

Comparing Various Voice Recognition Techniques

TUDOR BARBU

www.iit.tuiasi.ro/~tudbar

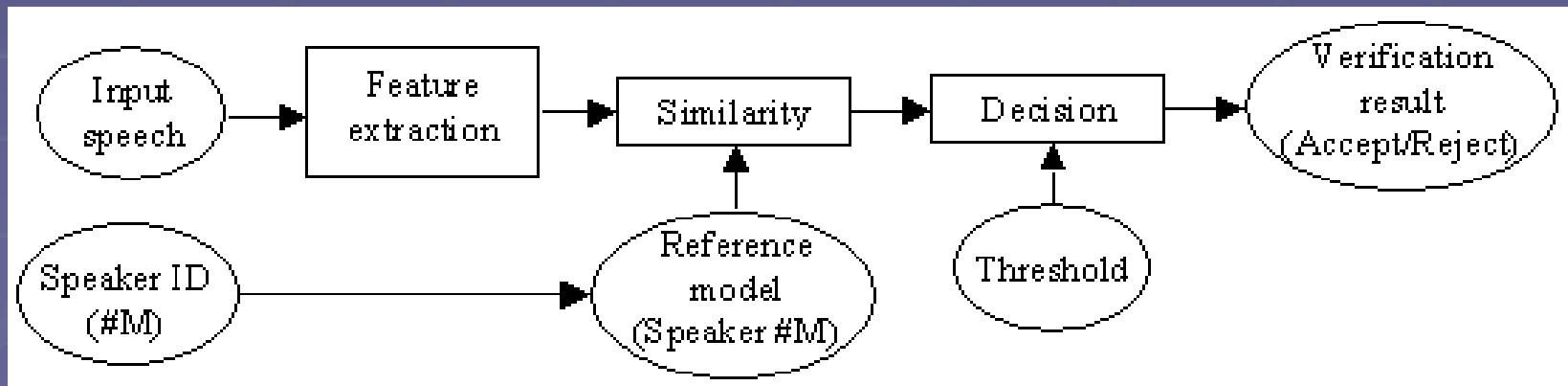
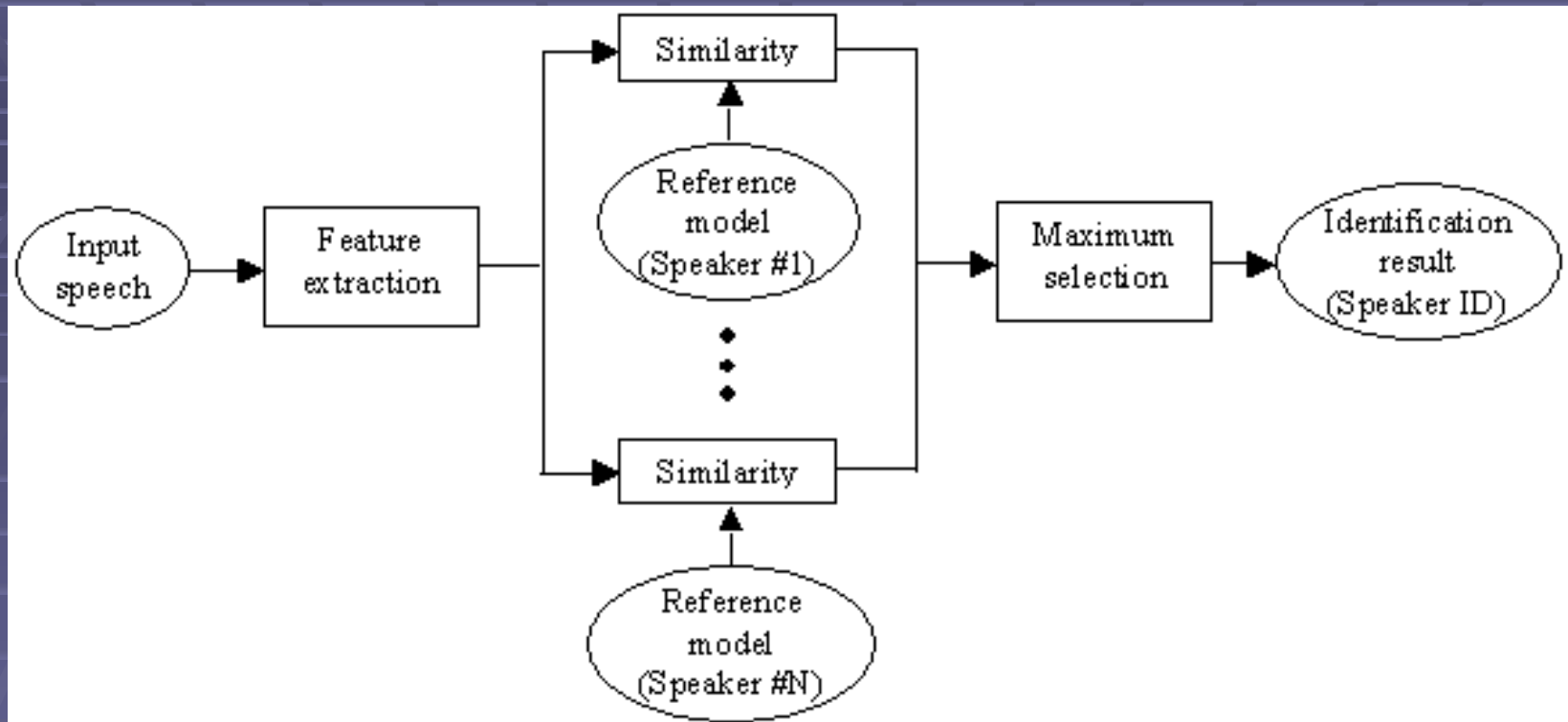
**Institute of Computer Science, Romanian Academy,
Iași branch**

