Aston University
BIRMINGHAM UK

Forensic Voice Comparison

(Automatic Speaker Recognition)

Phil Weber – BrumAI – 12 Aug 2021

UNIVERSITY OF THE YEAR 2020 The Guardian

Forensic voice comparison

1. The technology – how

2. Specific concerns in forensics – trust

3. Discussions in the context of AI – bias and humans

Dr. Phil Weber

- **Aston University**

Forensic Data Science Laboratory
Forensic Speech Science Laboratory

Aston Institute for Forensic Linguistics

THINK BEYOND DATA

Think Beyond Data
– free consultancy in AI & data analytics for SMEs in the UK West Midlands

Aston Forensic Data | Speech Science Laboratory

- New paradigm for the evaluation of forensic evidence
  - quantification of strength of evidence (likelihood ratio)
  - relevant data, quantitative measurement, and statistical models
  - validation under conditions reflecting those of the case under investigation
  - reduction of the potential for cognitive bias.

Forensic data science

https://www.aston.ac.uk/research/forensic-linguistics/data-science-laboratory
https://www.aston.ac.uk/research/forensic-linguistics/forensic-speech-science-laboratory

- Forensic Speech Science
- DNA
- Firearms
- ISO Standards Development
- Fingerprints
- Bloodstain Pattern Analysis
- Gait Analysis
- Cell Site Analysis
- Authorship Analysis
- Communication of Evidence

Context

To think about …

1. What would you prefer in court?
   – human expert or AI comparison of (your?) voice recordings?

2. If AI then what would you want from the AI?
   If human then what would you want from the human?

Image source StockMonkeys.com [http://www.stockmonkeys.com]

To think about …

Why is it hard?

1. Between-speaker and within-speaker variability

2. Variable-length recordings

3. Mismatch in recording conditions

What is the problem?

Automatic speaker recognition:

Classification

Speaker identification – is it speaker A or speaker B?

Speaker verification – are you speaker A?

What is the problem?

Forensic voice comparison:

- Courts make decisions, not forensic scientists

$$\Pr(Hypothesis \mid data) \in [0, 1]$$

vs

$$\mathrm{f}(data \mid Hypothesis) \in [0, \infty)$$

What is the problem?

Forensic voice comparison:

Weight of the evidence

- Compare two mutually-exclusive hypotheses using a likelihood ratio

Weight of the evidence

Forensic voice comparison:

Mutually-exclusive hypotheses – specific-source

- The **observed properties** of the voice on the questioned-speaker recording are more likely if it was produced by the **known speaker**.

- The **observed properties** of the voice on the questioned-speaker recording are more likely if it was produced by **some other speaker** selected at random from the relevant population.

Weight of the evidence

Forensic voice comparison:

Mutually-exclusive hypotheses – same-source

- The **observed properties** of the voice on the questioned- and known-speaker recordings are more likely if they were produced by the **same speaker** (selected at random from the relevant population).
- The **observed properties** of the voice on the questioned- and known-speaker recordings are more likely if they was produced by **different speakers** selected at random from the relevant population.

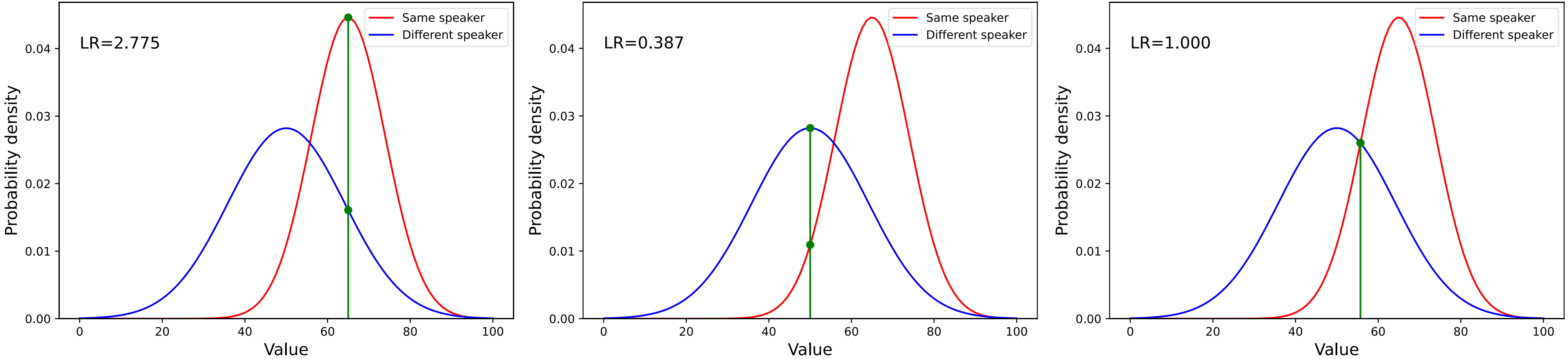
Likelihood ratio

- Likelihood ratio (LR)
  - likelihood of data given competing hypotheses

$$\frac{\mathrm{f}\,(\mathit{data} \mid \text{same [random] speaker})}{\mathrm{f}\,(\mathit{data} \mid \text{different [random] speaker})}$$

- Classifier
  - $\Pr(\text{known speaker} \mid \mathit{data})$ ***vs*** $\Pr(\text{different speaker} \mid \mathit{data})$

