E9 261

11-04-2016

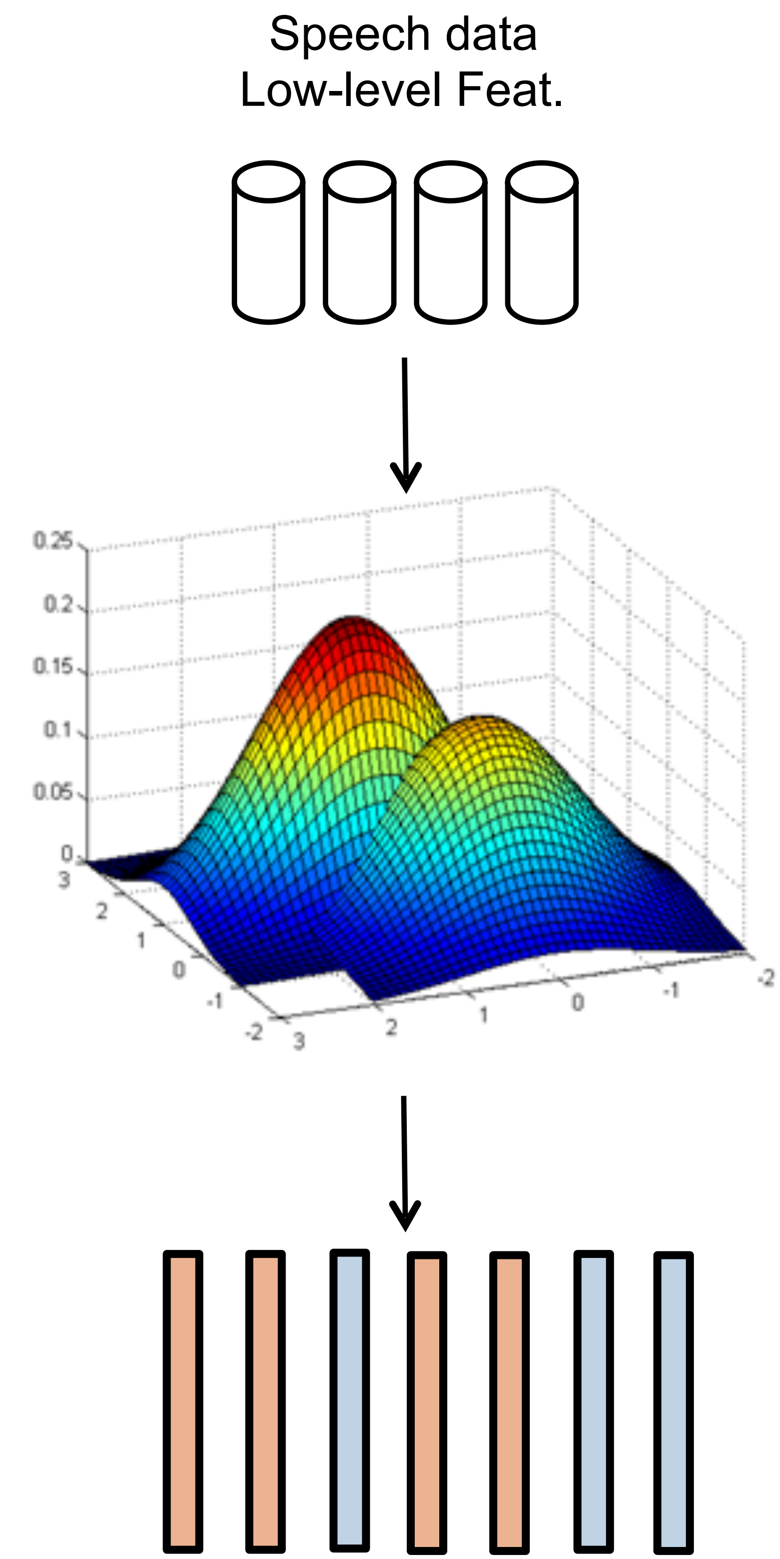
Data Driven Features

- Low-level features capture the acoustic signal information from the recording.
- For many applications, the statistical summary of the low-level features over the entire recording is useful.
  - Example, for speaker and language verification, these average statistic is a good representation and widely used.
  - Avoids dependency on the duration of the audio recording.

