Speech Classification

Lecture 5

A practical approach to feature extraction, speech modelling and common speech classification tasks

Alberto Abad

alberto.abad@tecnico.ulisboa.pt

Spoken Language Processing - **IST**, Spring 2024

Outline

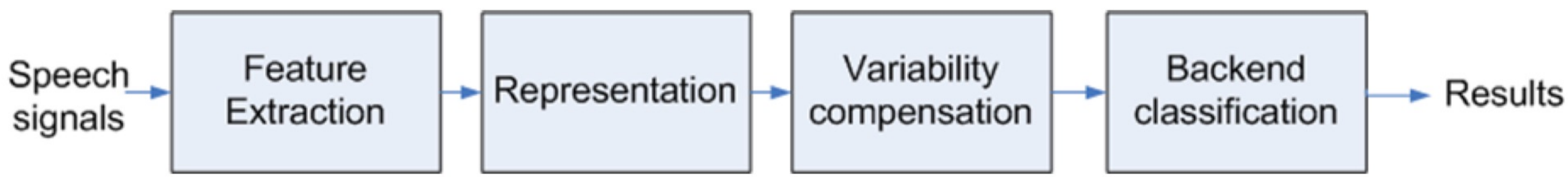
- Introduction to speech classification **[Lecture 4]**
- Feature Extraction **[Lecture 4]**
  - Type of features
  - Additional processing
  - Tools
- Modeling speech **[Lecture 4]**
  - Speech common models
  - Tools
- Case of study: Speaker Recognition **[Lecture 5]**
- Other speech classification task examples **[Lecture 5]**
- Lab assignment 2: Native Language Identification **[Lecture 5]**

PART III

CASE OF STUDY: SPEAKER RECOGNITION

Speaker recognition (SR)

- The typical processing pipeline of speaker recognition is very similar to any sequence-to-one class speech problem.



- Progress in this field has permitted achieving impressive results in certain tasks (super-human)
- Some of the key advancements are related with the development of methods to represent speaker information in a very compact way → speaker embeddings
- Other paralinguistic tasks have greatly benefit of the advancements in speaker recognition: language/dialect recognition, emotion, etc.

Speaker Recognition: Voice biometrics

- Biometric authentication paradigm:

What you possess/know
Token, password, etc.

→

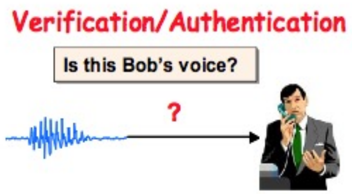
What you are (physiological)
DNA, fingerprints, iris, …

What you produce (behavioural)
Voice, signature, …

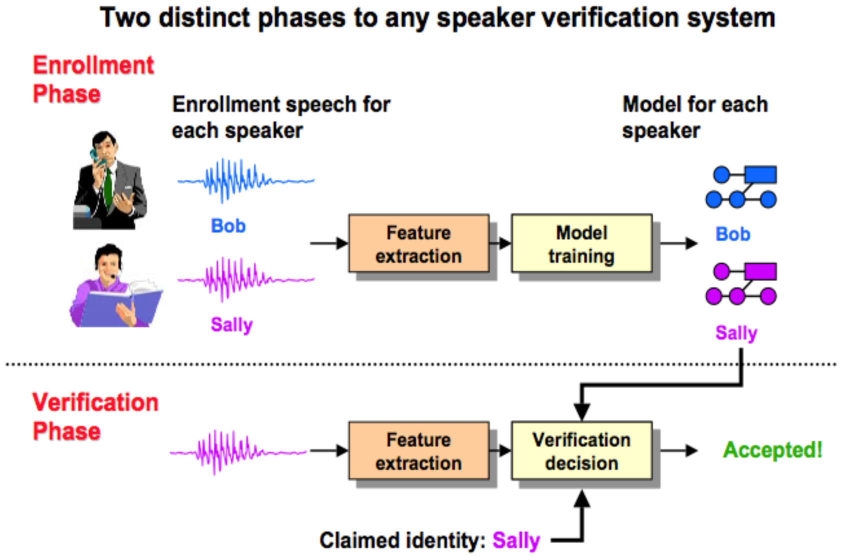
- Speech/voice is one form of biometric that carries lots of personal (identity) information:
  - Gender, age, accent, region, social class, illnesses (cold), style of speaking, mood, etc.
- Some advantages/particularities of voice:
  - It allows for remote authentication; Non intrusiveness; Low cost and wide availability; Ease of transmission, small storage space
- Caution:
  - Wrong finger-print idea, uniqueness.

Speaker recognition (SR) tasks

- Verification vs Identification



- Text-dependent vs text-independent
- Enrolment and test phases



SR evaluation measures

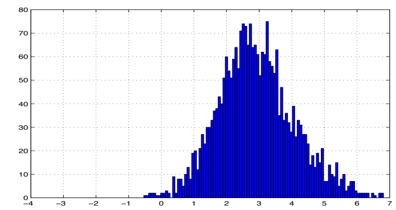
Trial definition

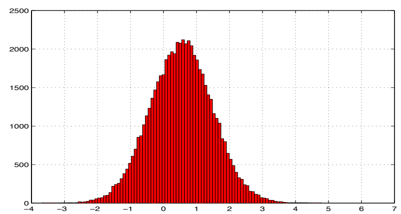
- Speaker verification tasks usually consist of a set of verification trials.
- **Test trials**: given a test segment, determine whether a given speaker is actually speaking
  - Target trials → The speaker is speaking in the test segment
  - Non-target/Impostor trials → The speaker is NOT speaking in the test segment
- Each trial (usually) requires two outputs:
  - Actual decision → True/false
  - Likelihood score → Confidence in decision

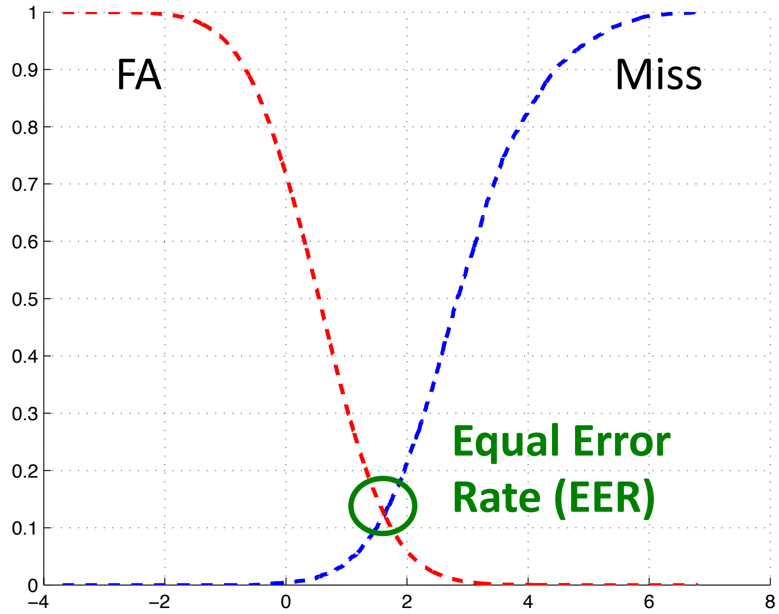
SR evaluation measures

Decision errors

- Two types of actual decision errors:
  - Missed detections ($P_{miss|target}$): Percentage of target trials rejected incorrectly
  - False Alarms ($P_{fa|impostor}$): Percentage of impostor trials accepted incorrectly



