Study on Speaker Recognition

Lantian Li

CSLT / RIIT

Tsinghua University

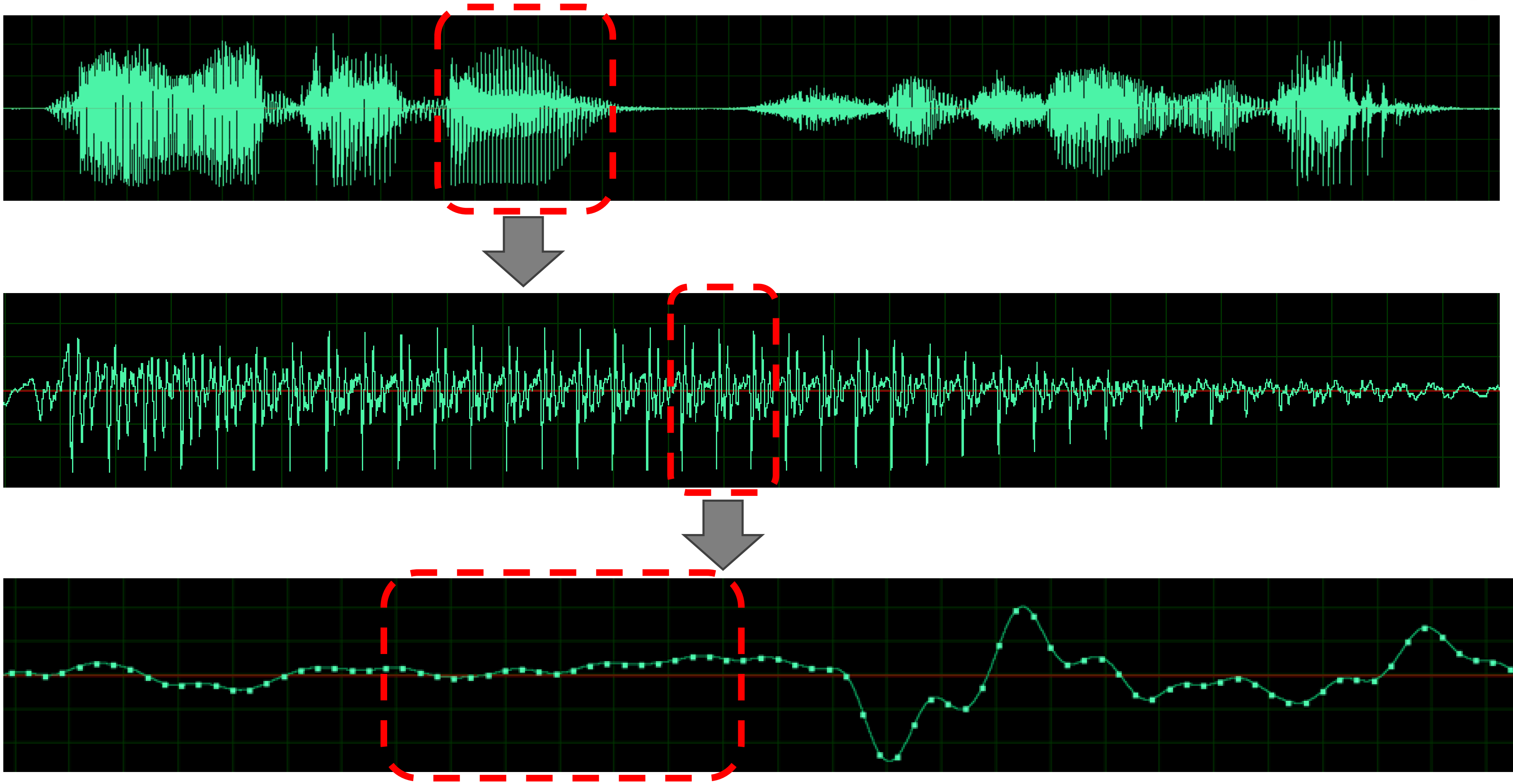
lilt13@mails.tsinghua.edu.cn

July 17, 2017

Outline

- Basic Concepts
- History of Speaker Recognition
  - Where does it come from
  - Three development stages
  - Comparison and combination
- Conclusions and Future work

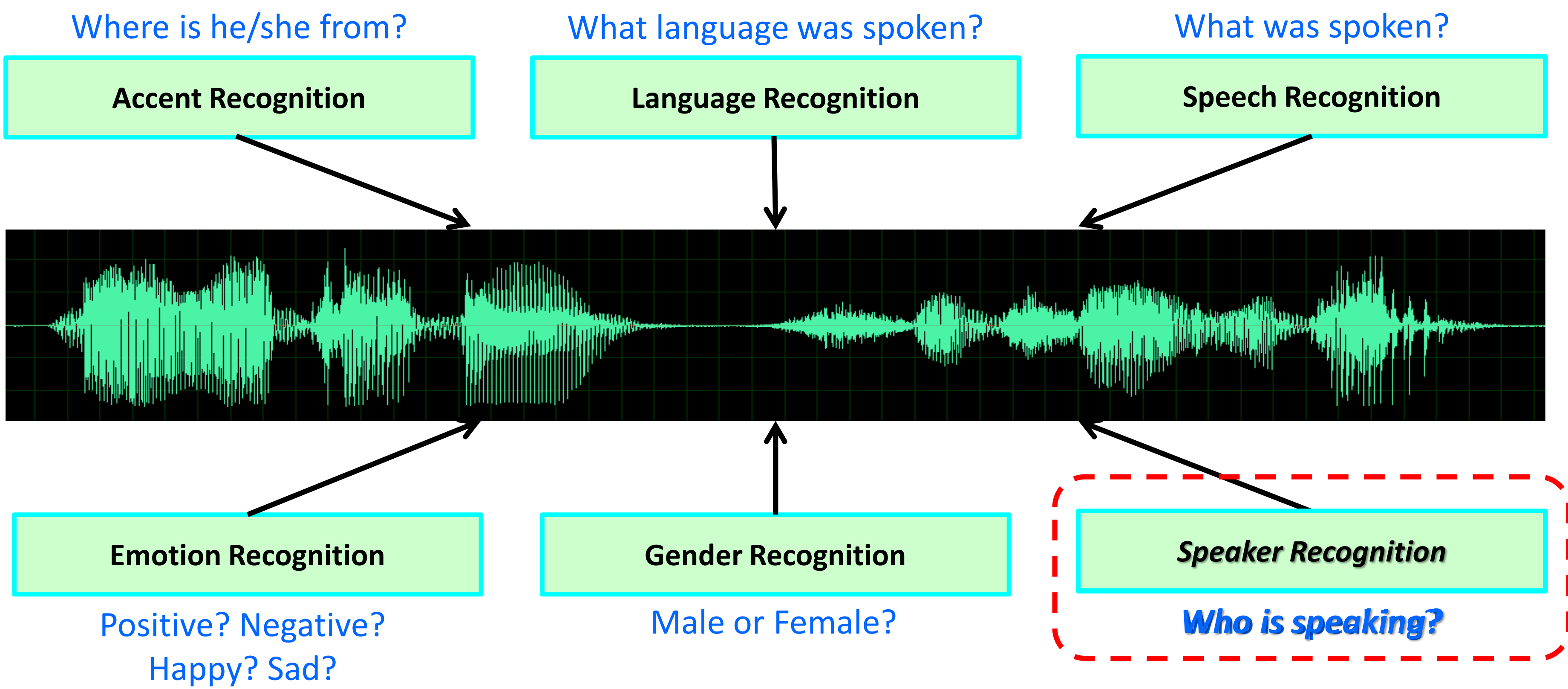
Speech signal



Short-time Fourier Transform (STFT)

Time domain → Frequency domain

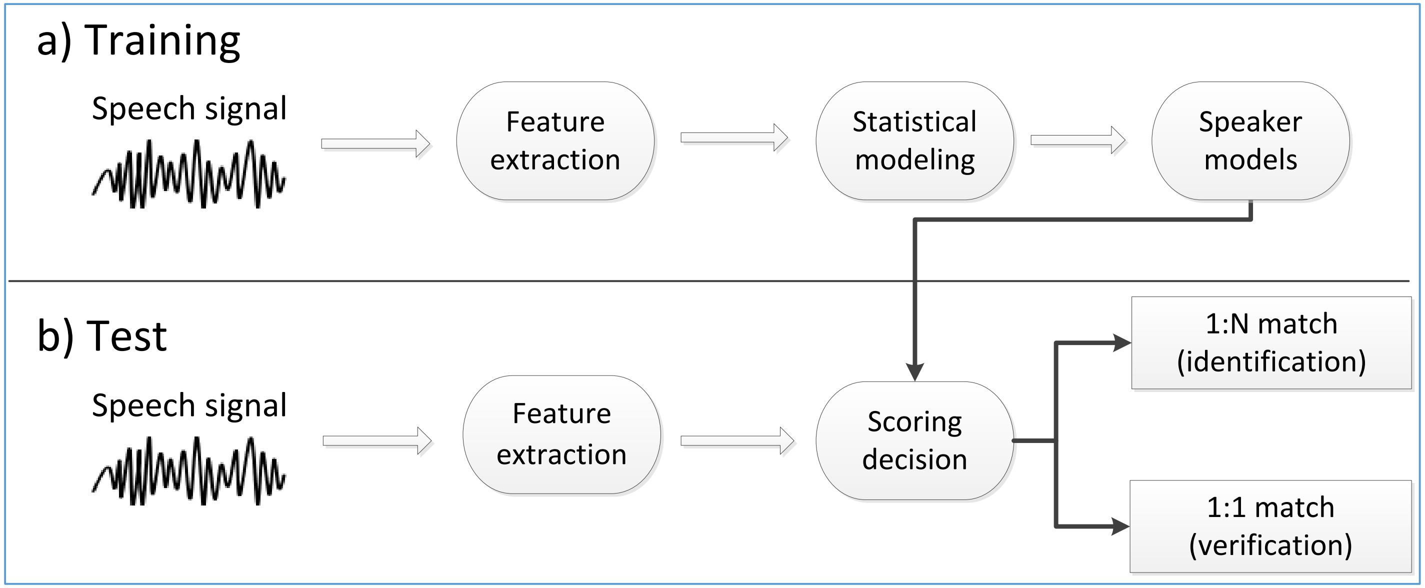
Rich information in speech signals



Mysterious and Fascinating: *simple* in form (one-dimensional vibration), *rich* in information.