Likelihood ratio



$$LR > 1 \qquad LR < 1 \qquad LR = 1$$

Evidence points to **same**-speaker

Evidence points to **different**-speaker

No conclusive evidence

Similarity and typicality are both important

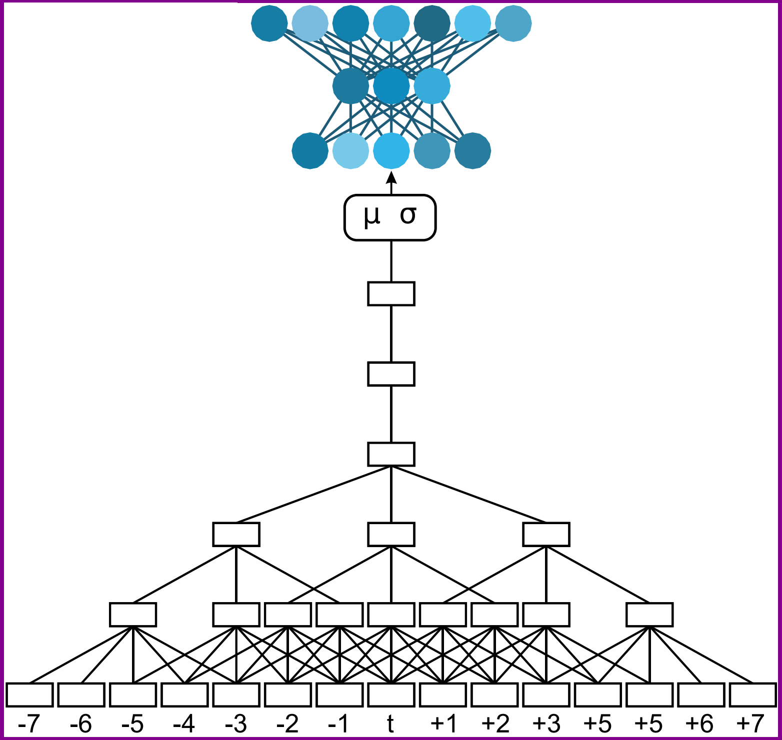
Aston University
BIRMINGHAM UK

The technology

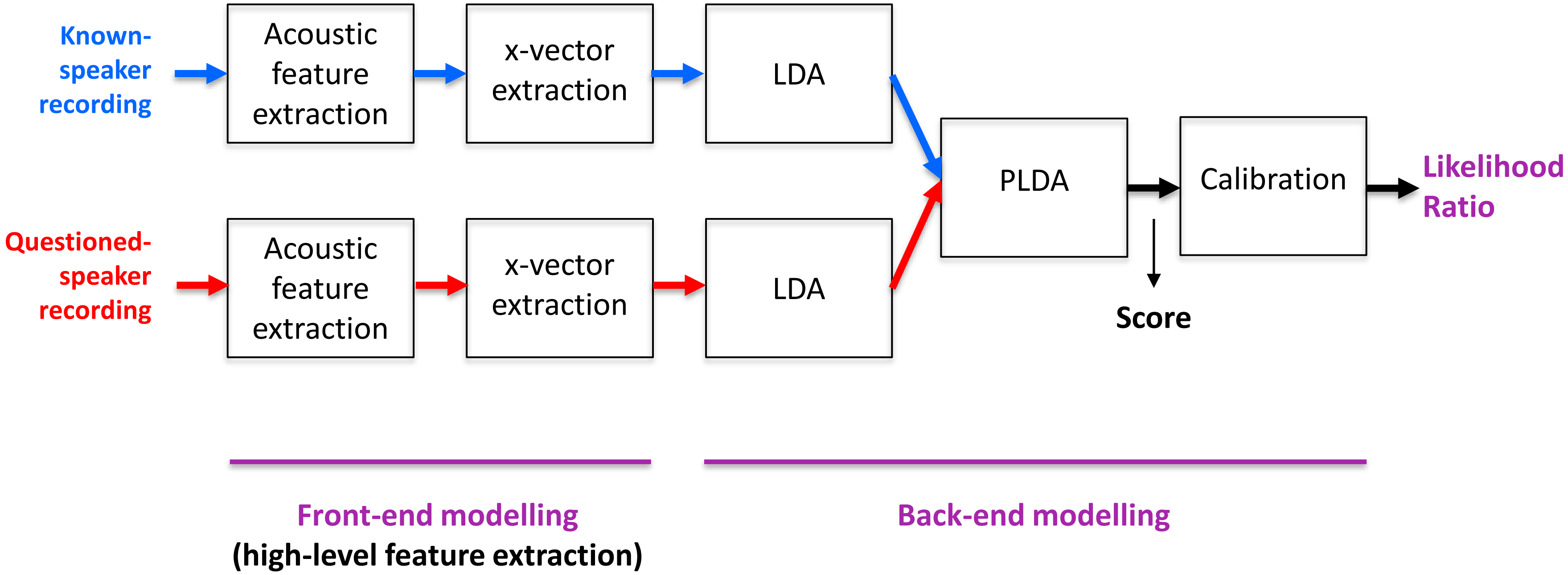
For automatic speaker recognition

and

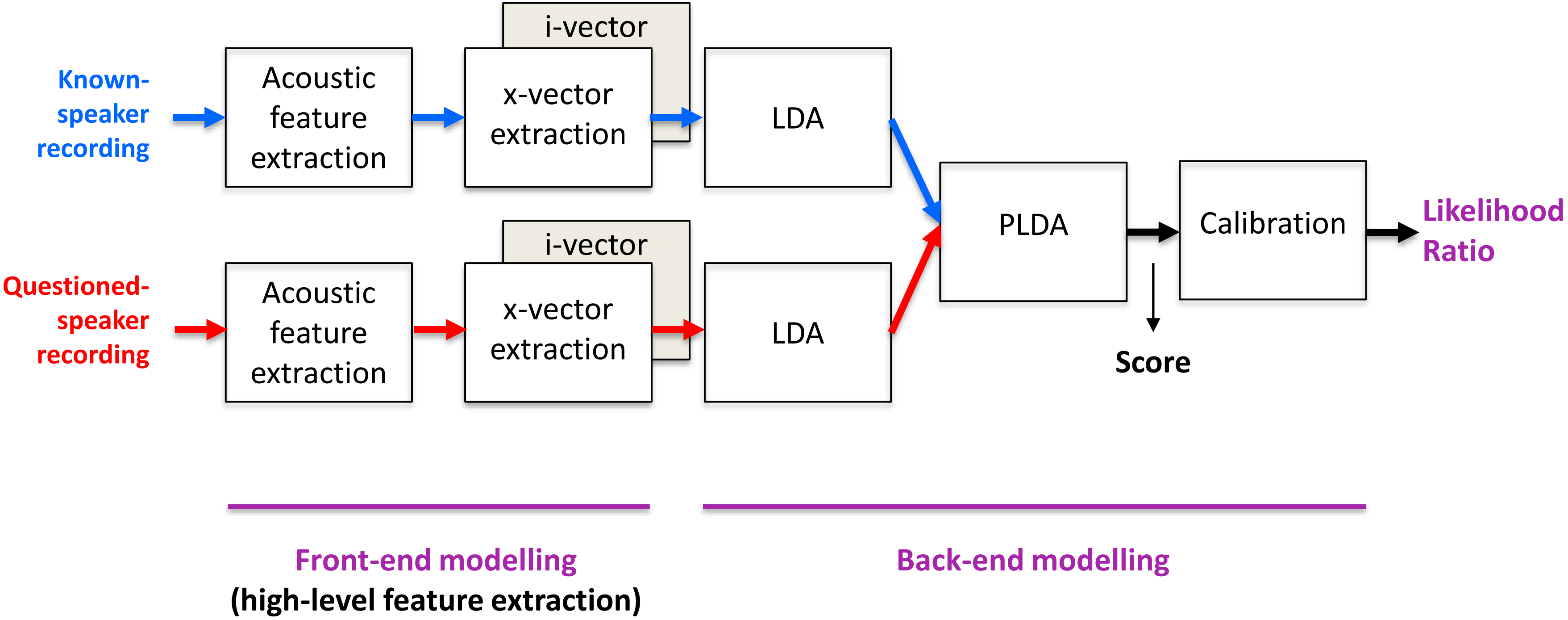
Forensic Voice Comparison



Machine learning pipeline (x-vector)



Machine learning pipeline (i-vector)



Machine learning pipeline (x-vector)

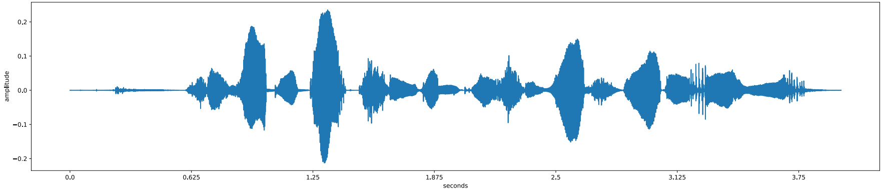
Key: everything we do

1. Enhance (make use of) between-speaker differences

2. Downplay (ignore) within-speaker differences

3. Remove effects of recording condition

Acoustic feature extraction (low-level features)



Digital sampling → Windowing → Fourier transform → Mel-frequency filtering → log energy → Discrete cosine transform → Mel Frequency Cepstral Coefficients

Audio recording

e.g 8000 Hz limits frequency resolution

short overlapping windows (25ms)

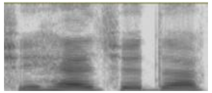
frequency analysis

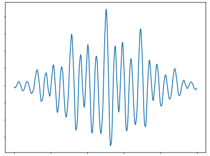
perceptually--inspired

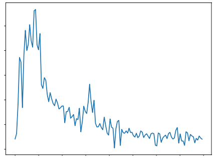
compress the dynamic range

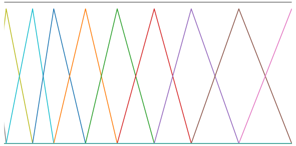
remove correlation

compact representation (12 dimensions)

