Accounting for a six year time difference between questioned and known speaker recordings in a forensic voice comparison case

Geoffrey Stewart Morrison
Finnian Kelly

Aston University

\[
\frac{p(E|H_p)}{p(E|H_d)}
\]

Forensic Evaluation

OxfordWaveResearch
Case

- Questioned-speaker recording made in 2011

- Known-speaker recordings made in 2017 (6 year time interval)

- Recordings of ~100 speakers sampled from relevant population (sample speakers):
  - multiple recordings of each speaker
  - hours to days apart

- Sample-speaker recordings used for:
  - training
  - testing (empirical validation under casework conditions)
  - 50-50 split + cross validation
Forensic voice comparison system

- Features: **MFCCs + deltas + double deltas**

- Feature domain mismatch compensation: **CMS**

- i-vector extractor: **UBM + T matrix (Vocalise pre-trained models)**

- i-vector domain mismatch compensation: **LDA trained using sample-speaker recordings (set 1)**

- i-vector to score model: **PLDA trained using sample-speaker recordings (set 1)**

- Score to likelihood ratio (calibration) model: **logistic regression trained using sample-speaker recordings (set 2)**
Problem

- Within-speaker variability generally increases with time interval between recordings

- Training and testing on sample-speaker recordings made hours to days apart will give misleadingly good results compared to if there were a 6 year time interval

- Likelihood ratio calculated for comparison of the questioned and known speaker recordings may be highly misleading
Observation

• In Kelly & Hansen (2016)
  – as time interval increased
  – same-speaker score values decreased

Solution

• Decrease the short-interval same-speaker score values so that they approximate the values expected for the longer target interval
  – similar to within source degradation (González-Rodríguez et al., 2006)


Training

- Multisession Audio Research Project (MARP) corpus
- Recordings from 46 male speakers
- Recorded approximately once every two months over a period of three years + one additional recording session approximately ten years after start
- **Different-speaker scores**: all possible session 1 v session 2 comparisons
- **Same-speaker scores**: all possible comparisons over all possible time intervals (excluding same-session comparisons)
Training

- Calculate trimmed mean for different-speaker score distribution
- Calculate trimmed mean for same-speaker score distribution for each time interval
- Fit weighted linear regression with exponential link function
  - distance between the same-speaker and different-speaker means
  - time interval
Application

- Read off distance between the ss and ds means at short interval
- Read off distance between the ss and ds means at longer target interval
- Calculate proportional reduction in the distance between the means
- Apply that proportional reduction to the casework same-speaker scores
  - the casework data do not have the same conditions as the MARP data or have the same absolute ds and ss score means, hence a proportional reduction in the difference between the ss and ds means is applied
  - applied to training and test ss scores, not to questioned v known speaker score
Testing

- Cross-validation on MARP data
- Leave out 2 month interval plus target interval
- Leave one speaker out

Results

- RMS error rate expressed as a percentage of the original distance between the different-speaker and same-speaker means: 4.51%
- Worst per-interval error (at 34 month interval): 8.53%
ss 2 months
ss 102 months
ss 2 months adjusted to 102 months
Additional work needed

- Validation work so far has been cross-validation on the MARP data
- Actual application is to case data that have conditions that differ from those of MARP
- Cross-database validation is needed
Thank You

http://geoff-morrison.net/