Speaker recognition

- **Speaker recognition** is the identification of a person from characteristics of *voices* (voice biometrics). It is also called **voiceprint recognition.** [wiki]



Advantages and applications

- Advantages of speaker recognition
    - Speech signal more acceptable by users
    - Data acquisition much easier (e.g. a mobile phone)
    - Remote authentication more convenient
- Application scenarios
    - Access control (e.g. voice lock)
    - Transaction authentication (e.g. remote payment)
    - Forensic analysis (e.g. police criminal detection)

Categories in application

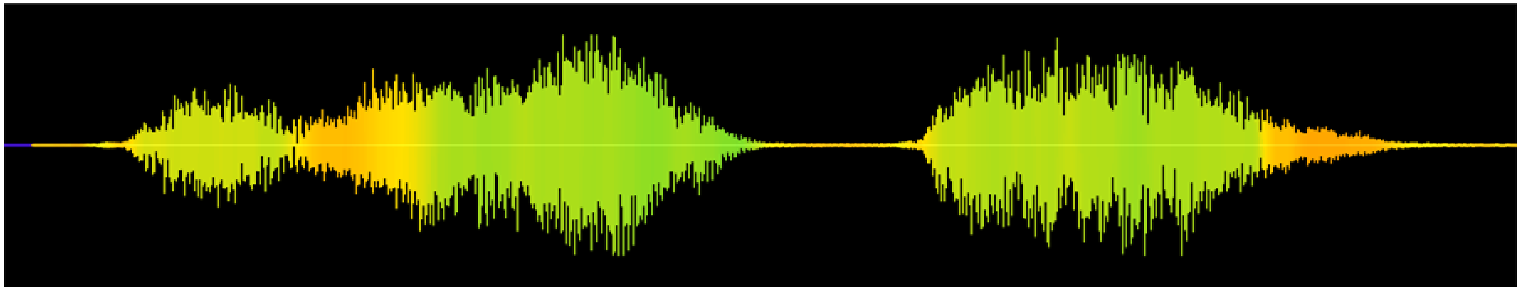
- Speaker Identification
  - Determining **which** identity in *a specified speaker set* is speaking. **1-vs-N**

Categories in application

- Speaker Verification
  - Determining **whether** a *claimed* identity is speaking. **1-vs-1**

Categories in application

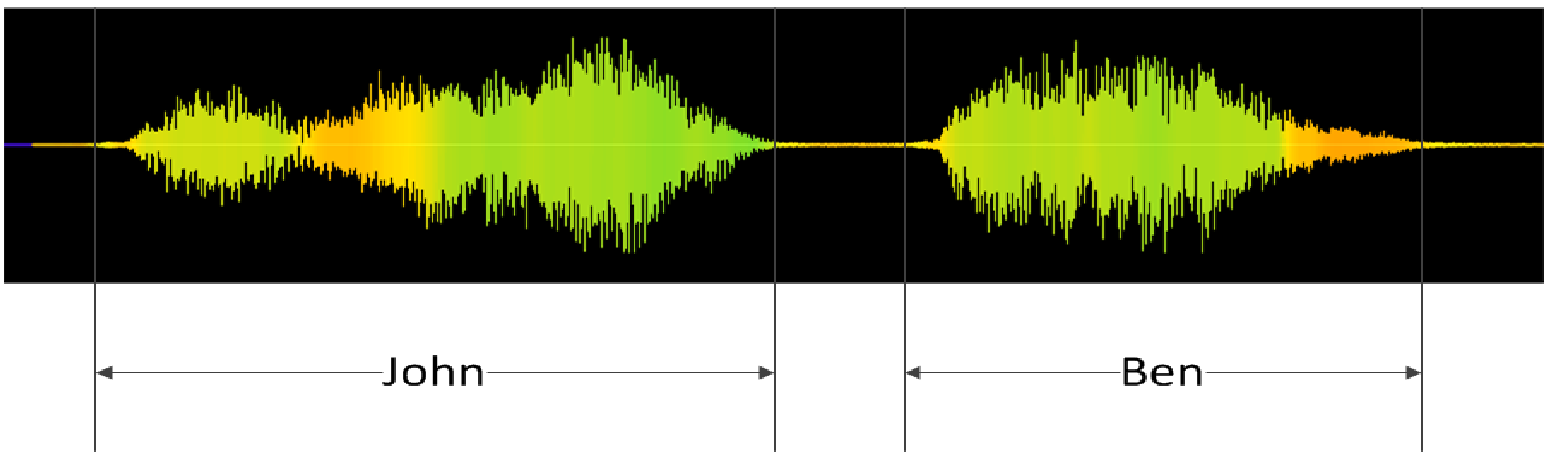
- Speaker Detection
  - Determining **whether** a *specified target speaker* is speaking.



This sentence is a conversation from John and Ben

Categories in application

- Speaker Tracking (Speaker Diarization)
  - Performing speaker detection as **a function of time**, giving the timing index of the specified speaker.



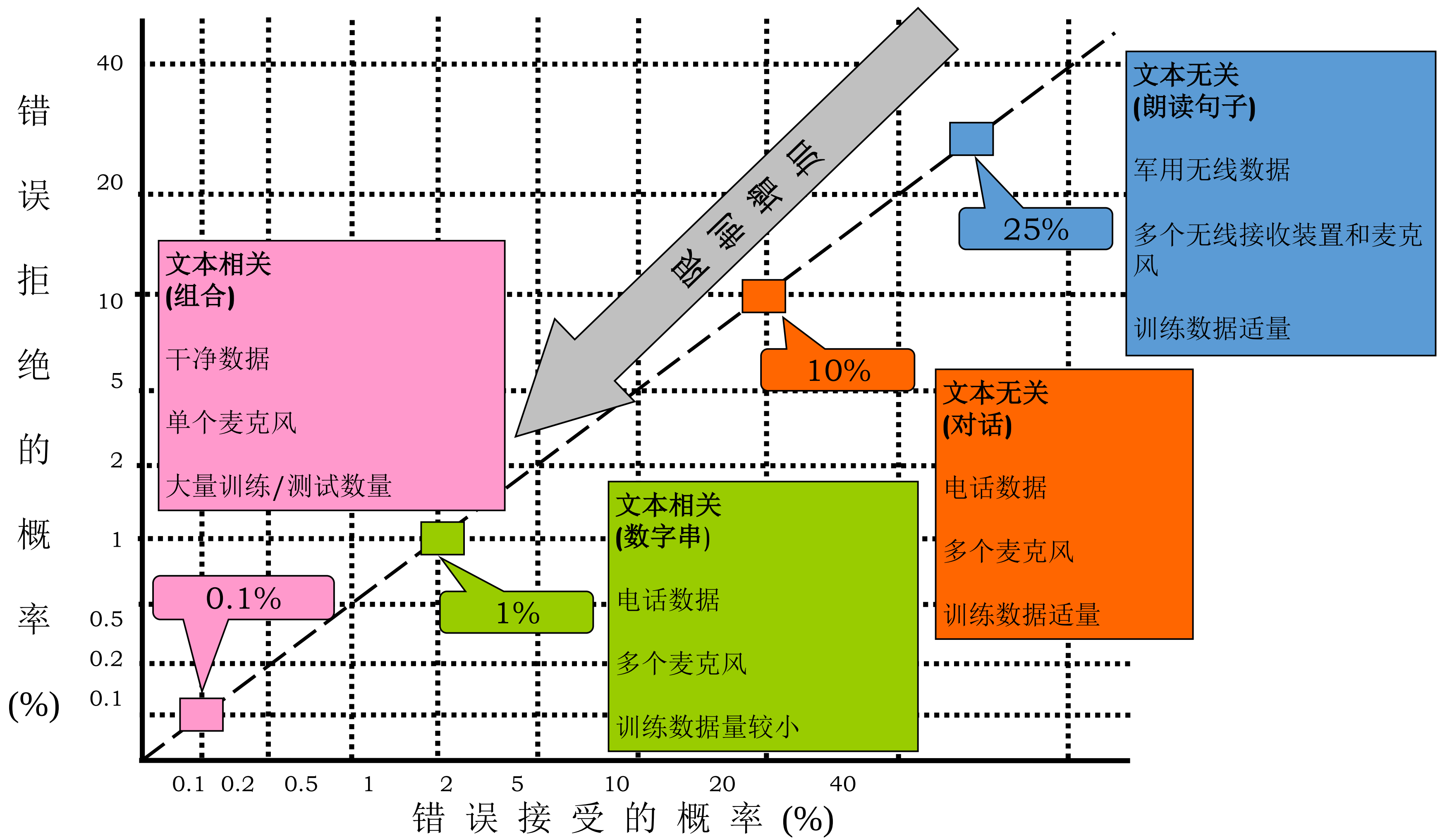
Categories in application

- Speaker Identification
  - Determining which identity in a specified speaker set is speaking. **1-vs-N**
- Speaker Verification
  - Determining whether a claimed identity is speaking. **1-vs-1**
- Speaker Detection
  - Determining whether a specified target speaker is speaking.
- Speaker Tracking (Speaker Diarization)
  - Performing speaker detection as a function of time, giving the timing index of the specified speaker.

Categories in text content

- Text-dependent
  - A pre-determined text for both training and test.
- Text-independent
  - No constraints on the text for training and test.
- Text-prompted
  - The text to speak is not fixed each time when use, but prompted by the system from a specific set of text.
  - Combined with ASR, it is regarded as a '***who spoke what***' task.

