Impact of mismatches in long-term acoustic features on different-speaker ASR scores

Chenzi Xu, Paul Foulkes, Philip Harrison, Vincent Hughes, Poppy Welch, Jessica Wormald, Finnian Kelly and David van der Vloed

IAFPA 2023

Netherlands Forensic Institute
Ministry of Justice and Security

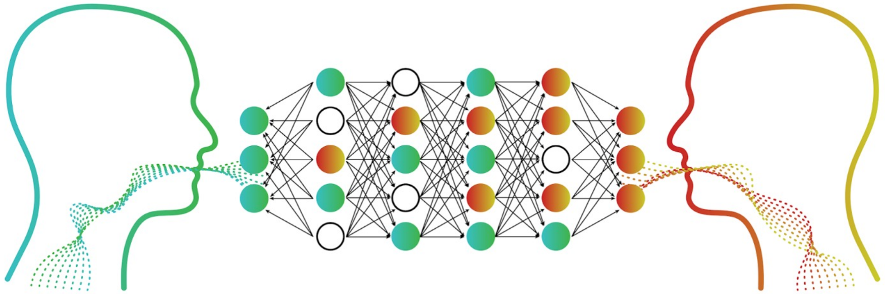
OxfordWaveResearch

UNIVERSITY of York

E·S·R·C
ECONOMIC & SOCIAL RESEARCH COUNCIL

Person-specific ASR: understanding the behaviour of individuals for applications of ASR

ESRC; 2022-25



Netherlands Forensic Institute
Ministry of Justice and Security

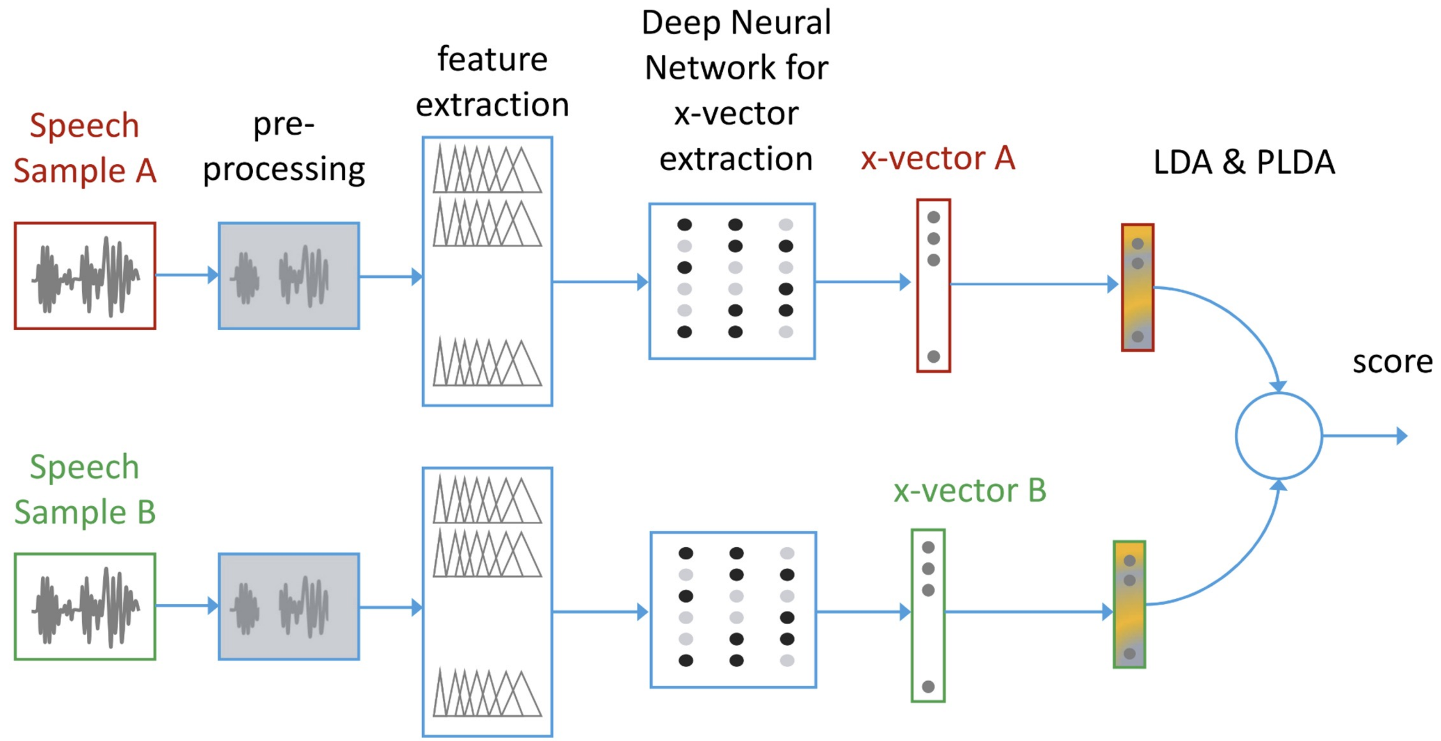
OxfordWaveResearch

- **Work Package 1:** Small-scale, analysis of controlled recordings produced by phoneticians. Systematic variation in vocal conditions (e.g. voice quality, accent guises)
- **Work Package 2:** Large-scale analysis of speakers from UK Government databases, involving 1000s of speakers. Identifying ‘problematic’ speakers and correlating performance with linguistic and demographic factors
- **Work Package 3:** What do we do about this? Developing solutions to issues raised in WPs1-2, via e.g. data augmentation, fusion with other features

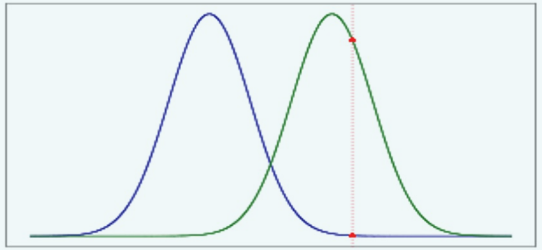
Outline

1. Background: the ASR pipeline
2. Motivations and Research Questions
3. Method
    a. Speech data: A subset of UK Government database
    b. Acoustic measurements
    c. Regression model
4. Analysis
    a. Mean DS scores for all speakers
    b. Effects of acoustic mismatches

1. Background: the ASR Pipeline



Calibration



$$LR = \frac{p(s\,|H_{ss})}{p(s\,|H_{ds})}$$

2. Motivations

ASR systems:

- Optimized to make accurate predictions for given data, **on average.**
  - Performance may **vary** across speakers or trials.
- Model **combined variations** in speaker, channel, content, duration, and other factors.
  - Challenges with **unseen** microphones, environments, speaking styles etc.
- May yield decisions **hard to interpret**.