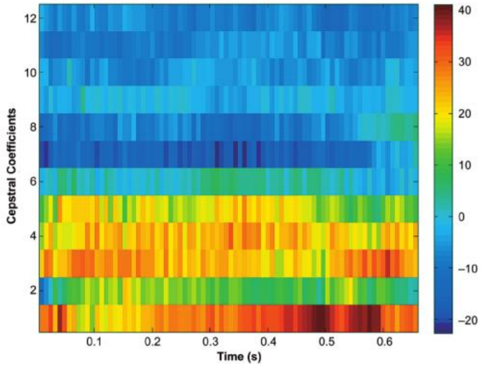




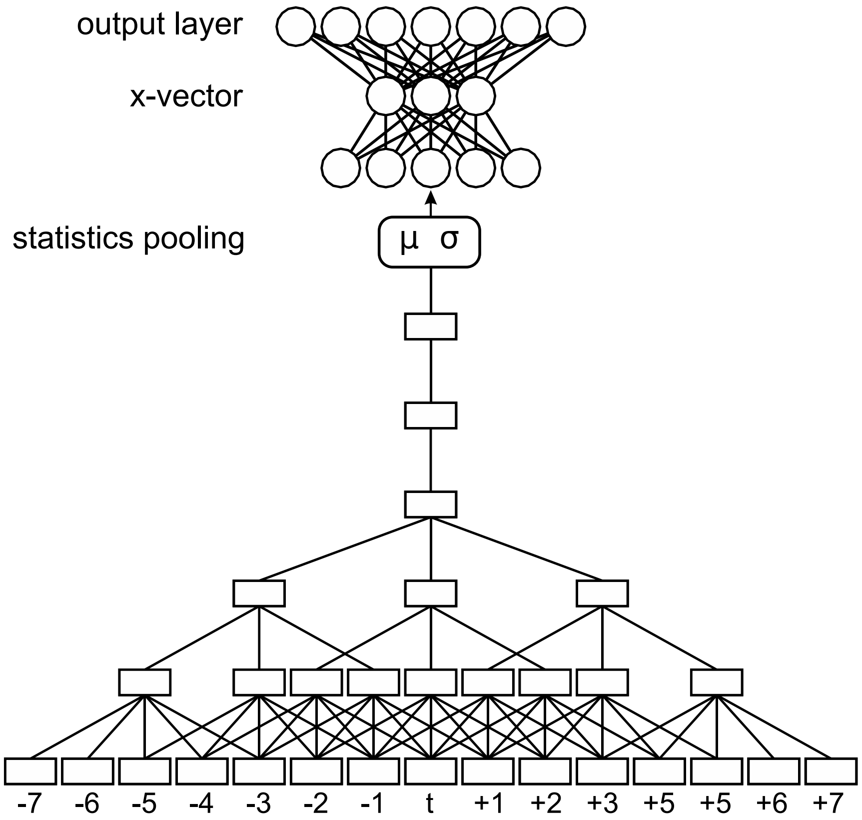




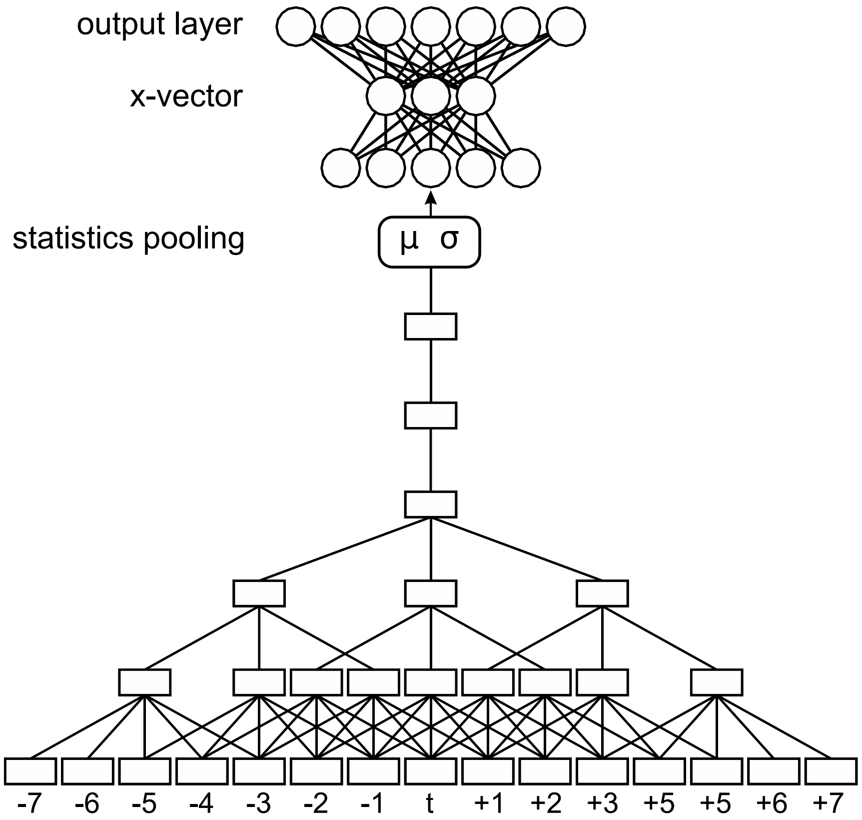
40 – 80d features



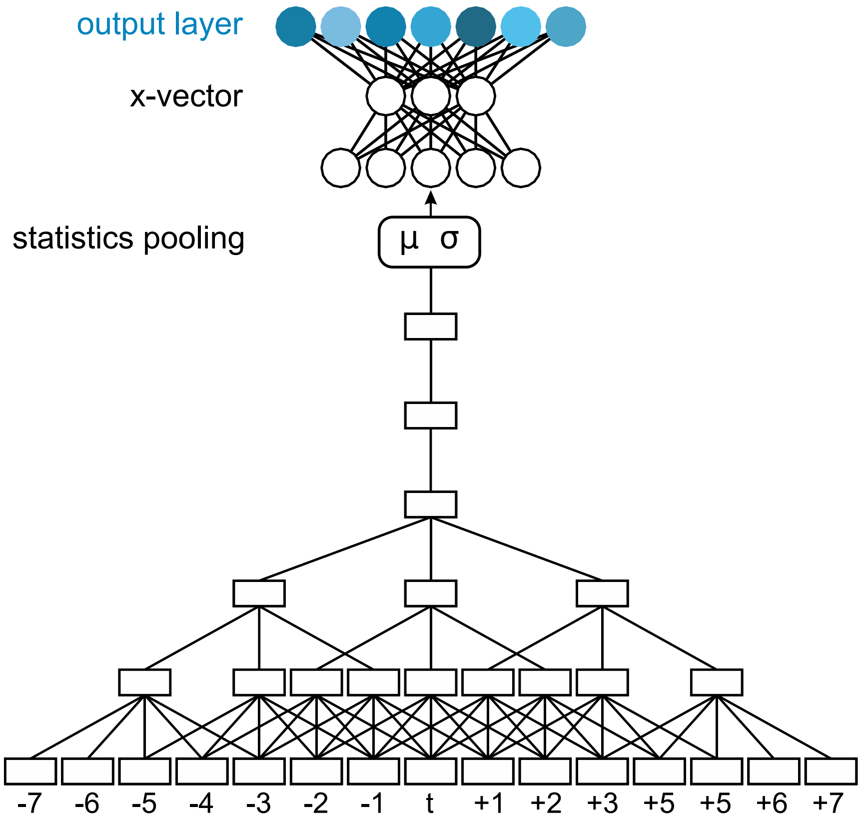
x-vector extraction – speaker (high-level) features