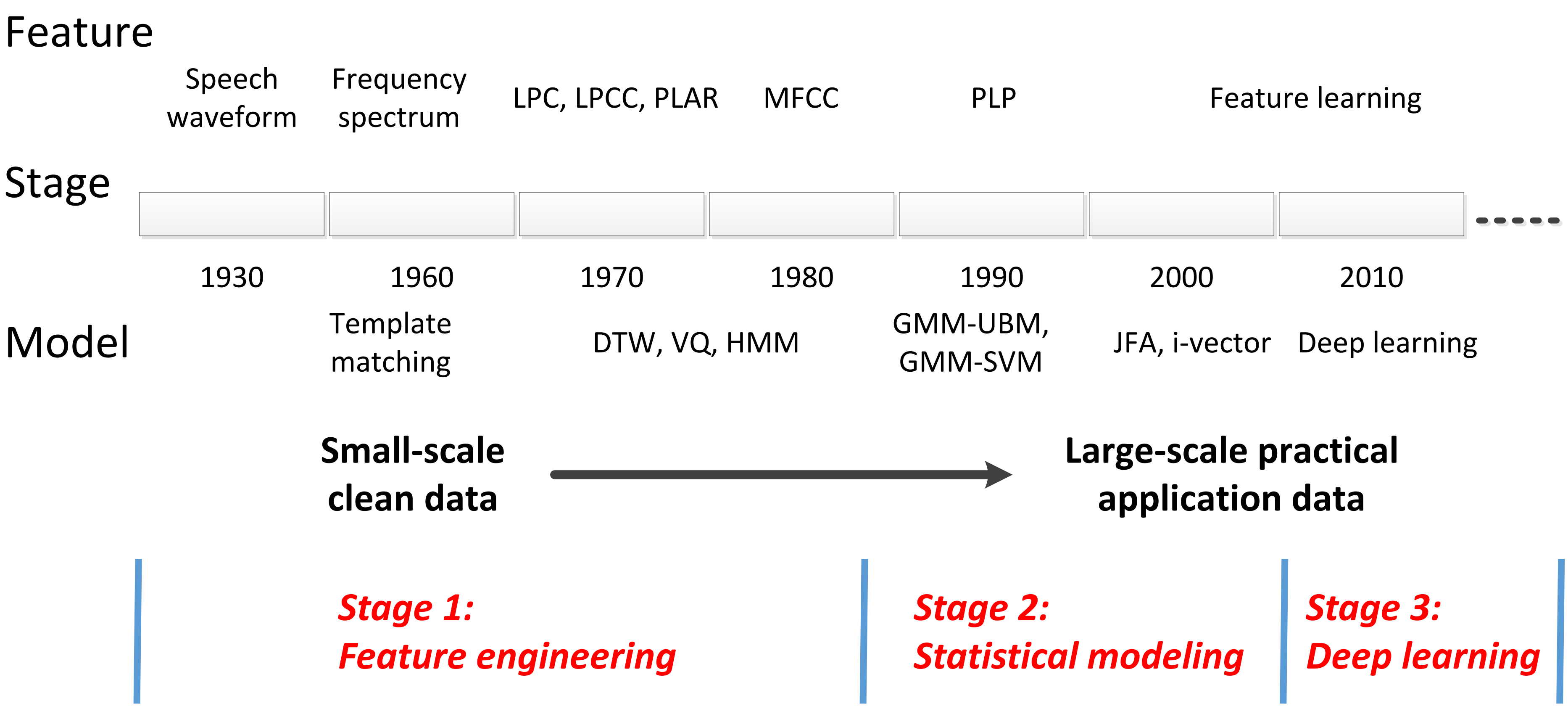
Categories in text content



History of speaker recognition

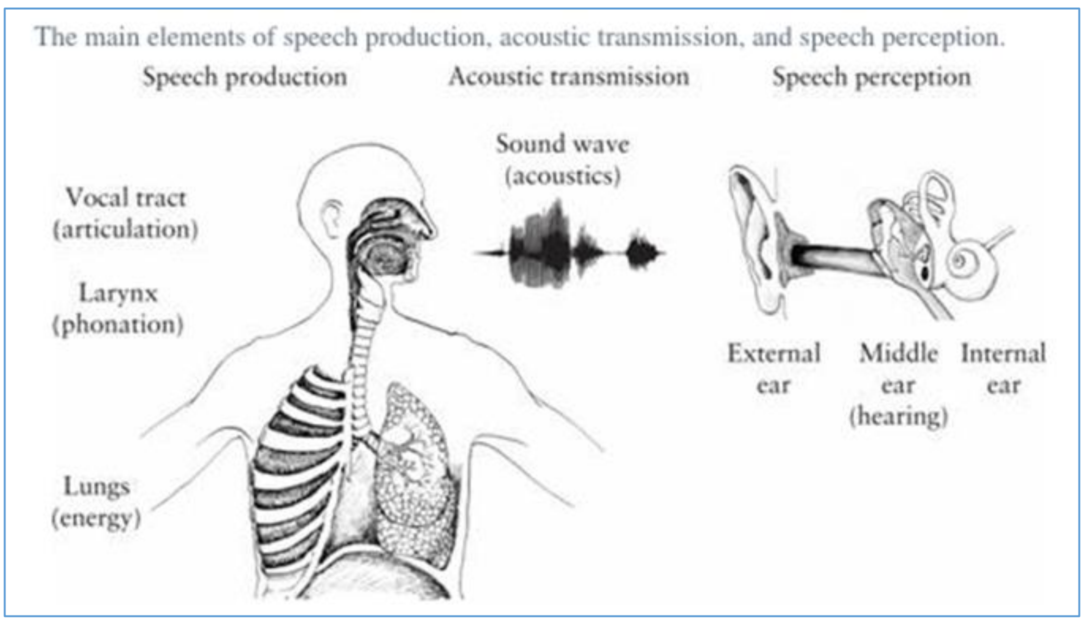
- A bird may be known by its song. (闻其声而知其人)
  - Auditory sense (听觉)
- Back to 1660s, people used the speech signal for identity authentication in the criminal detection (刑侦线索)
  - The trials on the death of Charles I. in Britain
- In 1930s, research on speaker recognition was started.
  - Human auditory perception (人耳听觉感知)
- Biological information research (生物信息研究) and computer information technology (计算机信息技术)

History of speaker recognition



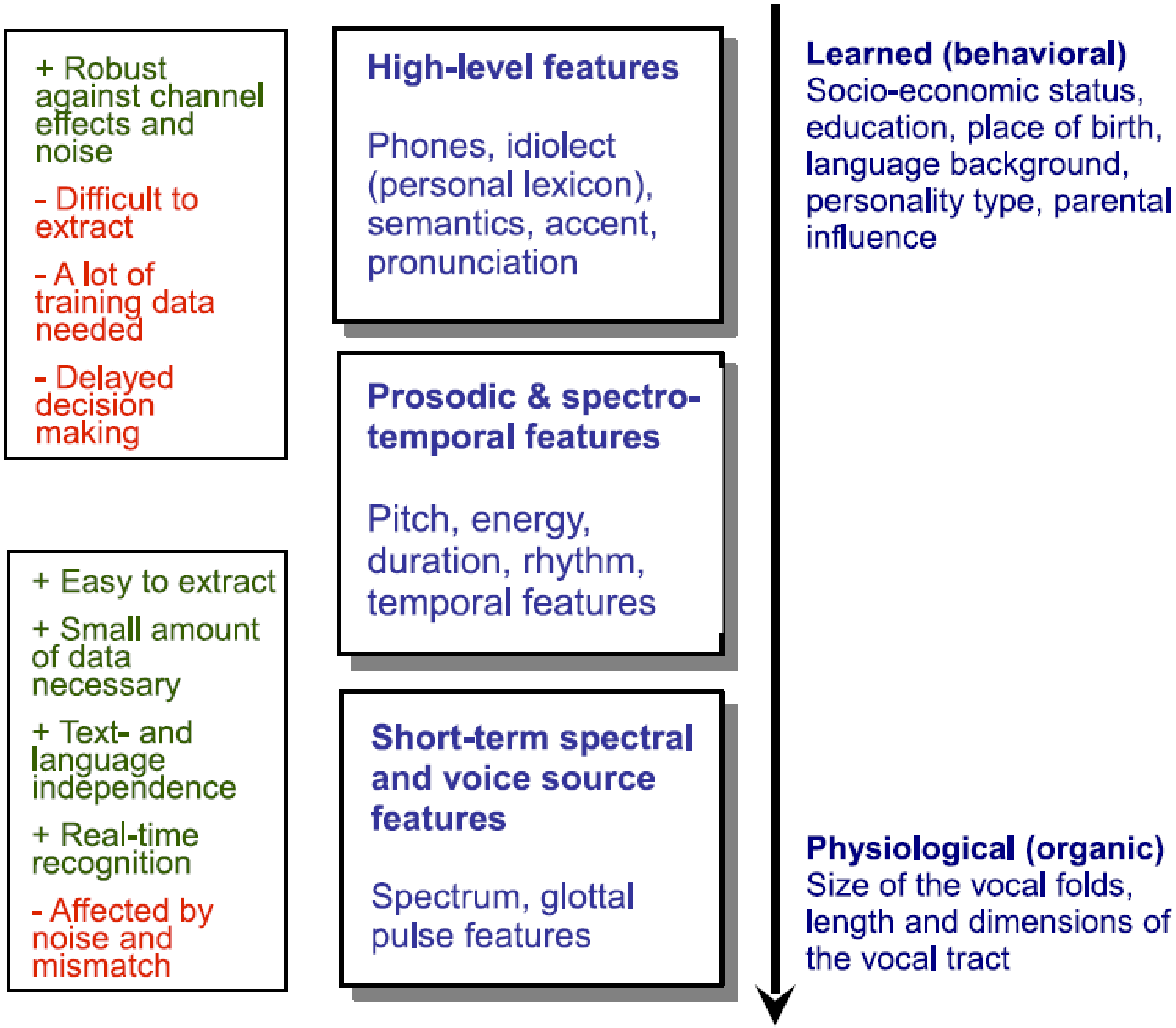
Stage 1: Feature engineering

- Goal: to discover features that are sensitive to speaker traits while invariant to other factors.
- Focus: *speech production* and *auditory perception*



Stage 1: Feature engineering

- Fi



后天学习的

:y

生来就有的

Stage 1: Feature engineering

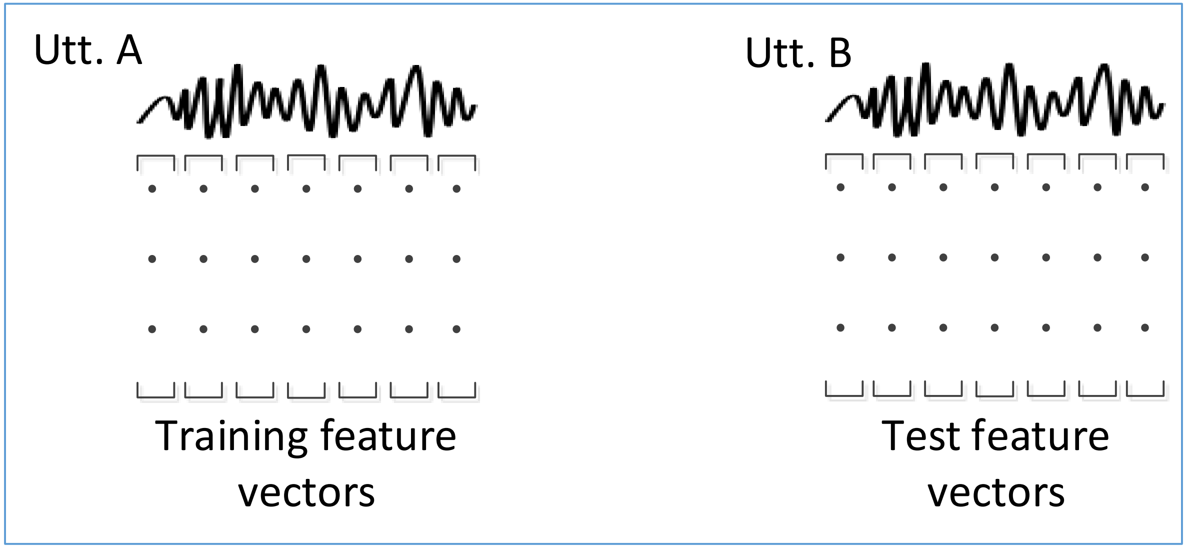
- Unfortunately, none of these features can be regarded as having caught the *fundamental* patterns of speakers.
- ***MFCC*** is still the most useful feature.
- Speaker modeling and evaluation metrics are too simple, and mostly applied in the *text-dependent* condition.
- Template (Nonparametric) models
    - VQ (Vector quantization)
    - DTW (Dynamic Time Warping)

Stage 2: Statistical modeling

- Classical speaker models
  - Template (Nonparametric) models
    - Vector quantization (VQ) model
    - Dynamic Time Warping (DTW)
  - Probabilistic (Parametric) models
    - Gaussian mixture model (GMM)
    - GMM-UBM / GMM-SVM
    - i-vector / JFA

Stage 2: Statistical modeling

- Template (Nonparametric) models
  - Training and test feature vectors are *directly* compared with each other.
  - The *distortion* between them represents their degree of *similarity*.



