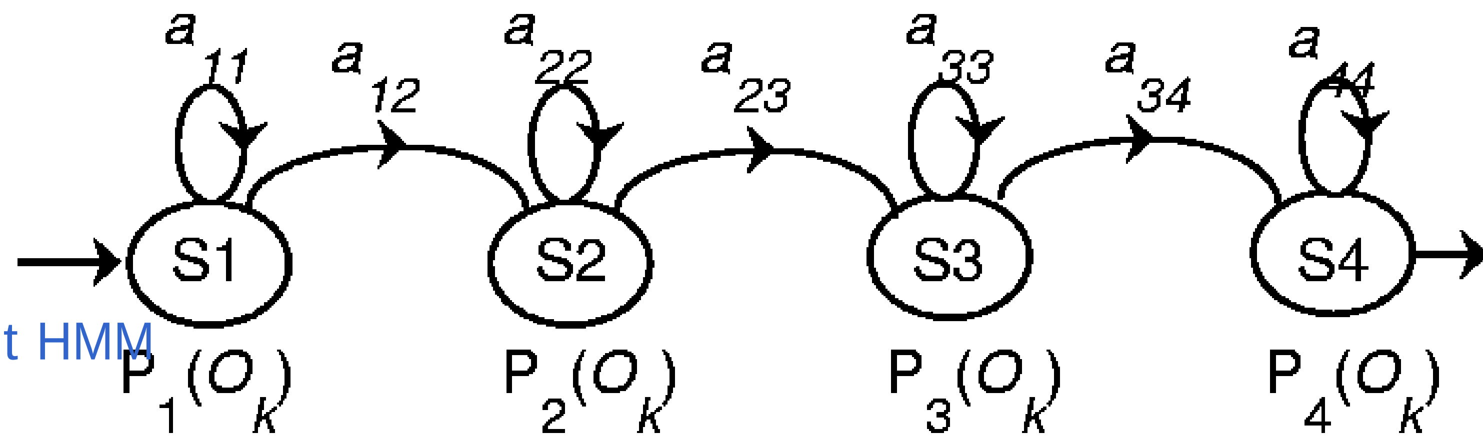
- Hidden Markov Models (HMMs)
  - Statistical - stochastic
  - Flexible
  - Text independent cases handled
  - Types
    - Continuous Density (CD) (real valued features)
    - Discrete (integer valued features - symbols)
    - SemiContinuous (SC) [Falavigna]
  - Model: prob. distributions of the feature vectors of the speaker’s utterances approximated by mixtures of Gaussians

Generative Models(2)

- Hidden Markov Models (HMMs)(2)
  - Topologies
    - Left-Right (LR) (self & right connections): attempts to catch the temporal structure of the speech & to link consecutive short-time observations together

      #states/unit(e.g. phoneme)