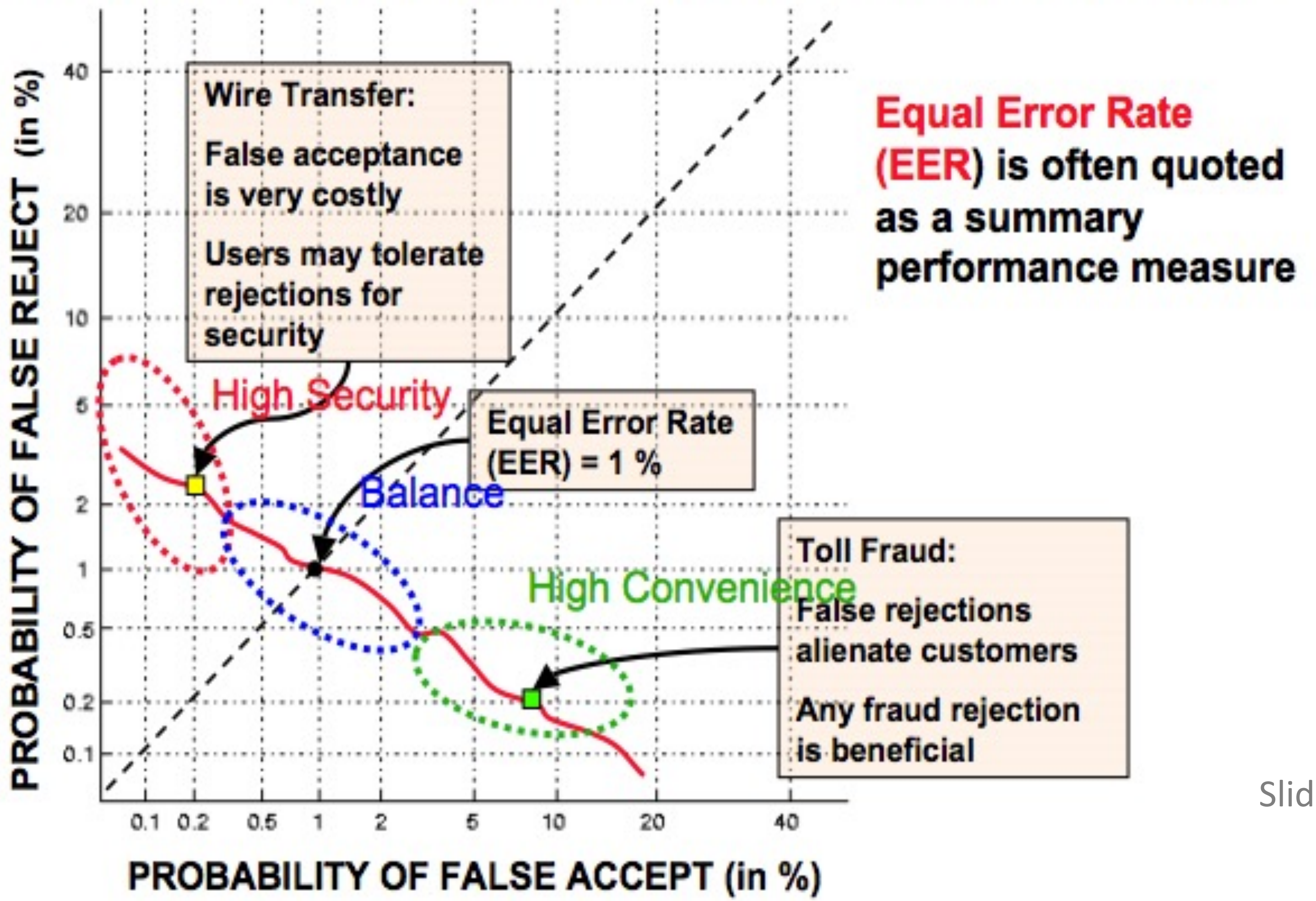




SR evaluation measures

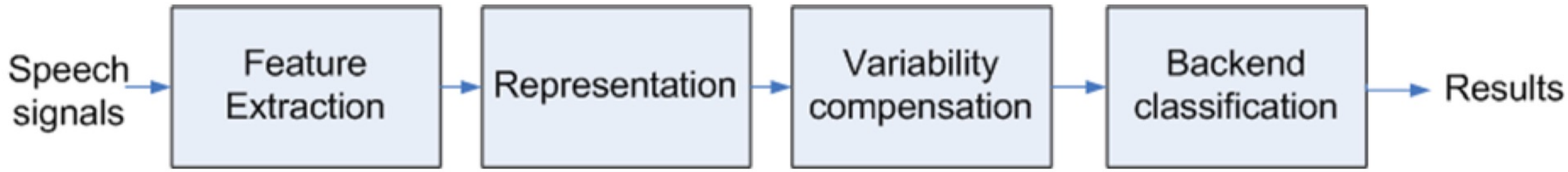
DET curve

- DET plots P_{miss} vs P_{FA} for every threshold (like ROC curves):
  - Axis follow normal distribution scale



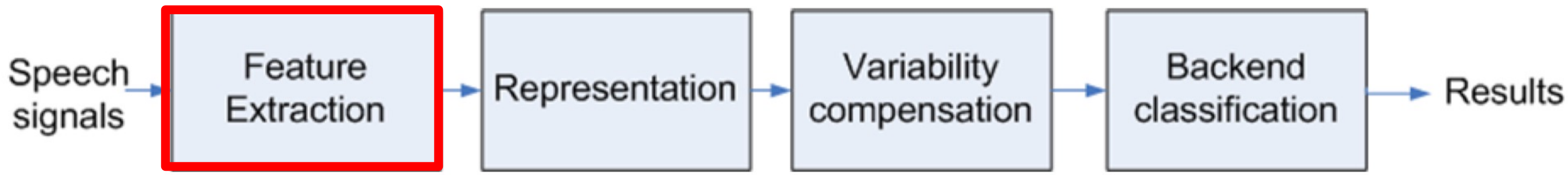
Slide after [1]

What are some of the key elements that contributed to SR advance?



- Significant contributions come from advancement in the four stages of the pipeline.
- Today we'll focus on how the modelling approaches evolved during the last years:
  - Focus on the two first stages; the latter are quite specific of speaker verification problems.

Speaker Recognition: Features



- Desirable attributes of features for automatic methods:
  - **Practical**
    - Occur naturally and frequently in speech
    - Easy to measure
  - **Robust**
    - Not change over time or affected by speakers' health
    - Not (very) affected by noise and channel
  - **Secure**
    - Not be subject to mimicry
- In practice,
  - No feature has all these attributes
  - Features derived from spectrum speech are the most successful

Speaker Recognition: Features

- Other typical features in speaker (or similar) tasks:
  - LPC, PLP, RASTA, SDC, etc.

P. Torres-Carrasquillo et al. “Approaches to language identification using gaussian mixture models and shifted delta cepstral features”. In: Proc. of ICSLP. 2002, pp. 89–92.

Speaker Recognition: Features

- Other typical features in speaker (or similar) tasks:
  - LPC, PLP, RASTA, SDC, etc.
  - NNET-based: bottleneck, tandem, PLLR/posteriors

Pavel Matejka et al. “Neural Network Bottleneck Features for Language Identification.” In: Proc. of Odyssey. 2014.
Ming Li and Wenbo Liu. “Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features”. In: Proc. of Interspeech. 2014.
Alberto Abad et al. “Exploiting Phone Log-Likelihood Ratio Features for the Detection of the Native Language of Non-Native English Speakers”, In: Proc. of Interspeech 2016

Speaker Recognition: Features

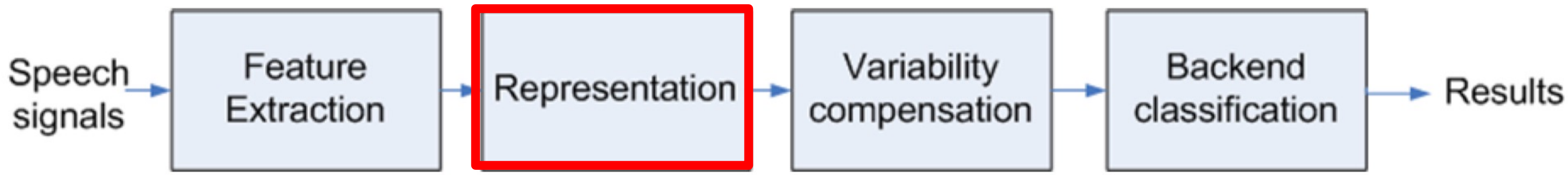
- Other typical features in speaker (or similar) tasks:
  - LPC, PLP, RASTA, SDC, etc.
  - NNET-based: bottleneck, tandem, PLLR/posteriors
  - CQCC, Modified Group Delay, etc.

Massimiliano Todisco, Hector Delgado, and Nicholas Evans. "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification". In: Computer Speech and Language 45 (2017).

Zhizheng Wu, Eng Siong Chng, and Haizhou Li. "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition". In: Proc. of Interspeech. 2012.

Maria Joana Correia, Alberto Abad, and Isabel Trancoso. "Exploiting magnitude and phase spectral information for converted speech detection". In: Proc. SLT 2014.

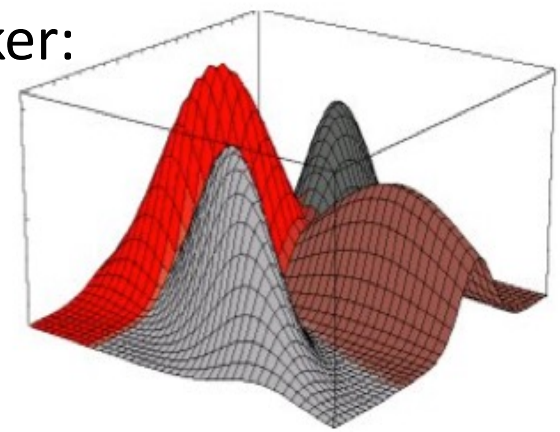
Speaker Recognition: Models



- **Speaker models** are used to represent the specific-speaker information in the feature vectors
- Several **different** modelling techniques have been applied:
  - Template matching (DTW for text-dependent)
  - Nearest neighbour
  - Neural networks
  - Hidden Markov Models
    - Single state HMM → **GMM**
  - Support vector machines
- Models provide some sort of score, reliability measure or **likelihood** for the target speakers

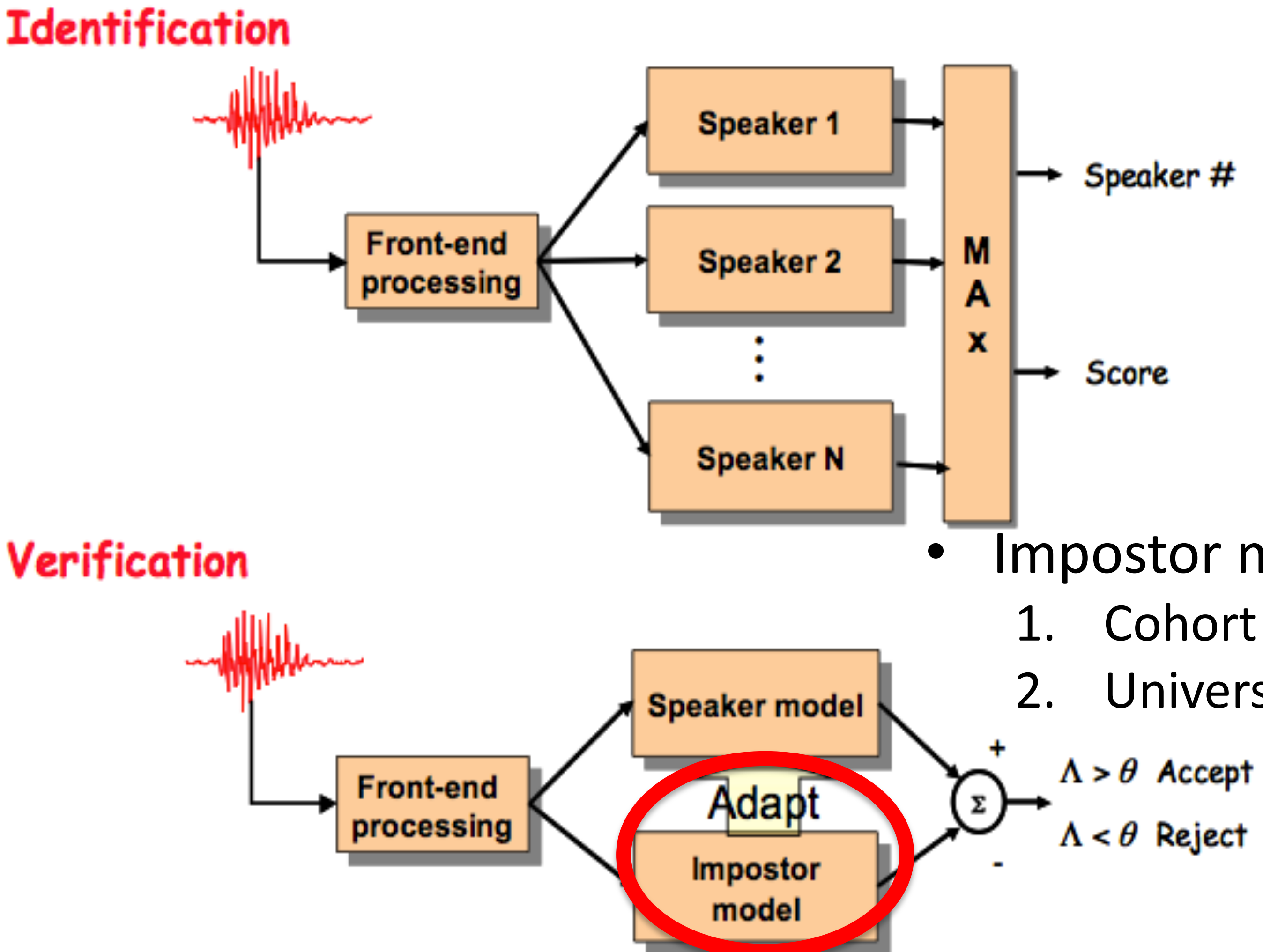
Gaussian mixture models (GMM)
GMM-ML

- Conventional **GMM-ML** approach:
  - Use cepstral features as front-end
  - In **train** phase:
    - Train a GMM model per target speaker:
      - Apply EM algorithm for ML estimation
  - In **test** phase:
    - Compute log-likelihoods for scoring:
      - Speaker ID → MAX(LL)
      - Speaker Verification → log-likelihood compared to a threshold or impostor model