Introduction

- **Voice (speaker) recognition** = operation that determine **who is speaking** on the basis of individual information included in the speech waves.
- **Speaker recognition methods:**
 - Text (speech) dependent voice recognition
 - Discriminate speakers on the basis of the same speech
 - Text (speech) independent voice recognition
 - Discriminate speakers on the basis of any speech



- Speaker recognition :
 - **Voice identification** => determine which registered speaker provides a given utterance.
 - **Voice verification** => accepting or rejecting the identity claim of a previously identified speaker.
- Identification stage:
 - MFCC based feature extraction
 - Supervised classification
- Verification stage:
 - Threshold – based approach



Mel cepstral feature extraction

- *Mel Frequency Cepstral Coefficients (MFCCs)* represent the dominant features used in speech and speaker recognition domains.
- A *short time* signal analysis is performed: divide into overlapping frames with 256 samples and overlaps of 128 samples
- A Hamming window, having 256 coefficients, is then applied to each resulted segment.

...

- Computing the MFCCs:
 - Take the FFT of a windowed signal.
 - Map the powers of the spectrum obtained above onto the [mel scale](#), using triangular overlapping windows.
 - Take the logs of the powers at each of the mel frequencies.
 - Take the [DCT](#) (Discrete Cosinus Transform) of the set of mel log powers, as if it were a signal.
 - The MFCCs are the amplitudes of the resulting spectrum.

...

- Each MFCC sequence = **acoustic vector**
- Derivation of MFCCs is applied
 - First derivative => delta mel cepstral coefficients = **DMFCC**
 - Second derivation = delta delta mel cepstral coefficients = **DDMFCC**
- Each DDMFCC acoustic vector => truncated at the first 12 coefficients
- Truncated DDMFCC acoustic matrix => **feature vector**



- Other possible speech feature vectors can be obtained from these DDMFCC matrices.
- They can be obtained computing some statistical values for the matrix columns: mean, standard deviation etc.
- Considering the mean of each 12-row DDMFCC acoustic vector => 1D feature vector
- DDMFCC feature vectors = different number of columns, based on signal length
- Euclidian distance cannot be used for these vectors.
- Special **non-Euclidian metrics** are needed.

A Hausdorff-derived Metric

- We propose a non-Euclidean metric that works for our speech feature vectors.
- It is able to measure the distance between matrices having different sizes (one common dimension).
- It derives from the Hausdorff metric for sets, defined as:

$$d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} \text{dist}(x, y), \sup_{y \in Y} \inf_{x \in X} \text{dist}(x, y) \right\}$$

X and Y being sets.

