



# IAFPA 2025

---

THE HAGUE, THE NETHERLANDS

BOOK OF ABSTRACTS

JULY 2025





**Universiteit  
Leiden**  
Centre for Linguistics



Netherlands Forensic Institute  
*Ministry of Justice and Security*



Immigration and Naturalisation  
Service  
*Ministry of Asylum and Migration*

## Welcome

Dear IAFPA friends,

As the former Rector of Leiden University, I am delighted to welcome you to our campus in The Hague. This is the 33rd Conference of the International Association for Forensic Phonetics and Acoustics, which highlights the strong position your interdisciplinary field has secured within the scientific landscape.

Leiden University is a university in two cities. One is in Leiden, the city where it all began exactly 450 years ago. Since its founding in 1575, Leiden has been a place where languages and cultures from around the world have been studied - a tradition that continues to this day. In recent decades, a second campus has been established in The Hague. The Hague is home to our government, ministries, the Dutch Supreme Court, the International Court of Justice, and the International Criminal Court. It also hosts numerous scientific and scholarly institutes, such as the Netherlands Forensic Institute (NFI), as well as government services, including the Immigration and Naturalisation Service (IND). Strangely enough, it is a city without its 'own' university.

In close collaboration with the city, Leiden University established a campus in The Hague in 1998, which now has almost 8,000 students. This facilitates the transfer of scientific knowledge to national and international organizations, while also allowing the university to learn from the questions and challenges posed by society.

I am especially pleased to write this foreword because I have known Tina Cambier-Langeveld, the President of your association, for many years. We are both members of Leiden Athletics and got to know each other while running. Over the years, I have witnessed the collaboration between NFI, IND and Leiden University flourish. When I stepped down as Rector in 2021, Tina, Willemijn Heeren, and I had a wonderful discussion about the importance of your scientific work in forensic phonetics: see <https://www.universiteitleiden.nl/en/news/2021/01/linguists-crimefighters-extraordinaire>

That conversation once again underscored the great significance of their work and your work.

I wish you all an inspiring scientific conference and, above all, a wonderful time in The Hague - and of course, a fantastic dinner at the beach, which I hear will be in Scheveningen. Interestingly, this word 'Scheveningen' - a pronunciation challenge for non-Dutch speakers - was once used as a shibboleth during the Second World War to distinguish native Dutch speakers from German spies.

Enjoy!

Prof. Carel Stolker  
Rector and Chair of Leiden University (2013-2021)

## Special Sessions and Workshop

### Special Session on Quality

- What the forensic sciences standard ISO 21043 has to offer to Forensic Phonetics and Acoustics  
*Meuwly, Didier* 9
- Quality Matters  
*Rhodes, Richard, Bryony Nuttall, Katherine Earnshaw and Megan Thomas* 10
- Bridging the gap: Expanding the scope of academic engagement to support practitioners  
*Wormald, Jessica, Vincent Hughes, and Joe Pattison* 12

### Special Session on Language Analysis in Asylum Cases

- Project CELIA: Common European Language Indication and Analysis  
*Cambier-Langeveld, Tina* 13
- Y-ACCDIST applied to Arabic  
*Brown, Georgina and Sam Hellmuth* 14
- Picture Naming Task in LADO  
*Gottschligg, Peter* 16

### Workshop

- Building your own LR system  
*Van Lierop, Stijn and David van der Vloed* 18

## Oral presentations

- Assessing the suitability of forensic authorship analysis methodologies for speech data  
*Tompkinson, James and Andrea Nini* 19
- Interpreting forensic transcriptions: from ambiguity to accuracy  
*Di Nunno, Christian* 20
- Toward an improved American English word list, with frequencies and transcription  
*Disner, Sandra F., Vincent J. van Heuven and Eileen Yang* 22
- Identifying Voice Super-Recognizers in Law Enforcement  
*Fröhlich, Andrea, Peter French, Meike Ramon and Volker Dellwo* 24
- Relevance of the distinctions score-based / feature-based and common source / specific source for forensic voice comparison  
*Jessen, Michael* 26
- Language proficiency as a useful index in predicting individual performance of trilingual speakers in cross-language forensic voice comparison using an automatic speaker recognition system  
*Cao, Grace W. L., Vincent Hughes, Justin J. H. Lo and Peggy Mok* 28
- The effect of a language mismatch: Strength-of-evidence in multilingual forensic speaker comparison  
*De Boer, Meike and Willemijn Heeren* 30

Forensic Analysis of the Singing Voice <i>Zuim, Ana Flavia, Dominic Watt, Geddy Warner</i>	<b>32</b>
Long-term formant measurement in casework: The more formants, the better? <i>Moos, Anja, Michael Jessen, Katharina Klug, and Almut Braun</i>	<b>34</b>
Between-speaker variability in information flow rate through temporal changes of spectral composition in speech signals <i>He, Lei and Bruce Wang</i>	<b>36</b>
Assessing the suitability of f0 estimators with respect to recording condition and voice quality <i>Klug, Katharina and Markus Niermann</i>	<b>38</b>
Properties of ENF interference in audio deepfakes <i>Schutten, Molly and Amelia Gully</i>	<b>40</b>
Empirical study on the application of forensic audio authentication evidence in Chinese courts (2016-2020) <i>Cao, Honglin and Danyang Li</i>	<b>42</b>
Towards an interpretation framework for forensic audio deepfake detection <i>Kelly, Finnian, Anil Alexander, Anna Bartle, Colleen Driscoll and Peter Milne</i>	<b>44</b>
FAUXDIO: An audio deepfake detector for law enforcement and forensics <i>Alexander, Anil, Linda Gerlach, Thomas Coy, Oscar Forth, Liam Lonergan and Finnian Kelly</i>	<b>46</b>
What can voice quality features do in detecting deepfake speech in forensic scenarios <i>Jintao, Kang, Jin Tian and Peng Cheng</i>	<b>48</b>
How Do Expert Witnesses Survive the Hot Seat? Discourse Strategies of Expert Testimony in Cross-Examination <i>Yang, Yuyao</i>	<b>50</b>
Teaching practices in Forensic Language Analysis <i>Tompkinson, James</i>	<b>52</b>
Thank you for watching: automatically evaluating transcriptions for hallucinations and missing meaning <i>Virji, Jadd and Finnian Kelly</i>	<b>54</b>
Discovery and retrieval of speakers from large unlabelled datasets using scalable clustering <i>Coy, Thomas, Finnian Kelly and Anil Alexander</i>	<b>56</b>
LiRI – voxplorer: an interactive dashboard to extract, visualise, and interact with feature-rich large speech corpora <i>De Luca, Alessandro, Srikanth Madikeri, and Volker Dellwo</i>	<b>58</b>
Variability in the performance of automatic speaker recognition systems across modelling approaches <i>Harrington, Lauren, Vincent Hughes, Philip Harrison, Paul Foulkes, Jessica Wormald, Finnian Kelly and David van der Vloed</i>	<b>60</b>