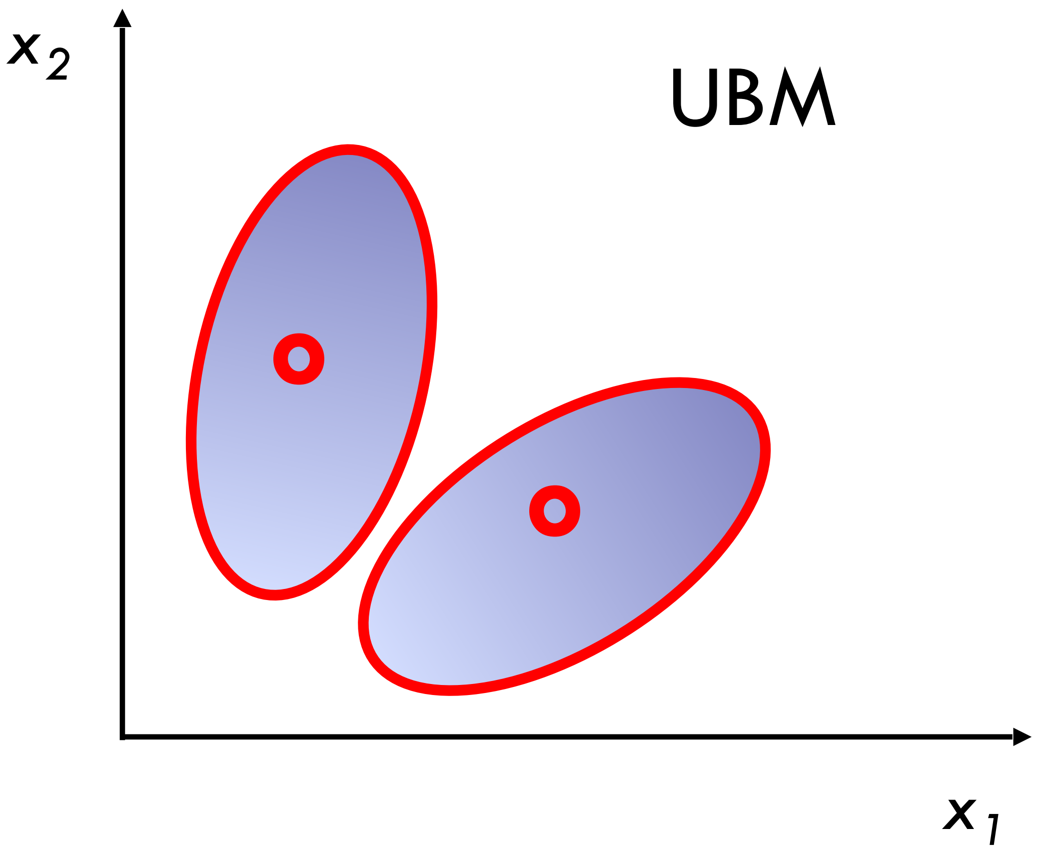
- This statistical summary can be derived from a universal background model (UBM).

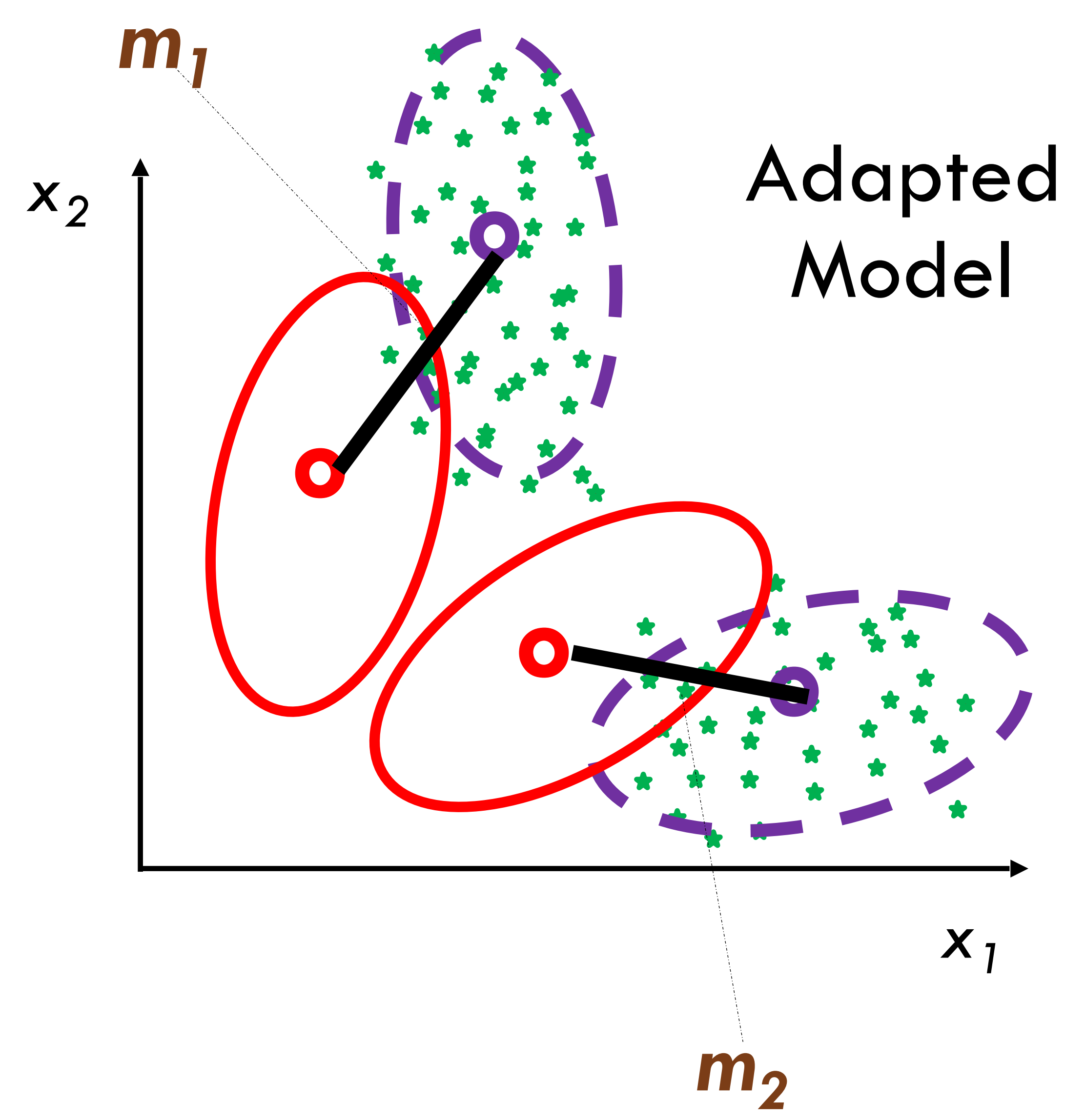
Overview of UBM Based Features

- Higher level features can be derived from lower level features by training an acoustic model. For example,
  - Derive low-level features like MFCC.
  - Training a Gaussian mixture model from a large number of speech recordings.
  - Aligning the low-level features with the GMM model.
  - Deriving model based features based on the alignment statistics.