      #Gaussian distributions(mixtures)/state



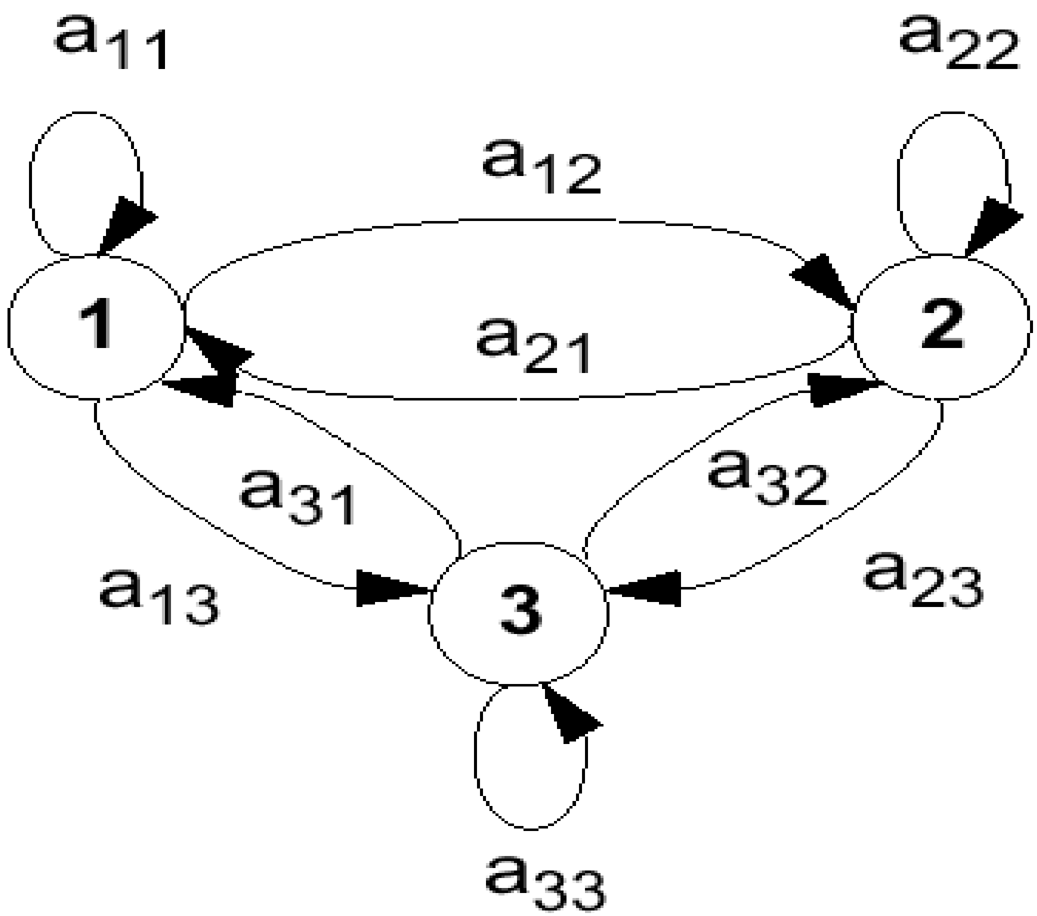
[Kumar]

Example of a left-right HMM

O_k : feature vectors

- Hidden Markov Models (HMMs)(3)
  - Topologies(2)
    - Ergodic (fully connected)

      -AR HMMs: the prob. distrib. associated with each state is estimated via an AR process [Bourlard]



[Picone]

Example of an ergodic HMM

Generative Models(4)

- Gaussian Mixture Models (GMMs)
  - Like single multi-Gaussian state HMMs
  - Uses a mixture of Gaussian densities to model the distribution of the feature vectors of each speaker
  - Local covariance info

Neural Networks (NNs)

- Feed-Forward Neural Networks
  - supervised learning
  - each speaker is modeled by processing results of his NN
  - when an identity is claimed the corresponding NN is consulted
  - positive/negative training (rivals)

Neural Networks (NNs)(2)

- Feed-Forward NNs(2)
  - Types [Haykin_bk]
    - Multilayer Perceptron (MLP): trained usually with the Back-Propagation (BP) algorithm
      - Error Correction Learning
      - Global optimization
    - Time Delay NNs (TDNNs)
    - Radial Basis Function (RBF) Networks [Lo]
      - Memory-Based Learning
      - Local optimization

Neural Networks (NNs)(3)

- Self Organizing Maps (SOMs) [Kohonen_bk]
  - unsupervised learning
  - method to form a topologically ordered codebook
  - speaker model: codebook
  - density of codevectors approaches the pdf of the input vectors during the training
  - like nonlinear projection of the feature space on the neural lattice
  - competitive (winner neuron) learning

NNs & Combined Methods

- DTW-SOM
  - associate an entire feature vector sequence, instead of a single feature vector, as a model with each SOM node (also DTW-LVQ) [Somervuo]
- Recurrent NNs (RNNs) [Shrimpton]
  - (self-or not) feedback
- Neural Tree Networks (NTNs)
  - hierarchical classifier that incorporates decision trees & NNs (e.g. 1 MLP NN per tree node)
- Combined methods [Genoud]

Support Vector Machines (SVMs)