Stage 2: Statistical modeling

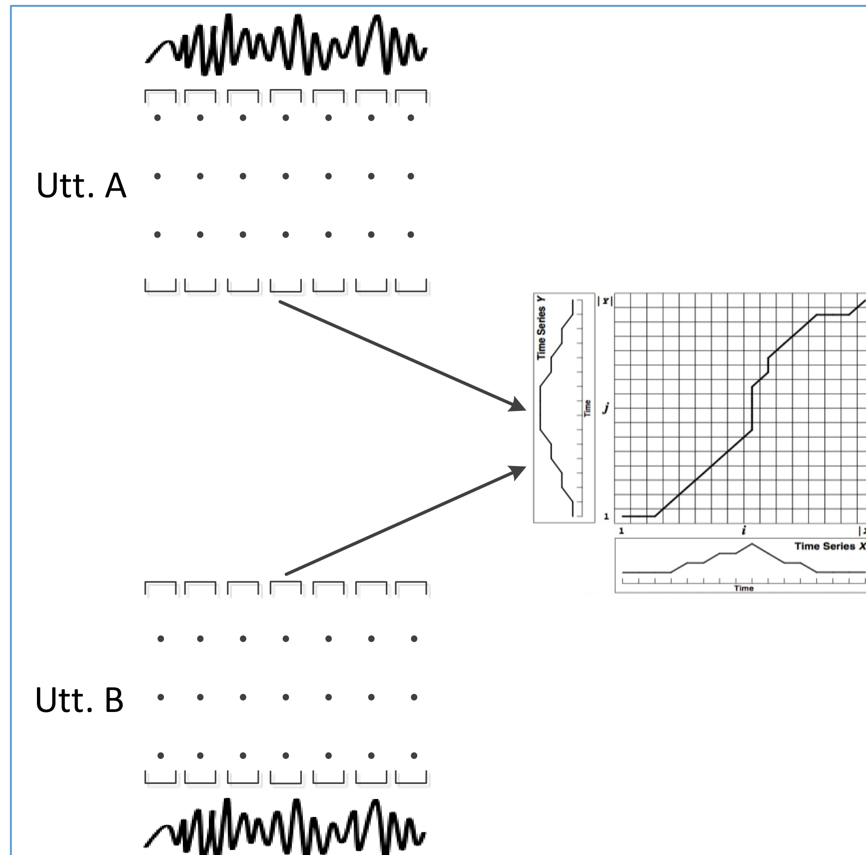
- Template models - *Vector quantization* (VQ)
  - Data compression / Computational speed-up
  - The *simplest* text-independent model
  - Clustering method such as *K*-means to construct a *codebook*. (Centroid model)
  - *Average quantization distortion*:

$$D_Q(\mathscr{X}, \mathscr{R}) = \frac{1}{T} \sum_{t=1}^{T} \min_{1 \leqslant k \leqslant K} d(\boldsymbol{x}_t, \boldsymbol{r}_k)$$

$\mathscr{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$: test feature vectors, $\mathscr{R} = \{\boldsymbol{r}_1, \ldots, \boldsymbol{r}_K\}$: enrollment feature vectors, $d(\cdot, \cdot)$: distance measure.

Stage 2: Statistical modeling

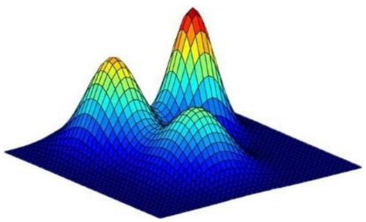
- Template models - *Dynamic Time Warping* (DTW)
  - Text-dependent model
  - Two variable-length temporal sequences
  - *Dynamic programming*
  - To search an optimal path with the lowest cost



Stage 2: Statistical modeling

- Probabilistic models - *Gaussian mixture model* (GMM)
  - A GMM:

$$p(\boldsymbol{x}|\lambda) = \sum_{k=1}^{K} P_k \, \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



  - The *k*-th Gaussian component:

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = (2\pi)^{-\frac{d}{2}} |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\}$$

  - *Maximum likelihood* (ML) estimation

$$\mathrm{LL}_{\mathrm{avg}}(\mathcal{X}, \lambda) = \frac{1}{T}\sum_{t=1}^{T} \log \sum_{k=1}^{K} P_k \mathcal{N}(\boldsymbol{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \xrightarrow{\textbf{\textit{EM}}}$$

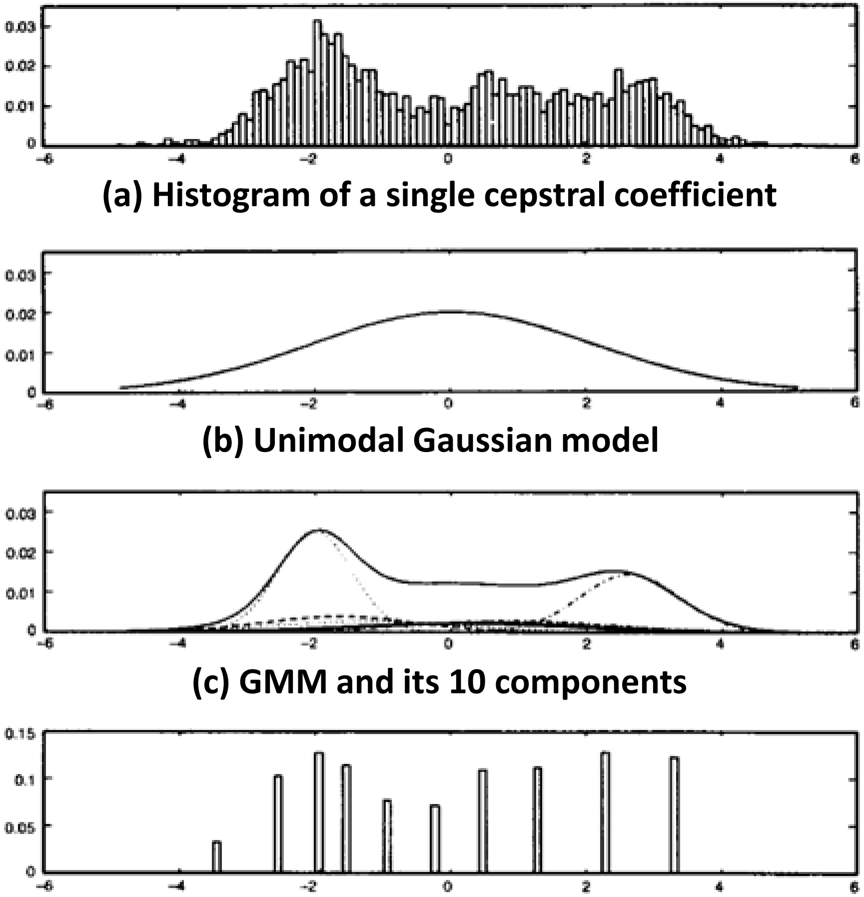
Mixture Weights: $\bar{P}_k = \frac{1}{T}\sum_{t=1}^{T} \underline{p(k|x_t, \lambda)}$

Means: $\bar{\mu}_k = \frac{\sum_{t=1}^{T} p(k|x_t, \lambda)\, x_t}{\sum_{t=1}^{T} p(k|x_t, \lambda)}$

Variances: $\bar{\sigma}_k^2 = \frac{\sum_{t=1}^{T} p(k|x_t, \lambda)\, x_t^2}{\sum_{t=1}^{T} p(k|x_t, \lambda)} - \bar{\mu}_k^2$

Stage 2: Statistical modeling

- Probabilistic models - *Gaussian mixture model* (GMM)



(a) Histogram of a single cepstral coefficient

(b) Unimodal Gaussian model

(c) GMM and its 10 components

(d) VQ centroid locations of a 10-element codebook

A linear combination of Gaussian functions.