Overview of i-vector Features





- The i–vector model is $\begin{bmatrix} m_1 \\ m_2 \end{bmatrix} = Vy$ where y is the i–vector

i-vector Feature Extraction

- A popular GMM based feature is the i-vector [Kenny, 2005]
- The GMM-UBM with C mixtures is typically trained with a EM algorithm on large number of recordings from a corpus.
- Let $\lambda = \{\pi_c, \mu_c, \Sigma_c\}$ denote the parameters of the GMM-UBM

$$p_\lambda(x) = \sum_{c=1}^{C} \pi_c N(x; \mu_c, \Sigma_c)$$

- Here, F is the dimension of μ_c and Σ_c is assumed diagonal F x F
- Let supervector $\boldsymbol{M}_0$ be the concatenation of μ_c for $c = 1..C$ with dimension CF x 1
- Let $\boldsymbol{\Sigma}$ be CF x CF block diagonal matrix with diagonal blocks $\Sigma_1 \ldots . \Sigma_C$

i-vector Feature Extraction

- Let $X(s)$ denote the low-level feature sequence for input recording with $X(s) = \{x_i^s, i = 1 \ldots H(s)\}$ where s denotes the recording index and i denotes the frame index, $H(s)$ denotes number of frames. Each $\boldsymbol{x}_i^s$ is a F dimensional feature vector.
- Let $\boldsymbol{M}(s)$ denote the CF x 1 supervector formed by the concatenation of means for the recording s.
- The i-vector model is

$$\boxed{\boldsymbol{M}(s) = \boldsymbol{M}_{\mathbf{0}} + \boldsymbol{V}\boldsymbol{y}(s)}$$

- $\boldsymbol{V}$ is of dimension CF x R known as total-variability matrix.
- The i-vector $\boldsymbol{y}(s)$ is of random vector of dimension R and assumed to be $N(\mathbf{0}, \boldsymbol{I})$

i-vector Model Estimation

- Outline of the iterative i-vector model estimation using EM algorithm (details of the proofs [Kenny, 2005]).
  - Step 1 – Finding the posterior distribution of the i-vector for the given the recording $X(s)$ and the current estimates of $\boldsymbol{V}$.

$$\boldsymbol{y}(s) = \underset{\boldsymbol{y}}{\operatorname{argmax}}\; p_\lambda(\boldsymbol{y} \,|X(s), \boldsymbol{V})$$

  This posterior distribution is a Gaussian and the mode is the mean.

  - Step II – Update the estimate of $\boldsymbol{V}$ using the entire set of recordings and the $s = 1 \ldots .S$ and the estimates $\boldsymbol{y}(s)$

$$\boldsymbol{V} = \underset{\boldsymbol{V}}{\operatorname{argmax}} \prod_{s=1}^{S} p_\lambda(X(s) \,|\, \boldsymbol{y}\,(s))$$

Course Summary (Second Half)

- Gaussian Mixture Models (GMM)
- Expectation Maximization (EM) Algorithm
- Dynamic Time Warping (DTW)
  - Dynamic Programming
- Hidden Markov Models (HMM)
  - Baum Welch Re-estimation
- Hybrid models using Deep networks (DNN)
  - Back Propagation Algorithm
  - Initialization and other considerations.

Applications

- Speech Recognition

Speech Recognition

- Decoding human speech automatically

Speech Recognition

- Decoding human speech automatically.

- Model the feature sequence using
  - Generative Models – Hidden Markov Models with GMMs
  - Discriminative Models – Neural Networks