Gaussian mixture models (GMM)

GMM-ML



- Impostor model approaches:
  1. Cohort of impostors
  2. Universal model

Slide after [1]

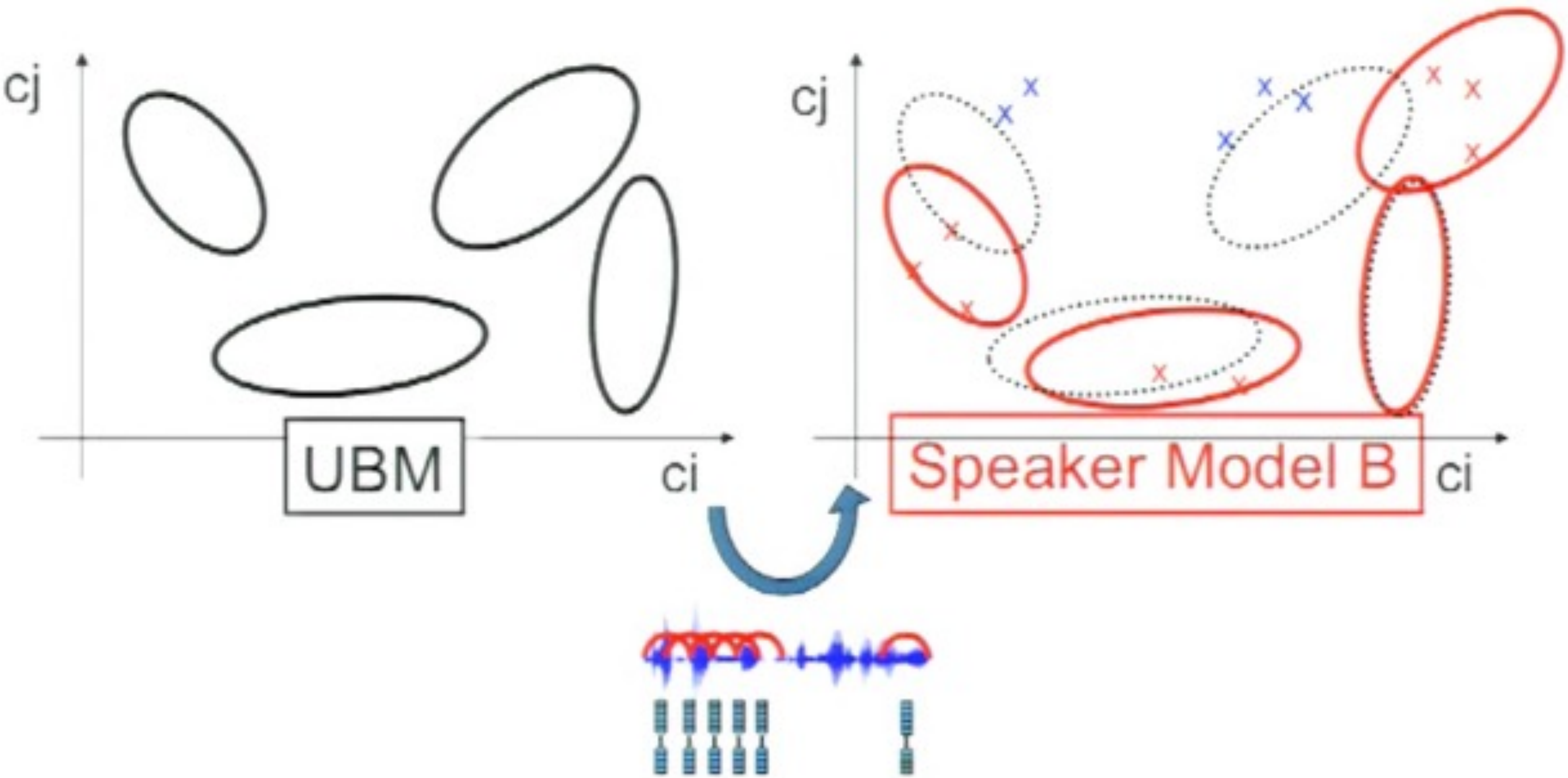
Gaussian mixture models (GMM)
GMM-UBM

- **GMM-UBM** approach:
  - Use cepstral features as feature extraction
  - In **train** phase:
    - Estimate the parameters of an UBM (Universal Background Model) with data from different speakers, channels, noise conditions, etc…
    - Adapt the UBM to each one of the target speakers:
      - Use MAP adaptation (usually only-means)
  - In **test** phase is like in previous GMM-ML approach.
  - **Advantages**
    - Needs less data,
    - permits updating only seen events,
    - keeps correspondence between means, allows fast scoring (top-M)

Douglas A. Reynolds, Thomas F. Quatieri, Robert B. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", In Proc: Digital Signal Processing 10(1-3): 19-41, 2000

Gaussian mixture models (GMM)

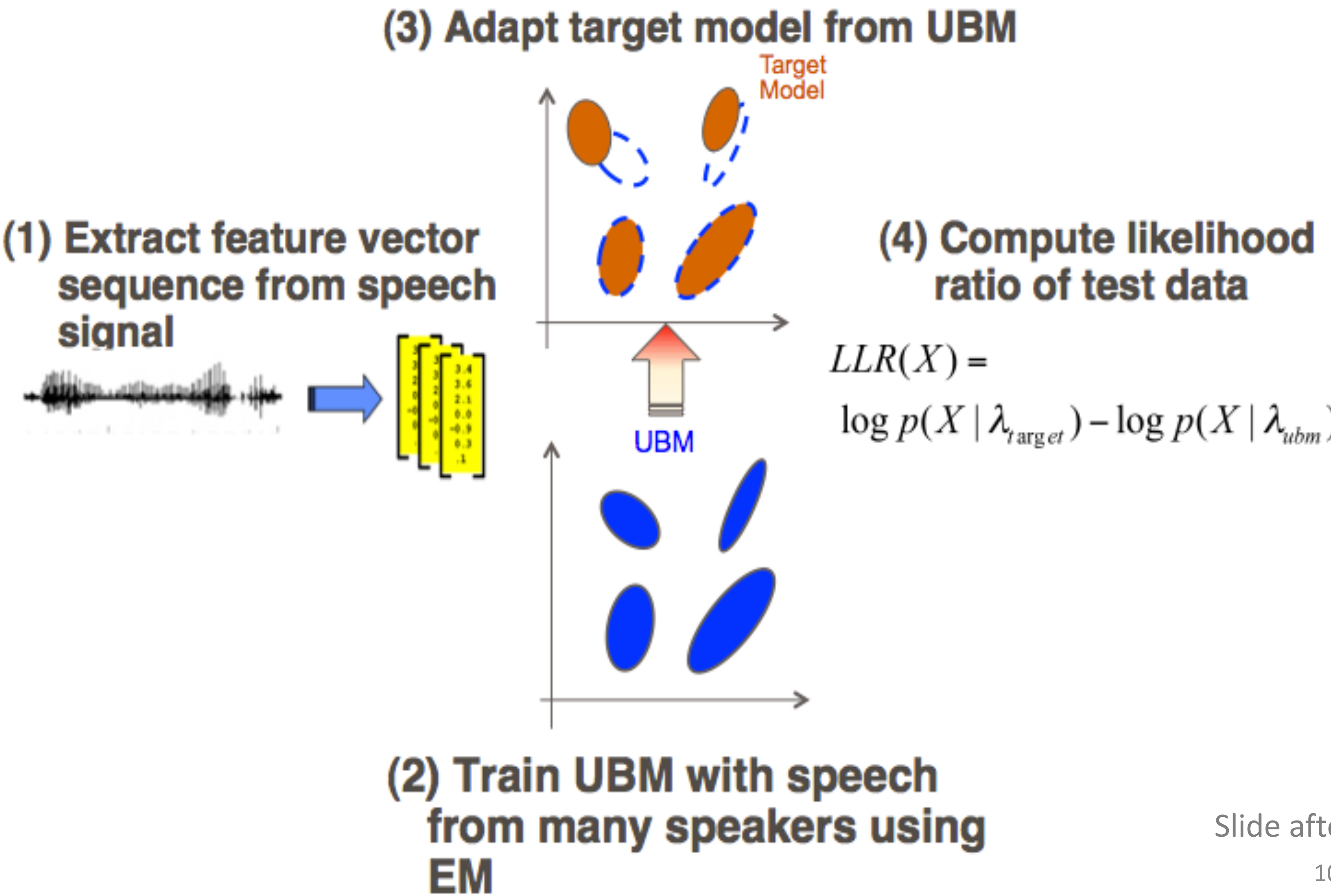
GMM-UBM



Slide after [3]

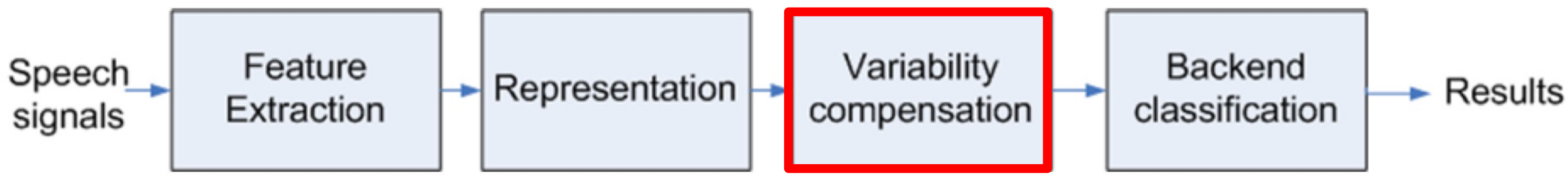
Gaussian mixture models (GMM)