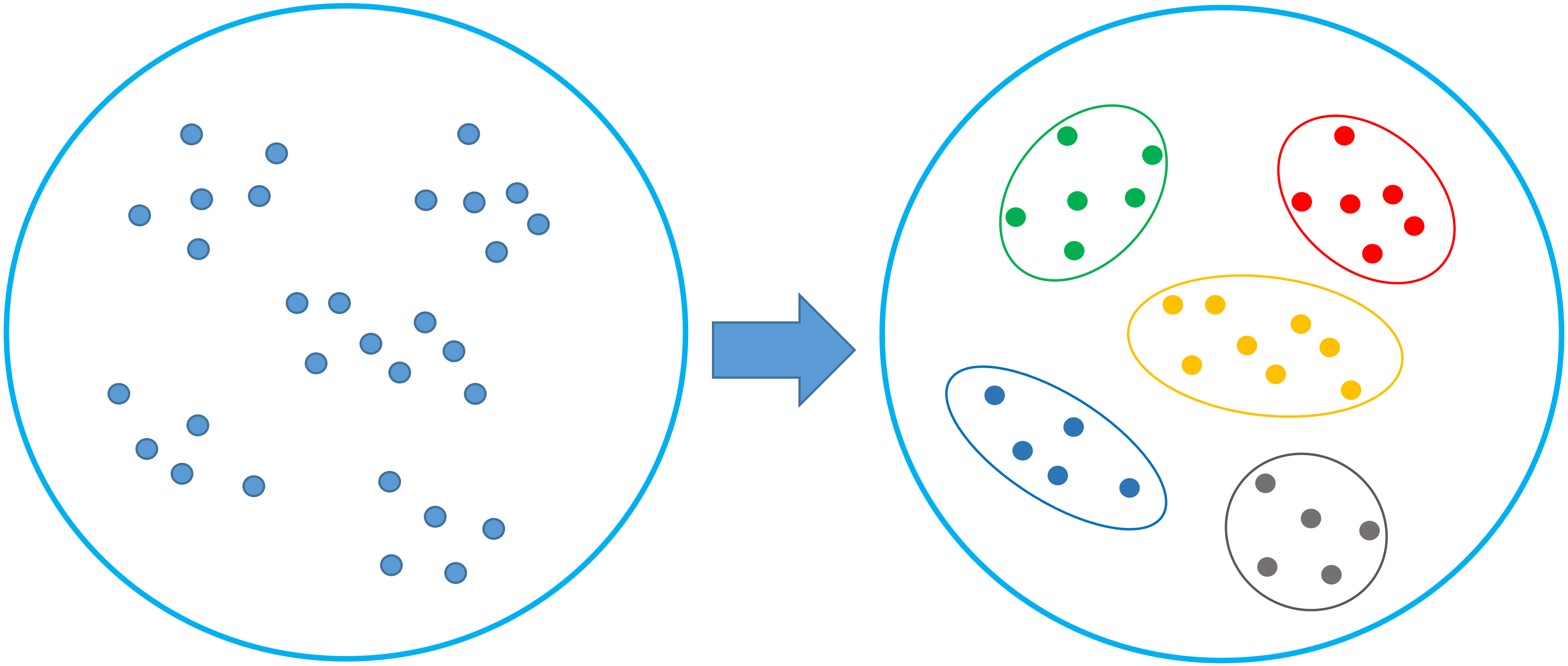
Smooth approximations to arbitrarily-shaped sample distributions.

A *hybrid* model between the unimodal Gaussian model and the VQ model.

Its components detail the multi-modal nature of the samples.

Stage 2: Statistical modeling

- Probabilistic models - *Gaussian mixture model* (GMM)

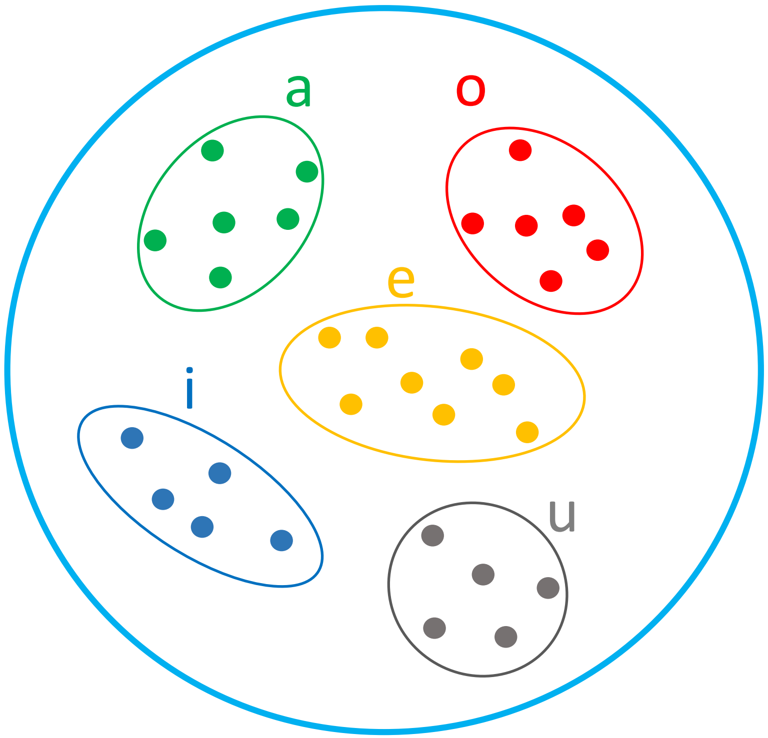


Acoustic space

Acoustic classes

Stage 2: Statistical modeling

- Probabilistic models - *Gaussian mixture model* (GMM)



Acoustic space

Underlying information

Each class represents a phonetic event

Mixture weight: the area size

Mean vector: the position

Covariance matrix: the elliptic shape

Stage 2: Statistical modeling

- Probabilistic models – GMM-UBM
  - Universal background model (UBM)
    - represents a speaker-independent model
    - reflects the common acoustic classes among humans
  - Maximum a posteriori (MAP): ***a linear transformation***
    - UBM --> speaker-specific GMM

$$\Pr(i \mid \boldsymbol{x}_t) = \frac{w_i p_i(\boldsymbol{x}_t)}{\sum_{j=1}^{M} w_j p_j(\boldsymbol{x}_t)}$$

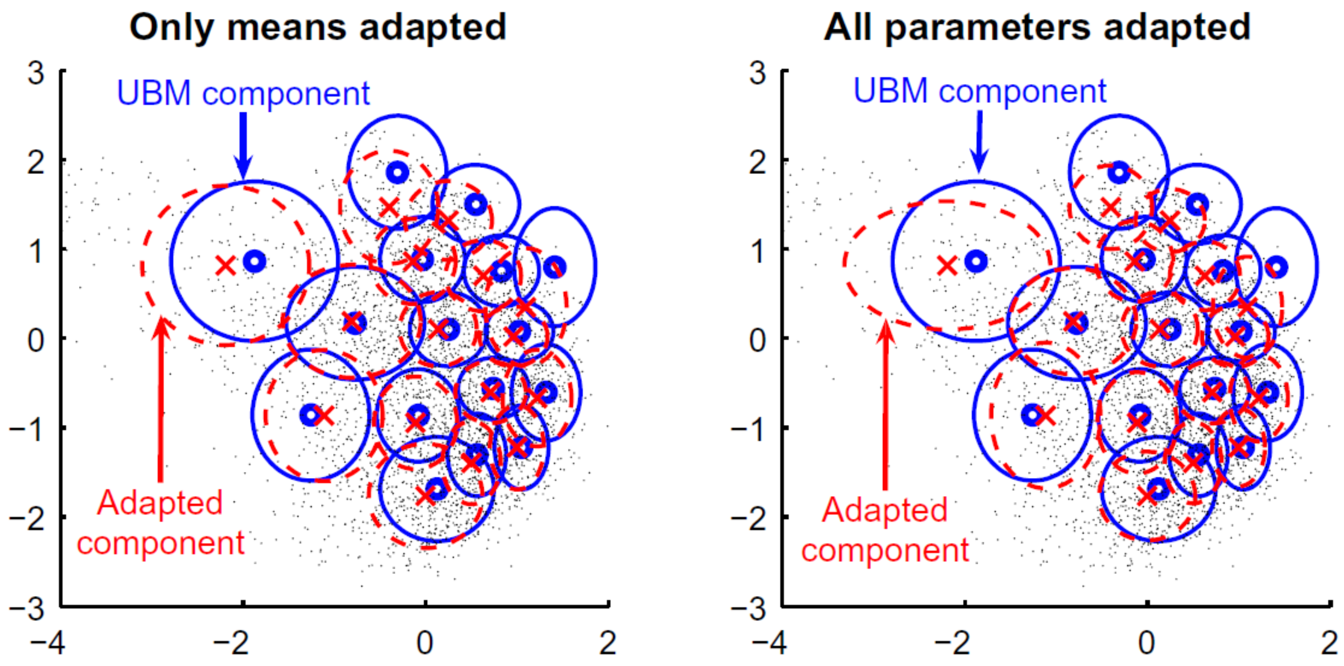
$$n_i = \sum_{t=1}^{T} \Pr(i \mid \boldsymbol{x}_t)$$

$$E_i(\boldsymbol{x}) = \frac{1}{n_i} \sum_{t=1}^{T} \Pr(i \mid \boldsymbol{x}_t) \boldsymbol{x}_t$$

$$\boxed{\begin{aligned} \hat{w}_i &= \left[\alpha_i^w n_i / T + (1 - \alpha_i^w) w_i\right] \gamma \\ \hat{\boldsymbol{\mu}}_i &= \alpha_i^m E_i(\boldsymbol{x}) + (1 - \alpha_i^m) \boldsymbol{\mu}_i \\ \hat{\boldsymbol{\sigma}}_i^2 &= \alpha_i^v E_i(\boldsymbol{x}^2) + (1 - \alpha_i^v)(\boldsymbol{\sigma}_i^2 + \boldsymbol{\mu}_i^2) - \hat{\boldsymbol{\mu}}_i^2 \end{aligned}}$$

Stage 2: Statistical modeling

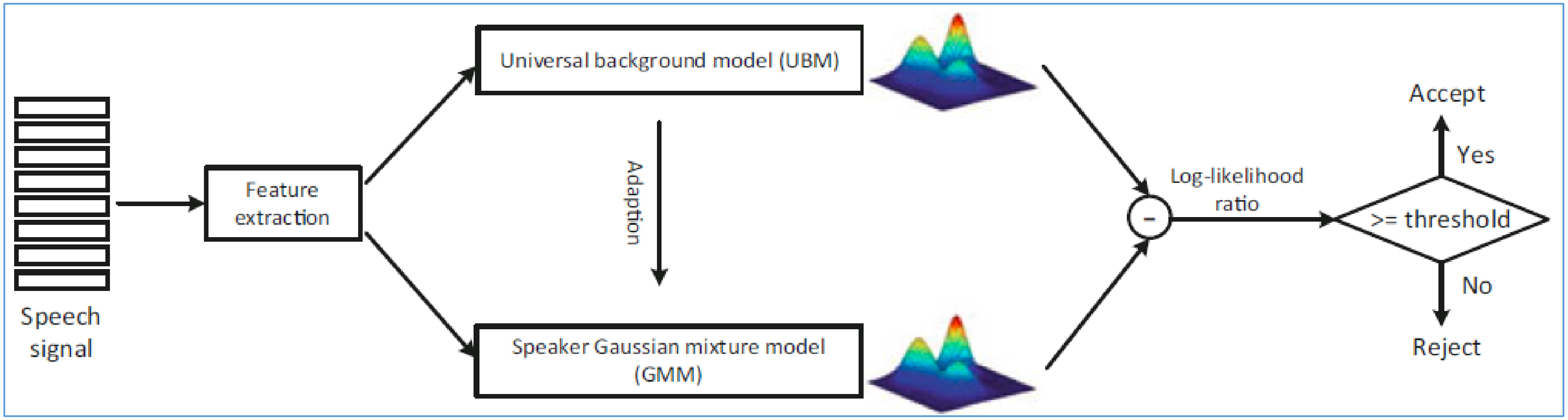
- Probabilistic models – GMM-UBM



Examples of GMM adaptation using *maximum a posteriori* (MAP) principle.
The solid ellipses and dashed ellipses are represented the UBM and speaker GMM.