## Poster sessions

### Session 1

Multiple Enrollments and Neural Back-End Modeling for Automatic Speaker Verification <i>Paulsson, Aron, Torbjörn Onshage, Greta Öhlund Wistbacka, Susanna Whitling and Andreas Jakobsson</i>	62
Engaging with Government and Police stakeholders regarding the use of speech technology in UK investigative interviewing <i>Wormald, Jessica, Lauren Harrington, James Tompkinson and Eloísa Monteoliva García</i>	64
Assessing the ability of deepfake voice clones to produce accent and context-specific phonological features <i>Gibb-Reid, Ben, Vincent Hughes and Jessica Wormald</i>	66
Perception of deception through prosodic information in 911 emergency calls <i>Plante-Hébert, Julien and Lucie Ménard</i>	68
The grandparent scam – a perceptual study <i>Schedel, Sara-Sophie and Gea de Jong-Lendle</i>	70
Developing a collection of mock police interviews for use in research and teaching <i>Tompkinson, James, Lotte Eijk, Sarah Knight and Eloisa Monteoliva-Garcia</i>	72
Spectral characteristics of sibilant fricative /s/ in voice disguise via age modification <i>Ghaffarvand-Mokari, Payam</i>	74
Validating the auditory-acoustic phonetic method for forensic speaker comparison <i>Hughes, Vincent, Lauren Harrington, Philip Harrison, Finnian Kelly, David van der Vloed and Richard Rhodes</i>	76
Towards a Transparent and Interpretable Strategy for Spoofed Speech Detection <i>Lins Machado, Carolina, Xin Wang, and Junichi Yamagishi</i>	78
Introducing the International Network for Forensic Transcription: initial plans and overall aims <i>Harrington, Lauren</i>	81
Voice quality across a speaker’s languages: investigating the case of L1 Persian – L2 English <i>Shalpush, Janan, Willemijn Heeren, and Niels O. Schiller</i>	82
Distinctiveness vs. Deception: Assessing the success of deepfake technologies for perceptually distinctive voices <i>Bradshaw, Leah</i>	84
Analysing the accuracy of the public’s perception on characteristics of synthetic voices and how the concept of “expertise” affects the accuracy of AI identification <i>Verry, Emily, Ben Gibb-Reid and Amelia Gully</i>	86
Sound Judgments: ASR Accuracy and Trust in ROTI Transcripts <i>Pepper, Lorimer, Paul Foulkes, and Lauren Harrington</i>	88
A new resource for Cantonese Forensic Voice Comparison <i>Wang, Bruce X., Shuming Huang, and Lei He</i>	90

Sent to Coventry: A Study of Accent Perception in British Midlands Varieties <i>Broadhurst, Erin, Paul Foulkes, and Ben Gibb-Reid</i>	92
Evaluating state-of-the-art generators and detectors of audio deepfake <i>Lee, Daniel Denian, Linda Gerlach, and Kirsty McDougall</i>	94
Reference Population Effects on Automatic Speaker Recognition Performance <i>Möller, Sophie, Andrea Fröhlich, Sarah Lim, Adrian Leemann, and Gea de Jong-Lendle</i>	96
Casework Conundrums <i>Thomas, Megan, Katherine Earnshaw, Bryony Nuttall, and Richard Rhodes</i>	98
<b>Session 2</b>	
Visualising latent representations: An interactive approach to improving the linguistic interpretability of MFCC-based phoneme models <i>Williams, Samantha</i>	100
Forensic phonetic analysis of spoofed European Portuguese speech <i>Almeida, Lina, Amelia Gully and Paul Foulkes</i>	102
The more the better: assessing formant-based features for speaker differentiation with random forest--a pilot study <i>Jintao, Kang, Gao Kai and Huang Wenlin</i>	104
Phonological Familiarity and Voice Discrimination: A Study on Quranic Reciters without Arabic Comprehension <i>Azad, Sadia, Elisa Pellegrino, Eleanor Chodroff, Volker Dellwo</i>	106
An exploration of the other-accent effect in Quebec and Hexagonal French <i>Plante-Hébert, Julien and Pamela Bautista-Boivin</i>	108
The effect of sample size on the evaluation of speaker discriminatory power of diphthong /ei/ <i>Cao, Honglin, Xuehui Li, Danlin Wang</i>	110
New Conclusion Framework for the Forensic Speaker Recognition methodology in the Hungarian Audio Forensics <i>Fejes, Attila</i>	112
Speech discernment using signal analysis technology <i>Anderson, Terese and Grandon Goertz</i>	114
Authorship analysis on speech data as a counter to AI voice cloning <i>Visser, Anneke</i>	116
A Pilot Acoustic Study of Mandarin /ʂ/ in Southern Min Speakers: Implications for Forensic Speaker Comparison <i>Lin, Jinjin, Paul Foulkes, and Vincent Hughes</i>	118
Accent Copycats: How Accurately Can Non-Linguists Mimic a Known Voice? <i>Cope, Amy, and Ben Gibb-Reid</i>	120
Exploring the interpretability of deep speaker-representations from a phonetic perspective <i>Deng, Guangmou</i>	121

DSER: Dialog-Structure-Aware Metric for Speaker Diarization Evaluation in Forensics <i>Fries, Tim, David Grünert, Alexandre de Spindler, and Volker Dellwo</i>	<b>123</b>
Interaction of linguistic contrast and speaker specificity: an investigation into F3 and vowel rounding <i>Baker, Annie, Eleanor Chodroff</i>	<b>125</b>
The influence of bilingualism on voice quality in cross-language voice comparison <i>de Graaf, D.J.</i>	<b>127</b>
CANDORspeech: A large-scale corpus of phonetically annotated conversational speech from dyadic online conversations with human quality control <i>Vyshnevetska, Valeriia, Alessandro De Luca, Nadine Lavan, Carolyn McGettigan, Gus Cooney, Andrew Reece, and Volker Dellwo</i>	<b>129</b>
Speaker characteristics of categorical data on filler particles in German <i>Ishihara, Shunichi, Michael Jessen, and Beeke Muhlack</i>	<b>131</b>

# What the forensic sciences standard ISO 21043 has to offer to Forensic Phonetics and Acoustics

*Didier Meuwly*

*Netherlands Forensic Institute, The Hague, The Netherlands*

*Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, The Netherlands*

*d.meuwly@nfi.nl, d.meuwly@utwente.nl*

The forensic sciences standard ISO 21043 [1] is a methodological and technical standard, recently published as world-wide International Standard by the ISO Technical Committee 272 (TC272). It applies to all forensic disciplines, including forensic phonetics and acoustics. This standard has been designed by forensic, legal, and standardisation professionals to ensure that forthcoming forensic services, products, and systems are safe, reliable, and consistently perform as intended. The aim of this contribution is to present what this standard has to offer to the forensic phonetic and acoustic community.

After an introduction and a review of the existing international standards applied in forensic science, the five-part structure of ISO 21043 is described: 1. Terms and definitions, 2. Recovery, 3. Analysis, 4. Interpretation and 5. Reporting. Then a selection of requirements and recommendations of this five-part standard are highlighted, and their implication for forensic phonetic and acoustic research, development, implementation, and practice is discussed (Meuwly, 2024).

The conclusion is that the adoption and use of ISO 21043 will help forensic institutes to reach accreditation and to enhance their existing quality management systems. It will also make it easier for researchers to develop, deliver, and implement more relevant, more reliable, and safer services, products, and systems to the forensic phonetic and acoustic examiners performing casework.

## References

ISO 21043 Forensic sciences: <https://www.iso.org/committee/4395817/x/catalogue/>

ISO 21043-1: <https://www.iso.org/standard/69732.html?browse=tc>

ISO 21043-2: <https://www.iso.org/standard/72041.html?browse=tc>

ISO 21043-3: <https://www.iso.org/standard/72040.html?browse=tc>

ISO 21043-4: <https://www.iso.org/standard/72039.html?browse=tc>

ISO 21043-5: <https://www.iso.org/standard/73896.html?browse=tc>

Meuwly, D. (2024). Implications of the forthcoming forensic sciences standard ISO 21043 for forensic biometrics. In *2024 12th International Workshop on Biometrics and Forensics (IWBF)*, IEEE, 1–6.