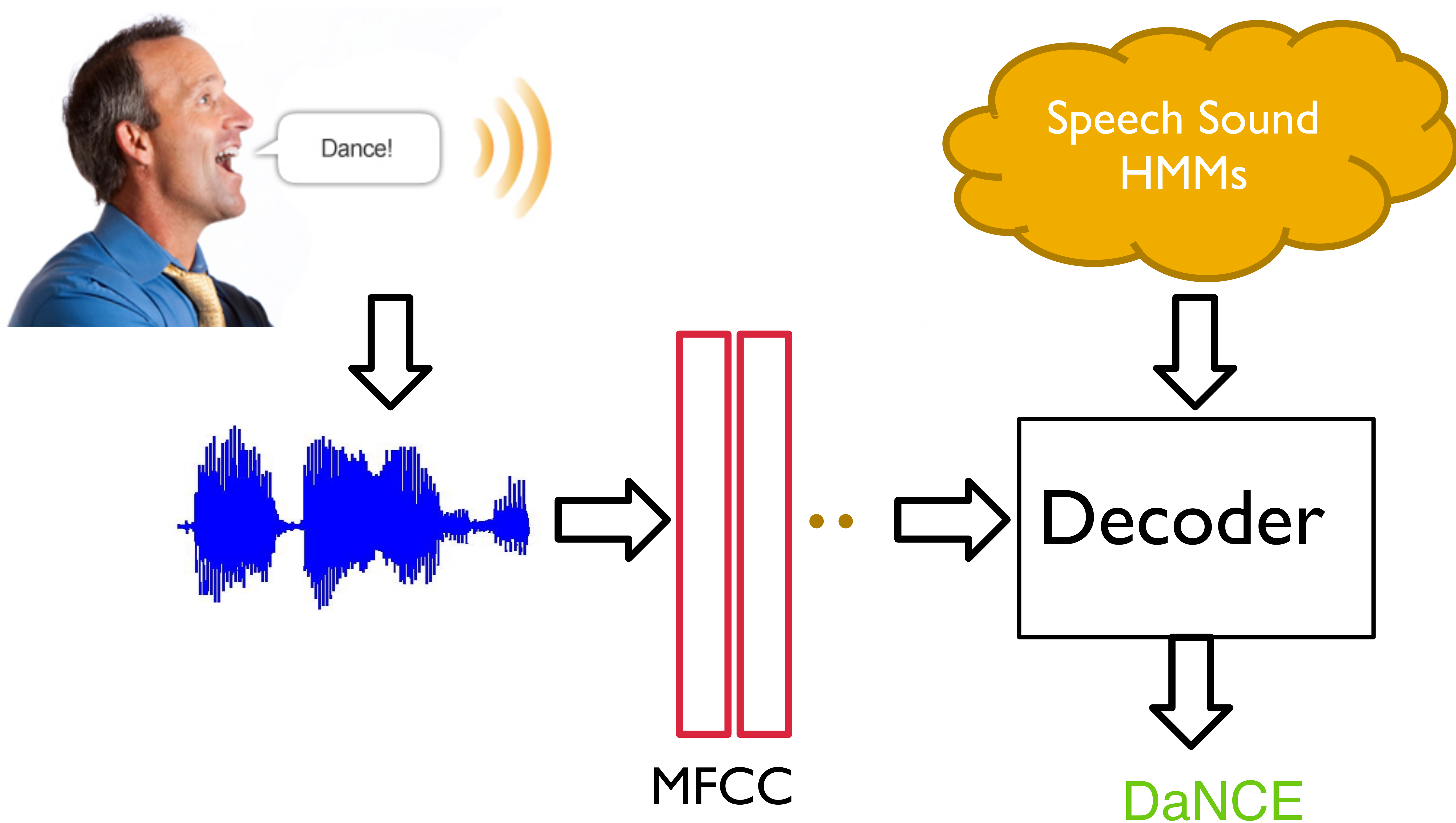
Siri

Cortana

Android

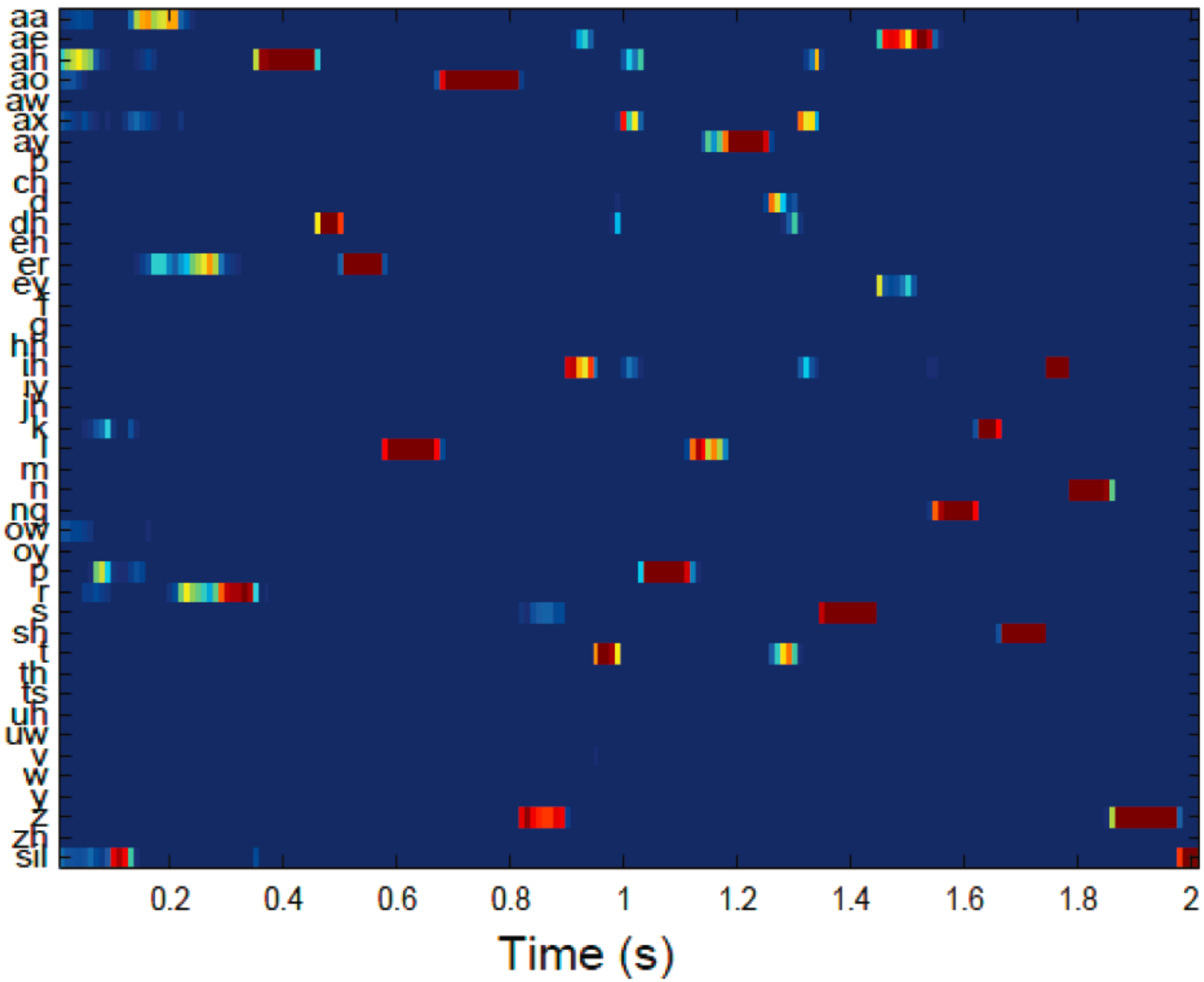
Echo

Speech Recognition with HMMs



Speech Recognition with DNNs

- Example of speech posteriogram



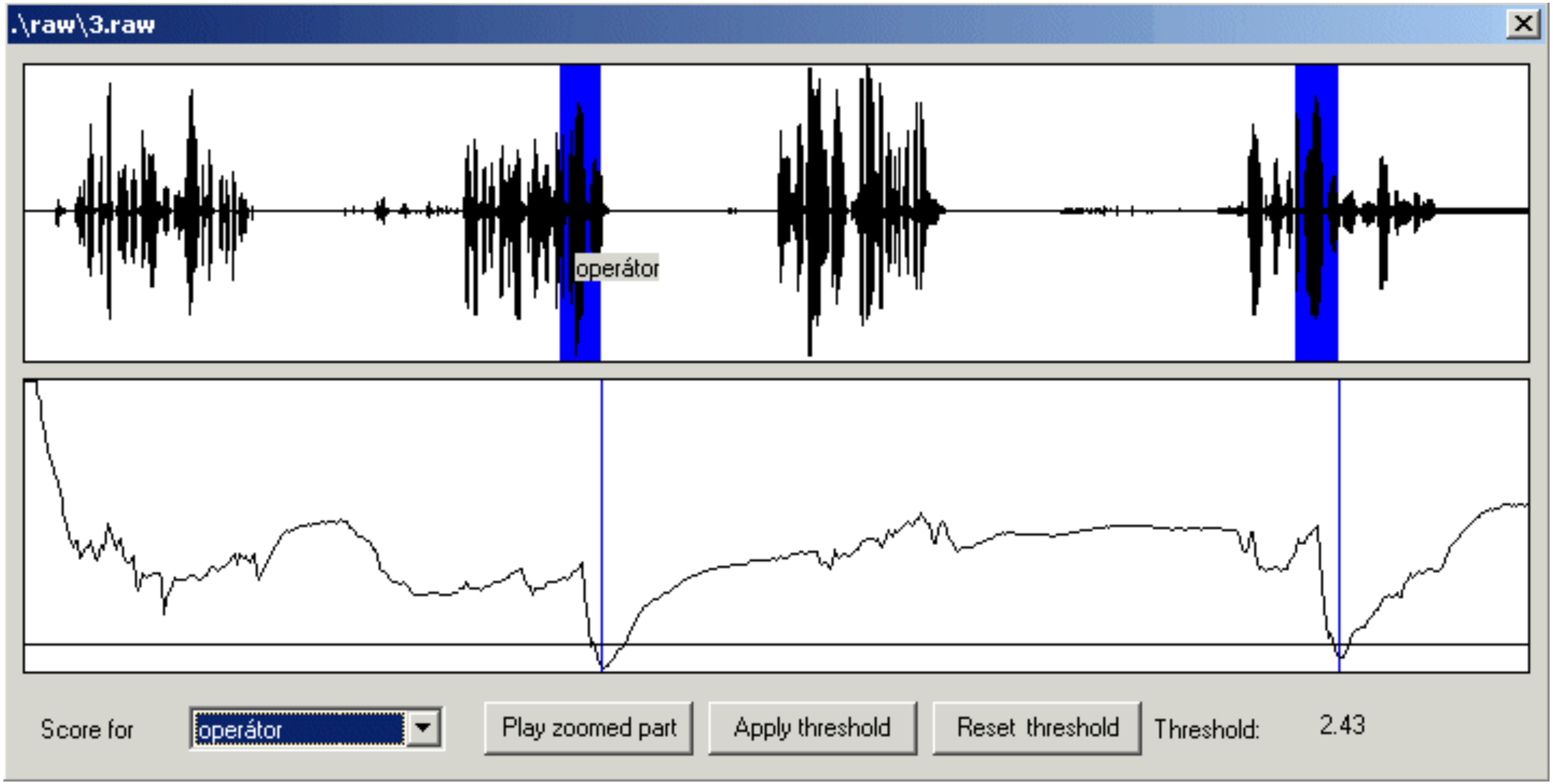
State of the art Recognizer

- MFCC/PLP features with speaker adaptation (like maximum likelihood regression) using GMM-HMMs
- Deep Models (Acoustic Models) -
  - DNNs
  - CNNs
  - RNNs (LSTMs)
- Language Models with n-grams or RNNs
- Decoding using Weighted Finite State Transducers (WFSTs).