Stage 2: Statistical modeling

- Probabilistic models – GMM-UBM
  - Recognition mode
    - The *log* likelihood ratio: $$\mathrm{LLR}_{\mathrm{avg}}(\mathscr{X}, \lambda_{\mathrm{target}}, \lambda_{\mathrm{UBM}}) = \frac{1}{T}\sum_{t=1}^{T}\{\log p(\boldsymbol{x}_t|\lambda_{\mathrm{target}}) - \log p(\boldsymbol{x}_t|\lambda_{\mathrm{UBM}})\},$$



The framework of a GMM-UBM system

Stage 2: Statistical modeling

- Probabilistic models – Factor analysis
  - JFA: $M = m + Vy + Ux + Dz$

Speaker supervector UBM Speaker space Session space Residual space

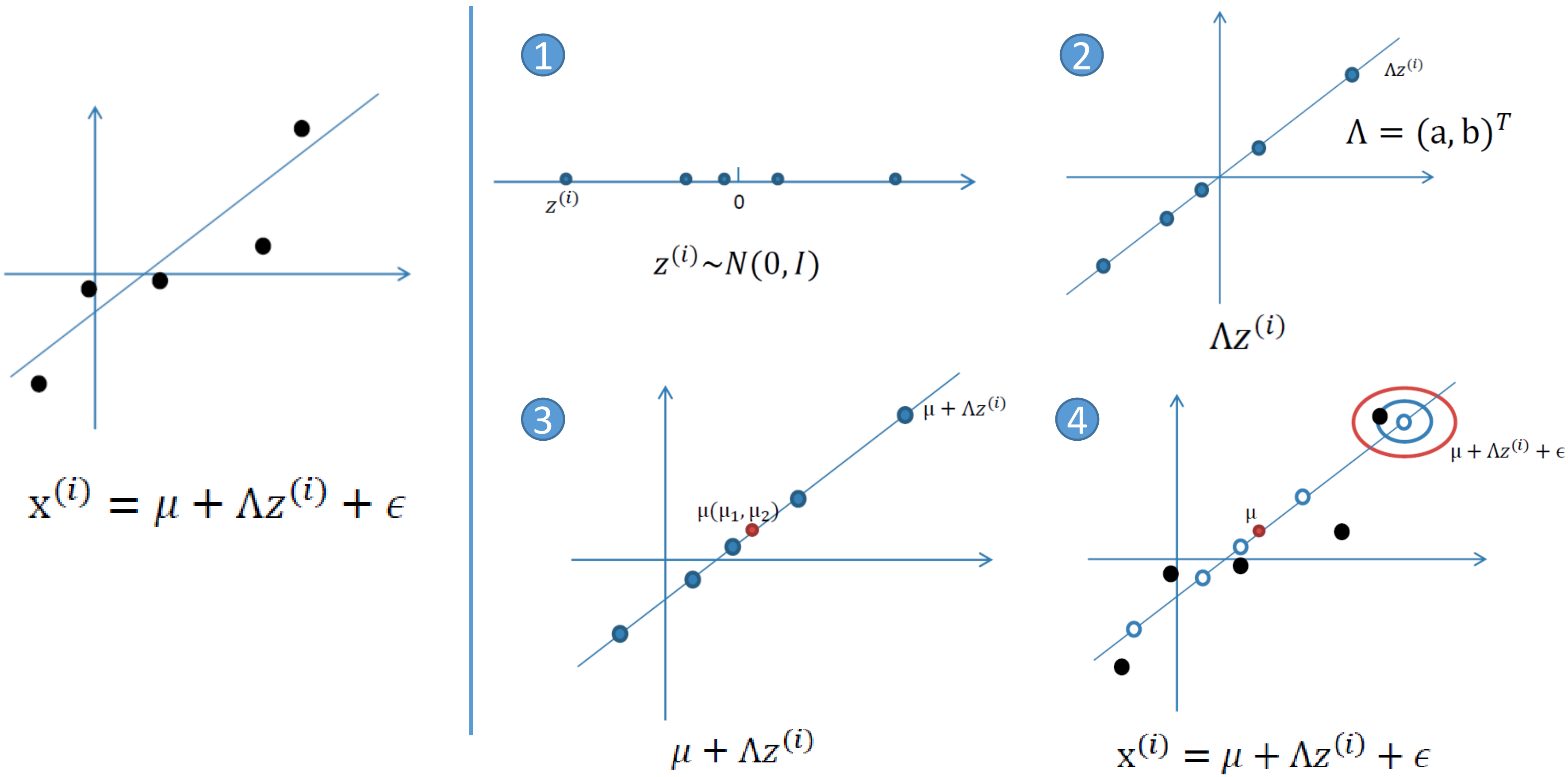
  - i-vector: $M = m + Tw,$

Total-variance space

    - Loading matrix T: EM procedure.
    - Total factor *w*: posterior probability $p(\mathbf{w}|\mathbf{X})$

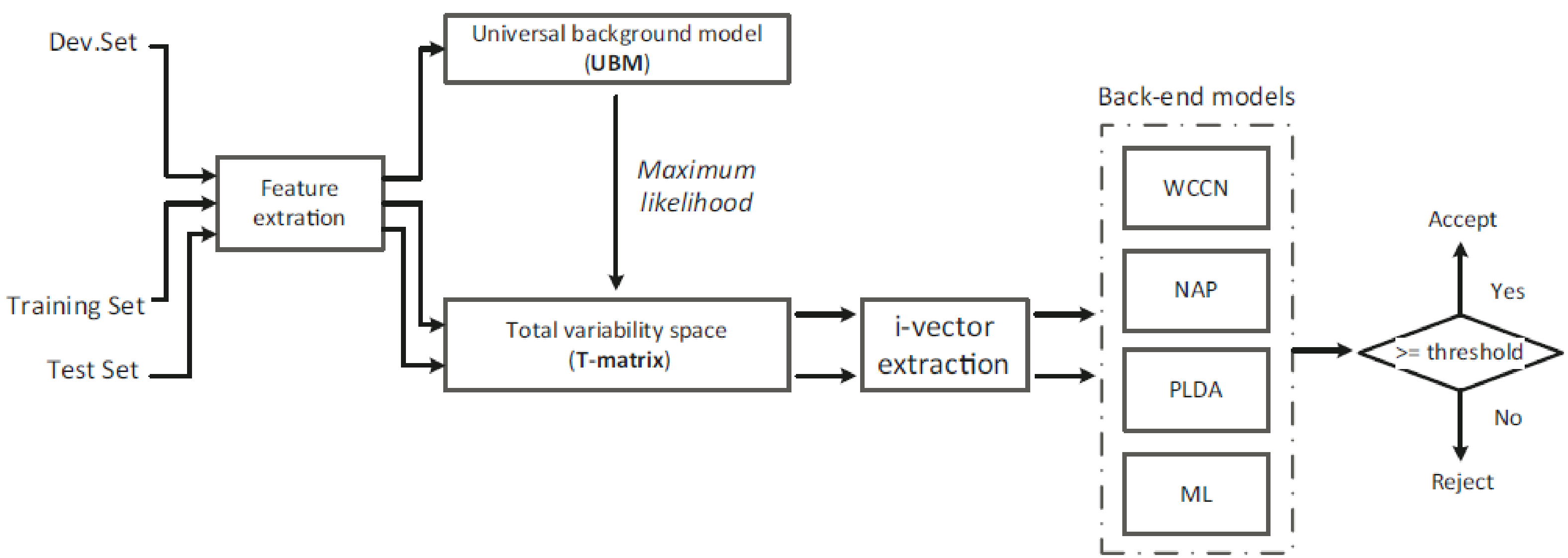
Stage 2: Statistical modeling

- Probabilistic models – Factor analysis



Stage 2: Statistical modeling

- Probabilistic models – Factor analysis
    - i-vector model



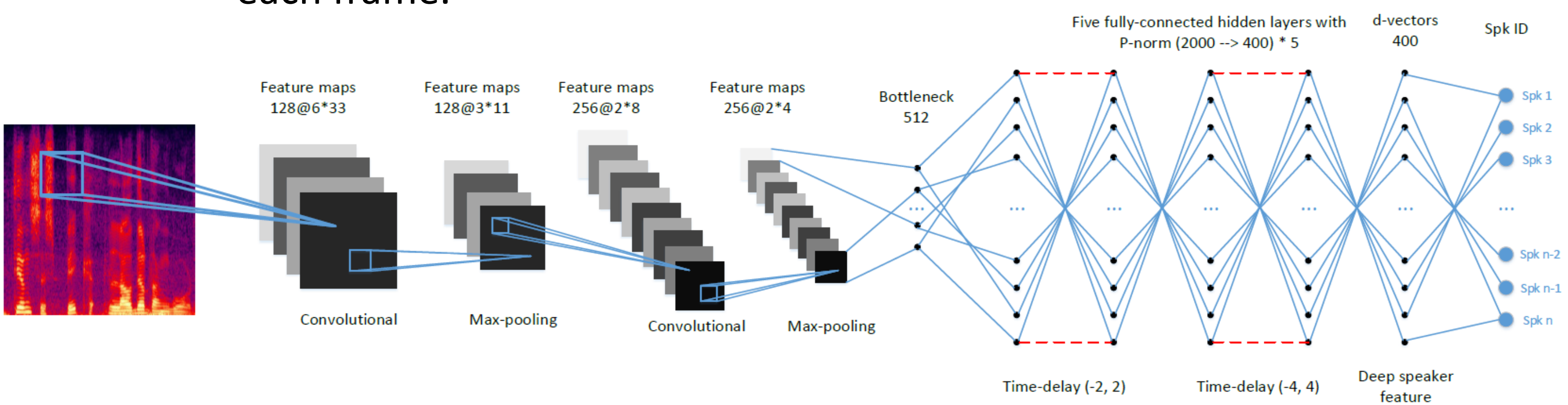
The framework of an i-vector system

Stage 3: Deep learning

- **Deep learning** is a branch of machine learning methods based on learning representations of data.
- The big *data* and the *BP* algorithms
- Two ***directions*** on deep speaker recognition
  - Deep feature learning
  - End-to-end learning