## Quality Matters

Richard Rhodes<sup>1,2</sup>, Bryony Nuttall<sup>1,2</sup>, Katherine Earnshaw<sup>1,2</sup>, and Megan Thomas<sup>1</sup>

<sup>1</sup>The Forensic Voice Centre, York, UK

richard.rhodes@forensicvoicecentre.com

<sup>2</sup>Department of Language and Linguistic Science, University of York, York, UK

A Quality Management System (QMS) is a documented series of systems, processes and ways of working developed to ensure that forensic casework is carried out validly, securely and in a controlled manner, by staff who are demonstrably competent to do so; it also ensures there is evidence of the above which can be assessed by an accreditation body. A quality-led approach ensures that organisations are open, self-critical and rigorous in assessing their own performance and their level of compliance with international standards (such as ISO17025), national standards (such as the UK FSR's Codes) and more local or field-specific guidelines (such as the ENFSI Best-Practice Manual for Forensic Speaker Comparison or the IAFPA Code of Practice). In the UK, accreditation to ISO17025 and compliance with the FSR Code - including operation with a formal QMS - are now statutory requirements for most types of forensic science; compliance will probably also become mandatory for forensic speech and audio analysis.

At the Forensic Voice Centre, we are developing and validating our QMS processes in line with these requirements. This involves reviewing metadata from real cases, and testing our staff, tools and methods using ground-truth data (casework-like materials where the ground-truth, such as *who said what* and *what was said*, etc. is known). This work is an integral part of validating the overall casework processes, and it also gives us insight into how certain tools, or approaches or methods operate in different circumstances. The purpose of this presentation is to share some of these results a) with other caseworkers to share validation results and discuss quality approaches and practices, and b) with researchers to contextualise and focus future research.

### 1) File format information: forensic recording metadata; format conversion ILC

The first stage in any case is often to take recordings made on a variety of devices and convert/transcode them to standard formats (e.g., PCM-WAV) which can be edited or played back in specialist software. For the purpose of validating these processes and tools in conditions which reflect those found in casework, we need to know what types of files are typically provided to forensic labs and we need to test how well these formats are converted by tools used in casework. We will also report on an inter-laboratory comparison (ILC) on file transcoding.

### 2A) Forensic voice comparison: screening level indications

In each case at the Forensic Voice Centre, recordings are screened to determine their suitability for analysis and to develop a case strategy - this information is given to the instructing party to assist them in deciding whether to proceed with the analysis. A screening process is adopted at a range of international forensic speech and audio laboratories and has been a discussion topic at a previous IAFPA conference (Earnshaw et al, 2024; Lim, 2024). During these discussions, we asked ourselves "how useful are these screening outcomes?" and so we started to record these screening outcomes and compare them with casework outcomes. We will present data on the correlations between these screening and case outcomes to establish how useful they are, and how they can be improved.

### 2B) Forensic voice comparison: reporting analyst and checker conclusions

One way to provide validation information for tests based on evaluative opinion is to 'demonstrate that [practitioners] can provide consistent, reproducible, valid and reliable results that are compatible with the results of other practitioners' (FSR Code v1 30.11.1: our emphasis). Conclusions are formed by both a reporting analyst and a checker for each voice comparison case at the Forensic Voice Centre. We will present the degree of concordance between these two conclusions and investigate the

types of cases where there is more or less agreement between reporting analyst and checker conclusions.

### 3) Forensic expert transcription: proficiency test (PT) results

We have developed resources for proficiency tests (PTs) for expert transcription. Using casework-like materials, we can compare transcripts against the ground truth of what was said to assess the performance of the process and to give individual analysts useful feedback on their transcriptions. We have also compared our transcripts with results from non-expert listeners (using results from Harrington, 2024) because this assists us in validating the overall transcription method: in order for expert evidence to be useful it needs to provide information which goes beyond the knowledge or ability of a jury, and by comparing expert transcripts with what normal people can hear or transcribe, we can quantify the value of expert approaches to transcription.

### 4) Audio Enhancement ILC

Another way of providing evidence that methods are carried out in a valid way that has consensus across the field is to carry out inter-laboratory comparisons (ILCs). These involve giving different organisations the same materials and instructions and reviewing their products or outcomes. We are working with FOR in Zurich and the University of York to create and run a series of ILCs focussed on audio enhancement with forensic casework-like recordings. Following a smaller pilot exercise, we intend to make these materials available to the IAFPA and ENFSI communities. These materials will also have ground truth data available for speaker identity and speech content, meaning they can be used for PTs, ILCs, or validation for speaker attribution and forensic transcription.

These are just some of the quality measures in development intended to meet the requirements set for our industry. We consider that the best approach to quality and validation is a collaborative one (see Rhodes, 2021) and that by sharing the time and creative thinking required to generate these tests, the community is much better equipped to meet these standards; in this spirit, we look forward to identifying more opportunities for future collaboration on analyst training, method development and other quality matters.

## References

- ENFSI-BPM-FSC-01 (2021, version 1) - *Best Practice Manual for the Methodology of Forensic Speaker Comparison* [URL: <https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf>]
- [UK] Forensic Science Regulator. (2023). Forensic science activities: Statutory Code of Practice (version 1 - March 2023; version 2 consultation draft also available from February 2024) - [URL: <https://www.gov.uk/government/publications/statutory-code-of-practice-for-forensic-science-activities>]
- IAFPA (2020). Code of Practice. [URL: <https://www.iafpa.net/the-association/code-of-practice/>]
- ISO/IEC 17025:2017 - International Standard: General requirements for the competence of testing and calibration laboratories. [URL - <https://www.iso.org/ISO-IEC-17025-testing-and-calibration-laboratories.html>]
- Earnshaw, K., Nuttall, B., Rhodes, R., & Thomas, M. (2024). Communicating Expert Evidence. Presentation; IAFPA Conference, Montreal.
- Harrington, L. (2024). Towards improving transcripts of audio recordings in the criminal justice system. (Doctoral dissertation, University of York).
- Lim, S. (2024). Initial Assessment of Suitability (IAS) for Forensic Voice Comparison: Learning from the Past. Presentation; IAFPA Conference, Montreal.
- Rhodes, R. (2021). Project proposal for IAFPA-led collaboration on method testing and validation. Presentation; IAFPA Conference, Marburg/Online.

## **Bridging the gap: Expanding the scope of academic engagement to support practitioners**

*Jessica Wormald, Vincent Hughes, and Joe Pattison*

*FoSS, Department of Language and Linguistic Science, University of York, UK*  
 {firstname.lastname}@york.ac.uk

How best to regulate forensic speech science has been an outstanding question in our field for decades (e.g. Rhodes & Cambier-Langeveld, in press). Practitioners strive to provide the best quality forensic evidence to the courts and endeavour to work within the regulatory framework of their specific jurisdiction. In the UK, the Forensic Regulator has statutory powers; a second version of the Code of Practice was released in March 2025 and is due to come into force in October 2025 (FSR, 2025). Speech and audio analysis is currently exempt from complying with the majority of the Code but this is subject to review in future versions.

A new unit at the University of York - Forensic Speech Services (FoSS) - began in September 2024 and is expanding the current offering of the Forensic Speech Science group at York beyond research and teaching. FoSS will be directly engaging with those working with speech and audio in law enforcement to support best practice and facilitate ongoing research and development. This includes forensic practitioners carrying out speech and audio analysis, but also includes those within this sector wanting to safely work with speech and audio (e.g. police). FoSS provides a unique opportunity to develop solutions from those who have expertise in this area but who are working outside of the confines of practice.