GMM-UBM



Slide after [1]

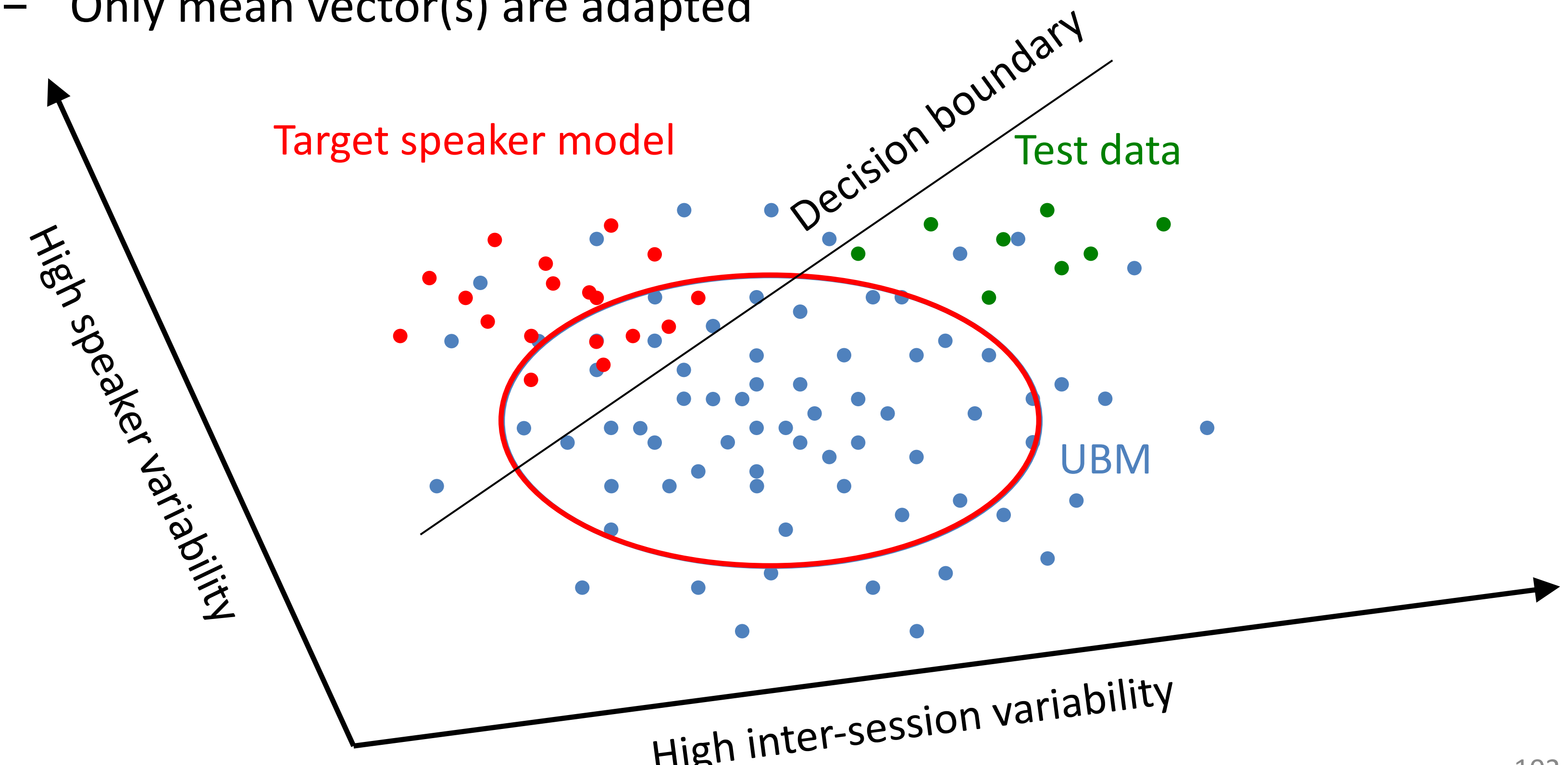
Robustness to channel missmatch



- Variability refers to changes in channel effects (and other) between training and successive detection attempts
- Session variability encompasses several factors
  - The microphones
    - Carbon-button, electret, hands-free, array, etc
  - The acoustic environment
    - Office, car, airport, etc.
  - The transmission channel
    - Landline, cellular, VoIP, etc.
  - The differences in speaker voice
    - Aging, mood, spoken language, etc.

Robustness to channel missmatch

- Relevance MAP adaptation example (GMM-UBM):
  - 2D features
  - Single Gaussian model
  - Only mean vector(s) are adapted

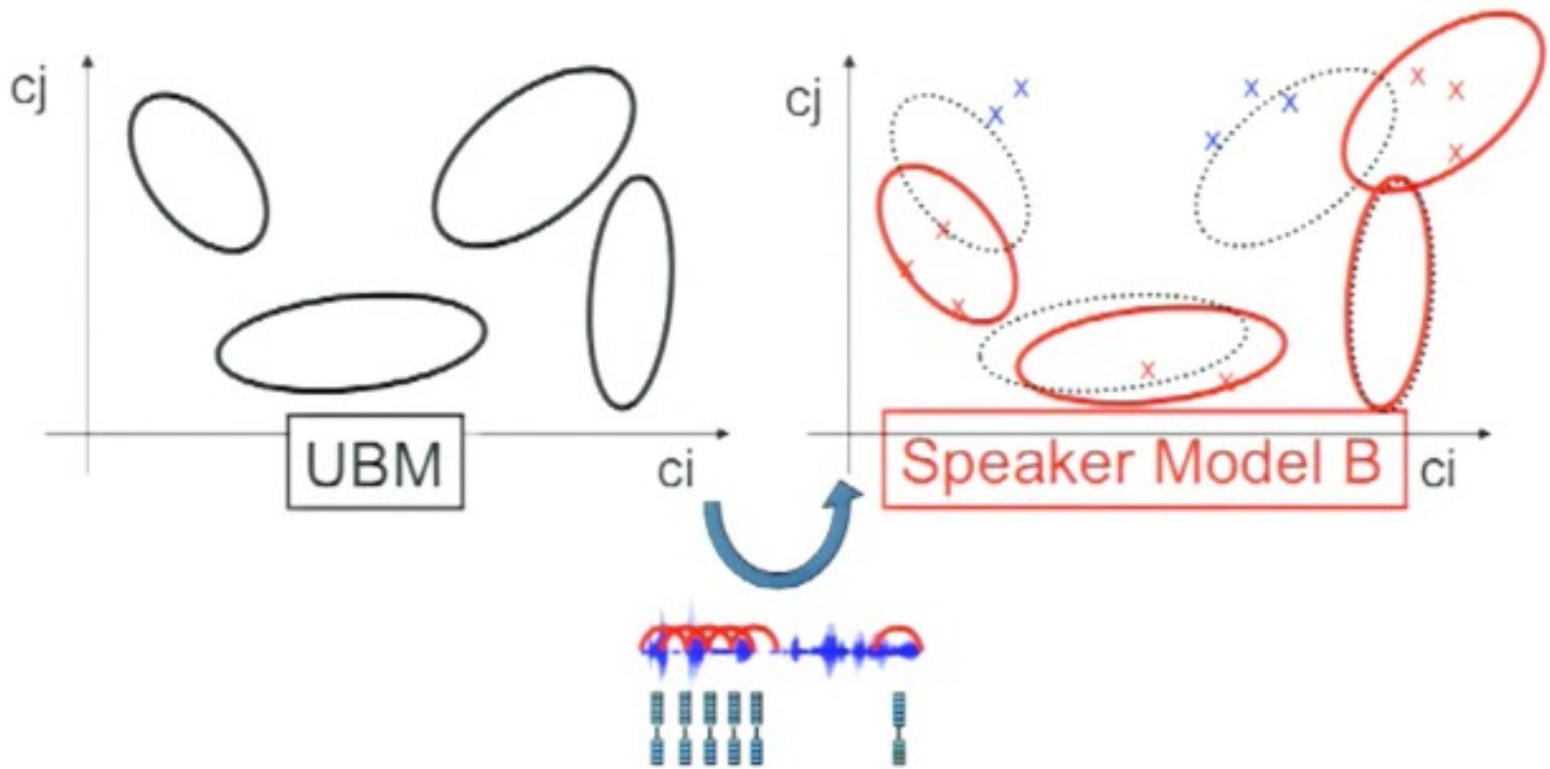


Robustness to channel missmatch

- The largest challenge to practical use of speaker recognition systems is channel/session variability
- Most of the research during the last decade focused on developing more robust systems to session variability:
  - Feature level
    - Normalization, robust speech enhancement, alternative features (high-level)
  - Model level
    - More robust models (GMM-SVM), compensation at high dimensional space (NAP), factor analysis and explicit channel modeling
  - Back-end/Score level
    - Score normalization (T-norm, Z-norm, etc.), calibration, fusion, etc.