Stage 3: Deep learning

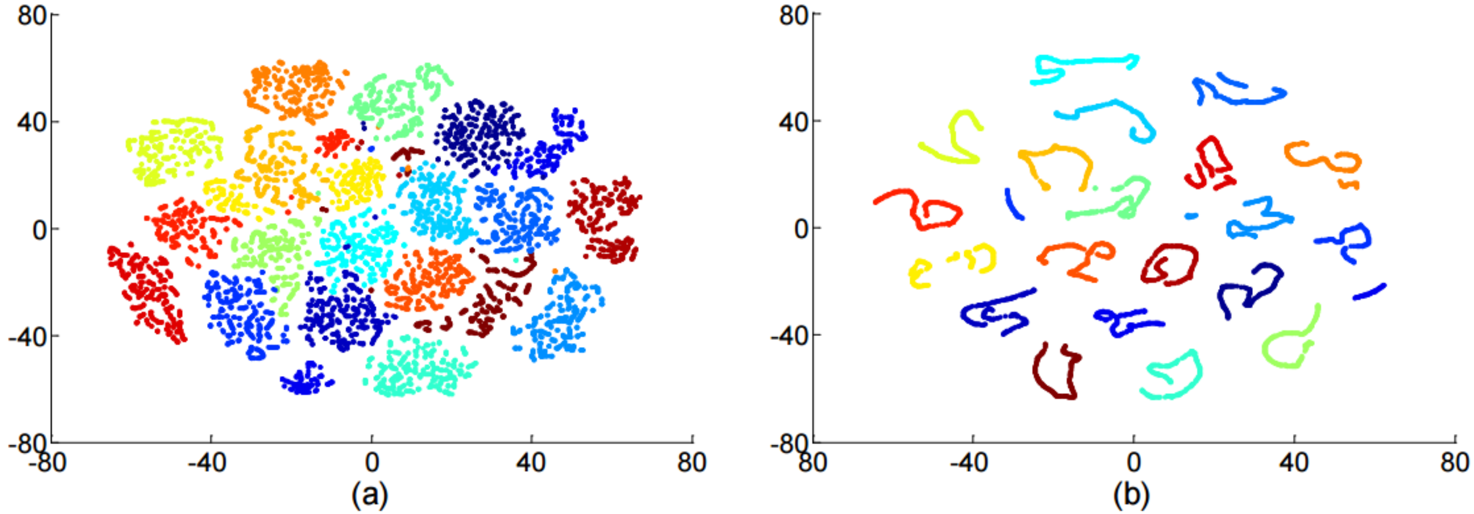
- Deep feature learning
  - Convolutional components: extract local discriminative patterns from the temporal-frequency space.
  - Time-delay components: increase the effective temporal context for each frame.



The CT-DNN structure used for deep speaker feature learning

Stage 3: Deep learning

- Deep feature learning

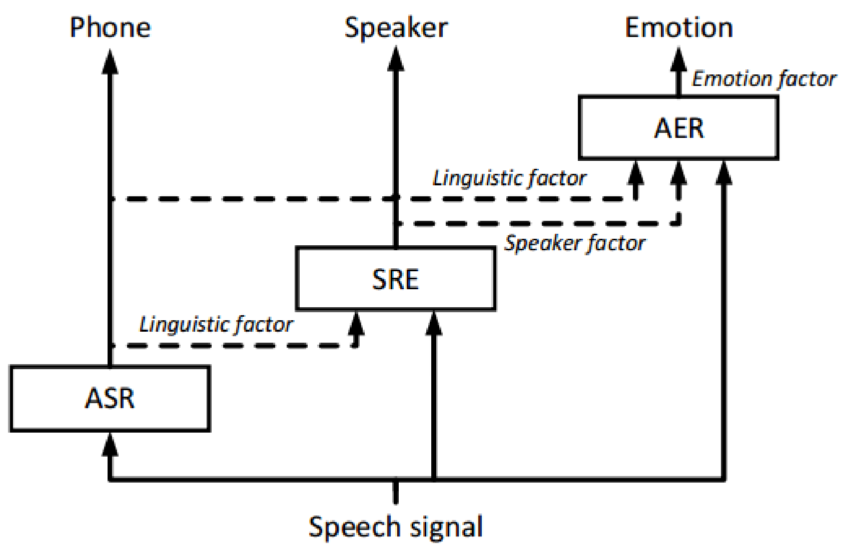


Deep features plotted by t-SNE. Each color represents a speaker.

- Short-time spectral patterns rather than long-term probabilistic patterns

Stage 3: Deep learning

- Deep speech factorization



The cascaded deep factorization approach

- Different from JFA, it is deep, non-linear and non-Gaussian.

Stage 3: Deep learning

- Spectrum reconstruction

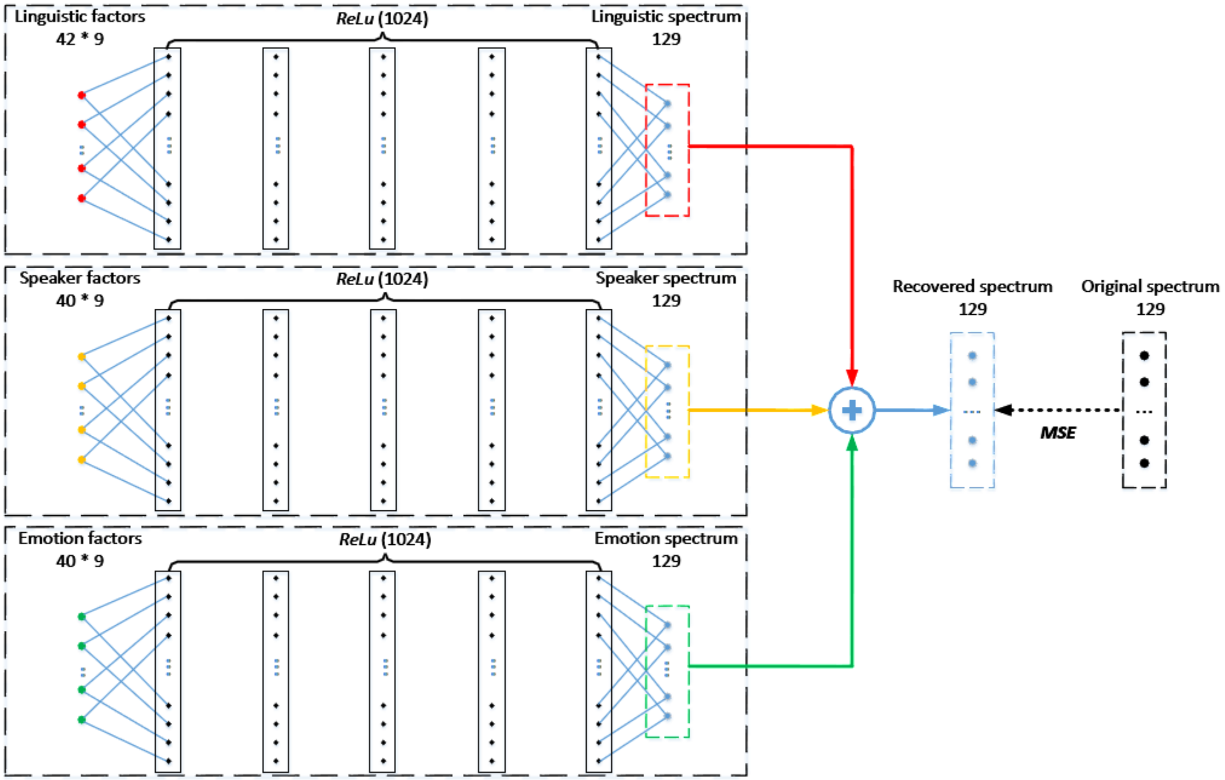
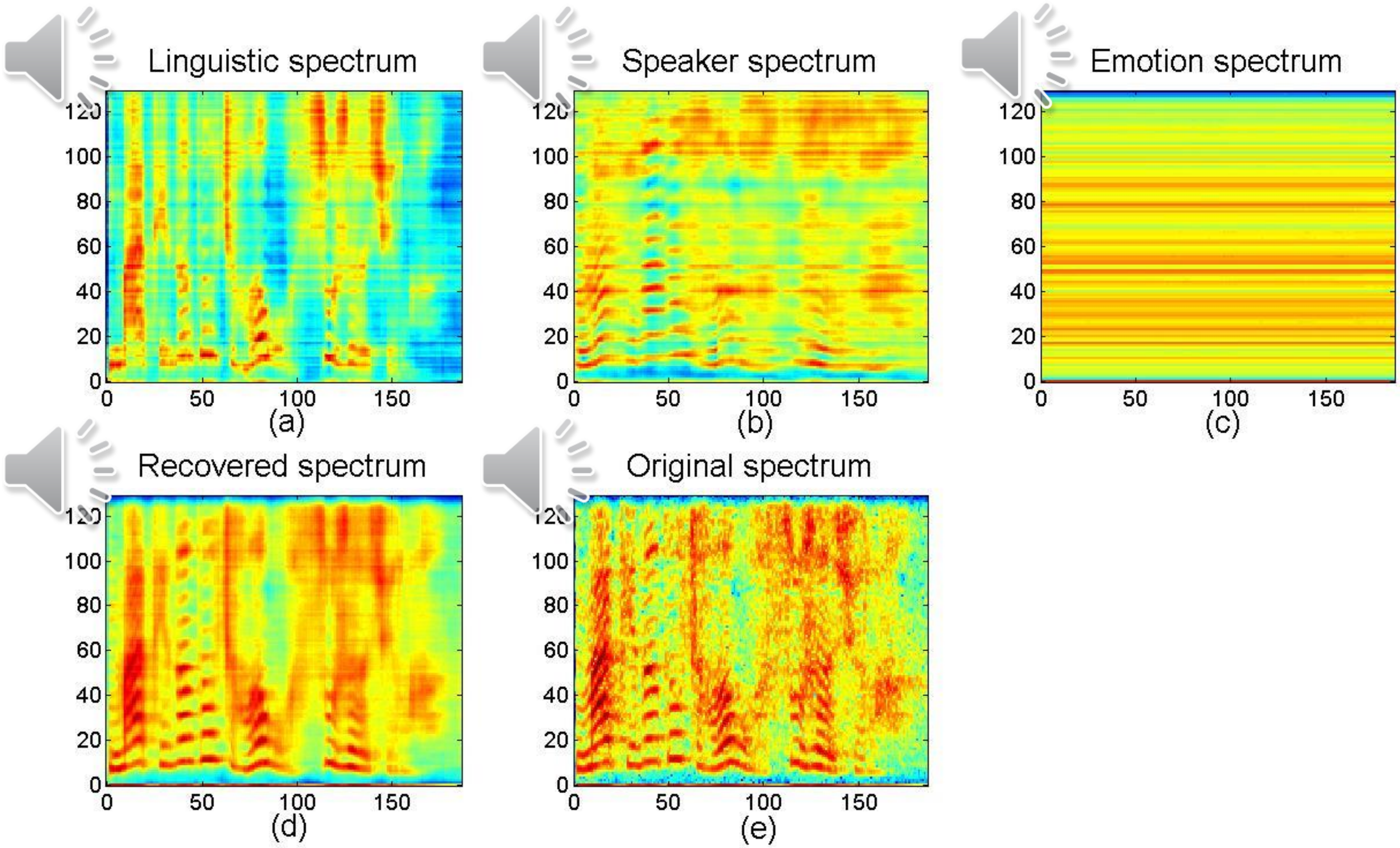


Figure 3: The architecture for spectrum reconstruction.