Applications

- Speech Recognition
  - Keyword Spotting

Keyword Spotting



State of the art Recognizer

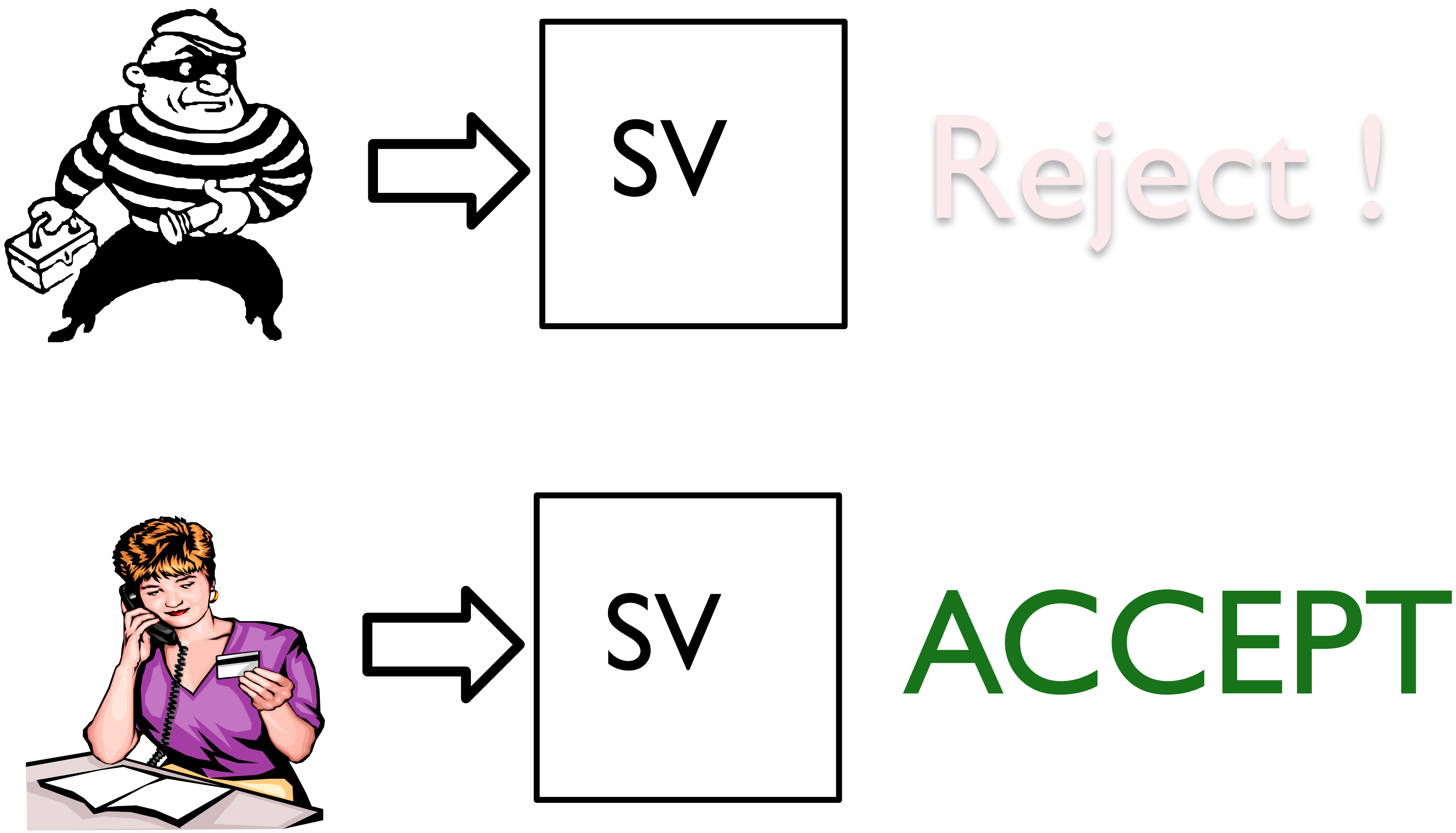
- Keyword Spotting - Done using the lattice (graph of all connected paths).
- Low-resource keyword spotting - using pattern matching techniques like DTW [Zhang, 2009, Jansen, 2013].

Applications

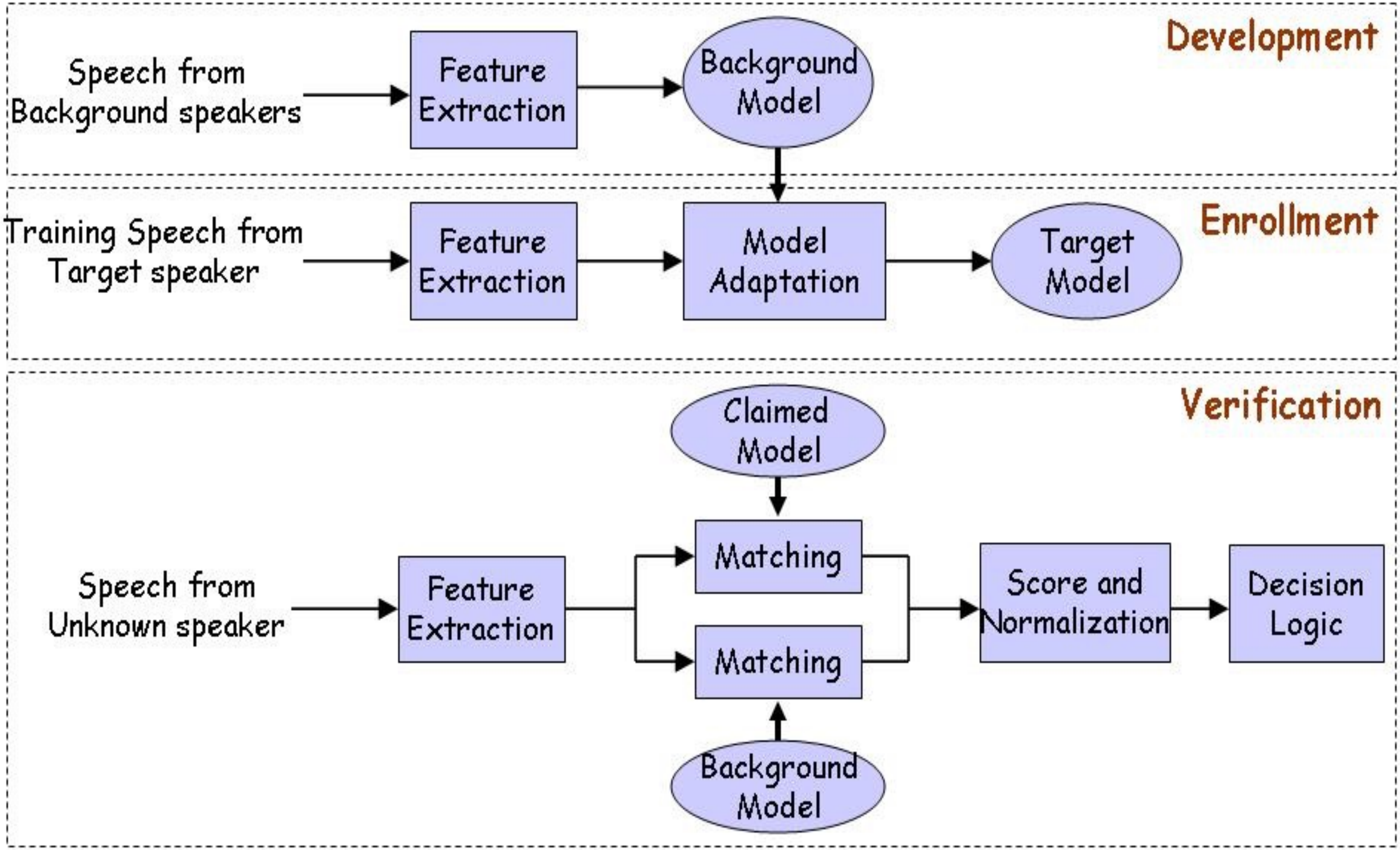
- Speech Recognition
- Speaker Verification