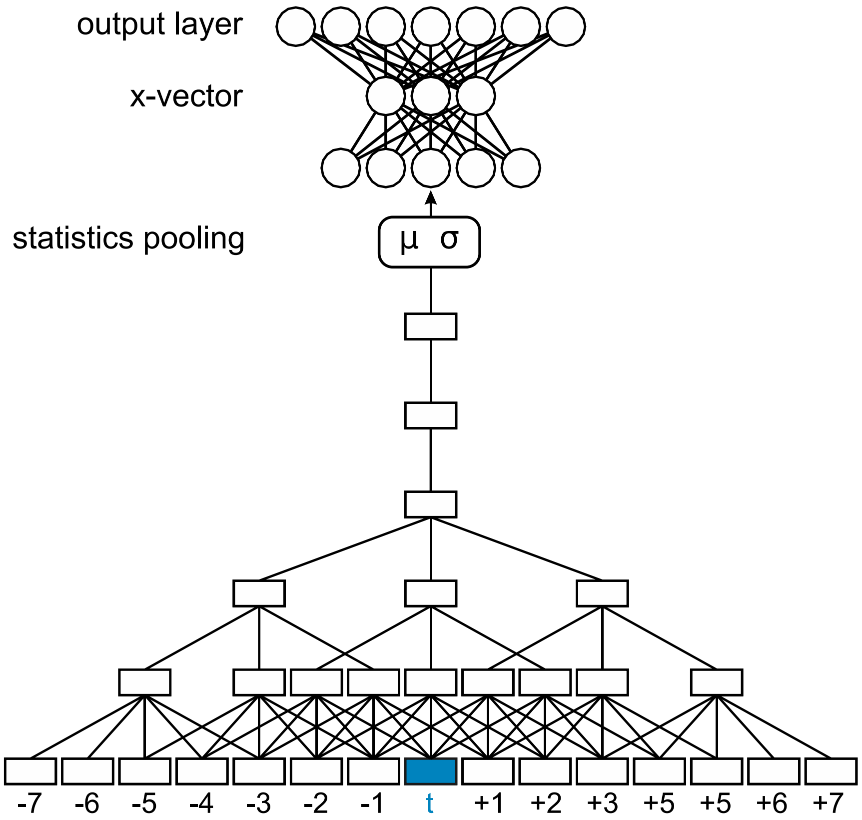
TDNN – Time-delay deep neural network



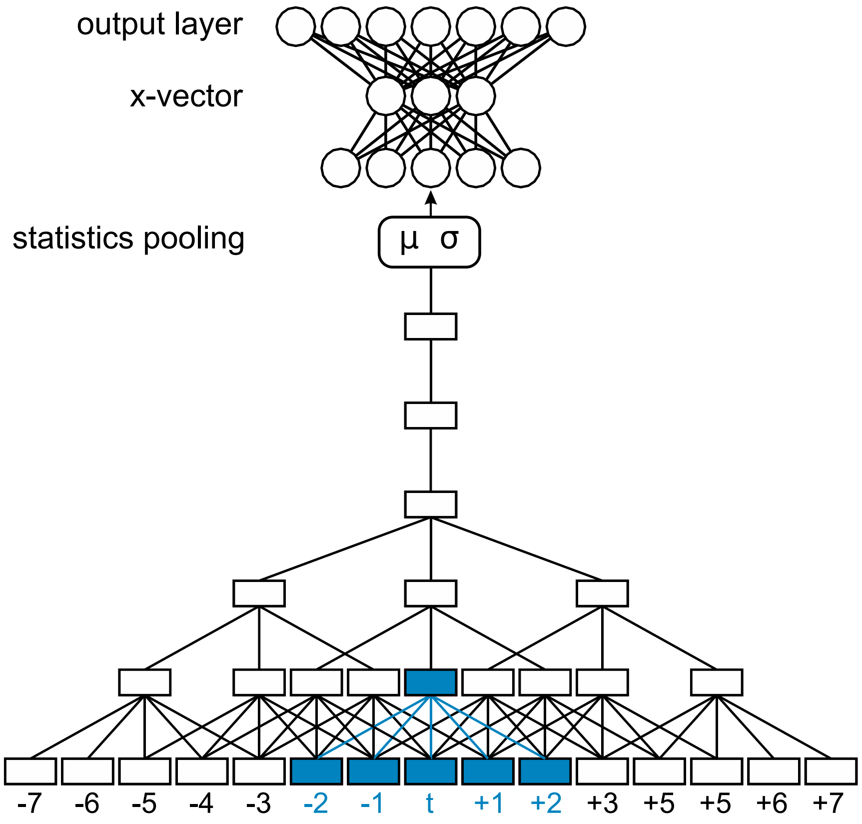
TDNN – classify speakers



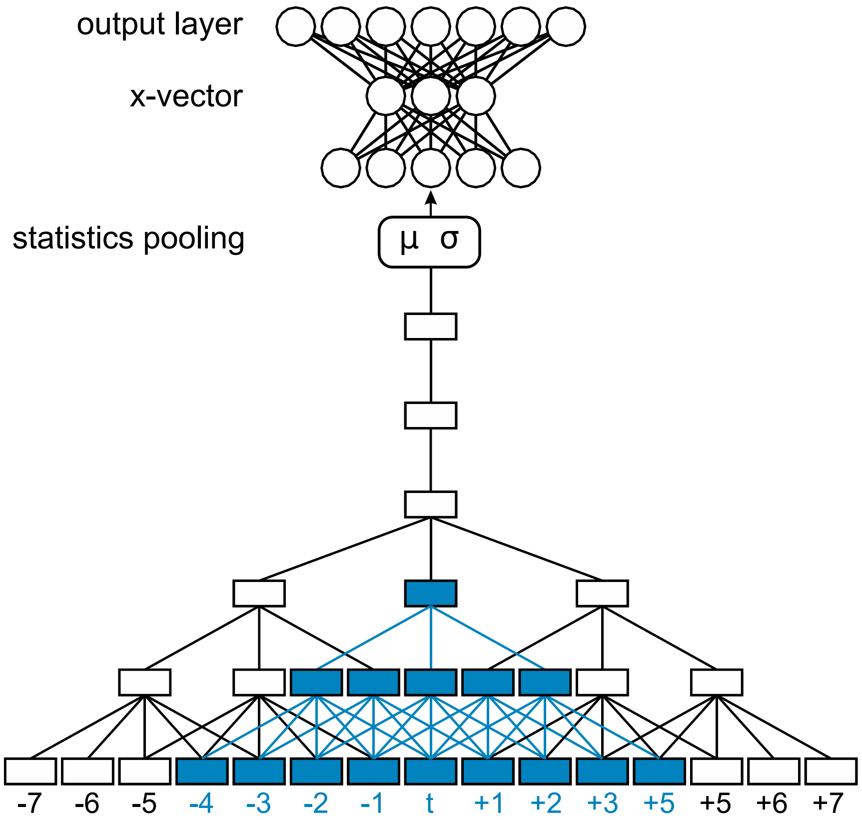
TDNN – input MFCC



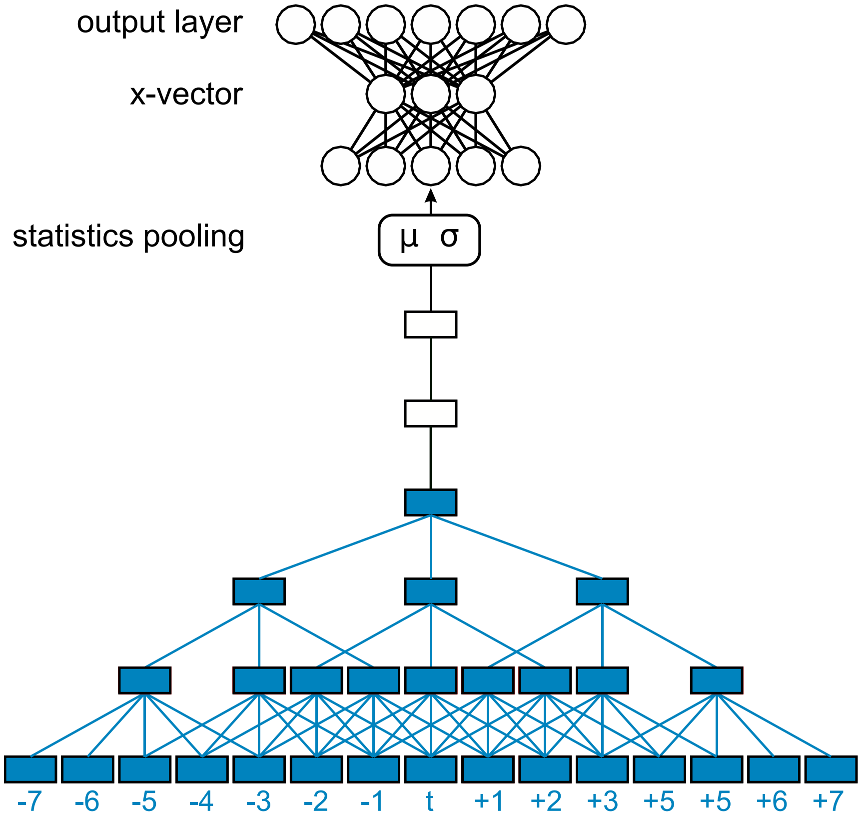
TDNN – aggregate over time



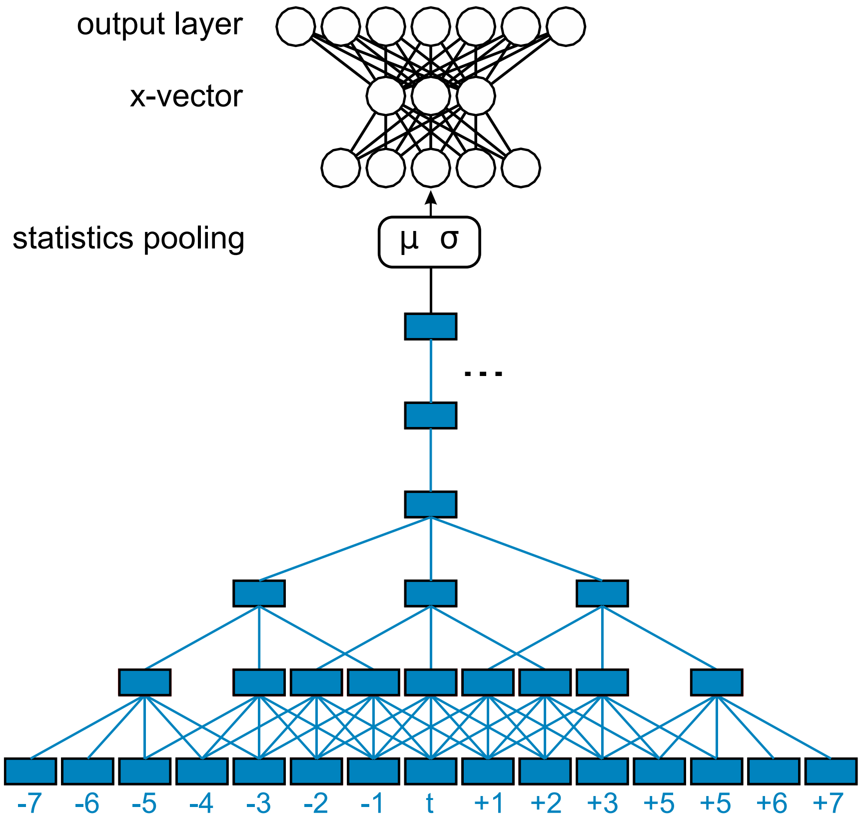
TDNN – aggregate over time



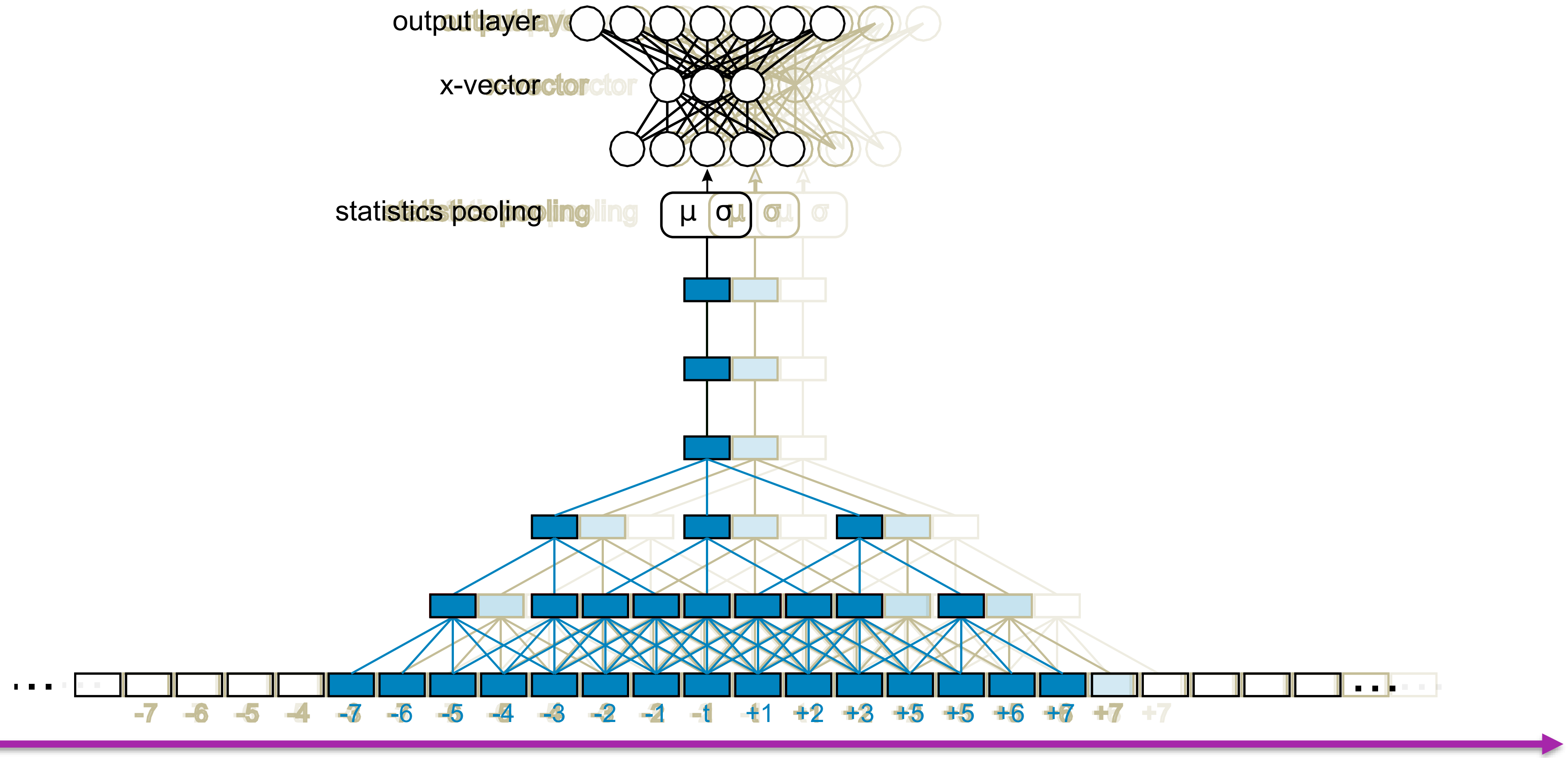