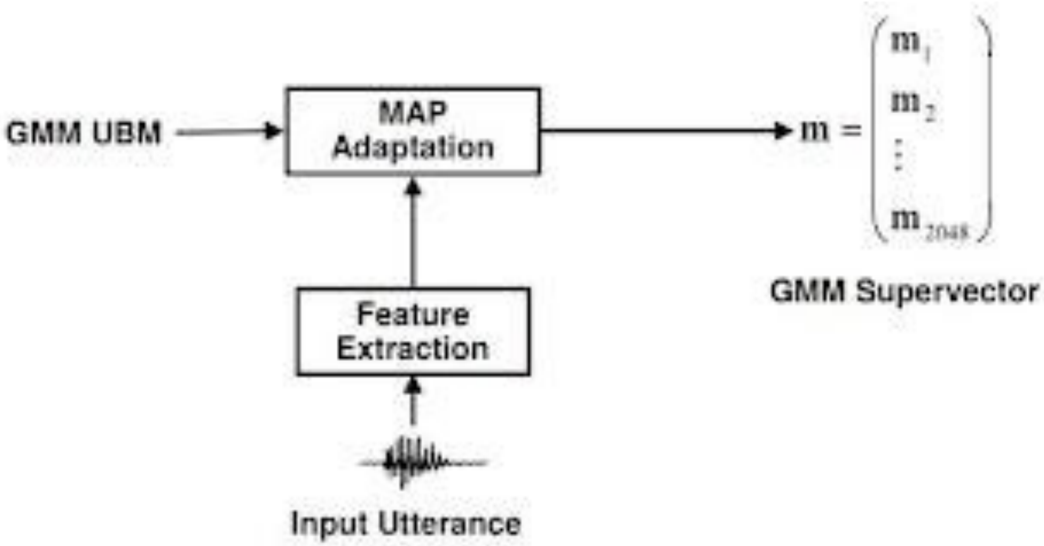
Improved modelling approaches

GMM-UBM: The supervector concept



Improved modelling approaches

GMM-UBM: The supervector concept



Typical dimensionality:

- M: number of components (512 -2048)
- F: feature dimensions (20-60)
- MF: ~20k-50k

MAP

$$\mathbf{m} = \mathbf{m}_{UBM} + \mathbf{D}\mathbf{z}_{sh}$$

D = Full rank diagonal matrix (relevance MAP)

$\mathbf{z}_{sh}$ = Full rank vector

- The supervector concept and its derivations had a **huge impact** in past decade:

1. As a kind of feature extraction for discriminative machine learning methods → GMM-SVM
2. As a tool for Factor Analysis derivation and session variability explicit modelling → JFA & i-vectors

Improved modelling approaches

Factor Analysis approaches: The i-vector

GMM-UBM (MAP) → $\mathbf{m = m_{UBM} + Dz_{sh}}$

- **D** diagonal full-rank
- $\mathbf{z_{sh}}$: speaker (and more) component

i-vectors → $\mathbf{m = m_{UBM} + Tw}$

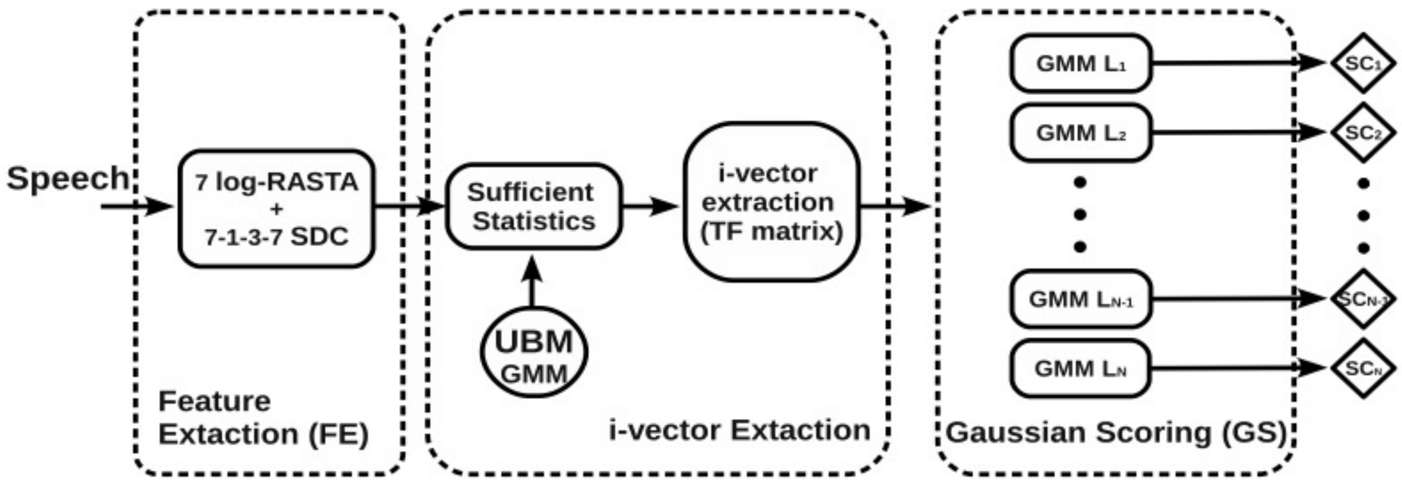
- **T** total variability subspace (low-rank)
- **w** variability (loading) factors, a.k.a i-vectors
  - ~400-600 dimensions
  - They contain all speaker and channel variability
  - It is used as a low-dimensional representation (on top of them other models can be trained)

N. Dehak et al. "Front-end factor analysis for speaker verification". In: IEEE Transactions on Audio, Speech, and Language Processing 19.4 (2011), pp. 788–798.

N. Dehak et al. "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification". In Proc Interspeech 2009.

The i-vector as a generalized speaker (**speech)** embedding

- The success of the i-vector paradigm was expanded to other **similar** tasks:
  - Language, dialect and native language



David Martínez et al. “Language recognition in ivectors space”, In Proc. of Interspeech 2011
Alberto Abad et al. “Exploiting Phone Log-Likelihood Ratio Features for the Detection of the Native Language of Non-Native English Speakers”, In: Proc. of Interspeech 2016

The i-vector as a generalized speaker (**speech**) embedding

- The success of the i-vector paradigm has been expanded to other **related** tasks:
  - AED, VAD, diarization, etc.

Z. Huang et al. “A blind segmentation approach to acoustic event detection based on i-vector”, In Proc. Interspeech 2013
E. Khoury and M. Garland, “I-Vectors for speech activity detection”, In Proc. Odyssey 2016
G. Sell and D. Garcia-Romero, “Speaker diarization with PLDA i-vector scoring and unsupervised calibration”, In Proc SLT 2014.

The i-vector as a generalized speaker (**speech)** embedding

- The success of the i-vector paradigm has been expanded to other **less related** tasks:
  - Age, emotion, cognitive load, etc.

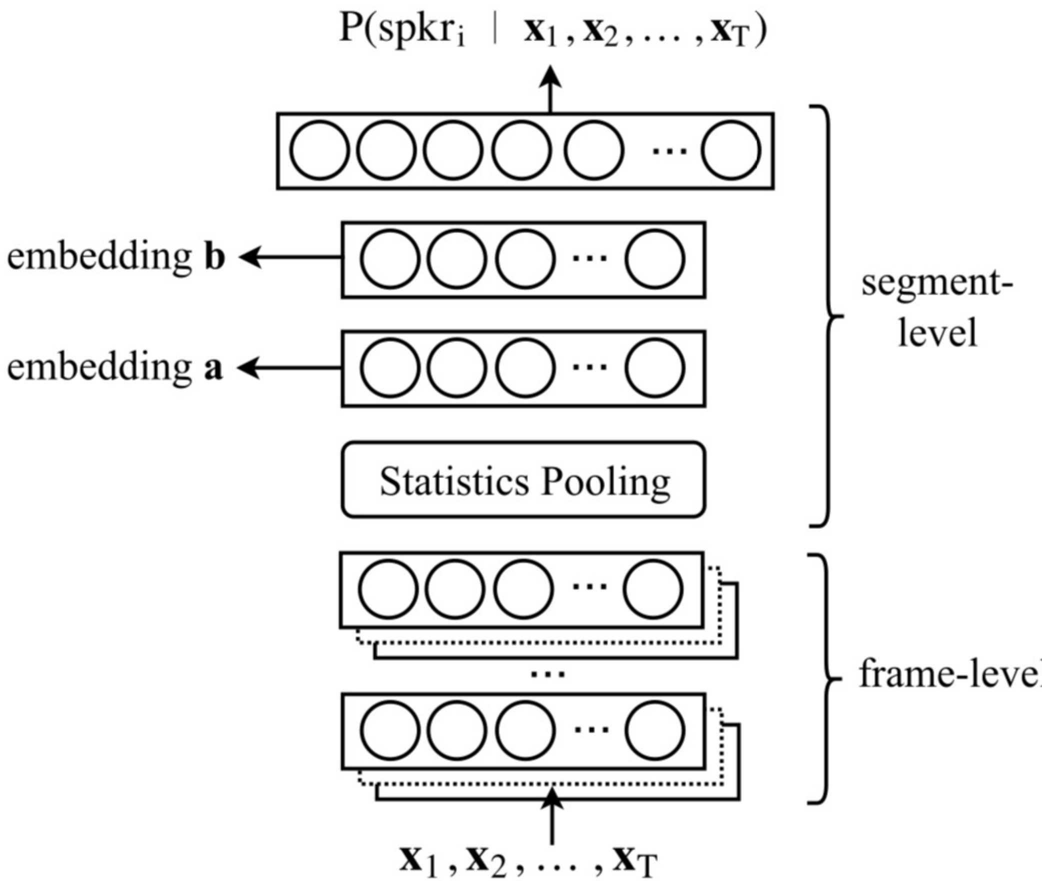
M .Bahari, M. McLaren and D. van Leeuwen, "Speaker age estimation using i-vectors", Engineering Applications of Artificial Intelligence, 34, 99-108, 2014
Xia, Rui, and Yang Liu. "Using i-vector space model for emotion recognition." In Proc. Interspeech 2012.
M.V. Segbroeck et al., "Classification of cognitive load from speech using an i-vector framework", In Proc. Interspeech 2014.

Improved modelling approaches
State of the art until 2017

- Until 2017:
  - i-vectors **extremely** successful:
    - Efforts (2015) of making of i-vectors more than a de-facto standard http://www.voicebiometry.org
    - Some SR evaluations do not rely (directly) on speech samples
      - The 2013-2014 SR i-vector Machine Learning Challenge: https://ivectorchallenge.nist.gov/evaluations/1
  - Deep learning **also** arrived to SR:
    - As a replacement of GMM-UBM in i-vectors
    - As features based on DNN

Improved modelling approaches

In 2018: welcome x-Vectors (bye bye i-vectors)!!



David Snyder, et al., "X-vectors: Robust DNN embeddings for speaker recognition", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Improved modelling approaches

In 2018: welcome x-Vectors (bye bye i-vectors)!!