We are facilitating a workshop in May 2025 specifically in relation to regulation. This workshop will bring together voices from across the field, to identify, discuss and work towards solutions to regulatory barriers in the UK context. Practitioners from across the UK have been invited, alongside members of the Forensic Speech and Audio working group which advises the regulator, and the Forensic Accreditation Specialist from UKAS. Additionally, we have invited a leading UK forensic scientist from a different discipline and international practitioners in a different jurisdiction all facing similar challenges.

At IAFPA, we will share the outcomes of the workshop. There are a number of known barriers to regulation, including how best to assess compliance, validation of methods, and practical issues relating to the UK forensic landscape (e.g. practitioners are working in private laboratories, often with small numbers of staff). The workshop will also explore other barriers. Solutions are as yet unknown; the workshop is being designed to facilitate flexibility, with practitioners and attendees deciding the priorities and associated solutions. These will be discussed at IAFPA and represent the start of an ongoing route to support practitioners in a meaningful and tangible way.

### **References**

- Forensic Science Regulator (2025). Draft Code of Practice 2025 (Version 2). Available online: [https://assets.publishing.service.gov.uk/media/67daba1e594182179fe0883b/E03313596+-+CoP+Forensic+Science+Regulator+2025\\_A4\\_v02\\_Web+Accessible.pdf](https://assets.publishing.service.gov.uk/media/67daba1e594182179fe0883b/E03313596+-+CoP+Forensic+Science+Regulator+2025_A4_v02_Web+Accessible.pdf)
- Rhodes, R. & Cambier-Langeveld, T. (forthcoming) Practitioner Standards. In McDougall, K., Nolan, F. & Hudson, T. (eds.) *Handbook of Forensic Phonetics*. Oxford University Press.

# Project CELIA:

## Common European Language Indication and Analysis

*Tina Cambier-Langeveld*  
*Immigration and Naturalization Service, the Netherlands*  
 Gm.cambier.langeveld@ind.nl  
 CELIA@ind.nl

Language analysis in the asylum process (LAAP) is used in some jurisdictions to investigate whether the language varieties spoken by an asylum applicant can support the claimed origin. For instance, an applicant claiming to be from the south of Somalia can provide support for this claim by speaking a south Somali variety, and an applicant claiming to be from Sierra Leone can prove this by speaking Krio plus another Sierra Leonean language (if not from the capital Freetown). Such linguistic evidence is used when an applicant has failed to convince the immigration service of their origin by other means. Often enough LAAP serves to help asylum seekers who have legitimate reasons for not being able to provide other evidence, or who had trouble with telling a fully consistent story (which can be due to trauma, cultural barriers or interpreting issues for example). In other cases, LAAP can serve as a negative indication, which – in combination with a lack of other positive evidence – can ultimately lead to an application being turned down.

Relatively little has been published on this particular forensic application of linguistics, and practitioners are not united in any forum. Cooperation between practitioners and/or agencies is hampered by confidentiality, commercial interests, and different views on how to best perform such analyses. Various aspects of these analyses, such as the collection of a speech sample from an asylum applicant, the analysis itself and the way in which results are reported, are prescribed by national jurisprudence and practical considerations, rather than by scientific best practices. Quality assurance is organized in-house, with little opportunity for external parties to gain insight or exert influence.

But there is good news to report. Project CELIA (Common European Language Indication and Analysis) is underway (2024-2027), seeking to establish new standards, to explore new methodologies, to develop peer-reviewed training modules and selection processes for experts, and to set up a pool of certified analysts. The project is a so-called Specific Action, co-funded by the Asylum, Migration and Integration Fund (AMIF) of the European Commission. The Commission is keen to have optimized and efficient tools developed for ‘fast-track’ language indications (possibly AI-based) and ‘full-track’ language analyses (by certified experts), which should eventually become available for migration offices in all EU+ Member States. The project welcomes input, feedback, contributions and peer review from external experts.

During this Special Session on Language Analysis in the Asylum Process, I will briefly introduce the current state of play in this field and I will present the specific aims of project CELIA. We will then zoom in on two subtopics within the project:

- 1) Further development and validation of the automatic accent recognition system Y-ACCDIST (see abstract by Georgina Brown and Sam Hellmuth)
- 2) The option of introducing targeted data elicitation techniques, inspired by methods in linguistic field research (see abstract by Peter Gottschligg)

The session will have room for questions, suggestions and thoughts from the audience.

### References

CELIA website: <https://ind.nl/en/celia-common-european-language-indication-and-analysis>

## Y-ACCDIST applied to Arabic

*Georgina Brown<sup>1</sup> and Sam Hellmuth<sup>2</sup>*

<sup>1</sup>*Department of Linguistics and English Language, Lancaster University, UK*  
g.brown5@lancaster.ac.uk

<sup>2</sup>*Department of Language and Linguistic Science, University of York, UK.*  
sam.hellmuth@york.ac.uk

Language Analysis in the Asylum Process (LAAP) entails a high volume of speech data from a relatively well defined set of accent/dialect varieties. In addition, it is possible to control the recording conditions of the audio samples obtained in this context. With these application parameters, the Y-ACCDIST automatic accent recognition system could present as a tool and opportunity to LAAP. Some of the more recent work that applied Y-ACCDIST to Arabic speech data (Brown and Hellmuth, 2022) has made Y-ACCDIST more relevant to the LAAP application, alongside the remote Arabic speech data collection efforts that supplement further development (Almbark et al., 2023).

Y-ACCDIST is distinct from neural network-based approaches to automatic accent and dialect identification. Although performance is generally on a par with other system architectures, Y-ACCDIST is routinely dismissed as a widely applicable approach because a transcription must accompany the audio samples that are processed by the system. A transcription enables the Y-ACCDIST system to first estimate the locations of the vowel and consonants within a speech sample so it can form a model of a speaker's accent based solely on the acoustic realisations of different phoneme-like units. While the practicality and resource barriers to providing transcriptions are obvious, the overall Y-ACCDIST approach presents qualities that are in tune with the spirit of the EU AI Act, making it a worthwhile research endeavour within the CELIA project.

Y-ACCDIST's accent modelling approach has been proposed as particularly 'explainable' (Brown et al., 2022) because the system explicitly targets vowel and consonant realisations from the outset of its processes. This differs from alternative AI-driven approaches which extract acoustic features from the audio samples and apply an extensive series of transformations to those features that encode all kinds of information (beyond accent/dialect information). The eventual result is a high-dimensional representation of the audio sample that does not immediately bear resemblance or mapping to linguistically identifiable features. Having said that, there are indeed now ways to infer which linguistic features these high-dimensional representations are likely to capture.

Implementing an AI-driven approach entails large volumes of training data, realistically amounting to hundreds of audio samples per accent/dialect group. Y-ACCDIST, by contrast, offers relatively low-resource training requirements. This in turn can lead to greater data quality assurances as it becomes more possible to train on smaller 'gold-standard' datasets, rather than on large unwieldy datasets that are notorious for their metadata errors and poor, unrepresentative data samples.

The Y-ACCDIST subproject within CELIA therefore aims to further develop and test these strengths and weaknesses of the system in light of the LAAP application. Firstly, it will review the transcription requirements of the Y-ACCDIST approach and evaluate whether and to what extent today's open-source automatic speech recognition systems could provide the required transcriptions. Secondly, the subproject will further test the low-resource data requirements by varying the levels of homogeneity in the training and test data to observe the risks and return of the Y-ACCDIST-Arabic system as a low-resource Arabic dialect identification solution.