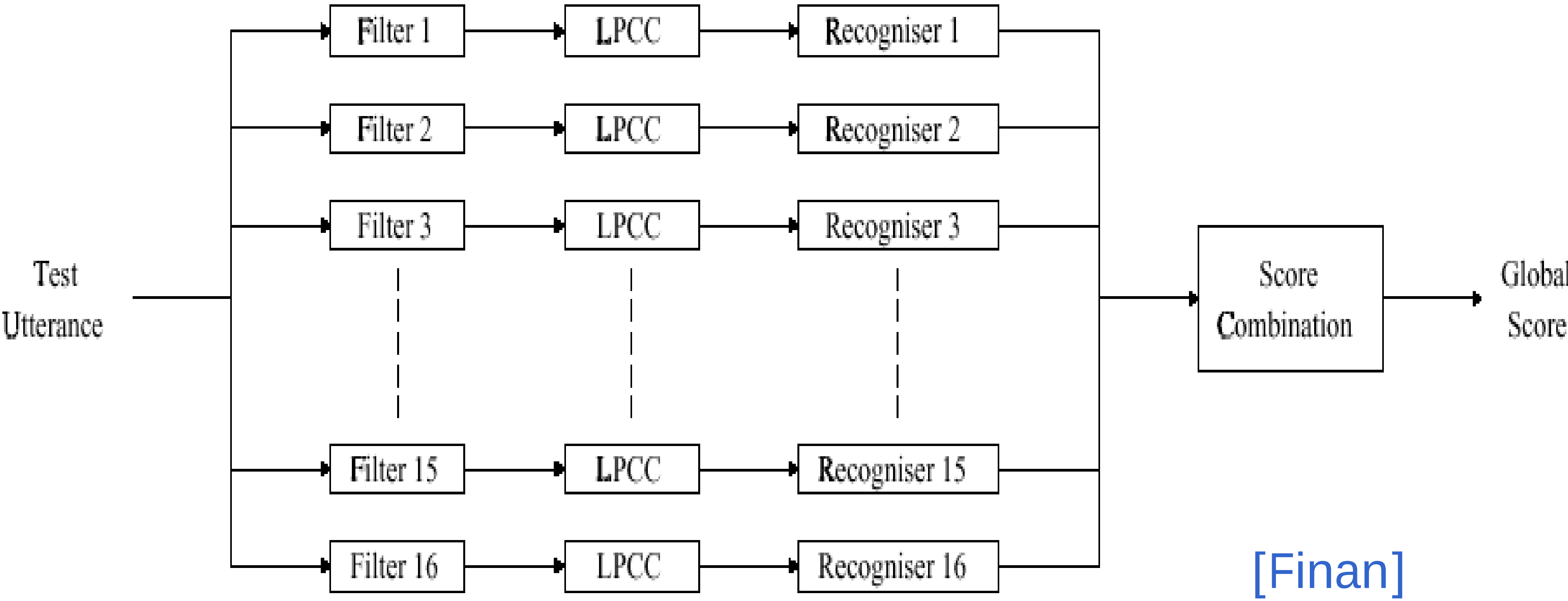
- [Cherkassky_bk]
- a combination of the most important examples (support vectors) is computed in a high dimensional space (kernel space)
- Learning by examples (supervised)
- Vapnik-Chervonenkis (VC) dimension: framework for the development of SVMs
- based on Structural Risk Minimization principle from statistical learning theory [Vapnik_bk]

Sub-band Processing Introduction

Matching - Modeling

- Speech signal split into band-limited channels (freq. ranges)

Block diagram of an LPCC-based sub-band processing system



[Finan]

Decision Making

- Decision Approaches
  - "Template"
  - Statistical & extensions
    - LLR (Log Likelihood Ratio)
    - Cohort/world model
- Threshold Setting
- Hypothesis Testing

Decision Approaches

- "Template" approach
  - threshold setting: based on inter- & intra-speaker scores/distances
  - comparison:

    test score<=threshold→acceptance [Fakotakis]
- Statistical approach [Bengio] [Bourlard]
  - S_c : speaker RV for identity *c* being claimed
  - U : utterance represented by feat. vectors
  - $\overline{S_c}$: other speakers RV

$$P(S_c \mid U) = \frac{P(U \mid S_c)P(S_c)}{P(U)}$$

Decision Approaches(2)

- Statistical approach(2)
  - Claim c is true if:

$$\frac{P(S_c|U)}{P(\overline{S_c}|U)} > 1 \Rightarrow \frac{P(U|S_c)}{P(U|\overline{S_c})} > \frac{P(\overline{S_c})}{P(S_c)} = \vartheta_c$$