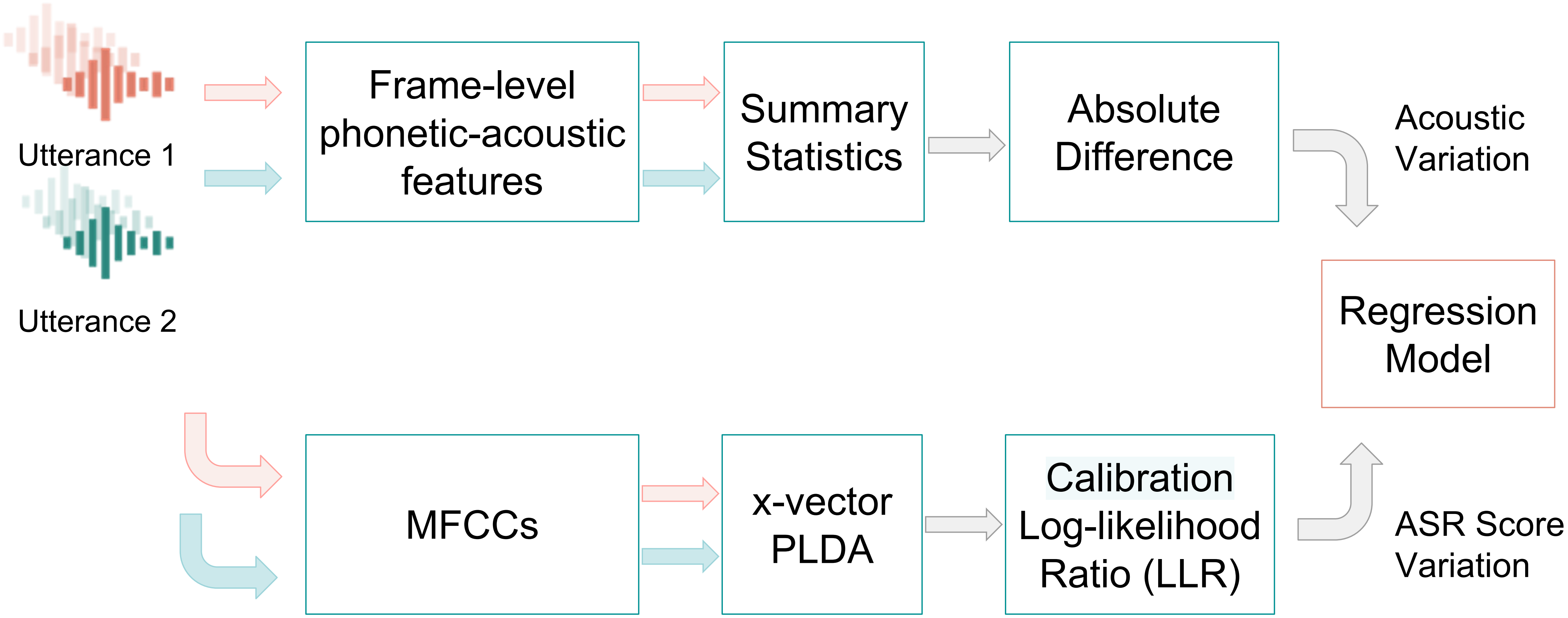
Forensic voice comparison is a high-stakes application: **Explainable decisions** are essential.

2. Research Questions

- Are there **systematic patterns** in ASR output depending on **acoustic** properties of speakers?

- How can scores be **explained** by differences in acoustic measures of compared speech?

3. Method Overview



3. Speech Data

Dataset

- **155 male** anglo speakers
- UK Government database (one recording per speaker)
- Mobile phone conversations (8kHz, single channel)
- **London** accent
- 3 age groups: 18-34 (65), 35-49 (59), over 50 (31)

Advantages:

- Forensically realistic input (all **spontaneous speech**)
- Limited variability in technical conditions (all **mobile phone**)
- Limited variability in accents (all **London** accent)

3. Speech Data

Dataset

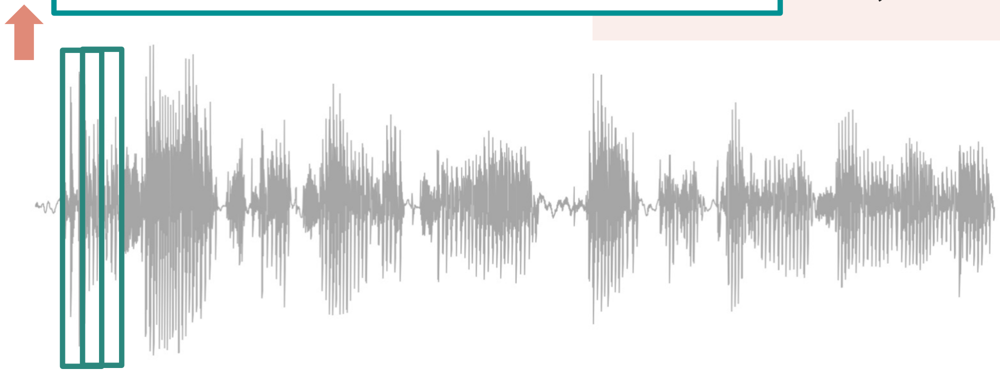
- **155 male** anglo speakers
- UK Government database (one recording per speaker)
- Mobile phone conversations (8kHz, single channel)
- **London** accent
- 3 age groups: 18-34 (65), 35-49 (59), over 50 (31)