			SITW Core			SRE16 Cantonese		
			EER(%)	$DCF10^{-2}$	$DCF10^{-3}$	EER(%)	$DCF10^{-2}$	$DCF10^{-3}$
4.1	Original systems	i-vector (acoustic)	9.29	0.621	0.785	9.23	0.568	0.741
		i-vector (BNF)	**9.10**	**0.558**	**0.719**	9.68	0.574	0.765
		x-vector	9.40	0.632	0.790	**8.00**	**0.491**	**0.697**
4.2	PLDA aug.	i-vector (acoustic)	8.64	0.588	0.755	8.92	0.544	0.717
		i-vector (BNF)	8.00	**0.514**	**0.689**	8.82	0.532	0.726
		x-vector	**7.56**	0.586	0.746	**7.45**	**0.463**	**0.669**
4.3	Extractor aug.	i-vector (acoustic)	8.89	0.626	0.790	9.20	0.575	0.748
		i-vector (BNF)	7.27	**0.533**	0.730	8.89	0.569	0.777
		x-vector	**7.19**	0.535	**0.719**	**6.29**	**0.428**	**0.626**
4.4	PLDA and extractor aug.	i-vector (acoustic)	8.04	0.578	0.752	8.95	0.555	0.720
		i-vector (BNF)	6.49	0.492	0.690	8.29	0.534	0.749
		x-vector	**6.00**	**0.488**	**0.677**	**5.86**	**0.410**	**0.593**
4.5	Incl. VoxCeleb	i-vector (acoustic)	7.45	0.552	0.723	9.23	0.557	0.742
		i-vector (BNF)	6.09	0.472	0.660	8.12	0.523	0.751
		x-vector	**4.16**	**0.393**	**0.606**	**5.71**	**0.399**	**0.569**

Table 2. Results using data augmentation in various systems. "Extractor" refers to either the UBM/T or the embedding DNN. For each experiment, the best results are **boldface**.

David Snyder, et al., "X-vectors: Robust DNN embeddings for speaker recognition", 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

Improved modelling approaches

In 2018: welcome x-Vectors (bye bye i-vectors)!!

- Exactly like for i-vectors, the success of x-vectors was expanded to other (more or less) **related** tasks:
  - Language, dialect and native language

David Snyder, et al., "Spoken language recognition using x-vectors", Odyssey 2018.

  - AED, VAD, diarization, etc.

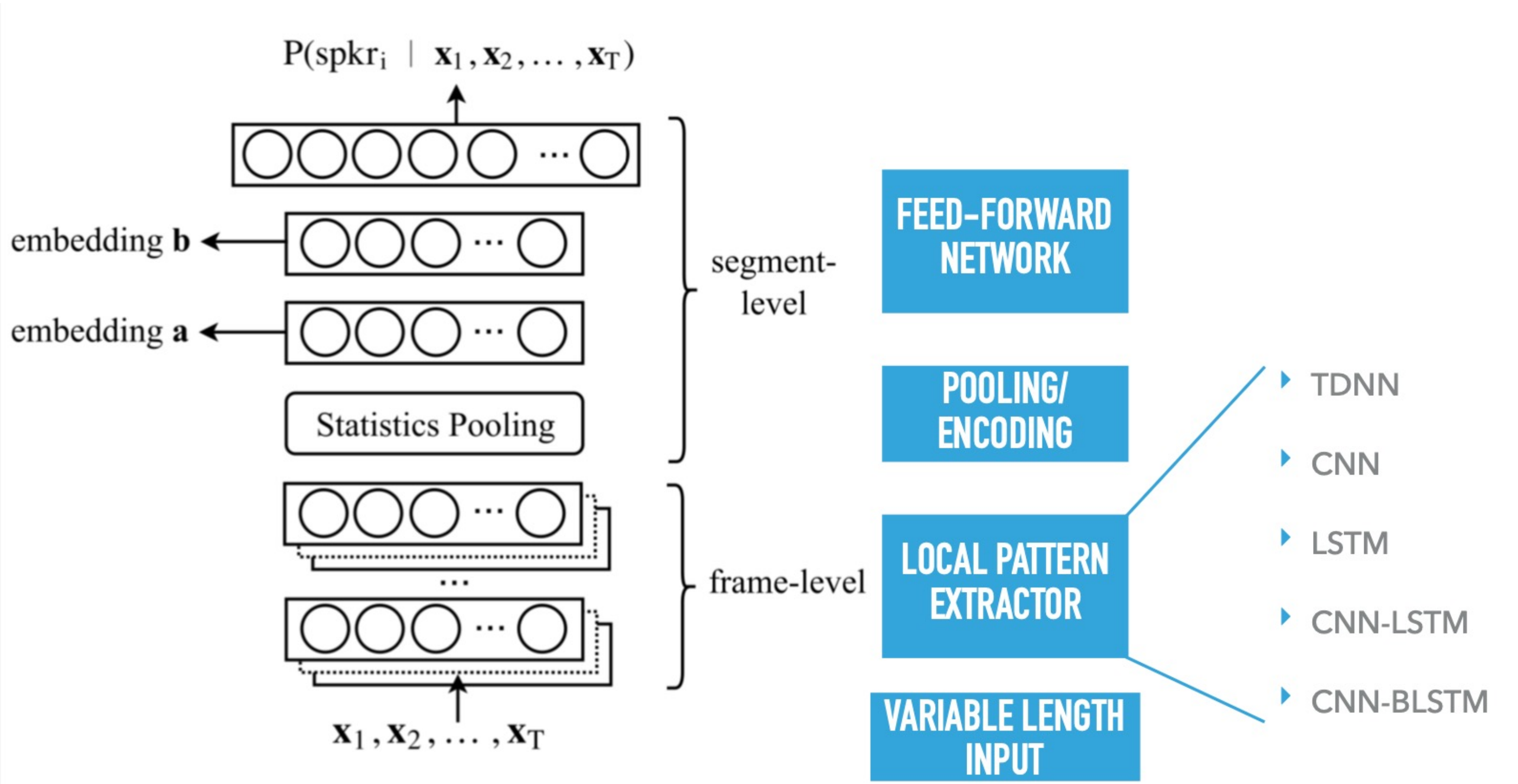
Zeinali, Hossein, Lukas Burget, and Jan Cernocky. "Convolutional neural networks and x-vector embedding for DCASE2018 acoustic scene classification challenge." In Proc. of DCASE Workshop, 2018.

  - Age, emotion, cognitive load, disordered speech, etc.

Botelho, Catarina, et al. "Pathological speech detection using x-vector embeddings." arXiv preprint arXiv:2003.00864 (2020).

Improved modelling approaches

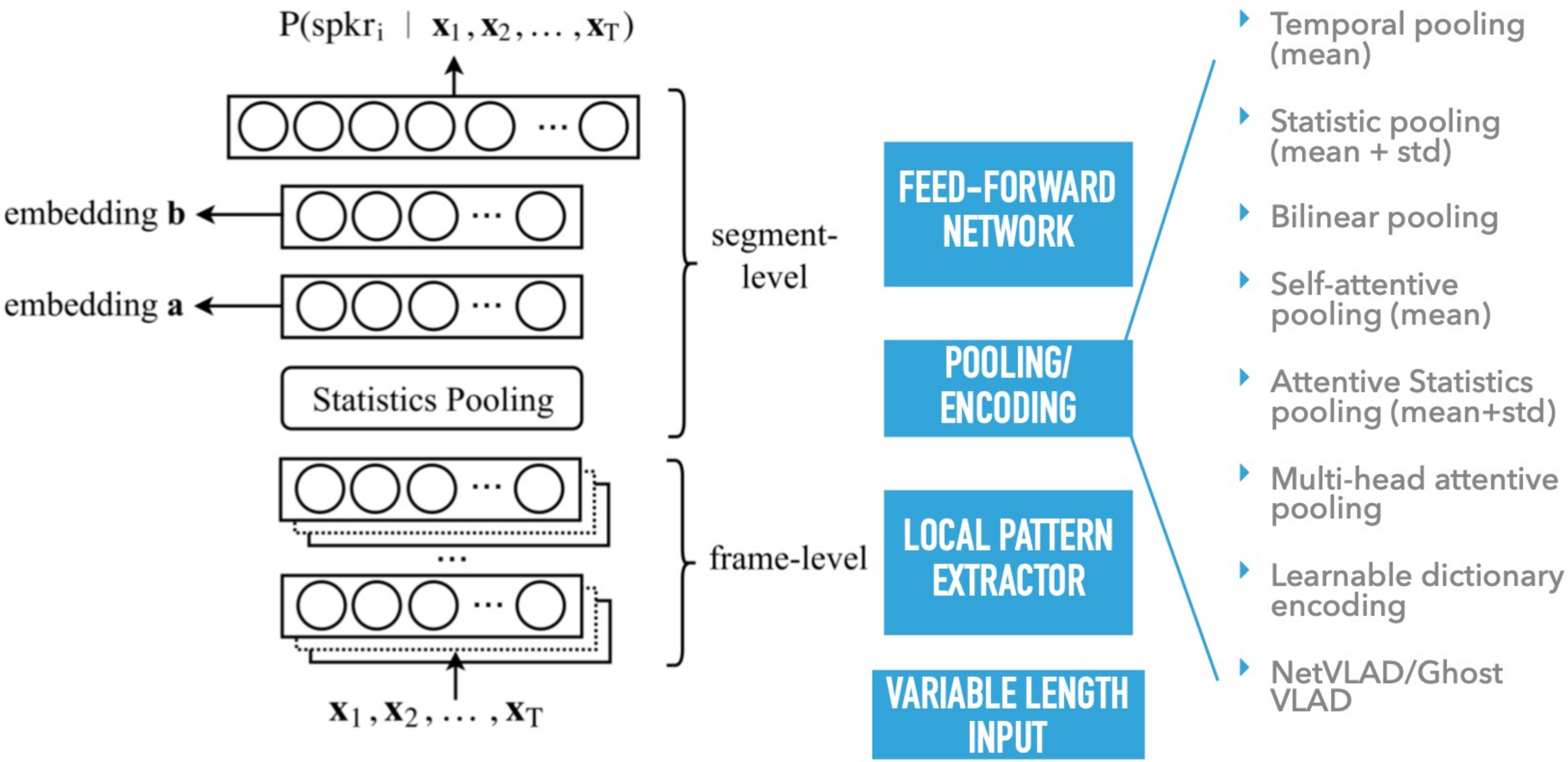
DNN architectures for speaker **(speech)** embedding extraction



Brecht Desplanques, et al., "Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification", in Proc. Interspeech, 2020.