  - ϑ_c : decision threshold usually found assuming Gaussian distributions for $P(U|S_c)$ and $P(U|\overline{S_c})$
  - → normalized likelihood - likelihood ratio
  - using logs:

$$\log\frac{P(U|S_c)}{P(U|\overline{S_c})} > \log\vartheta_c \Rightarrow \log P(U|S_c) - \log P(U|\overline{S_c}) > \Theta_c$$

  - → Log Likelihood Ratio (LLR)

Decision Approaches(3)

- Statistical approach(3)
  - $P(U \mid S_c)$: speaker dependent model
  - $P(U \mid \overline{S_c})$: normalization factor
  - cohort model $\overline{S_c} = S_{ch}$: group of selected speakers who are more competitive with the model of the claimed id
    - No well-established selection procedure
  - world model $\overline{S_c} = \Omega$: all other speakers
    - less computation & storage needed

Decision Approaches(4)

- Statistical approach extensions
  - If $y = \log P(U \mid S_c) - \log P(U \mid \overline{S_c}) - \Theta_c$
  - sign(y) gives the decision
  - Techniques:
    - Bayes Decision Rule (assumes prob.s perfectly estimated)
      - Minimizes Half Total Error Rate(HTER)

$$\mathrm{HTER} = \frac{\%\mathrm{FA} + \%\mathrm{FR}}{2}$$

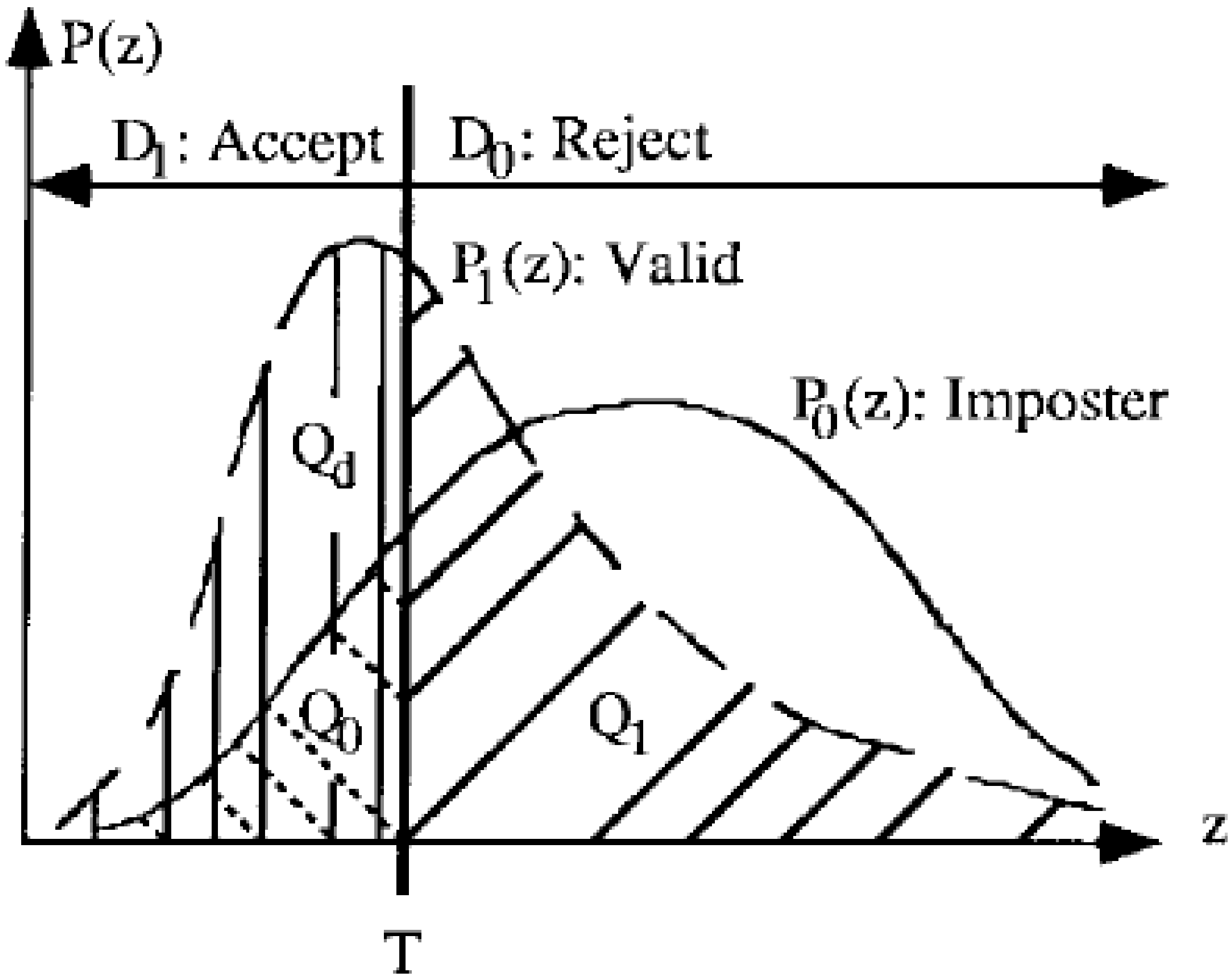
    - Linear Regression
    - SVM Regression