Calibration dataset

- **20 male** speakers
- GBR-ENG corpus (two recordings per speaker)
- Mobile phone conversations (8kHz, single channel)
- Both parents born in **London**
- Ages: 18-43

3. Acoustic Measurements

- Temporal features: Length of continuously voiced regions
- Voice Quality measures: HNR, Jitter, Shimmer
- Vocal tract related measures: F1, F2, F3, B1, B2, B3
- Glottal and energy measures: f0, loudness

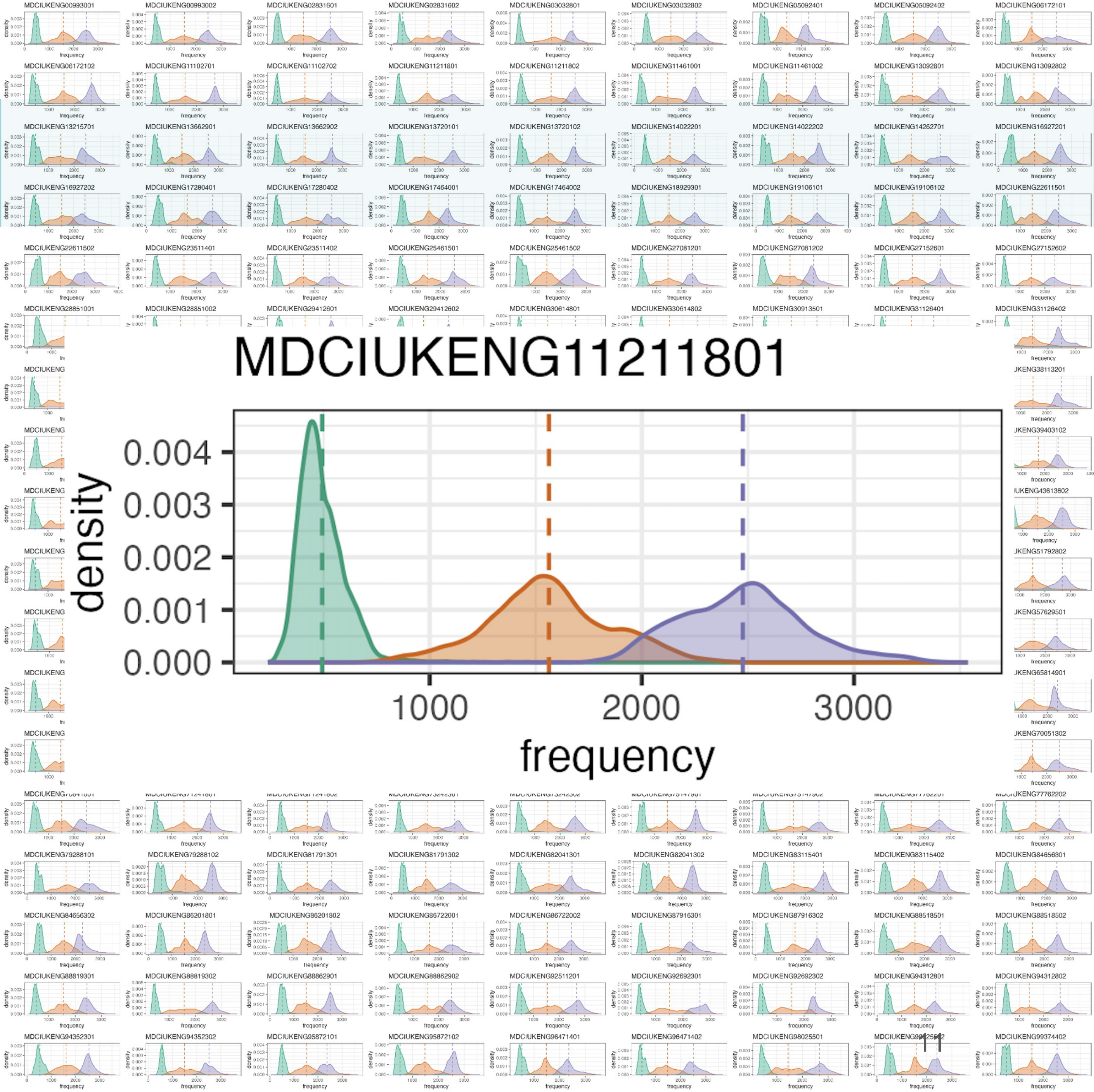


Acoustic distances

- Formants, bandwidths and f0 (all sonorant segments): **Praat**
- Others: **OpenSMILE**
- Two summary statistics (**mean** and **SD**) of these features
- **Standardisation** of means and SDs (z-score)
- Distance computation: **absolute difference** between standardised means and SDs

3. F0 and Formants

- Praat algorithm, implemented in Python
- Made use of forced alignment information (Montreal Forced Aligner)
- f0: 40-300 Hz
- Visual inspection: Long-term formants (F1, F2, F3) distribution plot for each speaker



3. Regression Model

Linear Mixed Effects Model (in lme4 syntax)

$$\underbrace{\text{Calibrated DS LLRs}} \sim \underbrace{\text{Acoustic Distances}} + \underbrace{(1|\text{speaker1}) + (1|\text{speaker2})}$$

Response Variable:

- The higher the LLR, the more confident the ASR system is that the speaker identities agree

Predictors/Fixed Effects:

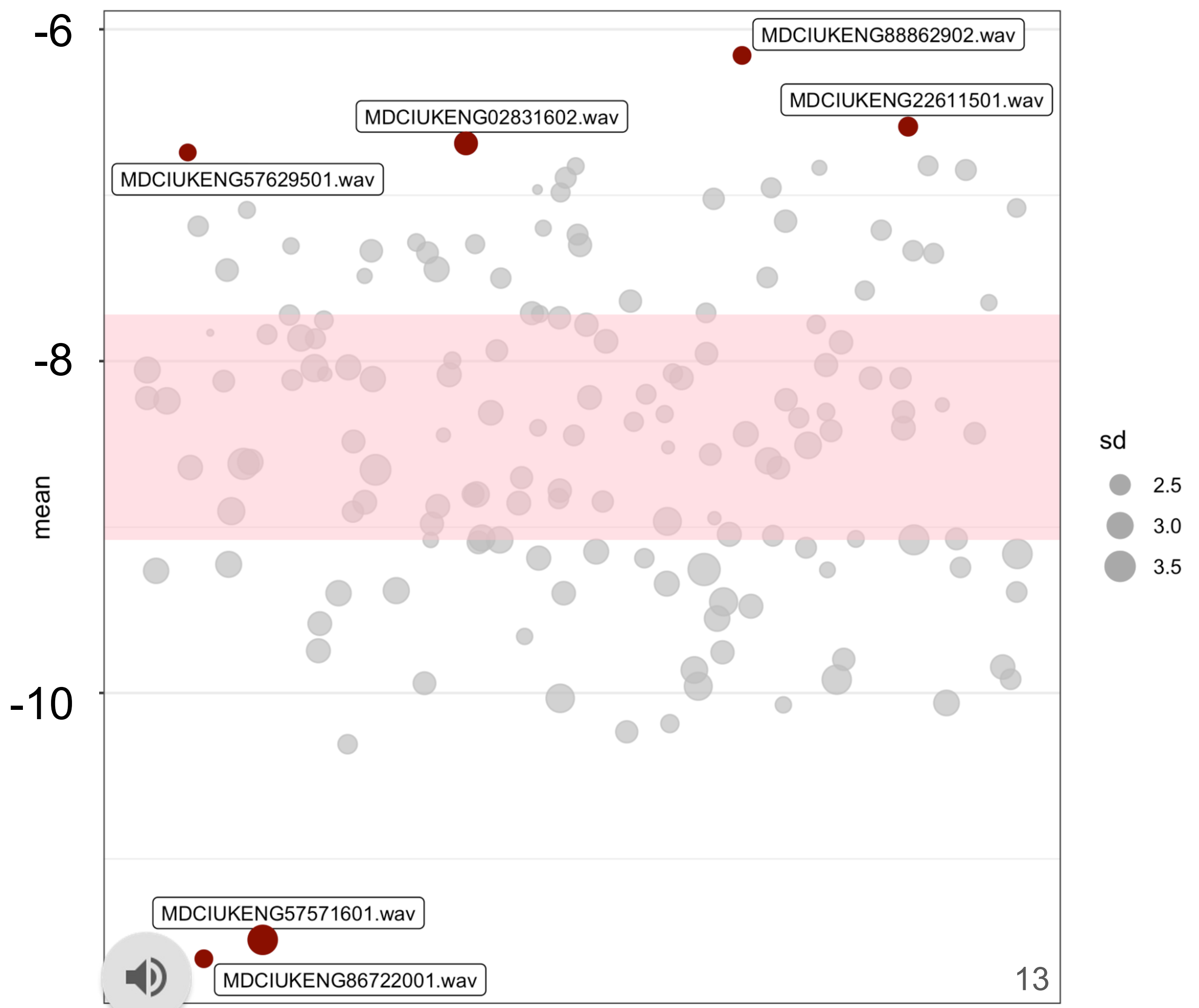
- Absolute difference of standardised means and SDs of 12 acoustic features
- The lower the distances, the more acoustically similar the two utterances are

Random Effects:

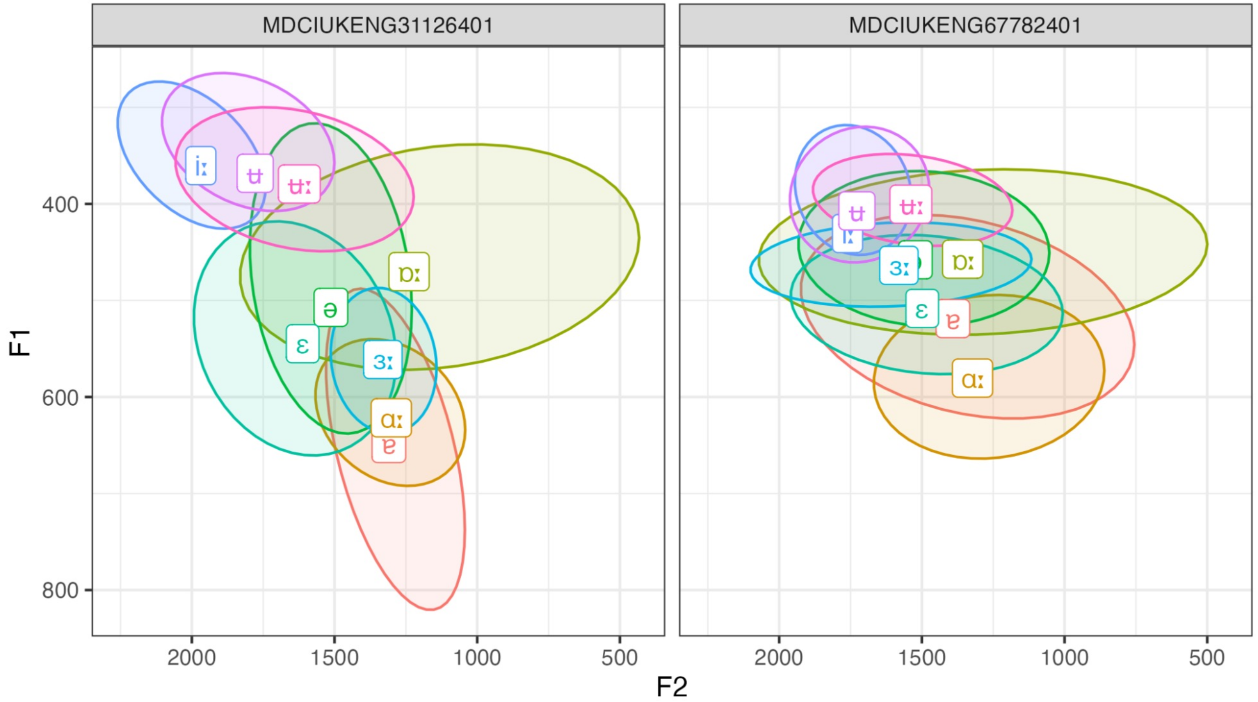
- Per-speaker *random intercept* as a variable with zero mean and unknown variance.
- Explicitly model the group structure: the same speaker appears multiple times