### References

Almbark, R., Hellmuth, S. & Brown, G. (2023). Working with Public Involvement Coordinators to support remote collection of high quality audio speech data. *Laboratory Phonology*. 14. DOI: <https://doi.org/10.16995/labphon.10541>

- Brown, G., Franco-Pedroso, J. & González-Rodríguez, J. (2022). A segmentally informed solution to automatic accent classification and its advantages to forensic applications. *International Journal of Speech, Language and the Law*. 28. 201-232.
- Brown, G. & Hellmuth, S. (2022). Computational modelling of segmental and prosodic levels of analysis for capturing variation across Arabic dialects. *Speech Communication*. 141. 80-92.

## Picture Naming Task in LADO

*Peter Gottschligg*  
*Vienna*

As an independent linguistic expert, I have been providing language analysis for various European authorities and courts since 1999. In interviews conducted by myself, picture naming often plays a central part, be it to test a competence in a particular language, including lesser documented languages, or else, to test a competence in a particular variety of a language.

A picture-naming task can constrain the referential content of a response in an effective way. Instead of “waiting” for the interviewee to use e. g. certain dialectal expressions or dialectal pronunciations during a free interview, such vocabulary and pronunciations can be elicited in much easier ways, e. g. by a picture-naming task. Of course, there are concepts that are less suited to be represented by picture stimuli and there are pictures which allow for a too wide range of interpretations to be useful.

Elicitation will normally focus on items and features which have already been documented as being peculiar to certain dialectal varieties. This will facilitate the subsequent analysis of the data and the development of arguments leading up to the conclusions of the report. It will provide instances of otherwise rarely used dialectal expressions and generate more instances of certain dialectal forms or patterns that could be rare or even absent in speech samples generated by unstructured interviews.

Picture naming can be used in a monolingual situation. It helps to reduce the effects of accommodation to the interviewer’s dialect of the language.

Pictures can and do usually trigger spontaneous responses, comments and explanations by the interviewee, they can also be used as a starting point to engage the interviewee in a discussion of various aspects of life in his country of origin.

Pictures can be used to test whether the interviewee has knowledge of elements specific to the cultural or natural environment he claims to have lived in, or else, to another environment, he has possibly lived in.

Pictures can also be used or be interpreted as stimuli in an association test, which can yield interesting results, e.g. when the interviewee mistakes an object he can be assumed to know for another object which, given the environment he claims to have lived in, he is unlikely to be familiar with.

Picture naming, in the context of language analysis, is neither a psychological test nor a research tool to be used under tightly controlled conditions. Reaction time is, within limits, irrelevant, errors can be corrected and the further responses can be elicited and constrained by further questions, by reminding the interviewee to answer with expressions and forms of the dialect he is assumed to speak and/or by telling the interviewee, that he has not given the expected answer. This gives an advantage to authentic speakers who have lived for longer periods outside their “original” speech community. On the other hand, it will provide and solidify negative evidence in cases, in which the interviewee does not speak the language or dialectal variety he can be expected to be a speaker of.

### Select Bibliography on Elicitation, Picture Naming and Dialect Performance

- Bassiouney, R., Ed. (2017) *Identity and Dialect Performance. A Study of Communities and Dialects*. Routledge.
- Behnstedt, P. & Woidich, M. (2005) *Arabische Dialektgeographie. Eine Einführung*. Brill.
- Bochnak, M. R. & Matthewson, L. Eds. (2015) *Methodologies in Semantic Field Work*. Oxford University Press.

- Bouquiaux, L. & Thomas, J M. C., Eds. (1976) *Enquête et description des langues à tradition orale. 3 Volumes*. SELAF.
- Bucholtz, M. (2003) Sociolinguistic nostalgia and the authentication of identity. *Journal of Sociolinguistics*, 7 (3), pp. 398-416.
- Casad, E. H. (1997) Language assessment tools: Uses and limitations. In: Pütz, M. (Ed.) *Language Choices. Conditions, constraints, and consequences*. Amsterdam. John Benjamins Publishing Company (pp. 253-273)
- Glaser, W. R (1992) Picture naming. In: *Cognition* No. 42 (pp. 61-105).
- Kiese-Himmel, C. (2005) *AWST-R. Aktiver Wortschatztest für 3- bis 5-jährige Kinder-Revision. Manual*. Hogrefe.
- Newman, P.& Ratliff, M., Eds. (2001) *Linguistic Fieldwork*. Cambridge University Press.
- Sachs-Hombach, K. (2013) *Das Bild als kommunikatives Medium. Elemente einer allgemeinen Bildwissenschaft*. Herbert von Halem Verlag.
- Schiesser, A. (2020) *Dialekte machen. Konstruktion und Gebrauch arealer Varianten im Kontext sprachraumbezogener Alltagsdiskurse*. De Gruyter.

## Workshop: Building your own LR system

*Stijn van Lierop and David van der Vloed*

*Netherlands Forensic Institute, The Hague, The Netherlands*

`{s.van.lierop|d.van.der.vloed}@nfi.nl`

In this workshop, you will solve a case using data science! We will introduce you to the basic steps of an LR system. Next, through a code notebook, you will be given case scores and a set of scores of background data, all from automatic speaker recognition software.

Working with that data yourself, you will experience the full pipeline from case scores to case LRs, including calibration and validation.

To get a grasp of what's happening, we will guide you through the process, explain all concepts and help you to visualize the data. You will get a feel for what impact some of the design choices may have on the final case LR. Hopefully, by the end, we are able to compare our results and to reflect from an insider's perspective on the suitability of this strictly data driven method in real casework.

The goal is not to just keep clicking buttons and come up with some opaque end result. On the contrary, we are only happy when we have put you in a position that you are in control of it all and the LR system becomes a trusty tool that you know how to appreciate.

Although you will encounter some Python code in this workshop, a background in coding is not necessary. We are happy to have a diverse range of skill sets in the audience. The code will be simple, mostly pre-made, heavily commented with explanations, and we will happily assist you whenever needed. Bring your laptop and Google account. If you don't have any one of those: no problem, we will try and team you up with someone else.

# Assessing the suitability of forensic authorship analysis methodologies for speech data

James Tompkinson<sup>1</sup> and Andrea Nini<sup>2</sup>

<sup>1</sup>*Department of Language and Linguistic Science, University of York, UK*  
james.tompkinson@york.ac.uk

<sup>2</sup>*Department of Linguistics and English Language, University of Manchester, UK*  
andrea.nini@manchester.ac.uk

The development of new analytical methods and frameworks which could be integrated into forensic speaker comparison (FSC) work is a core focus for research in forensic speech science. In this paper, we explore the applicability of methods that have been used in forensic authorship analysis (FAA) to speech data. Our work has two main areas, 1) whether methods borrowed from authorship analysis can be used to analyse discrete phonetic variables using a likelihood-ratio based framework and 2) whether the embedding of auditory phonetic analysis with “higher order” features (Gold and French 2011) such as lexis, grammar and morphology, which are frequently considered in FAA tasks, can be used for speaker comparison.

Our work builds on research by Sergidou et al. (2023), who showed that frequent words did have some speaker discriminatory power, and argued that this could be useful in FSC casework. We expand this work to examine how phonetic variation can be incorporated into such a framework. We analysed transcribed speech data from a random sample of 30 speakers from the West Yorkshire Regional English Database (Gold 2020) across two different speaking styles (Task 1 and Task 2), using two well-known authorship analysis methods which incorporate the likelihood ratio (LR) framework: Cosine Delta (Ishihara 2021) and Phi n-gram tracing (Nini 2023). We applied these methods to transcripts which had been adapted to represent a range of phonetic features - vocalised hesitation markers, syllable-initial realisations of /θ/, intervocalic word-medial /t/, syllable-initial /l/ and realisations of the -ing suffix - to assess 1) whether algorithms used in FAA are similarly effective on phonetic feature sets of this kind and 2) whether the combination of “higher-order” linguistic features with segmental phonetic analysis would achieve greater speaker discriminatory power.