TDNN – aggregate over time



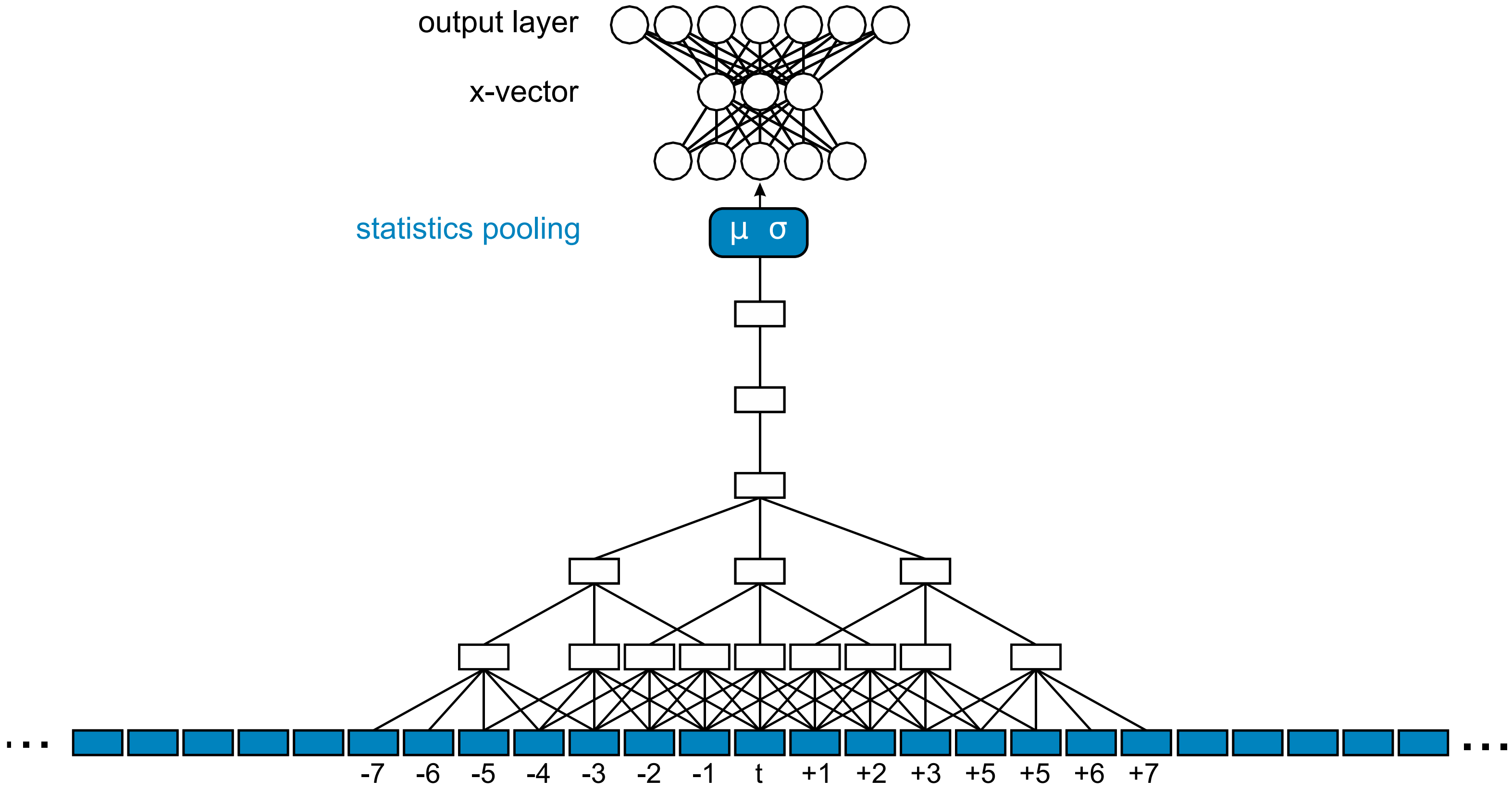
TDNN – aggregate over time



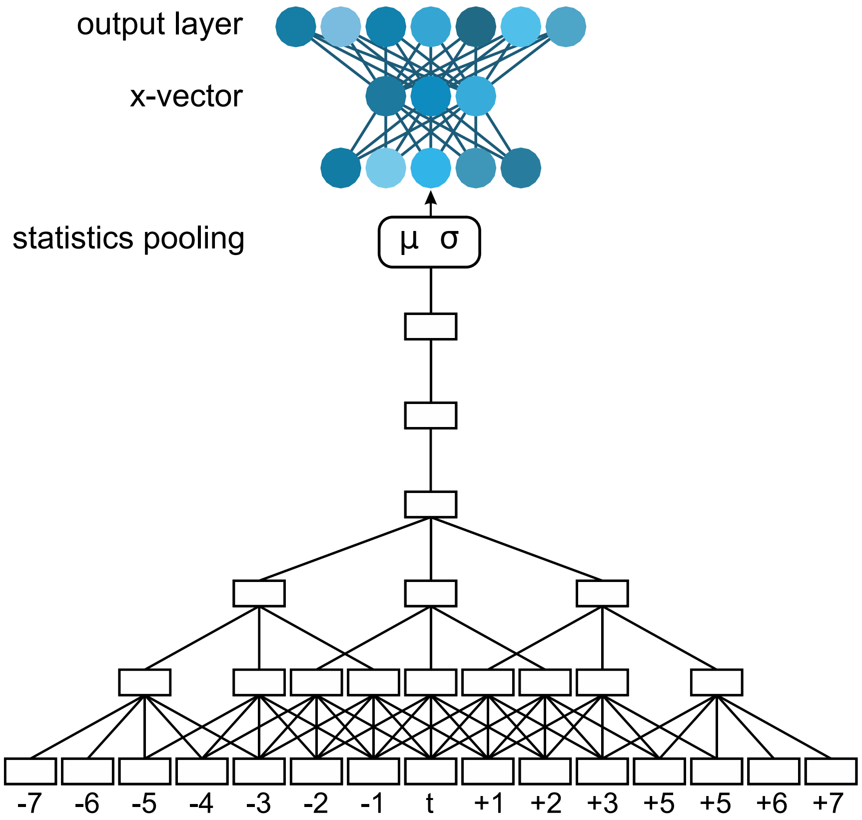
TDNN – pass over recording



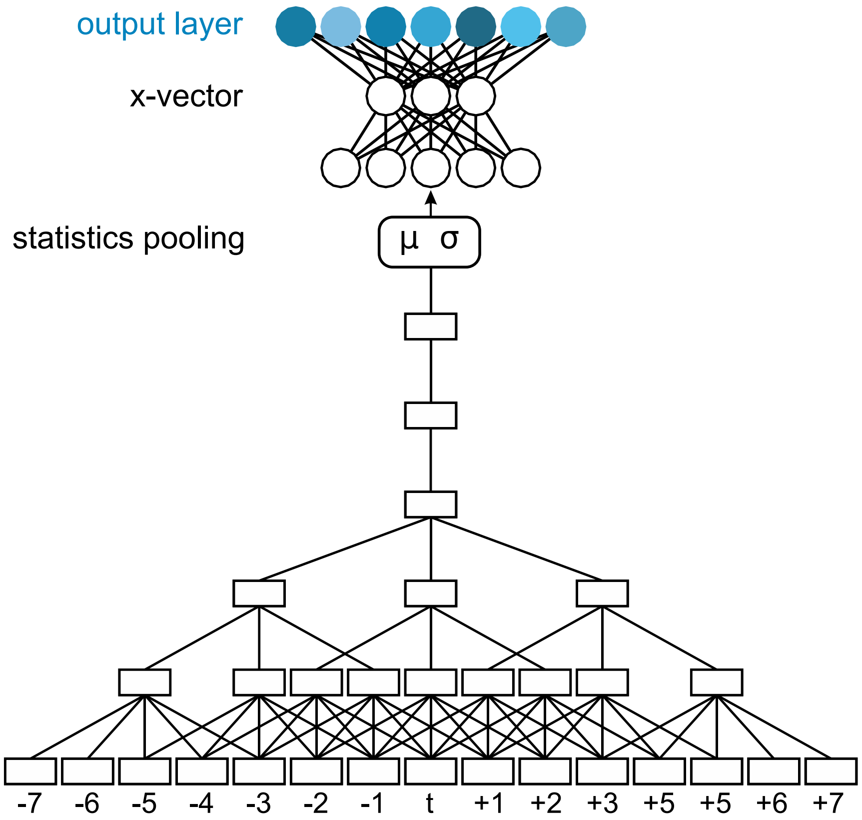
TDNN – pooling layer



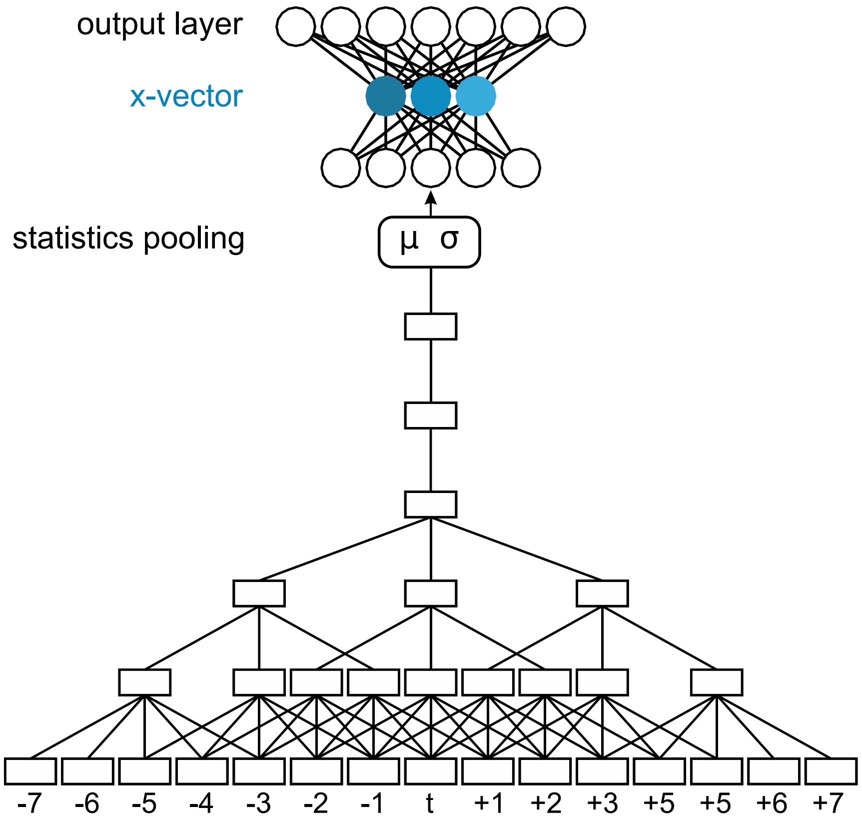
TDNN – segment-level processing



TDNN – classify speakers

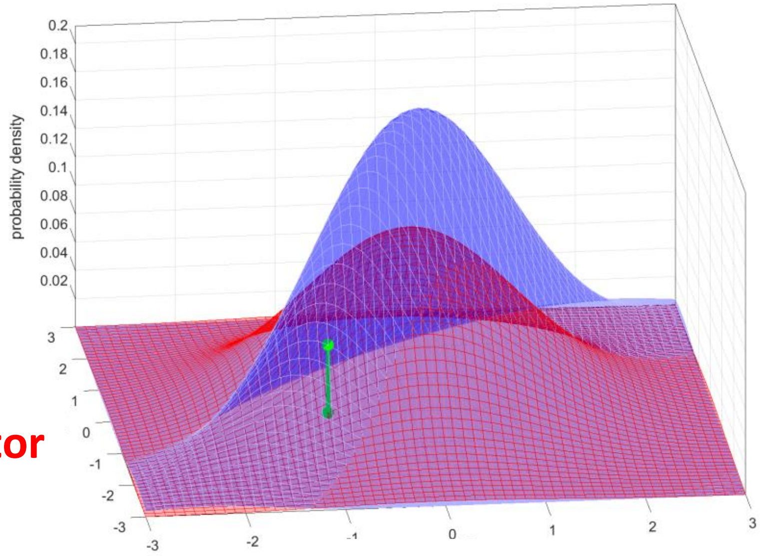


TDNN – x-vector bottleneck