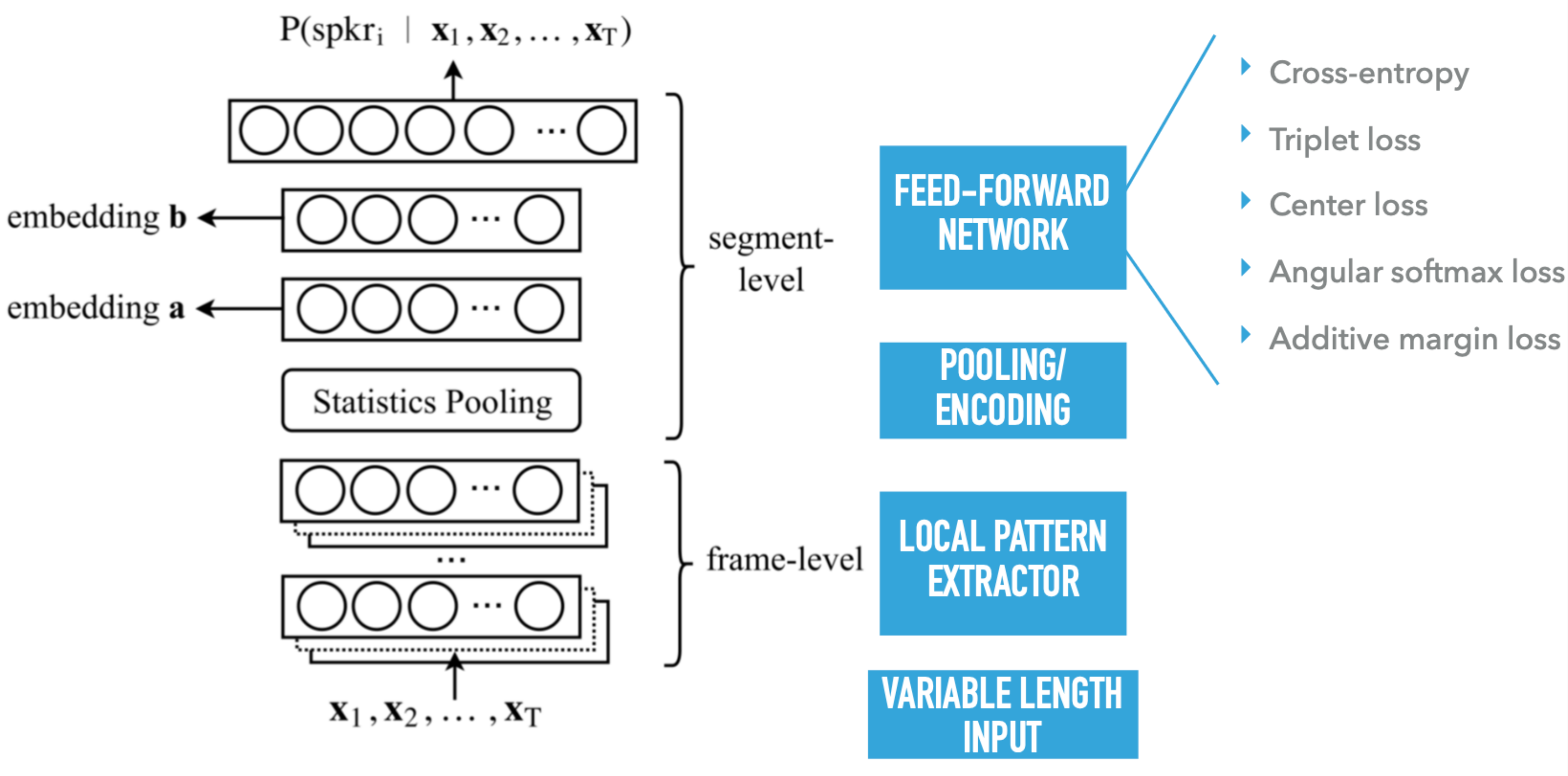
Improved modelling approaches

DNN architectures for speaker **(speech)** embedding extraction



Improved modelling approaches

DNN architectures for speaker **(speech)** embedding extraction



Tools for model-based FE: SpeechBrain

SpeechBrain SpeechBrain is an open-source and all-in-one speech toolkit based on PyTorch.

```
import torchaudio
from speechbrain.pretrained import EncoderClassifier

signal, fs = torchaudio.load('sample.wav')

# standard x-vectors
classifier = EncoderClassifier.from_hparams(source="speechbrain/spkrec-xvect-voxceleb", savedir="pretrained_models/spkrec-xvect-voxceleb")

x_vec_embeddings = classifier.encode_batch(signal)


# ecapa xvectors
classifier = EncoderClassifier.from_hparams(source="speechbrain/spkrec-ecapa-voxceleb")

ecapa_embeddings = classifier.encode_batch(signal)
```

PART IV

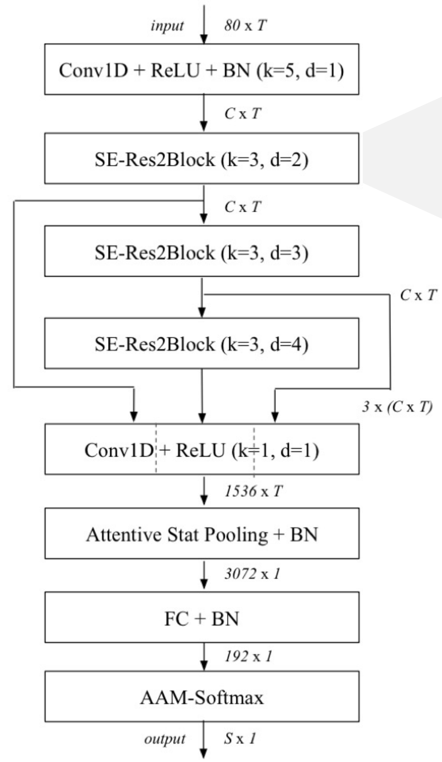
OTHER SPEECH CLASSIFICATION TASK EXAMPLES

State-of-the-art: Speaker Verification

ECAPA-TDNN

- Deep residual convolutional neural network
- Attentive statistics pooling
- Additive Angular Margin loss.

< 2% Equal Error Rate (EER).



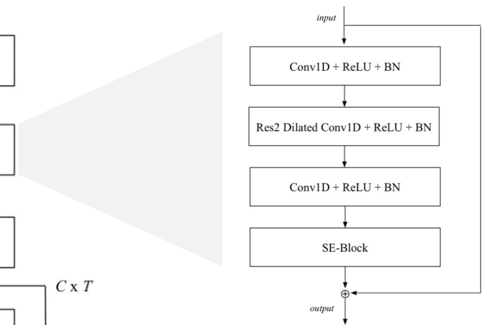


Figure 1: *The SE-Res2Block of the ECAPA-TDNN architecture.*

Figure 2: *Network topology of the ECAPA-TDNN. We*

Brecht Desplanques, et al., ”Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification”, in Proc. Interspeech, 2020.

State-of-the-art: Speaker Diarization

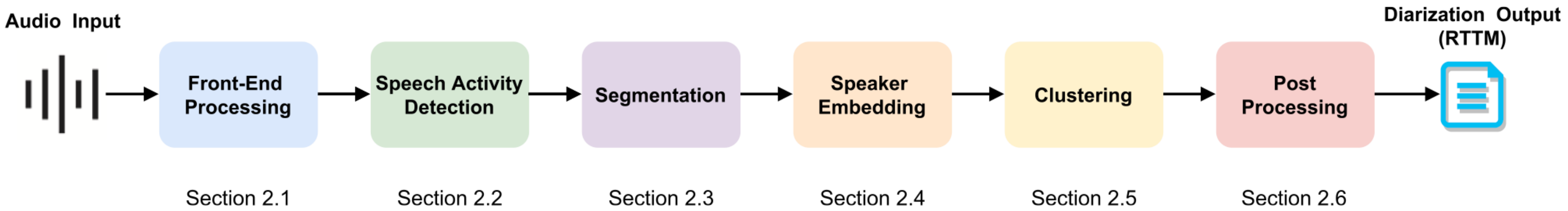


Fig. 1. Traditional speaker diarization system.

"Who spoke when?" in a recording with multiple speakers

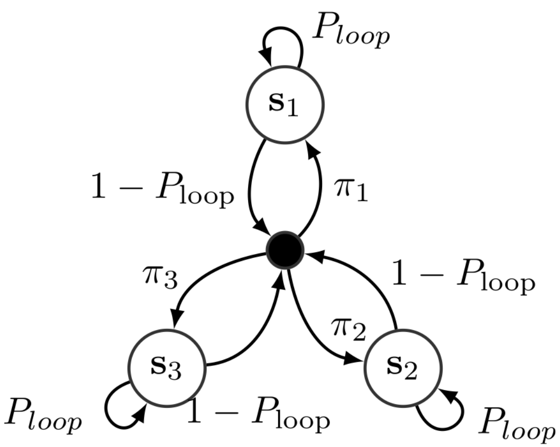
- DIHARD Challenge: https://dihardchallenge.github.io/dihard3/
- *pyannote*: Python library for Speaker Diarization https://huggingface.co/pyannote/speaker-diarization

Tae Jin Park, et al., " A review of speaker diarization: Recent advances with deep learning", in Computer Speech and Language, Volume 72, 2022, 101317, ISSN 0885-2308.

State-of-the-art: Speaker Diarization

VBx system

- Variational Bayes Hidden Markov Model with x-vectors.
- Each time frame is represented by an x-vector.
- An HMM as the one in the figure is used to align the sequence of x-vectors to each state.
- States are initialised with agglomerative hierarchical clustering.



~5% Diarization Error Rate (DER) w/ forgiveness collar and ignoring overlapped regions.
~20% DER w/o forgiveness collar and scoring overlapped regions.

Federico Landini, et al., ”Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks”, in Computer Speech and Language, Volume 71, 2022, 101254.