Our findings support previous research which has suggested that methods used to discriminate between authors can be usefully applied to transcribed speech data. We find that Cosine Delta and N-gram tracing are both effective in performing speaker comparison on transcribed speech data. In addition, our results show how a logistic regression calibrated Cosine Delta using the consonant phonetic features alone already offers valuable information. The analytical framework for this project, where phonetic information is embedded in transcripts and then subjected to authorship analysis techniques using the likelihood ratio paradigm, could potentially be used as a way of systematically evaluating auditory phonetic variables within a likelihood-ratio approach even when the phonetic features are discrete.

## References

- Gold, E. (2020). WYRED - West Yorkshire Regional English Database 2016-2019. [data collection]. UK Data Service. SN: 854354, DOI: 10.5255/UKDA-SN-854354
- Ishihara, Shunichi. 2021. Score-based likelihood ratios for linguistic text evidence with a bag-of-words model. *Forensic Science International*. Elsevier 327. 110980.
- Nini, A. (2023). *A Theory of Linguistic Individuality for Authorship Analysis*. Elements in Forensic Linguistics. Cambridge University Press.
- Sergidou, E. K., Scheijen, N., Leegwater, J., Cambier-Langeveld, T., & Bosma, W. (2023). Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150, 1-8.

# Interpreting forensic transcriptions: from ambiguity to accuracy

*Christian Di Nunno*

*PhD Candidate in Linguistics, University for Foreigners of Siena (Siena - Italy)*

c.dinunno@dottorandi.unistrasi.it

This study examines the interpretative complexity of forensic transcription in the Italian judicial context, where the absence of formal training programs for transcribers has led to significant errors with serious legal consequences. The research begins with an analysis of real judicial cases in which transcription inaccuracies—often caused by the lack of linguistic and phonetic expertise—have compromised investigations and verdicts.

A foundational component of the study is a questionnaire administered to members of Italian judicial police tasked with transcribing intercepted communications during preliminary investigations. Results highlight an alarming lack of methodological consistency and professional preparation, particularly regarding the transcription of dialects, spontaneous speech, and ambiguous content. Participants were also asked to transcribe selected audio excerpts and reflect on previously completed transcriptions. These tasks revealed widespread inconsistencies and interpretative distortions, underscoring the urgent need for standardized transcription protocols in Italy.

The empirical core of the study is a corpus of approximately ten hours of intercepted speech—including both telephone and environmental recordings—collected from multiple criminal proceedings and representing a range of diatopic varieties of Italian. A detailed analysis focuses on the transcriptional treatment of spoken features such as pauses, overlaps, hesitation markers, discourse particles, and prosodic contours. Particular attention is given to the recurrent loss or misrepresentation of these elements in written form, a process which erases critical interpretative cues. One of the central findings is the systematic suppression of prosodic features in highly sensitive portions of the corpus, especially in segments coded as “incriminated” or “cryptic”, where intercepted speakers intentionally reduce volume and fundamental frequency ( $f_0$ )—an acoustic behavior that diverges from the well-known Lombard effect and may serve to avoid detection or reduce intelligibility.

Starting from these observations, the study incorporates tools from acoustic phonetics, pragmatics, and conversation analysis to explore how prosodic and supra-segmental traits can indicate levels of intentional opacity, deception, or emotional inhibition. Without overstepping the boundaries of legitimate interpretation, the research cautiously investigates whether and how vocal cues can reflect emotional and physiological states (e.g., stress, fear, or controlled speech). These phenomena are approached with scientific restraint, in line with the study’s fundamental principle: maintaining transcriptional accuracy while avoiding speculative interpretations that could compromise legal objectivity.

The project also addresses the broader issue of layout variability and the absence of a shared and unified symbolic system for transcription in Italy: currently, layout, symbols, and formatting vary widely across cases, creating confusion for judicial authorities who must interpret heterogeneous transcriptions. Such inconsistencies hinder judicial readability and interpretation. The study therefore proposes a clear, standardized, and prosodically informed transcription model that can enhance both legibility and evidentiary reliability. It also explores the challenges posed by dialect to Italian translation - particularly salient issue in the linguistically fragmented Italian context - the marginalization of non-standard variants, and the effects of priming and contextual bias on transcription accuracy. Strategies are proposed to maintain fidelity to the source speech while avoiding ambiguous or overly interpretative renderings.

Ultimately, this research aims to delineate the boundaries of the “*interpretative margin*” in forensic transcription—the threshold beyond which interpretation becomes ambiguous or unreliable. Through an interdisciplinary and data-driven approach, it proposes practical, scientifically grounded solutions to reduce error, increase transparency, and establish a more equitable and linguistically aware framework for the transcription and analysis of intercepted speech in judicial settings. This research contributes to a more transparent, consistent, and equitable use of transcription as linguistic evidence in the Italian judicial system.

## References

- Bellucci, P. (2005). *A onor del vero. Fondamenti di linguistica giudiziaria*. UTET.
- Coulthard, M., & Johnson, A. (2007). *An Introduction to Forensic Linguistics: Language in Evidence*. Routledge.
- Fraser, H. (2014). Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science. *Language and Law/Linguagem e Direito*, 1(2), 5–21.
- Orletti, F., & Mariottini, M. (2017). *Forensic Communication in Theory and Practice. A Study of Discourse Analysis and Transcription*. Cambridge: Cambridge Scholars Publishing.
- Romito, L. (2014). *Manuale di Linguistica Forense*. Bulzoni Editore.

# Toward an improved American English word list, with frequencies and transcription

*Sandra F. Disner<sup>1</sup>, Vincent J. van Heuven<sup>2,3,4</sup>, and Eileen Yang<sup>1</sup>*

<sup>1</sup>*Department of Linguistics, University of Southern California, Los Angeles, USA*  
{sdisner;eileenya}@usc.edu

<sup>2</sup>*Leiden University Centre for Linguistics, Leiden, The Netherlands*

<sup>3</sup>*Doctoral School of Multilingualism, University of Pannonia, Veszprém, Hungary*

<sup>4</sup>*Fryske Akademy, Leeuwarden, The Netherlands*

v.j.j.p.van.heuven@hum.leidenuniv.nl

## Summary

A list of the ~4000 most frequent words in American English (COCA corpus) – lemmatized, without homographs, and augmented by phonetic transcriptions – has been developed to facilitate court cases involving orthographic (lookalike) and phonetic (soundalike) word similarity (e.g., trademarks or medical drugs), and possibly provide validation of black-box approaches.<sup>1</sup>

## Background

In legal disputes over undue similarity between competing trademarks, US courts do not recognize black box approaches such as those that have been developed in speech technology. Instead, judges insist on transparent and understandable quantification of similarity between trademarks (brand names). Acoustic measurements of actual speech (or visual pattern recognition) are not an option. Instead, we have to rely on the symbols in a phonetic transcription or in the orthography of words. Transparent measures have been proposed to quantify orthographic (look-alikeness) and phonetic (sound-alikeness) similarity, such as the number of shared symbol bigrams (or trigrams). It is unclear, however, how the results of such computations are to be interpreted. For instance, is 15% shared phone trigrams a small or a large difference? To answer this, we need to know the distribution of similarity (or distance, its complement) between (a representative sample of) all pairs of words in the lexicon. This distribution should be an indispensable tool in legal disputes over trademark similarity (and product names in general). This forensic application of linguistics/phonetics that has received relatively little attention so far.