4. Overview: Mean DS LLRs of London Speakers

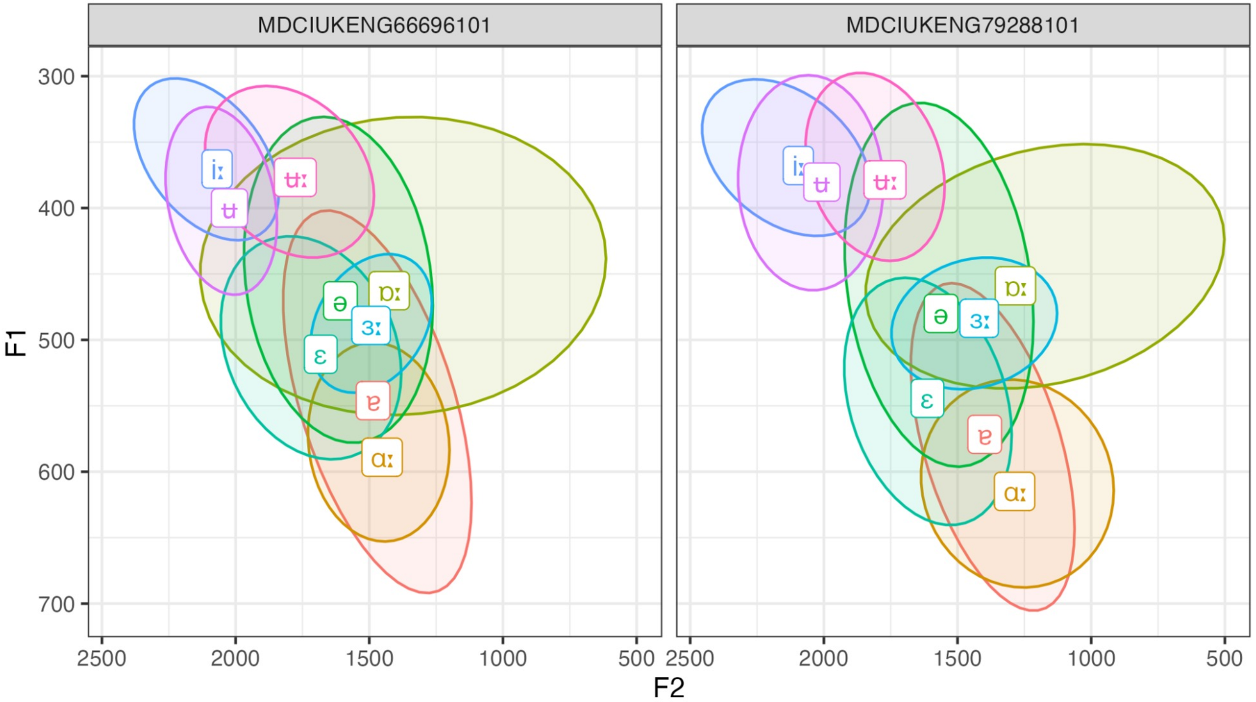
- Bayesian calibration with Jeffreys non-informative priors
- DS C_{llr}: **0.0152**
- 0.15% of the pairs (18/11935) had a positive calibrated score (i.e. contrary-to-fact support to a same-speaker decision)
- Suspicious pairs / voice twins:



4. Formant Frequencies Mismatches



LLR = −18.93



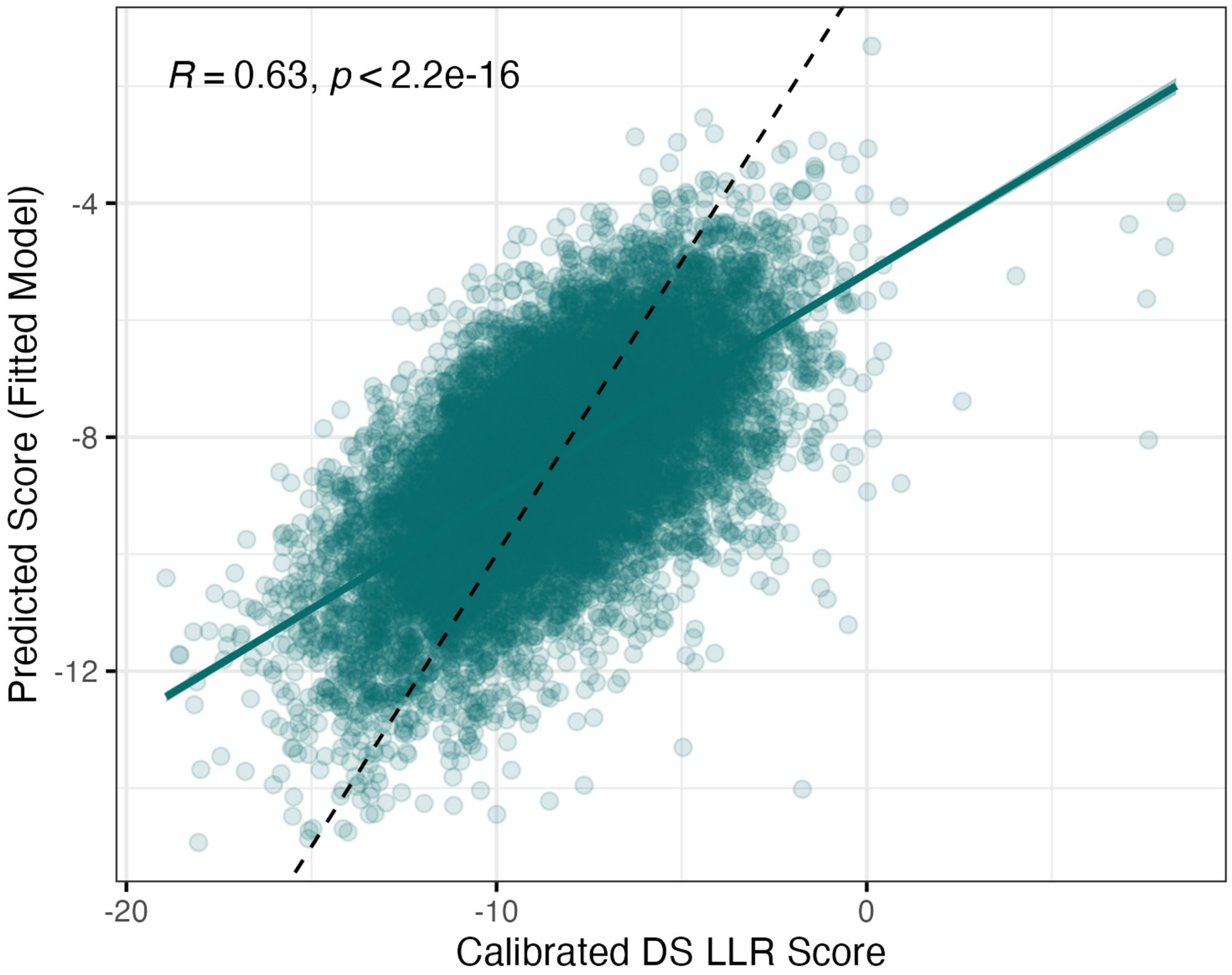
LLR = 8.36

4. Mixed Effect Model as a whole

How good is the mixed effect model at explaining the variation in LLR score?

- Pearson Correlation between between the model fitted values and the LLR scores: ***r* = 0.63**
- Marginal R^2 = 0.243 / Conditional R^2 = 0.472

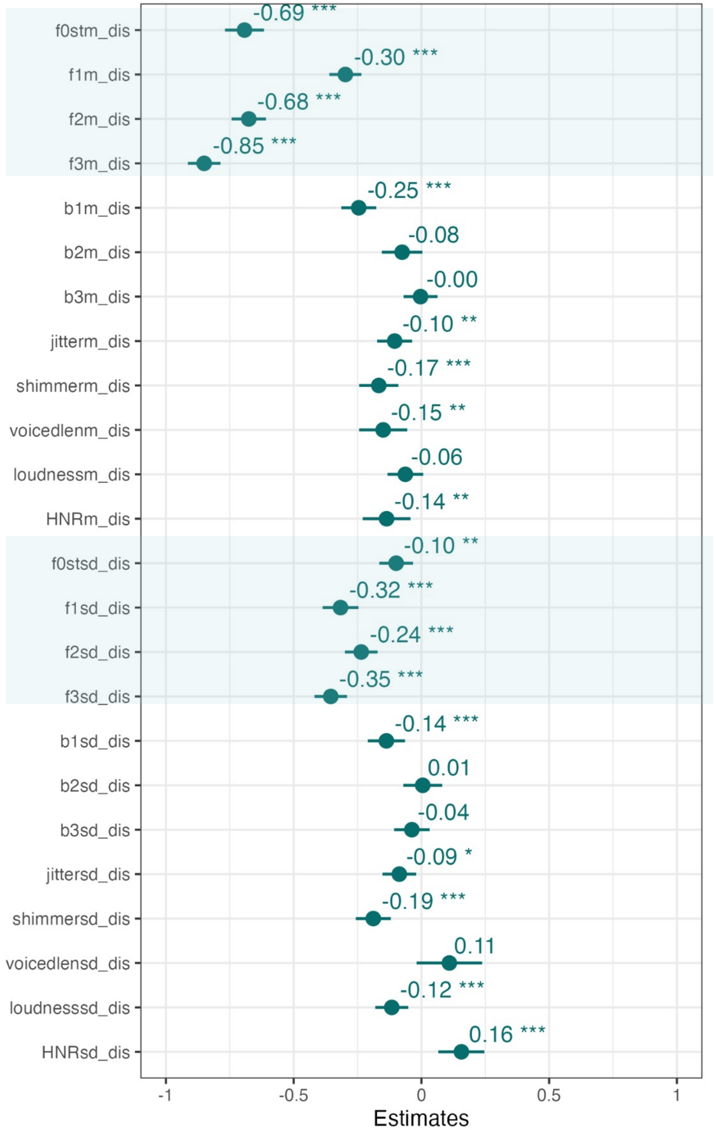
⟶ The presence of unmodelled effects



4. Effects of Acoustic Mismatches

Statistically significant fixed effects:

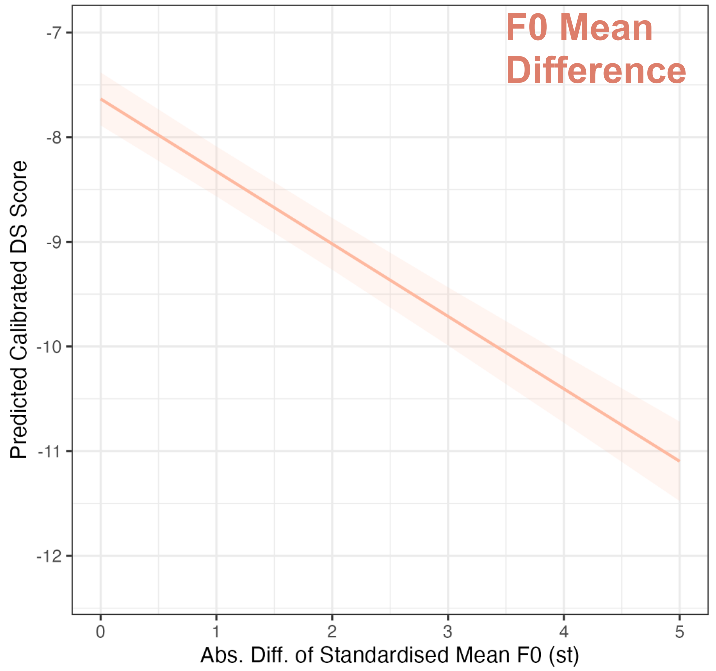
- Long-term average difference: **F3, F0, F2, F1,** B1, Shimmer, length of continuously voiced regions, HNR, Jitter
- Long-term SD difference: **F3, F1, F2**, Shimmer, HNR, B1, Loudness, **F0**, Jitter
- Most coefficients are **negative**: the larger the acoustic distance, the lower the calibrated DS LLRs.

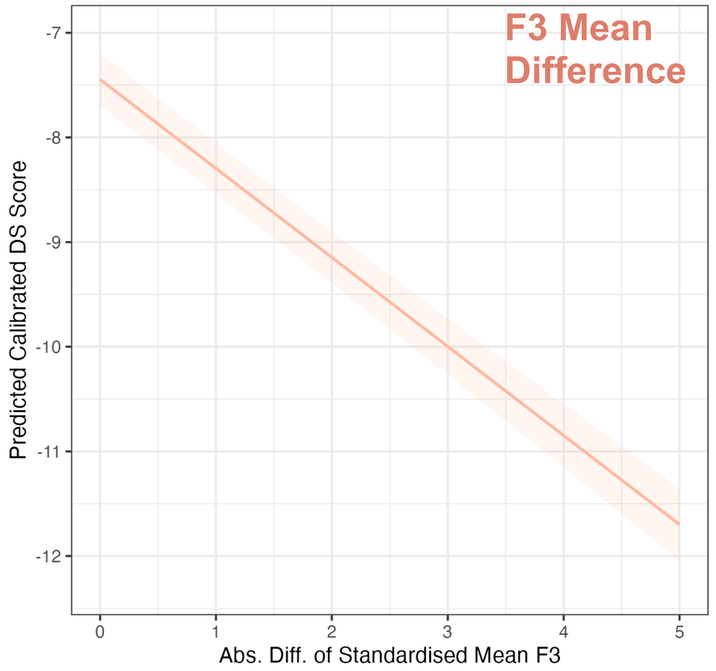


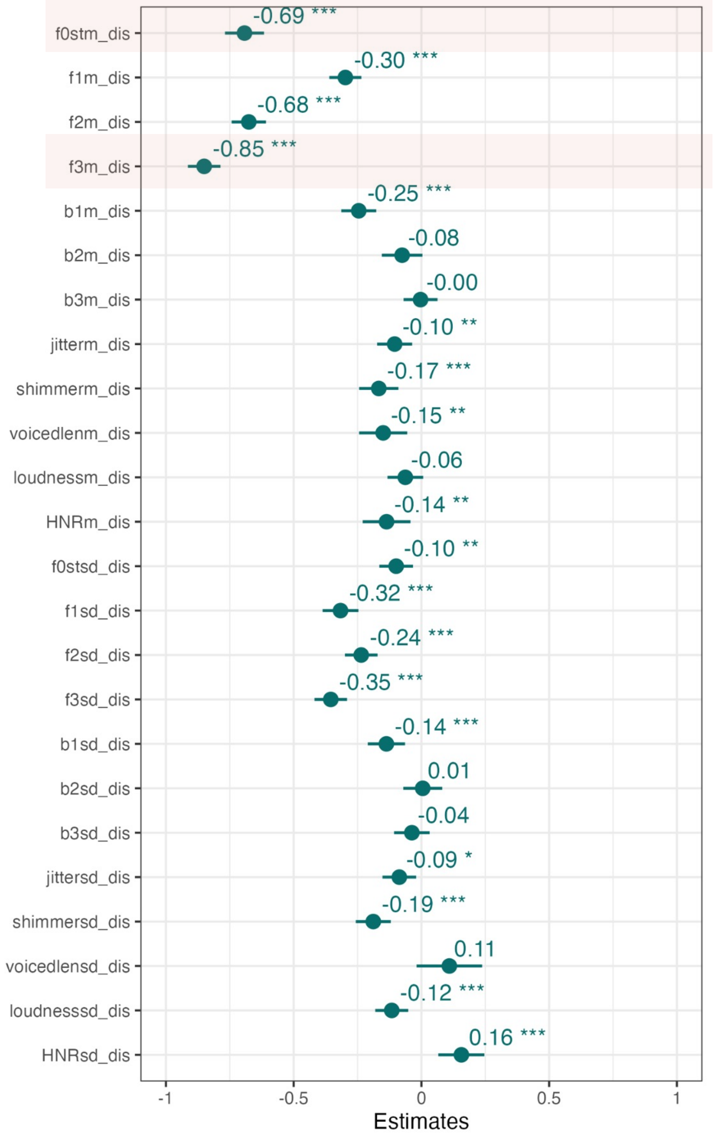
F0 and Formants (mean)

F0 and Formants (SD)

4. Effects of Acoustic Mismatches







- **Negative**: a larger mean F0 / F3 difference predicts a lower calibrated DS LLR.
- Changing standardised mean F0 5 units predicts the LLR score to go down by 3.5 units.