...

- We consider matrices instead of sets:
 - $A=(a_{ij})_{n \times m}$, $B=(B_{ij})_{n \times p}$
- We transform the previous distance formula and obtain:

$$d(A, B) = \max \left\{ \sup_{1 \leq k \leq p} \inf_{1 \leq i \leq m} \sup_{1 \leq j \leq n} |b_{ik} - a_{ij}|, \sup_{1 \leq i \leq m} \inf_{1 \leq k \leq p} \sup_{1 \leq j \leq n} |b_{ik} - a_{ij}| \right\}$$

- Function d satisfies the metric properties:
 - Positivity: $d(A,B) \geq 0$
 - Symmetry: $d(A,B) = d(B,A)$
 - Triangle inequality: $d(A,B) + d(B,C) \geq d(A,C)$
- Distance d is not a Hausdorff metric anymore
- Useful tool in speech feature vector classification.

Text-dependent Voice Recognition

- The system performs the described DDMFCC – based feature extraction
- $\{S_1, \dots, S_n\}$ = set of speech signals to be recognized
- $S_i \Rightarrow V(S_i)$ = feature vector
- Supervised classification \Rightarrow we propose a **minimum mean-distance classifier**:
 - Training set : same-speech vocal utterances provided by *registered speakers* (1 to N)
 - **Feature vector set** \Rightarrow the corresponding feature vectors.

$$\{ \{V(s_1^1), \dots, V(s_{n(1)}^1)\}, \dots, \{V(s_1^N), \dots, V(s_{n(N)}^N)\} \}$$

...

- The system introduces each input vocal sequence in the class of the speaker corresponding to the smallest mean distance between the feature vector of the input signal and the speaker's vectors:

$$n_i = \arg \min_{j \in [1, M]} \frac{\sum_{k=1}^{n(j)} d(V(S_i), V(s_k^j))}{n(j)}, \quad \forall i \in [1, n]$$

- Each input speech => closest speaker identification
- **Speaker verification** – we propose a threshold based approach

...

- **Automatic threshold detection method:** consider the overall maximal distance between any two feature vectors belonging to the same training feature subset, as a threshold:

$$T = \max_{i \leq N} \max_{k \neq t} \sum_{k=1}^{n(i)} d(V(s_k^i), V(s_t^i))$$

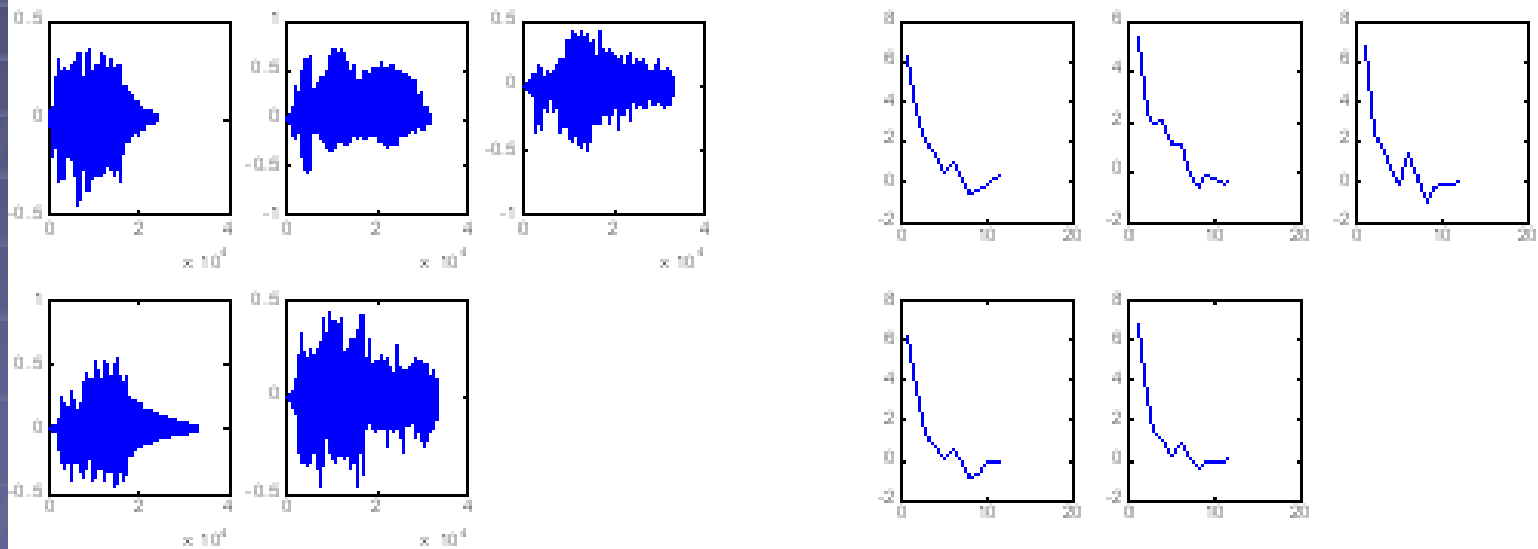
- Each mean distance computed in each voice class is compared with this threshold T : if it is greater than it, a vocal utterance must be rejected as ***unknown speaker***.