## Word list

We extracted the 5000 most frequent word lemmas (no inflections) from the Frequency Dictionary of Contemporary American English (Davies & Gardner, 2010), which is based on the Corpus of Contemporary American English (COCA, Davies, 1990-2023, 2009). Part of Speech (PoS) and frequency information (stratified by genre) was copied from the source but phonetic transcriptions had to be imported from the CMU pronouncing dictionary for American English.<sup>2</sup> We then reduced the set to only the nouns, adjectives, main verbs and adverbs (except those derived from adjectives). Lemmas with transparent derivational affixes were reduced to their base form (and frequencies added together) if the affix had no effect on the pronunciation (including the stress pattern) of the base. These modifications were partly done automatically and partly by hand (mainly by the third author). The final list of 3927 lemmas is now available with IPA transcriptions (when homographic noun/verb/adjective/adverb pairs and triplets are collapsed into single lemmas, 3455 entries remain).

## Applications

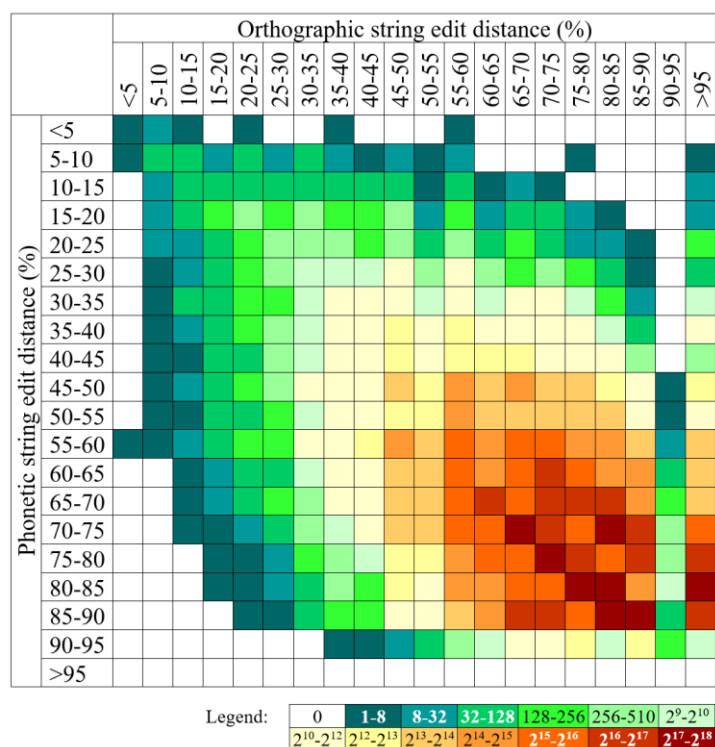
<sup>1</sup> We gratefully acknowledge the financial support by the IAFPA, which allowed the third author to work on this project.

<sup>2</sup> <http://www.speech.cs.cmu.edu/tools/lextool.html>

For our talk, we will compile a dataset comprising all unique pairs of words from the lemma list with similarity/distance measures:

- Number/percentage of shared letter and phone bigrams and trigrams
- Plain and Phonetic feature-weighted Levenshtein distance for orthographic and phonetic representations (using LED-A, Heeringa et al., 2023)

We will present heatmaps (Figure 1, computed on a precursor of the present lemma list – with CMU pronunciation but inclusion based on token frequencies in British English) of orthographic versus phonetic similarity using the same metric for both. Similar work for Dutch confirms that orthographic and phonetic distance are only weakly correlated ( $r^2 = .33$ , Torenbosch & Van Heuven, 2023). Clearly, the assumption made by many courts that phonetic and orthographic similarity are practically identical is wrong – even more so for English with its capricious letter-to-sound relationships.



**Figure 1.** Phonetic Levenshtein string edit distance plotted against orthographic distance of 4.5 million word pairs ( $r^2 = .21$ ) drawn from the 3000 most frequently used content word lemmas in the British National Corpus (with CMU US transcriptions imported). In our talk, we will present similar data but then based on the new American lemma selection. For frequency brackets see legend.

## References

- Davies, M. & Gardner, D. (2010). *A Frequency Dictionary of Contemporary American English*. Routledge.
- Davies, M. (1990-2023). *Corpus of Contemporary American English*. <https://www.english-corpora.org/coca/>
- Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, 14(2), 159–190. <https://doi.org/10.1075/ijcl.14.2.02dav>
- Heeringa, W., Van Heuven, V. & Van de Velde, H. (2023). *LED-A: Levenshtein Edit Distance App* [computer program]. <https://www.led-a.org/>
- Torenbosch, J.R. & Van Heuven, V. J. (2023). Visuele en auditieve overeenstemming van merknamen: fabels en feiten [Visual and auditory similarity of trademarks: fables and facts]. *Intellectuele Eigendom en Reclamerecht*, 39(5), 295–307. <https://www.researchgate.net/publication/375379700>

# Identifying Voice Super-Recognizers in Law Enforcement

Andrea Fröhlich<sup>1,2,3</sup>, Meike Ramon<sup>3</sup>, Peter French<sup>2</sup> and Volker Dellwo<sup>2</sup>

<sup>1</sup>Zurich Forensic Science Institute, Switzerland

<sup>2</sup>Department of Computational Linguistics, University of Zürich

<sup>3</sup>Applied Face Cognition Lab, Bern University of Applied Sciences

andrea.froehlich@uzh.ch

In 2009, Russell et al. identified "Super-Recognizers" (SR) with exceptional face processing abilities. Since their discovery, SR have been the subject of many research projects and were successfully deployed in law enforcement (Mayer & Ramon, 2023; Ramon & Rjosk, 2022). Until today, it remains largely unknown whether a similar phenomenon exists for voices (Jenkins et al., 2021). Identifying voice super-recognizers (VSR) could greatly benefit law enforcement, where they could efficiently post-process machine-based results (Fröhlich et al., 2023).

When examining *general* voice processing abilities, researchers previously described the highest-ranking participants as potential VSR (Aglieri et al., 2017; Humble et al., 2022; Schäfer & Foulkes, 2023). However, reliably identifying VSR requires challenging and sensitive assessment tools (Schäfer, 2023). To address this, we developed a challenging voice identity processing test aimed at detecting high performers and propose a novel approach for identifying VSR, supported by normative data from law enforcement practitioners.

## Defining voice super-recognition skills

Given the lack of definition for VSR, we propose a working definition based on the approach used to identify face SR (Ramon, 2021) and insights from casework (Ruch et al., 2023): VSR are individuals who demonstrate a naturally occurring high proficiency for voice identity processing that is consistent, i.e., observable across different tests (voice discrimination and voice sorting). Their performance remains robust even under challenging conditions that mirror real forensic casework, including short or low-quality stimuli, mismatched audio quality, and variations in speaking style.