Speaker Verification

- Verify the identity of a speaker



Speaker Verification



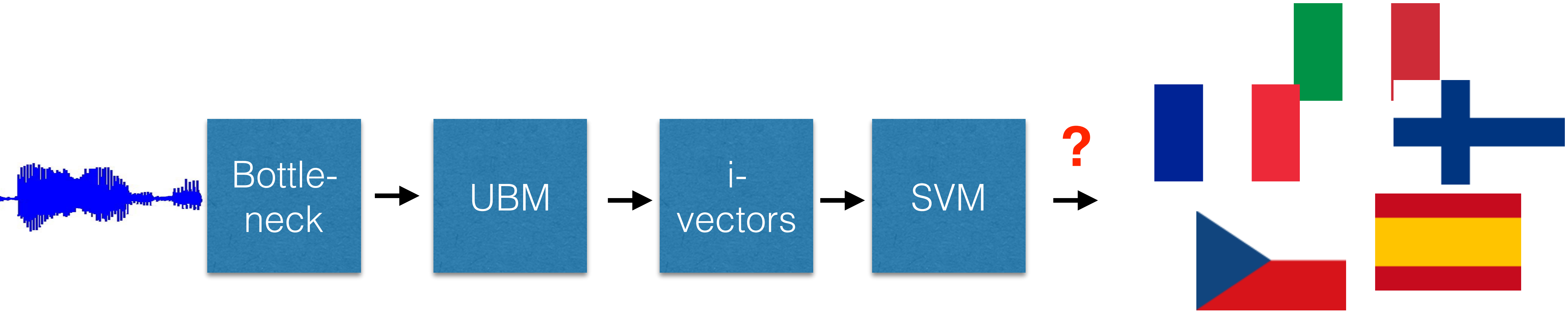
State of the art Speaker Recognition System

- Building GMM-UBM
- i-vector extraction
- Scoring with probabilistic linear discriminant analysis (PLDA) models.

Applications

- Speech Recognition
- Speaker Verification
- Language Identification

Language Identification System

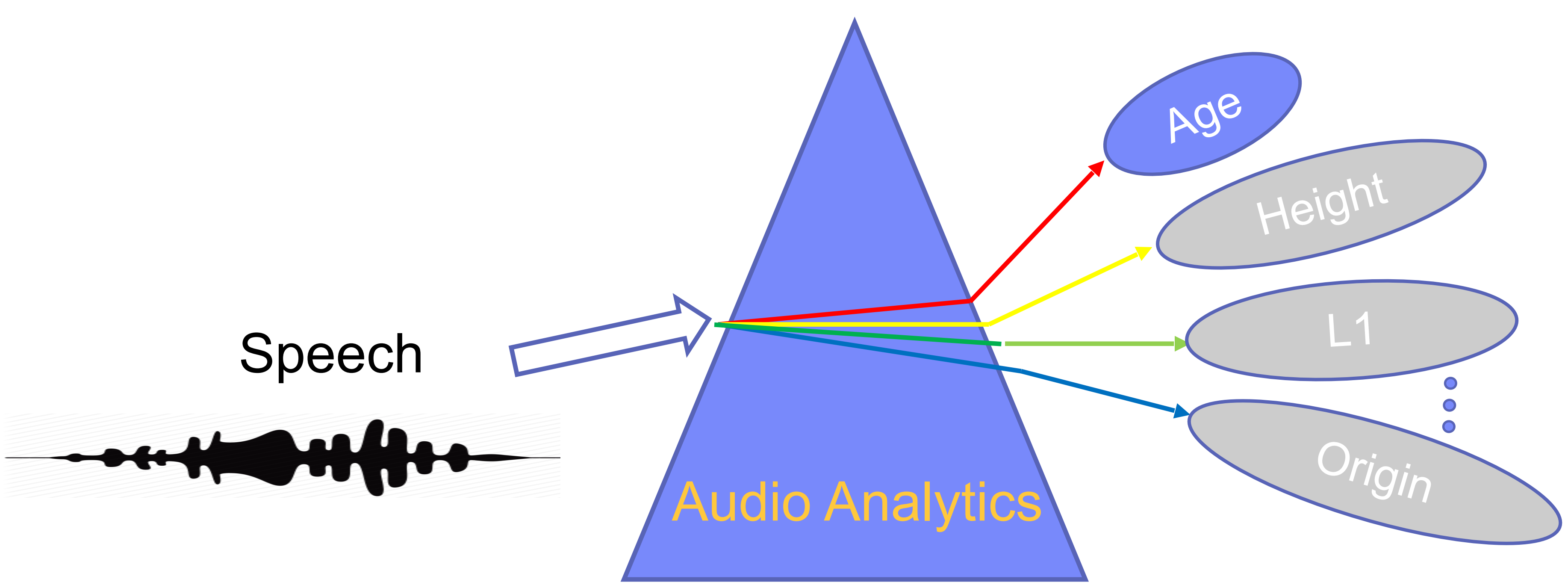


Applications

- Speech Recognition
- Speaker Verification
- Language Identification
- Audio Analytics
- Speech Activity Detection