...

- Experiments: numerous test on various dataset
- Satisfactory results have been obtained
- High recognition rate = approximately 85 %.
- We tested both 2-D and 1-D DDMFCC based speech feature vectors.
- 1D feature vectors = consider the mean of the DDMFCC acoustic matrix = average of each column.



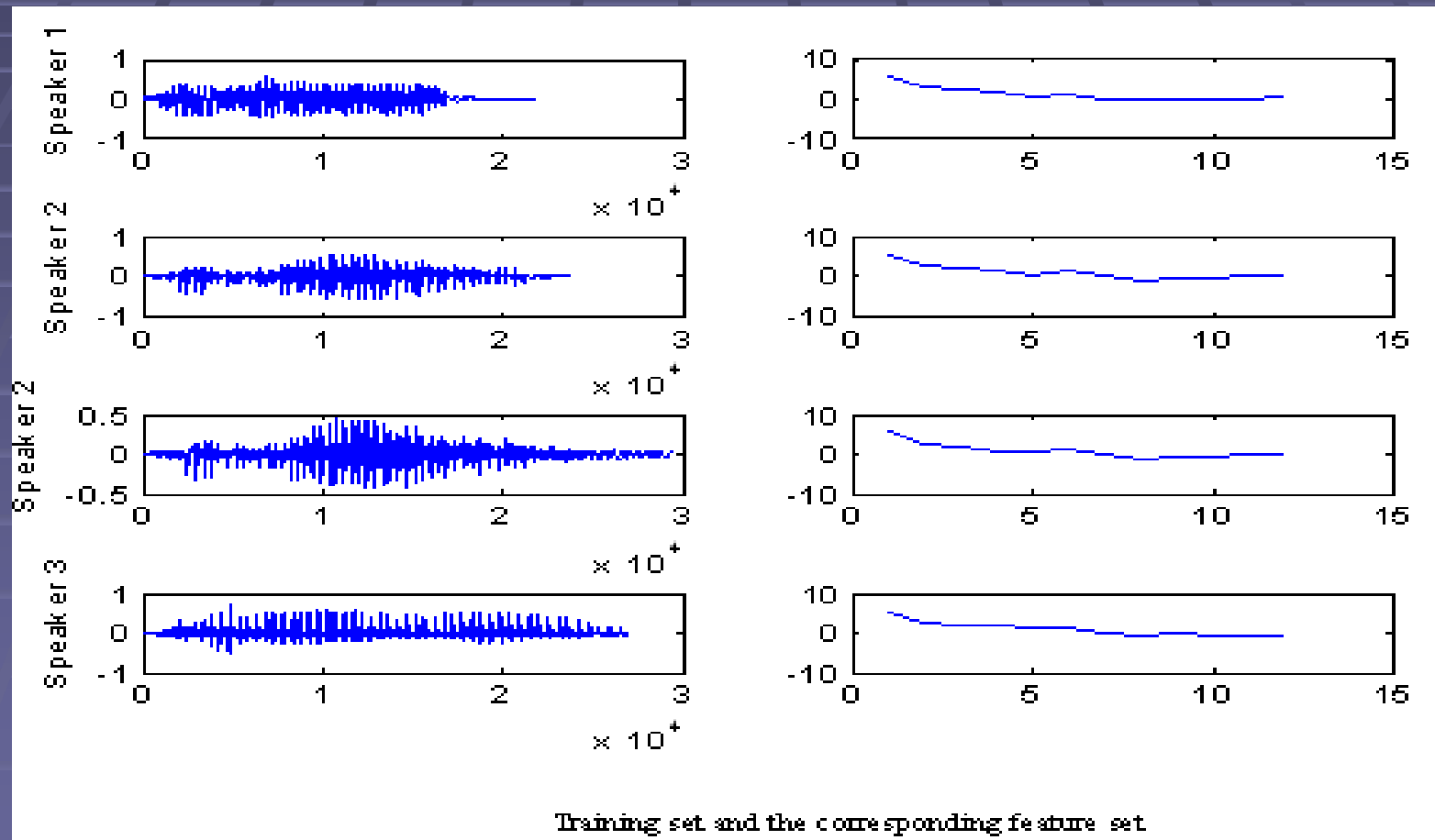
- Example:
 - 5 input vocal utterances – speech *Hello!*
 - 3 registered speakers