Threshold Setting

- speaker dependent
  - |P| thresholds: $\vartheta_c, \quad \mathrm{c} = 1,\ldots,|\mathrm{P}|$
- speaker independent
  - 1 threshold: ϑ
- leave one (client *o*) out
  - |P|*|P| thresholds: $\vartheta_{co}, \quad c = 1,\ldots,|P|, \quad o = 1,\ldots,|P|$
- a priori: computed on training set (enrollment data) [Lindberg]
- a posteriori: computed on test set (obtained during actual use of the system)

Hypothesis Testing

Valid & impostor densities



[Campbell]

Hypothesis Testing(2)

Decision Making

Probability terms & definitions

Performance Probabilities	Decision D	Hypothesis H	Name of Probability	Decision Result	
Q_0	1	0	Size of test "significance"	Type I error	False acceptance or alarm
Q_1	0	1		Type II error	False rejection
$Q_d = 1 - Q_1$	1	1	Power of test		True acceptance
$1 - Q_0$	0	0			True rejection

[Campbell]

Performance Evaluation

- Accuracy
  - FAR (False Acceptance Rate)
  - FRR (False Rejection Rate)
  - EER (Equal Error Rate)
  - ROC (Receiver Operating Characteristics)
- Resources Requirements
  - CPU
  - memory, disk

Accuracy

- Error %s
  - FAR (False Acceptance Rate): Prob. of false acceptance
    - Estimate: $\dfrac{\#\,\text{false acceptances}}{\#\,\text{false claims}}$
  - FRR (False Rejection Rate): Prob. of false rejection
    - Estimate: $\dfrac{\#\,\text{false rejections}}{\#\,\text{true claims}}$
  - Values for FAR & FRR are adjusted by changing the threshold values: ⇲ FAR vs. ⇲ FRR

Accuracy(2)

- Error %s(2)
  - EER (Equal Error Rate): operating point where FAR≈FRR
  - Choice of 2 subsequent operating points to approximate the EER value

$$\mathrm{EER} = \frac{\mathrm{FRR}_k \cdot \mathrm{FAR}_{k+1} - \mathrm{FRR}_{k+1} \cdot \mathrm{FAR}_k}{(\mathrm{FAR}_{k+1} - \mathrm{FAR}_k) - (\mathrm{FRR}_{k+1} - \mathrm{FRR}_k)},$$

$$\mathrm{FAR}_{i+1} \geq \mathrm{FAR}_i \wedge \mathrm{FRR}_{i+1} \leq \mathrm{FRR}_i, \quad \forall i$$