Probabilistic linear discriminant analysis (PLDA)

Attempt to calculate a likelihood ratio

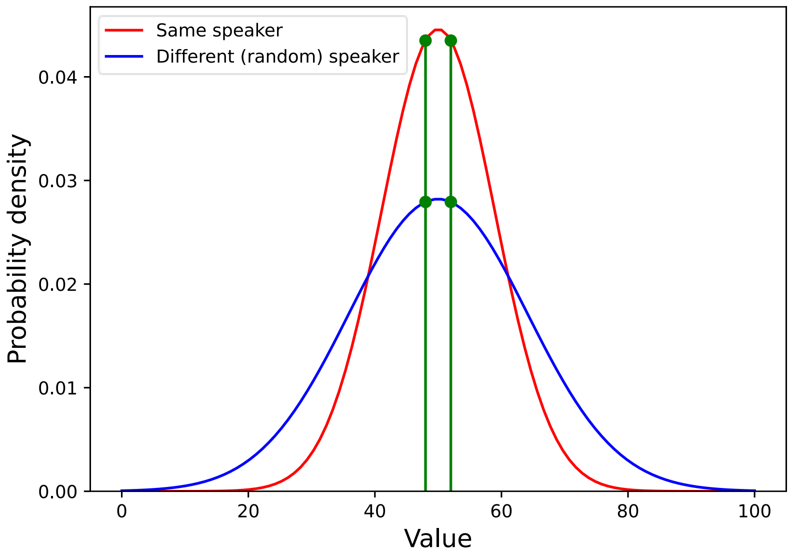


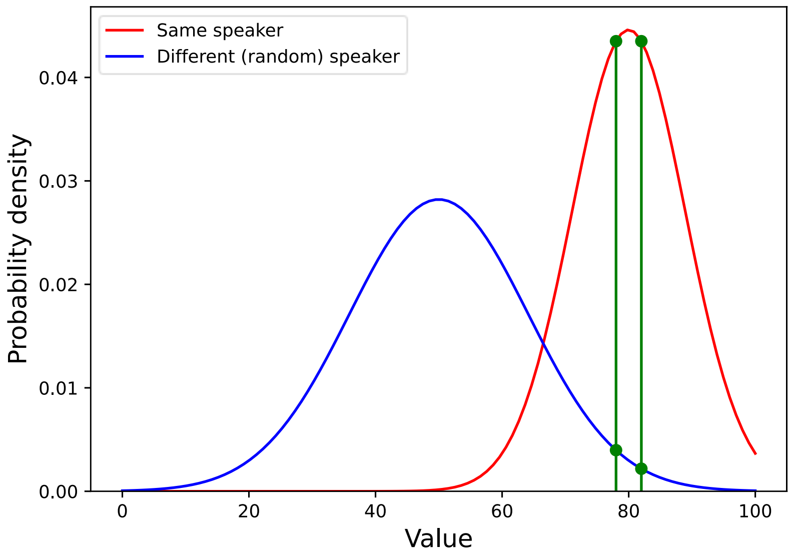
Known speaker **x-vector**

questioned speaker **x-vector**

Morrison, G. S.; Enzinger, E.; Ramos, D.; González-Rodríguez, J. & Lozano-Díez, A. Banks, D. L.; Kafadar, K.; Kaye, D. H. & Tackett, M. *(Eds.)