Input speech signals and their vocal feature vectors



- Training set:





- Mean distance values table:

	Speaker 1	Speaker 2	Speaker 3
Speech input 1	0.8512	1.0957	1.3115
Speech input 2	0.7495	0.9350	0.2556
Speech input 3	1.5512	1.1123	1.5452
Speech input 4	1.6914	1.4814	1.5013
Speech input 5	1.2527	1.4456	1.5123

- Recognition result: Speaker 1 \Rightarrow $\{S_1, S_4\}$, Speaker 2 \Rightarrow $\{S_3\}$, Speaker 3 \Rightarrow $\{S_2\}$, Unregistered Speaker \Rightarrow $\{S_4\}$.

Text-independent Voice Recognition

- These systems involve impressing volumes of training data
- Useful for not cooperative subjects, like those in the surveillance systems
- Most successful speech-independent recognition techniques:
 - **Vector Quantization (VQ)** = parametric methods, using codebooks
 - **Gaussian Mixture Models (GMMs)** = non - parametric methods, using K Gaussian distributions.

...

- These methods use 1D feature vectors, we propose 2D speech feature vectors.
- Each signal $S \Rightarrow V(S) = DDMFCC$ truncated acoustic matrix
- The described speaker feature extraction is applied.
- Sequence to be recognized: $\{S_1, \dots, S_n\} \Rightarrow$ not characterized by the same speech.
- A similar minimum mean distance supervised classifier is used.

...