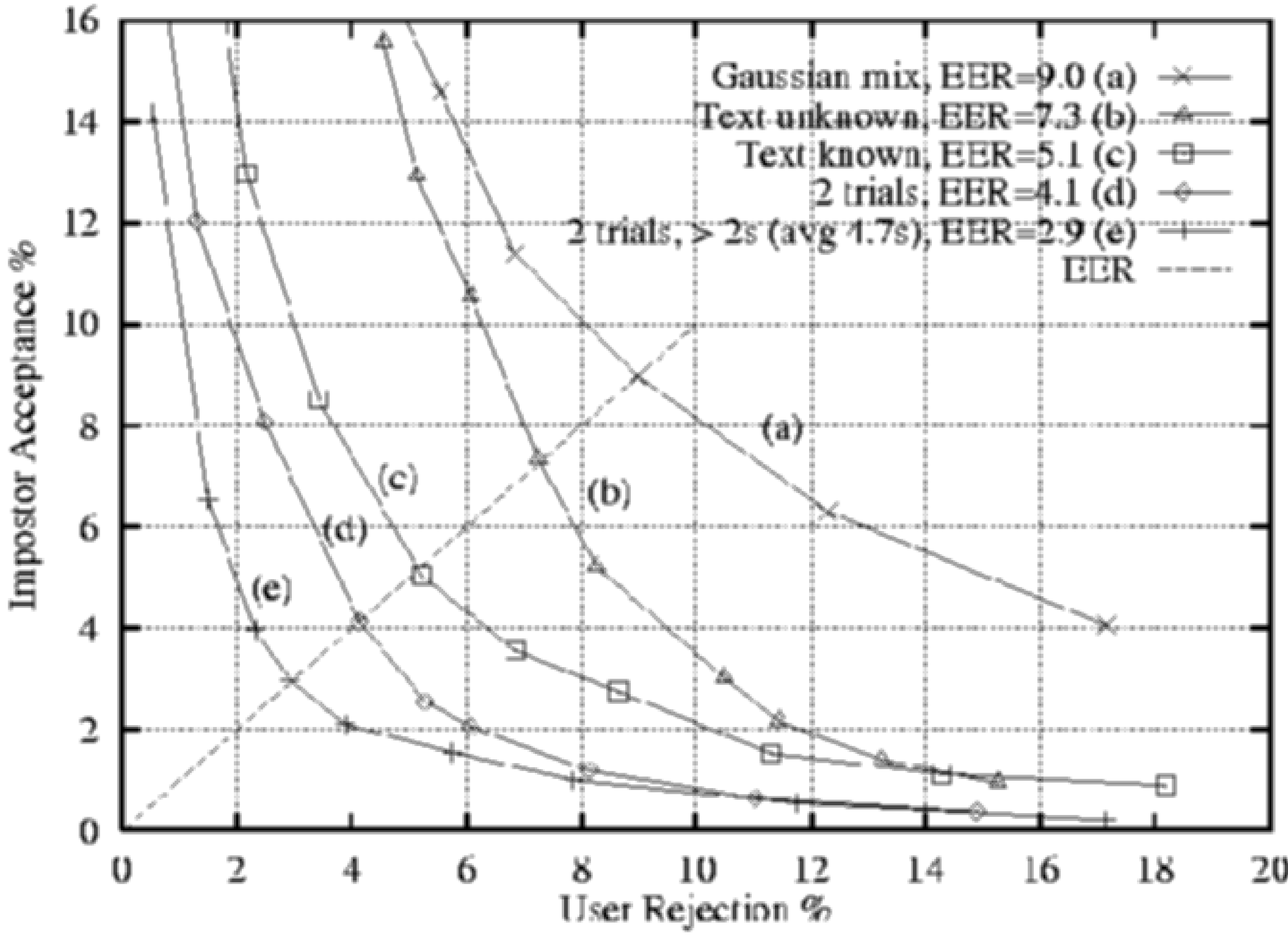
$$\mathrm{FAR}_k \leq \mathrm{FRR}_k \wedge \mathrm{FAR}_{k+1} \geq \mathrm{FRR}_{k+1}$$

  - MDE (Minimum Decision Error): operating point where $\mathrm{FRR} \approx 10 \cdot \mathrm{FAR}$

Accuracy(3)

- Graphs

ROC (Receiver Operating Characteristics) curve: Plot of different operating points (FRR vs. FAR values). Called also DET (Detection Error Tradeoff) plot [Gauvain]



- Quantities
  - #speakers correctly/wrongly verified

Resources Requirements

- CPU time
  - Training
    - Feature creation
    - Modeling
    - Threshold setting
  - Testing (verification throughput)
    - Feature creation
    - Matching
- Memory-disk storage
  - Speech database, Features, Models, Thresholds

Experimental Results

- Parameters
- EERs

Parameters

Text dependent – Fixed vocab.: Digits 0-9 in French or Spanish
→|V|=10 |P|=37 (M2VTS database)

Discrete utterance speech flow

#sessions(shots)/speaker=5, the 5th is for testing→|S|=4

#phrases/session=1 (0-9 utterance)