Audio Analytics

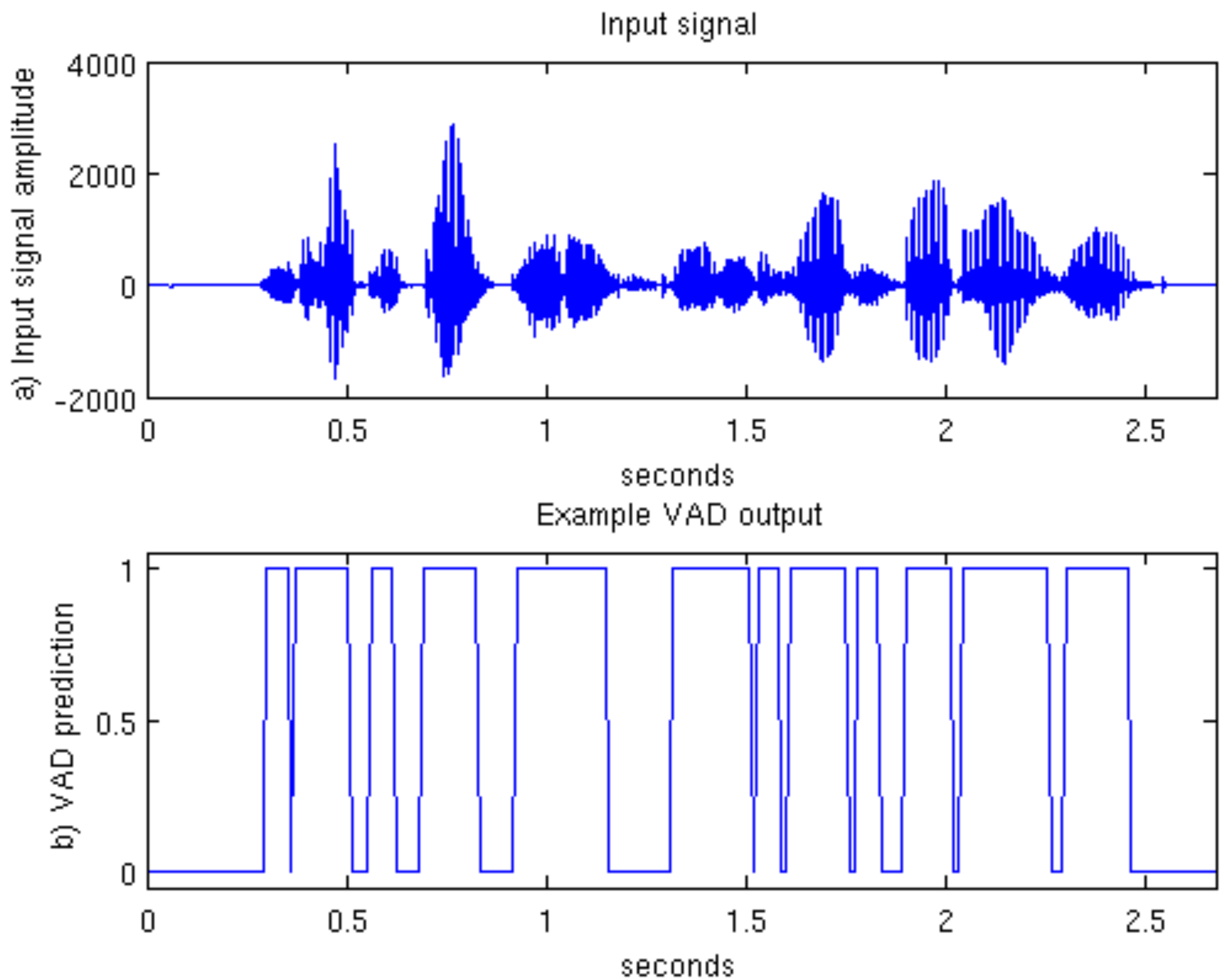
- **Speech** is a unique physiological signal that contains both linguistic and paralinguistic information.
- It also carries useful information about the environment (production and transmission medium)
- **Audio analytics:** automatic extraction of paralinguistic content from speech input.
- State-of-art systems use i-vector features and some regression/classification methods.



Applications

- Speech Recognition
- Speaker Verification
- Language Identification
- Audio Analytics
- Voice/Speech Activity Detection

VAD/SAD



Input signal
a) Input signal amplitude
4000
2000
0
-2000
0
0.5
1
1.5
2
2.5
seconds
Example VAD output
b) VAD prediction
1
0.5
0
0
0.5
1
1.5
2
2.5
seconds

State of the art SAD

- Generative Model based
  - Building class specific GMMs for speech, noise, silence.

- Discriminative models
  - Using DNNs/CNNs.

- Smoothing the frame based estimates using a Viterbi decoding algorithm.

Research Directions

- Unsupervised and semi-supervised learning
  - Low resource scenarios.
  - Information extraction from small audio snippets.

- Robustness to noise
  - Additive noise, reverberation, non-linear channel noises.

- Understanding Deep Models
  - Uncovering the representation.
  - Links with biology ?

Questions & Feedback