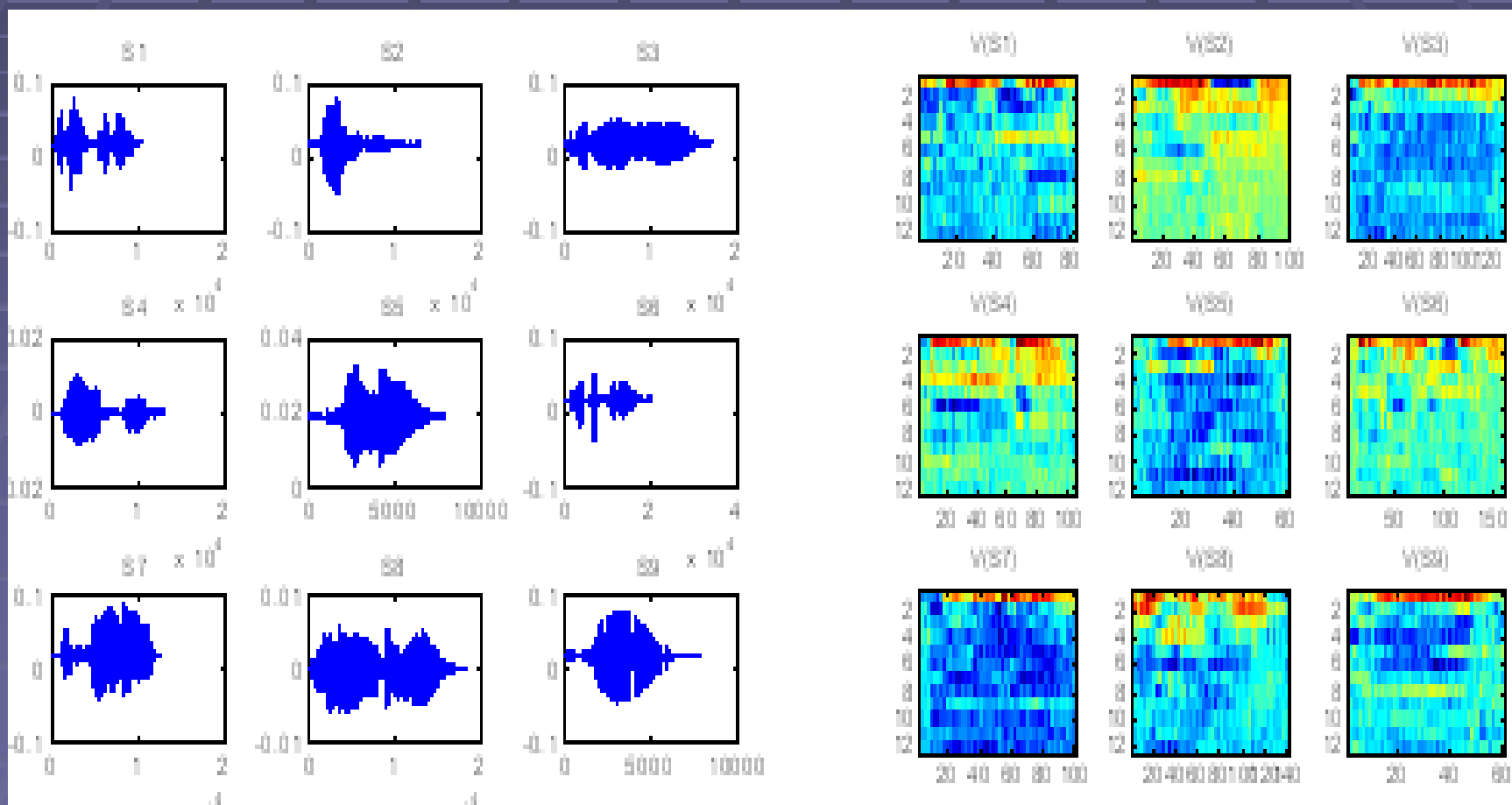
- A different training set is used => all the training vocal utterances have the same speech = large spoken utterance, containing all the English language phonemes.
- Each registered speaker generates this speech several times.
- Each input vocal sequence in the class of the speaker corresponding to the smallest mean distance.
- Threshold-based voice verification => threshold value is set empirically.

...

- Experiments:
 - We performed many experiments on a lot of speech datasets.
 - A satisfactory speaker recognition rate is obtained (80%).
- Example:
 - 3 registered speakers
 - 9 input vocal sequences
- Their feature vectors are computed