Benchmark of SSL models

SUPERB (**S**peech processing **U**niversal **PER**formance **B**enchmark) is an online benchmark for several speech tasks (recognition, detection, semantics, speaker, paralinguistics and generation)

https://superbbenchmark.org/

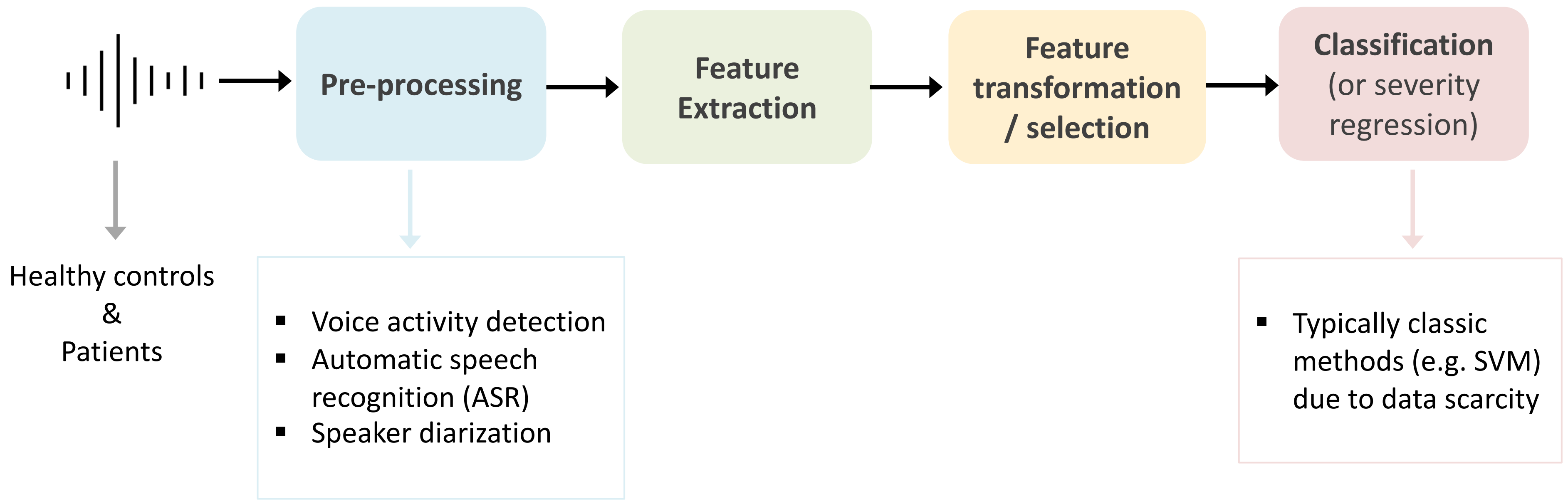
SUPERB

Table Scatter Chart Radar Chart

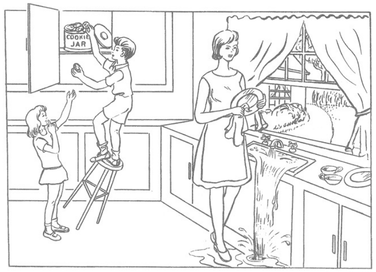
* The four columns (1)~(4) correspond to the macs calculated with short, medium, long, longer bucket respectively
* Params = Parameter shared without fine-tuning

Method	Name	Description	URL	Params ↓	MACs ↓	(1) ↓	(2) ↓	(3) ↓	(4) ↓	Rank ↑	Score ↑	KS ↑	IC ↑	PR ↓	ASR ↓	ER ↑	QbE ↑	SF-F1 ↑	F-CER ↓	SID ↑	SV ↓	SD ↓
WavLM Large	Microsoft	M-P + …		3.166e+8	4.326e+12	3…	6…	1…	2…	25.8	1145	97.86	99.31	3.06	3.44	70.62	8.86	92.21	18.36	95.49	3.77	3.24
WavLM Base+	Microsoft	M-P + …		9.470e+7	1.670e+12	1…	2…	4…	8…	24.05	1106	97.37	99	3.92	5.59	68.65	9.88	90.58	21.2	89.42	4.07	3.5
WavLM Base	Microsoft	M-P + …		9.470e+7	1.670e+12	1…	2…	4…	8…	20.95	1019	96.79	98.63	4.84	6.21	65.94	8.7	89.38	22.86	84.51	4.69	4.55
data2vec Large	CI Tang	Maske…		3.143e+8	4.306e+12	3…	6…	1…	2…	20.8	949	96.75	98.31	3.6	3.36	66.31	6.28	90.98	22.16	76.77	5.73	5.53
LightHuBERT…	LightHu…	Once-f…		9.500e+7	-	-	-	-	-	20.1	959	96.82	98.5	4.15	5.71	66.25	7.37	88.44	25.92	80.01	5.14	5.51
HuBERT Large	paper	M-P + VQ		3.166e+8	4.324e+12	3…	6…	1…	2…	19.15	919	95.29	98.76	3.53	3.62	67.62	3.53	89.81	21.76	90.33	5.98	5.75
data2vec-aqc…	Speech…	Maske…		9.384e+7	1.657e+12	1…	2…	4…	8…	19.05	935	96.36	98.92	4.11	5.39	67.59	6.65	89.39	22.88	59.87	5.82	4.84

State-of-the-art: Automatic Disease Detection



State-of-the-art: Automatic Disease Detection



Example: Detection of Alzheimer's Disease

Pre-processing → **Feature Extraction** → **Feature transformation / selection** → **Classification** (or severity regression)

Healthy controls
&
AD Patients

- Automatic speech recognition (ASR)
- ....

- Acoustic features (e.g. silence-to-speech ratio)
- Linguistic features (e.g. content density; coherence)
- Embeddings from pre-trained models

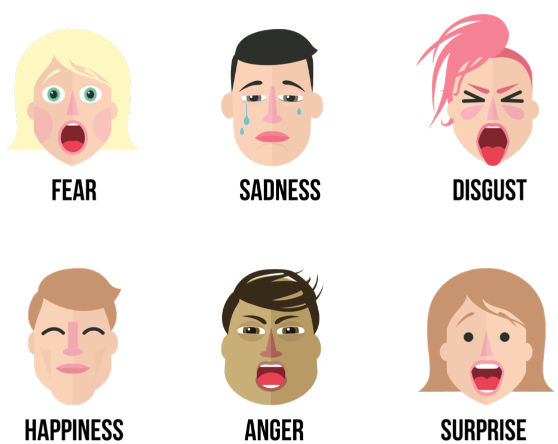
- ADReSS Challenge (2020) / ADReSSo Challenge (2021) / ADReSS-M Challenge (2023)

Hecker, P., et al., "Voice Analysis for Neurological Disorder Recognition–A Systematic Review and Perspective on Emerging Trends", in Frontiers in Digital Health, 4, 2022.
Pompili, A., Rolland, T., & Abad, A., "The INESC-ID multi-modal system for the ADReSS 2020 challenge", in Interspeech, 2020.

State-of-the-art: Speech Emotion Recognition

Emotion

Ekman's six universal emotions

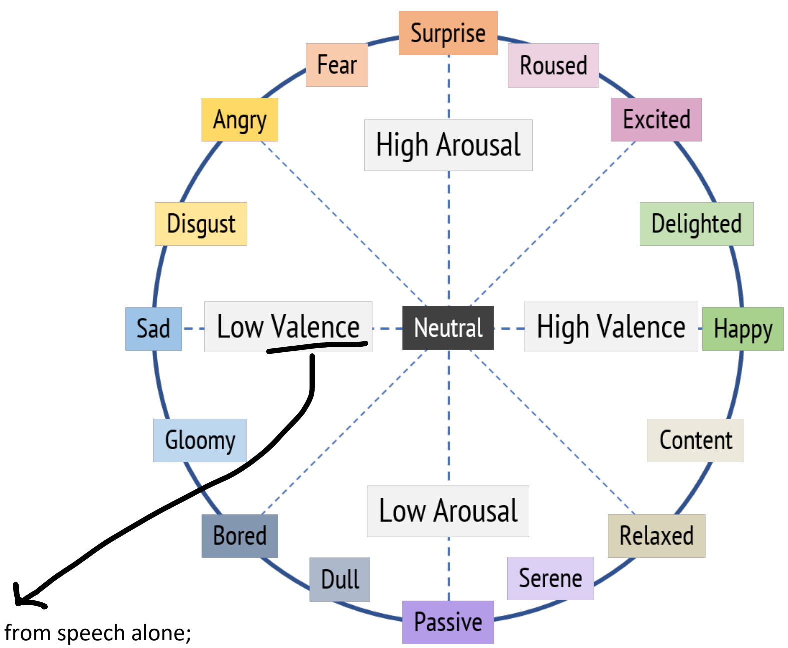


- 6-class classification

Table 2.2: Summary of traditional acoustic correlates of emotions.

	mean F0	F0 range	F0 variability	downward F0 contours	mean Energy	high freq. Energy	speech rate
Anger	↑		↑	↑	↑	↑	↑
Fear	↑	↑				↑	↑
Sadness	↓	↓		↑		↓	↓
Joy	↑	↑	↑		↑	↑	↑
Disgust	↑ / ↓						

Emotional state model



hard to identify from speech alone;
better when combined with text

- separate classification of valence and arousal, based on intensity levels

State-of-the-art: Speech Emotion Recognition

- 2D CNN LSTM:
  - IEMOCAP – speaker-dependent accuracy: 89.16%; speaker-independent: 52.14%
  - EmoDB – speaker-dependent accuracy: 95.83%; speaker-independent: 95.89%

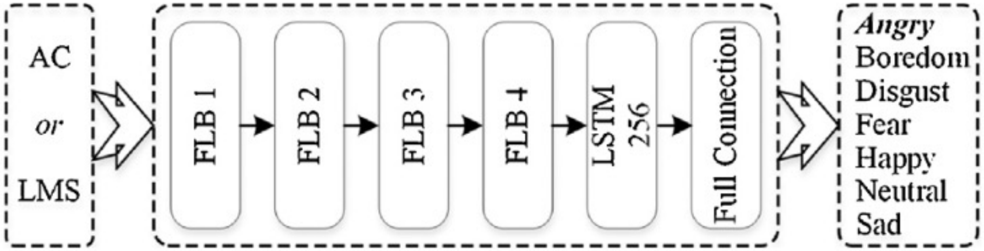


Fig. 5. Block diagram of the overall architecture of the designed 1D and 2D CNN LSTM networks. For brevity, audio clip and log-mel spectrogram are abbreviated as AC and LMS.

Jianfeng Zhao, et al., "Speech emotion recognition using deep 1D & 2D CNN LSTM networks", Biomedical Signal Processing and Control 47 (2019): 312-323.

PART V
LABORATORY ASSIGNMENT 2

References and additional materials

PART I: Feature extraction

https://speechprocessingbook.aalto.fi/Representations/Representations.html

https://speechprocessingbook.aalto.fi/Recognition/Voice_activity_detection.html

PART II: Modeling tools for speech

https://speechprocessingbook.aalto.fi/Modelling_tools_in_speech_processing.html

PART III: Speaker recognition

https://speechprocessingbook.aalto.fi/Recognition/Speaker_Recognition_and_Verification.html

PART IV: Other speech recognition tasks

https://speechprocessingbook.aalto.fi/Recognition_tasks_in_speech_processing.html