Phrase duration~6sec

$F_s = 48\text{KHz}$ Proc. Freq.=12KHz

$\alpha_{pe} = 0.95$ $N = 360\,(30\text{ms})$ $M = 240\,(20\text{ms})$

Window type: Hamming

Coefficients: LPCCs $N_{LPCC} = 12$ $N_{LPC} = 12$

Liftering-weighting: $c_{w-LPCC}(l;m)$

Parameters(2)-EER

Matching method: DTW

d_N : Euclidean d_T : Type 4 $d_B = d_N \times d_T$

$\Delta iS = 10, \Delta jS = 10, \Delta iE = 10, \Delta jE = 10$ $W = 30$

Local path constraint: Sakoe & Shiba (b)

Decision approach: Template

Threshold setting: leave one out

|P|(client left out).|P-1|(rest clients as claimants).|S|(shot left out for claiming-testing)=5328 client claims

|P|(client left out as impostor).|P-1|(claims of the impostor as one of the rest clients).|S|(shot left out for claiming)=5328 impostor claims

EER(avg)$\in$ [0.6569%,1.5390%] (FAR_1=1.5390% >FRR_1=0.6569%)

EER(avg)=[EER(1|234)+EER(2|134)+EER(3|124)+EER(4|123)]/4

Parameters(3)-EER

Shot 4 left out, shot 5 used for testing:
|P|.|P-1|=1332 client & 1332 impostor claims
EER(5|123)=2.7027%

Difference:

Coefficients: MFCCs $N_{MFCC} = 12$

$$N_{FFT(mel)} = 512 \qquad N_{filters(mel)} = 40$$

EER(avg)=4.1817%
EER(5|123)=5.4054%

References

- [Bengio] S. Bengio and J. Mariéthoz, *Learning the Decision Function for Speaker Verification*, IDIAP Research Report, 2001
- [Bimbot] F. Bimbot, I. Magrin-Chagnolleau and L. Mathan, *Second-Order Statistical Measures for Text-Independent Speaker Identification,* Speech Communication, vol. 17, pp. 177-192, 1995
- [Bourlard] H. Bourlard and N. Morgan, *Speaker Verication: A Quick Overview*, IDIAP Research Report, 1998
- [Campbell] J.P. Campbell Jr., *Speaker Recognition: a Tutorial*, Proc. of the IEEE, vol. 85, no. 9, pp. 1437-1462, 1997
- [Cherkassky_bk] V. Cherkassky and F. Mulier, *Learning from data: Concepts, Theory, and Methods*, Wiley, New York, 1998
- [Deller_bk] J.R. Deller, J.G. Proakis and J.H. Hansen, *Discrete-time Processing of Speech Signals*, Macmillan, New York, 1993
- [Fakotakis] N. Fakotakis, E. Dermatas, G. Kokkinakis, *Optimal Decision Threshold for Speaker Verification*, in Signal Processing III: Theories and Applications, editor: I.T. Young et al., pp. 585-587, Elsevier Science Publishers B.V. (North Holland), 1986

References(2)

- [Falavigna] D. Falavigna, *Comparison Of Different HMM Based Methods For Speaker Verification* (citeseer)
- [Finan] R.A. Finan, R.I. Damper and A.T. Sapeluk, *Improved Data Modeling for Text-Dependent Speaker Recognition Using Sub-Band Processing* (citeseer)
- [Gauvain] J. Gauvain, L. Lamel and B. Prouts, *Experiments with Speaker Verification over the Telephone*, Eurospeech’95, pp. 651-654, 1995
- [Genoud] D. Genoud, F. Bimbot, G. Gravier and G. Chollet, *Combining Methods to Improve Speaker Verification Decision*, Proc. of ICSLP'96, vol. 3, pp. 1756-1759, 1996
- [Haykin_bk] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan, New York, 1995
- [Hermansky] H. Hermansky and N. Morgan, *Rasta Processing of Speech*, IEEE Trans. on Speech and Audio Processing, vol. 2, no. 4, pp. 578-589, 1994

References(3)