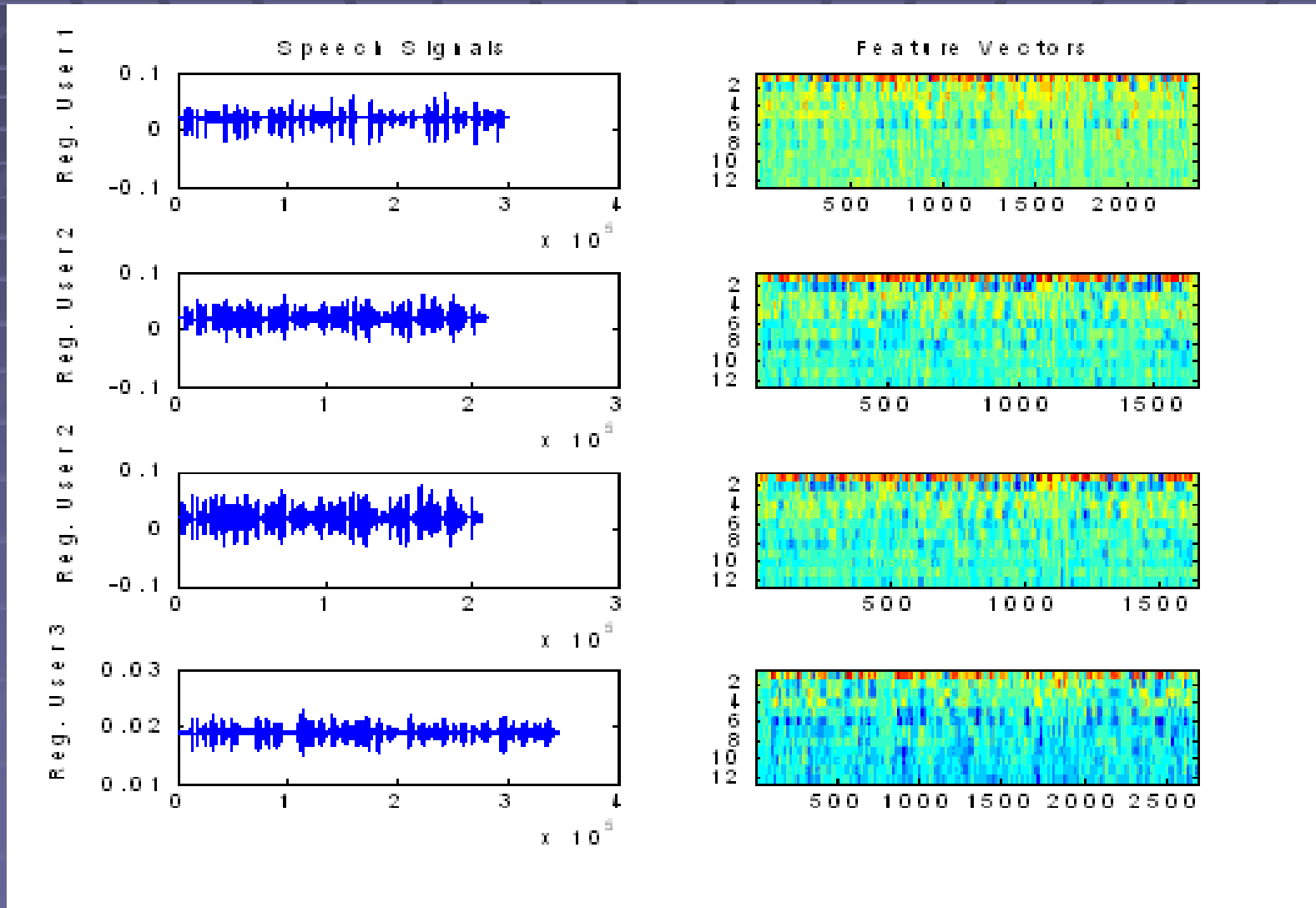


- Input utterances and feature vectors:





- Training signals and feature set :



...

- The mean distance values:

	Speaker 1	Speaker 2	Speaker 3
Speech input 1	5.6678	3.4676	6.2476
Speech input 2	4.1606	7.2581	6.4327
Speech input 3	5.9543	3.8976	4.3671
Speech input 4	6.0946	4.7342	2.9857
Speech input 5	10.6853	9.7366	8.6545
Speech input 6	5.1522	5.7855	6.3879
Speech input 7	7.5031	4.8871	10.8775
Speech input 8	8.7624	7.9964	5.8976
Speech input 9	3.2465	8.0245	4.9082

- Recognition result: Speaker 1 \Rightarrow $\{S_2, S_6, S_9\}$, Speaker 2 \Rightarrow $\{S_1, S_3, S_7\}$, Speaker 3 \Rightarrow $\{S_4, S_8\}$ and Unregistered Speaker \Rightarrow $\{S_5\}$.

Conclusions

- Voice recognition => important domain, making possible to use the speakers' voice to verify their identity and control access to various **services** (banking by telephone, database access services, information services, voice mail, security control for confidential information areas, and remote access).
- We approached the both sub-domains of speaker recognition: text-dependent and text-independent.
- Main contributions:
 - The proposed feature extraction,
 - Supervised classifier
 - The provided metric
 - Threshold computation approach

Thank You!

- Any Questions?