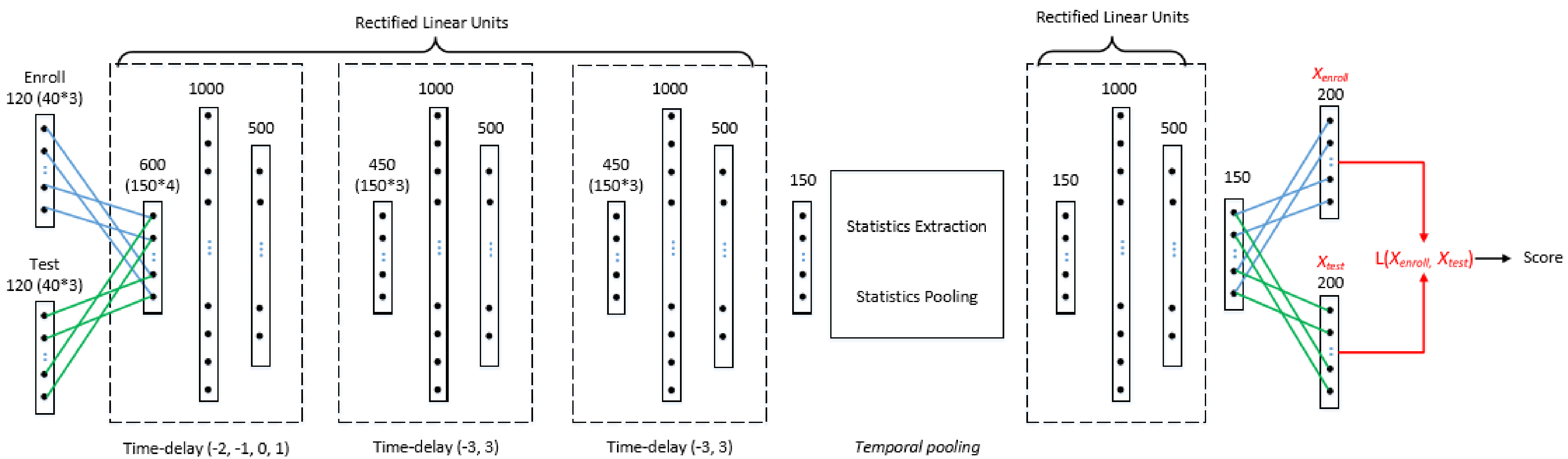
Stage 3: Deep learning

- Spectrum reconstruction



Stage 3: Deep learning

- End-to-end learning
    - A whole black box



The DNN structure of the end-to-end learning system

Stage 3: Deep learning

- Deep feature learning and end-to-end learning
  - Different in model structure
    - Front-end and back-end
    - End-to-end
  - Different in training objectives
    - Speaker identification
    - Speaker verification
  - Different in training scheme
    - one-hot style
    - pair-wised style

Comparison and combination

- Comparison the EER results of three system performances.

EER(%) RESULTS OF THE THREE SV SYSTEMS.

		EER%	
Systems	Scoring	C(4-4)	C(40-4)
i-vector	Cosine	16.96	4.81
	LDA	10.95	3.30
	PLDA	8.84	3.39
Deep feature	Cosine	10.31	4.01
	LDA	**7.86**	**2.39**
	PLDA	13.01	5.24
End-to-end	-	9.85	4.59

Comparison and combination

- GMM i-vector and DNN i-vector

(a) (b)

- Frame posteriors are computed from ***GMM***.
- Each Gaussian component represents a region / class for stats computation.
- ***Unsupervised*** clustering.

- Phone posteriors are derived from ***DNN***.
- Each sub-phonetic categories (senones) of DNN represents a region / class.
- ***Supervised*** discriminative training.

Diagram of the (a) GMM- and (b) DNN-based statistics computation

Comparison and combination

- Joint Training for speech and speaker recognition

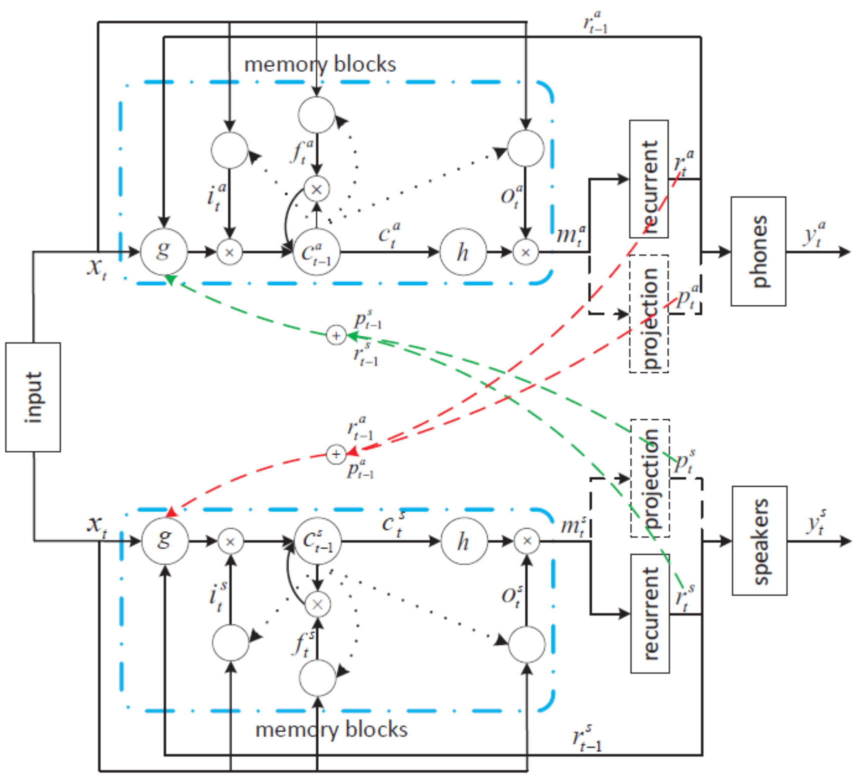
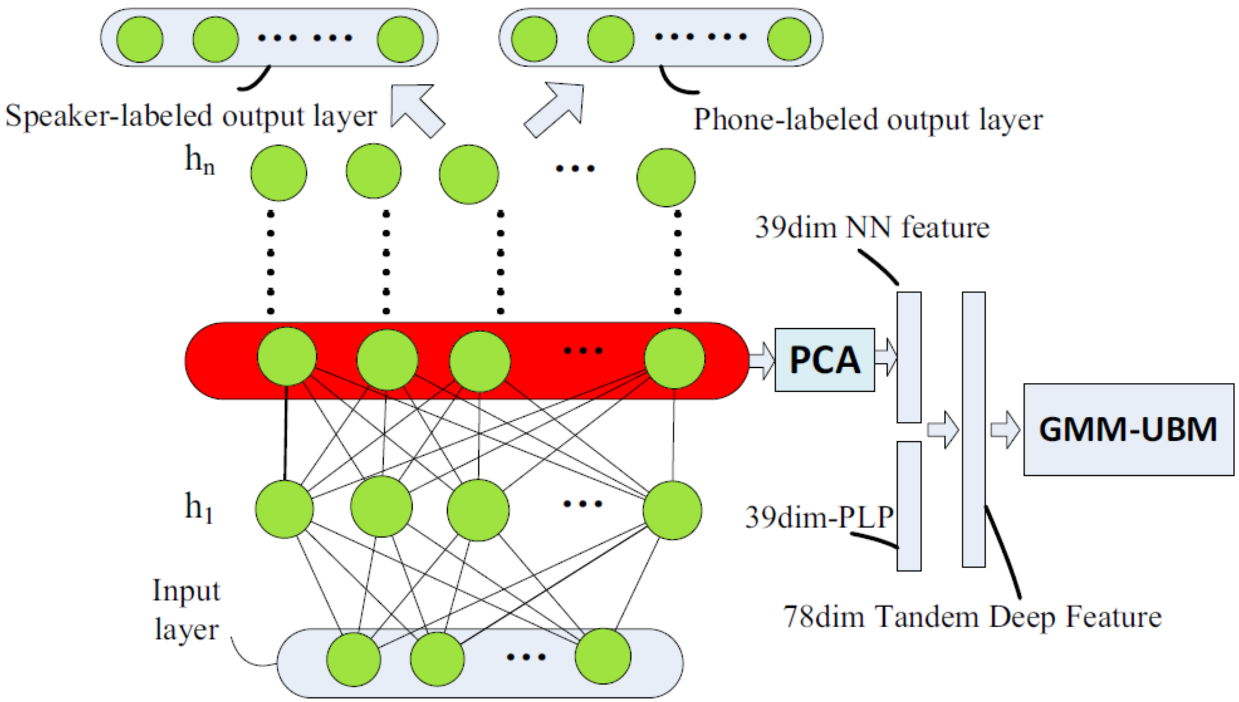


Table 3: *Joint training results.*

Feedback Info.		Feedback Input				ASR WER%	SRE EER%
r	p	i	f	o	g		
						7.41	1.84
√		√				7.05	0.62
√	√	√				6.97	0.64
√			√			7.12	0.66
√	√		√			7.24	0.65
√				√		7.26	0.65
√	√			√		7.28	0.59
√					√	7.11	0.62
√	√				√	7.11	0.67
√		√	√	√		7.06	0.66
√	√	√	√	√		7.23	0.71
√		√	√	√	√	7.05	0.55
√	√	√	√	√	√	7.23	0.62

Comparison and combination

- Deep speaker feature and statistical model



Conclusions and Future work

- Long history
- Combination of the past and the present
- Questions:
  - The fundamental feature of speaker traits
  - A powerful speaker model

Thank you

http://lilt.cslt.org/