Take-home Message

Inter-speaker acoustic mismatches are negatively correlated with ASR scores.

- **F0 and formant frequencies**-related mismatches (both mean and SD) have the greatest explanatory power in LLR scores.
- The **average F3 difference** is individually the most important feature: usually most sensitive to the tip of the tongue and lip rounding.
- First formant bandwidth (B1), Jitter, and Shimmer-related mismatches (both mean and SD) also contribute to explain the LLR scores.

→ Ultimately help towards enhancing explainability to ASR system

Questions and Comments

{chenzi.xu | paul.foulkes | philip.harrison |

vincent.hughes | poppy.welch | jessica.wormald}@york.ac.uk

@ChenziAmy

chenzixu.rbind.io

Netherlands Forensic Institute
Ministry of Justice and Security

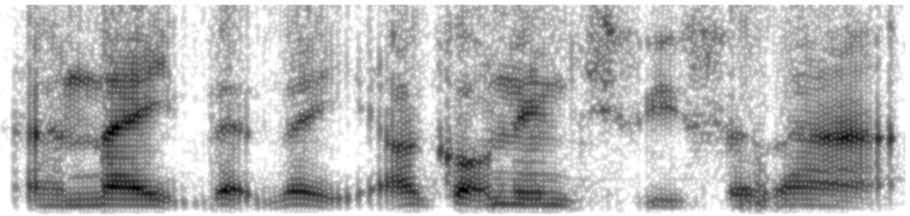
OxfordWaveResearch

UNIVERSITY of York

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Automatic Forced Alignment

Audio recording

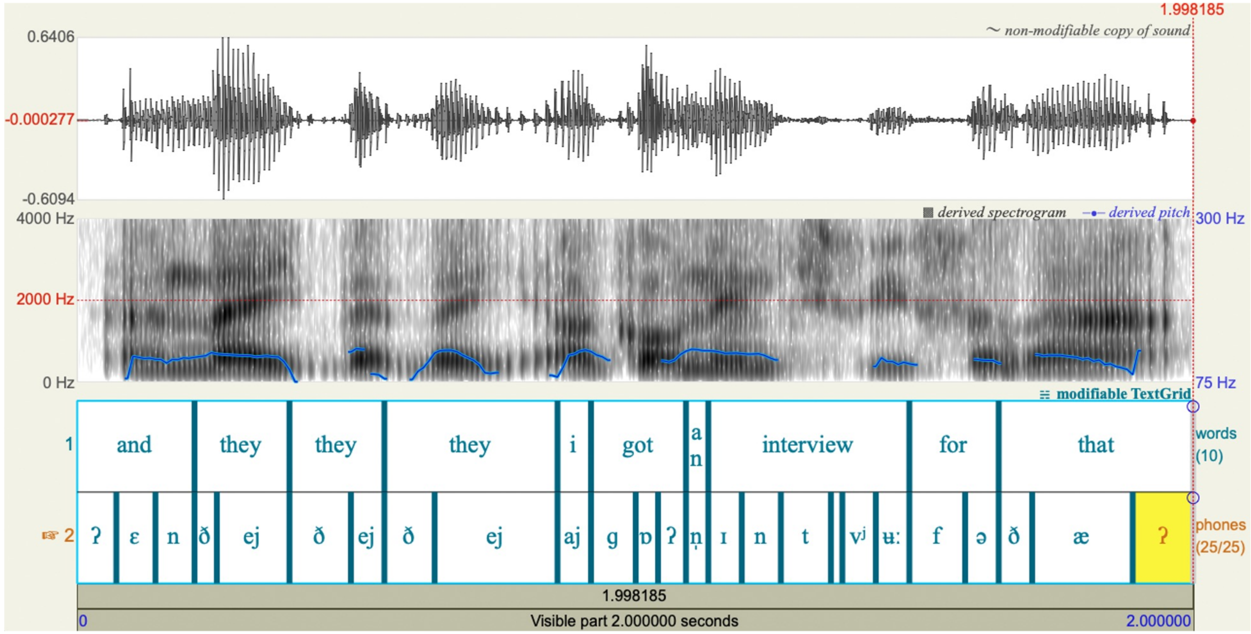


Acoustic vectors

Acoustic models

Pronunciation dictionary

Words	Phonemes
that	**ð æ ʔ**
they	**ð ej**
interview	**ɪ n t ə vʲ ʉː**



Evaluation:
Likelihood score

Transcript

…and they they they I got an interview for that…

Data Cleaning Workflow

Sanity Check → **Metadata Management** → **Automatic Forced Alignment** → **File Renaming and Organisation**

Sanity Check

- ✓ Unique identifier (duplicates)
- ✓ Total file number by corpus
- ✓ Any missing audios
- ✓ Any missing transcripts
- ✓ Any exceptionally short audios
- ✓ Any problematic timestamps in the transcripts

sanche.py

Metadata Management

- ✓ Gather various spreadsheets
- ✓ Use consistent formats
- ✓ Encode questionnaire
- ✓ Aggregate the metadata of all corpora

metadata.ipynb

Automatic Forced Alignment

- ✓ Organise working directory
- ✓ Set up Montreal Forced Aligner (MFA)
- ✓ Generate input Textgrids from transcripts
- ✓ Trace Out-of-Vocabulary items (OOVs) and fix typos
- ✓ Update pronunciation dictionary
- ✓ MFA alignments with multiple sets of parameters
- ✓ Evaluation of outputs

mfa_align.job

File Renaming and Organisation

- ✓ Generate new filenames using metadata
- ✓ Format: corpus code, participant number, session, repetition, speaking condition, and microphone type, separated by "_"

rename.py

pasr-forced-alignment
pasr-ho-documentation

Mean DS Scores (Vowels-only)

- Bayesian calibration with Jeffreys non-informative priors
- DS C_{llr}: **0.3301**
- 10% of the pairs (1178/11925) had a positive calibrated score (i.e. contrary-to-fact support to a same-speaker decision)