* Statistical models in forensic voice comparison, 20, *Boca Raton, FL: CRC,* **2020**, 451-497,
http://handbook-of-forensic-statistics.forensic-voice-comparison.net/

PLDA – Similarity and typicality





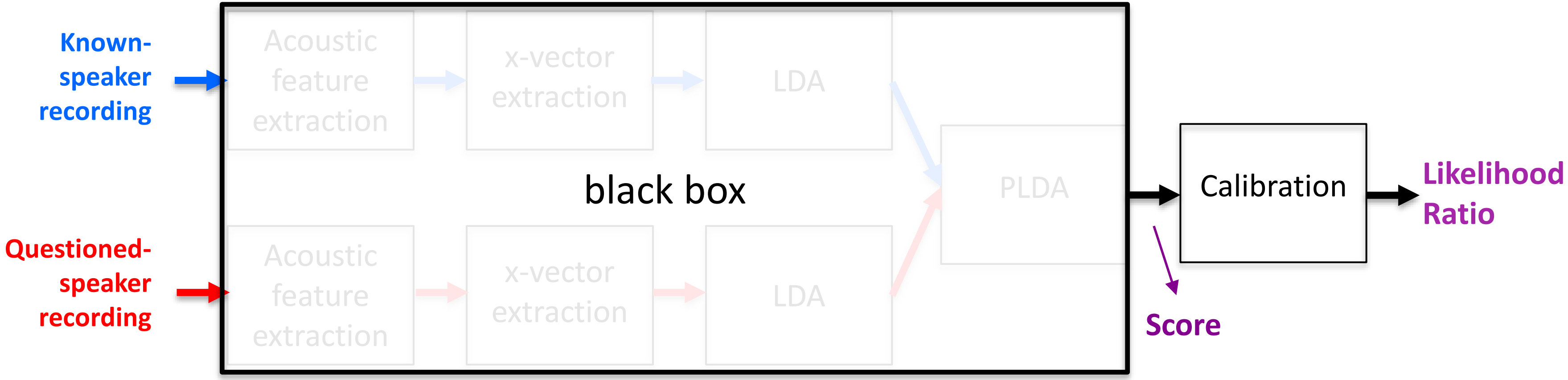
voices are similar

but very common

voices are similar

but very atypical

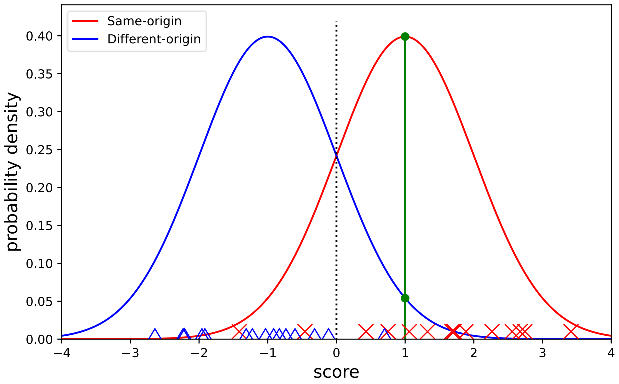
System calibration



Front-end modelling
(high-level feature extraction)