- [Jain_bk] A. Jain, R. Bolle and S. Pankanti, editors, *Biometrics: Personal Identification in Networked Society,* Kluwer Academic Publishers, Boston, MA, 1999
- [James] D. James, H. Hutter and F. Bimbot, *CAVE -- Speaker Verification in Banking and Telecommunications* (citeseer)
- [Jin] Q. Jin and A. Waibel, *Application of LDA to Speaker Recognition* (citeseer)
- [Kohonen_bk] T. Kohonen, *Self-Organizing Maps,* Springer-Verlag, Berlin, Germany, 2nd edition, 1997
- [Konig] Y. Konig, L. Heck, M. Weintraub and K. Sonmez, *Nonlinear Discriminant Feature Extraction for Robust Text-Independent Speaker Recognition*, Proc. of RLA2C'98 (Speaker Recognition and Its Commercial and Forensic Applications), 1998
- [Koolwaaij] J.W. Koolwaaij and L. Boves, *A New Procedure for Classifying Speakers in Speaker Verification Systems*, Proc. of Eurospeech'97, pp. 2355-2358, 1997
- [Krishnan] M. Krishnan, C. Neophytou and G. Prescott, *Wavelet Transform Speech Recognition using Vector Quantization, Dynamic Time Warping and Artificial Neural Networks*, 1994

References(4)

- [Kuitert] M. Kuitert and L. Boves, Speaker Verification with GSM Coded Telephone Speech, Proc. of Eurospeech'97, vol. 2, pp. 975-978, 1997
- [Kumar] N. Kumar, *Investigation of Silicon Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, PhD thesis, Johns Hopkins University, 1997
- [Lindberg] J. Lindberg, J.W. Koolwaaij, H.-P. Hutter, D. Genoud, M. Blomberg, F. Bimbot and J.-B. Pierrot, *Techniques for a priori Decision Threshold Estimation in Speaker Verification*, Proc. of RLA2C'98, 1998
- [Lo] T.F. Lo and M.W. Mak, *A New Intra-Frame and Inter-Frame Cepstral Processing Method for Telephone-Based Speaker Verification*, Int. Workshop on Multimedia Data Storage, Retrieval, Integration and Applications, pp. 116-122, 2000
- [Mammone] R.J. Mammone, X. Zhang and R.P. Ramachandran, *Robust Speaker Recognition*, IEEE Signal Proc. Magazine, vol. 13, no. 5, pp. 58-71, Sep. 1996
- [Milner] B. Milner, *Inclusion of Temporal Information into Features for Speech Recognition*, Proc. of ICSLP'96, pp. 256-259, 1996

References(5)

- [Morgan] N. Morgan and B. Gold, *Speech Analysis and Synthesis Overview*, Lecture, Univ. of California Berkeley, 1999
- [Nedic] B. Nedic and H. Bourlard, *Recent Developments in Speaker Verification at IDIAP*, IDIAP Research Report, 2000
- [Picone] J. Picone, *Fundamentals of Speech Recognition: A Short Course*, Mississippi State Univ., 1996
- [Picone2] J. Picone, *Signal Modeling Techniques in Speech Recognition*, Proc. of the IEEE, vol. 81, no. 9, pp. 1215-1247, 1993
- [Rabiner_bk] L. Rabiner and B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, 1993
- [Shrimpton] D. Shrimpton and B.D. Watson, *Comparison of Recurrent Neural Network Architectures for Speaker Verification*. Proc. of the Fourth Australian International Conference on Speech Science and Technology, pp. 460-464, 1992

References(6)

- [Somervuo] P. Somervuo, Speech Recognition using Context Vectors and Multiple Feature Streams, Helsinki University of Technology, Faculty of Electrical Engineering, 1996
- [Vapnik_bk] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, 1998
- [Vergin] R. Vergin, D. O'Shaughnessy and A. Farhat, *Generalized Mel Frequency Cepstral Coefficients for Large-Vocabulary Speaker-Independent Continuous-Speech Recognition*, IEEE Trans. on Speech and Audio Processing, vol. 7, no. 5, pp. 525-532, 1999
- [Wassner] H. Wassner, G. Maitre and G. Chollet, *Speaker Verification : a Review*, Technical Report, IDIAP, 1996
- [Weingessel] A. Weingessel, *Speech Recognition* (citeseer)
- [Young] S. Young, *Large vocabulary continuous speech recognition*, IEEE Signal Proc. Magazine, vol. 13, no. 5, pp. 45-57, 1996