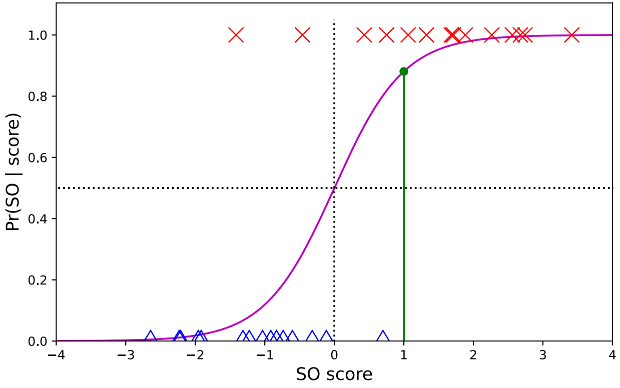
Back-end modelling

System calibration



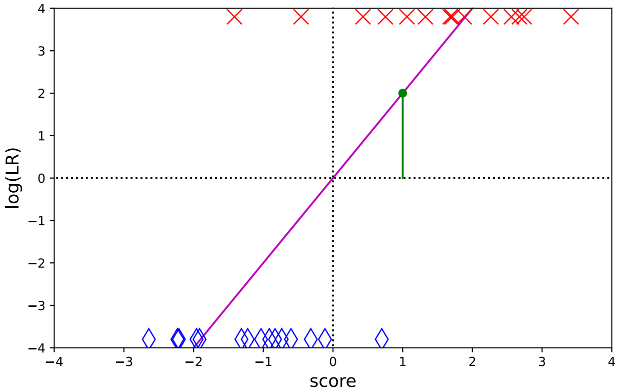
Same-score vs different score

Same-variance Gaussians



Logistic regression

$$\Pr(\text{same}|\text{score}) = \left(\frac{f(\text{same})}{f(\text{same}) + f(\text{different})}\right)$$



Score → Log LR mapping

Shift and scale

Specific concerns in forensics

Trust

Image source StockMonkeys.com [http://www.stockmonkeys.com]

Forensic voice comparison

- **Critical**
  - Court must be able to trust the output.
  - The likelihood ratio must mean what it says.
- Implications for data and processes

Historically … subjectivity

- Historical issues – spectrogram reading, voiceprint, auditory-phonetic, acoustic-phonetic, …
- Pseudoscience – bitemarks
- Over-reliance on "experts", training, procedures, faulty statistics
- "Identification" *vs* strength of evidence **(CSI *et al.*)**
- Confusion of the role of the expert and of the court
- Faulty processes introducing cognitive bias
- …

Forensic voice comparison … objectivity

- **Critical**

  1. Use of data

  2. Calibration

  3. Avoidance of cognitive bias

  4. Validation of the system under the conditions of the case

Critical – 1. use of data

- Must use relevant data
  - estimating the different-speaker (defence) hypothesis
  - to estimate typicality
- Must **train** (and/or adapt) with relevant data
  - Relevant population
  - Relevant recording conditions

  - **Subjective decisions**
  - **Collect or simulate data**

Critical – 2. calibration

- You might calibrate someone's prediction (weather, football, lottery, …)
- Treat the whole system as a black box
- Train a parsimonious model (few parameters)
- **Known** same-source, different-source **pairs**
  population and conditions reflect the case

Critical – 3. avoidance of cognitive bias

- report the strength of evidence : likelihood ratio
- move the human as early in the pipeline as possible
- separation of duties
- careful pipeline for processing case data

Critical – 4. black-box validation

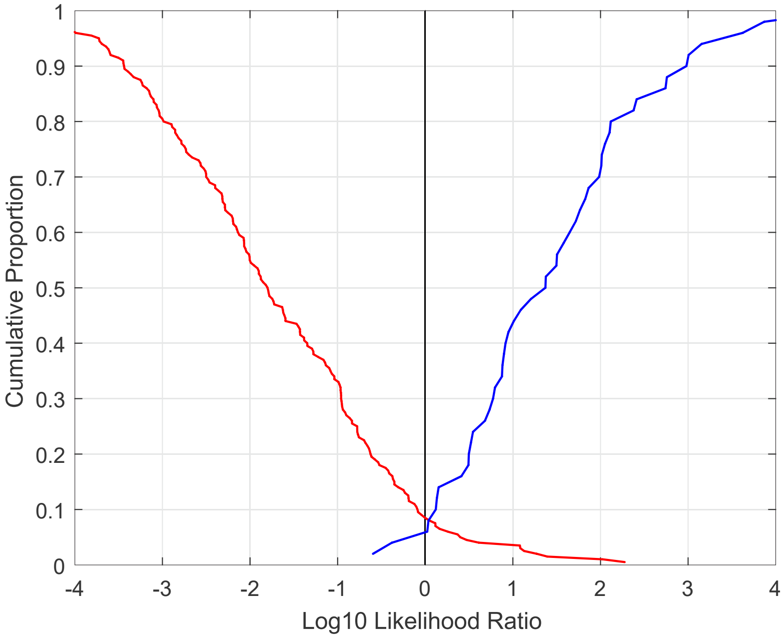
visualise and measure performance

on

known data

same-source pairs

different-source pairs



$$C_{llr} = \frac{1}{2}\left(\frac{1}{N_{so}} \sum_{i=1}^{N_{so}} \log_2 \left(1 + \frac{1}{LR_{so_i}} \right) + \frac{1}{N_{do}} \sum_{j=1}^{N_{do}} \log_2 \left(1 + LR_{do_j} \right) \right)$$

Aston University
BIRMINGHAM UK

Discussions in the context of AI

Bias and humans

Bias in AI

- There’s no intelligence here **(?)**

- All the **intelligence** in the **system** comes from the human
  - data (population) selection
  - training
  - Interpretation

- But society has valid concerns about bias in AI systems
- (and interpretability)

Bias in AI

- There's no intelligence here **(?)**

- All the **problems** in the **system** come from the human
  - data (population) selection
  - training
  - Interpretation

- But society has valid concerns about bias in AI systems
- (and interpretability)

Bias


facebook


© © AFP/GETTY IMAGES


STOP


CODED BIAS
full frame
hotdocs
sundance
SFFILM FESTIVAL
SXSW 2020

Interpretability

“If you can’t explain it simply, you don’t understand it well enough.”

Albert Einstein

“You can't depend on an AI system you don't understand.”

Interpretable AI

https://www.interpretable.ai/

Keys?

Data?

Validation?

Calibration?

Training and process?

Keys?

Data?

- Who selects it?
- When is it selected?
- Is it appropriate?
- How much is needed?
- …

Validation?

Calibration?

Training and process?

Keys?

Data?

Validation?

- Demonstrate performance to engender trust
- Is it always possible?
- Does interpretability then matter?
- …

Calibration?

Training and process?

Keys?

Data?

Validation?

Calibration?

- We could calibrate the human!
- Is it always possible?
- Can it correct for bias?
- Does interpretability then matter?
- …

Training and process?

Keys?

Data?

Validation?

Calibration?

Training and process?

- How to use the system – garbage in = garbage out
- How to interpret the results
- How to avoid (cognitive) bias
- …

Keys?

Data?

Validation?

Calibration?

Training and process?

These are all human factors in our (development of, use of) AI

Aston University
BIRMINGHAM UK

The end

Thank you!

Phil Weber – p.weber1@aston.ac.uk

Research Fellow – Aston University

Forensic Speech Science Laboratory (FSSL)

Forensic Data Science Laboratory (FDSL)

Aston Institute for Forensic Linguistics (AIFL)

Computer Science