## Methods

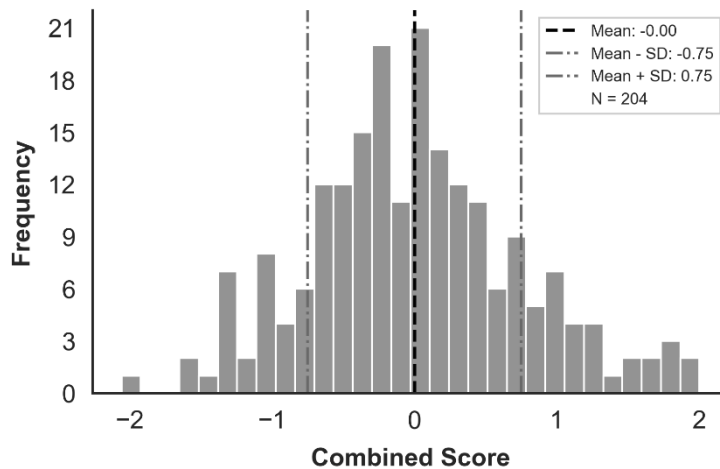
To simulate real-world deployment, we ensured a high task difficulty by using stimuli from the Pool corpus (Jessen et al., 2005), which includes varied speaking styles and recording qualities, and by reducing stimulus length to 1.2 seconds. Between-speaker trials were selected using a validated method combining automatic speaker recognition and F0-delta information to identify similar-sounding voices (Fröhlich et al., 2023). The resulting test battery — two unfamiliar voice perception tests (identity discrimination and sorting) — was run with Swiss law enforcement personnel (N=204) as a self-paced online experiment using the Gorilla Experiment Builder (Irvine et al., 2020).

## Results and Conclusions

We developed two highly challenging voice processing tests: a discrimination task (mean accuracy = 63.37%, SD = 4.84) and a sorting task (mean accuracy = 66.87%, SD = 6.34). In accordance with previous findings (Johnson et al., 2020), no significant correlation was found between the two tasks, suggesting that voice sorting and discrimination rely on different underlying skill sets. A combined overall voice processing score was therefore calculated based on performance in both tasks (Figure 1).

Following the proposed definition of VSRs and drawing on approaches from the visual domain, we aimed to identify individuals who consistently perform well across multiple test batteries. Two participants, scoring in the top 10% on both tasks, were identified. Detailed analysis of our top performers is underway and will be presented at the conference.

Furthermore, as recent findings on face SR indicate that their exceptional abilities might result from a perceptual processing advantage that is not exclusive to faces (Jenkins et al., 2021; Nador et al., 2025), we have also initiated testing face SR (N=11) for unfamiliar voice perception.



**Figure 1.** Distribution of the combined voice processing score (higher = better).

## References

- Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, *49*(1), 97–110.
- Fröhlich, A., Dellwo, V., French, P., & Ramon, M. (2023). Automatic speaker recognition-based development of challenging speaker discrimination tests. *Proceedings of the 20th International Congress of Phonetic Sciences*.
- Humble, D., Schweinberger, S. R., Mayer, A., Jesgarzewsky, T. L., Dobel, C., & Zäske, R. (2022). The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices. *Behavior Research Methods*.
- Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder.
- Jenkins, R. E., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology*, *35*(3), 590–605.
- Jessen, M., Koster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. *International Journal of Speech, Language and the Law*, *12*(2), 174–213.
- Johnson, J., McGettigan, C., & Lavan, N. (2020). Comparing unfamiliar voice and face identity perception using identity sorting tasks. *Quarterly Journal of Experimental Psychology*, *73*(10), 1537–1545.
- Mayer, M., & Ramon, M. (2023). Improving forensic perpetrator identification with super-recognizers. *Proceedings of the National Academy of Sciences*, *120*(20), e2220580120.
- Nador, J. D., Uittenhove, K., Gordillo, D., & Ramon, M. (2025). Super-recognizers, or Su-Perceivers? Insights from Fast Periodic Visual Stimulation (FPVS) EEG. *OSF Preprints*.
- Ramon, M. (2021). Super-recognizers – A novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809.
- Ramon, M., & Rjosk, S. (2022). *beSure? – Berlin Test for Super-Recognizer Identification: Part I: Development*. Verlag für Polizeiwissenschaft.
- Ruch, H., Fröhlich, A., & Lim, S. (2023). Clustering a large number of unknown voices. *Proceedings of the 31st IAFPA Conference*.
- Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic Bulletin & Review*, *16*(2), 252–257.
- Schäfer, S., & Foulkes, P. (2023). Towards a screening test for earwitnesses. *The International Journal of Speech, Language and the Law*, *30*(2), 234–267.
- Schäfer, S. (2023). *The individual lay listener as a variable in speaker individualisation tasks* (Doctoral dissertation, University of York).

# Relevance of the distinctions score-based / feature-based and common source / specific source for forensic voice comparison

*Michael Jessen*

*Department of Text, Speech and Audio, Bundeskriminalamt, Germany*

Michael.Jessen@bka.bund.de

The distinctions score-based / feature-based as well as common source / specific source have been used and discussed increasingly in the recent literature on forensic science in general. The first distinction is about two different manners in which the evidence  $E$  in the LR formula  $p(E|H1)/p(E|H2)$  is understood. In the feature-based manner,  $E$  are features, often multidimensional in nature, which are extracted from (or described in) both the questioned and the known items, as well as from a relevant population. The feature content is then compared, taking into account similarity and typicality (while considering how mismatch might have a warping effect on similarity and typicality). In the score-based manner,  $E$  are degrees of similarity (aka scores) between the questioned and the known items. Typicality is not considered in score-based analysis, although in automatic speaker recognition there may be a previous feature-based stage (see Bolck et al. 2009, 2015; Meuwly et al. 2017; Neumann et al. 2021; Vergeer 2023; Leegwater et al. 2024 about this characterisation of the score-/feature-based distinction, including further details).

The second distinction is about different manners in which the hypotheses  $H1$  and  $H2$  are expressed. In a specific source analysis,  $H1$  says that the questioned item originates from the known source (suspect material), whereas  $H2$  says that the questioned item originates from another source in a relevant population. In a common source analysis,  $H1$  says that questioned and known items have the same source within a given relevant population and  $H2$  that questioned and known items have a different source within that relevant population (Ommen and Saunders 2018; Vergeer 2023).

In this contribution, the two distinctions are applied to forensic voice comparison. When the distinctions are projected into a four-way matrix, it seems that all four possible combinations are instantiated by one or more voice comparison approaches. A possible classification of the different approaches is shown in Table 1. A given approach may fit into more than one of the combinations, especially the auditory-acoustic one; this is not reflected in the table.

	Common source	Specific source
Feature-based	FASR with i/x vectors before calibration but after PLDA;  Semiautomatic speaker recognition using MVKD	Methods/features of the auditory-acoustic approach;  FASR and semiautomatic speaker recognition using the GMM/UBM technique
Score-based	FASR with i/x vectors after calibration;  Holistic listening	FASR using the scoring method proposed by Meuwly/Drygajlo/Alexander (same-speaker score distribution using several suspect recordings)

**Table 1.** Four-way matrix based on the distinctions score-based / feature-based and common source / specific source and the classification of different forensic voice comparison approaches. [FASR = Forensic Automatic Speaker Recognition; PLDA = Probabilistic Linear Discriminant Analysis; MVKD = Multivariate Kernel Density method, see Rose 2006; GMM/UBM = Gaussian Mixture Model/Universal Background Model; for the lower right segment see Meuwly & Drygajlo 2001 and Drygajlo et al. 2003; for further explanation Morrison et al. 2021]

The motivations for the classifications made in Table 1 will be presented. A problem seems to emerge when evidence from several approaches is combined in a case analysis and the hypotheses are not expressed in the same way, specifically some are expressed in common source others in specific source manner. This would be in conflict with the principle that combination of different LR's from different pieces of evidence require the hypotheses to be identical. A way is shown as to how ultimately the same hypotheses can be used throughout, thereby proposing a solution to this combination problem.

## References

- Bolck, A., Weyermann, C., Dujourdy, L., Esseiva, P. & van den Berg, J. (2009). Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic Science International*, 191, 52–51.
- Bolck, A., Ni, H. & Lopatka, M. (2015). Evaluating score- and feature-based likelihood ratio models for multivariate continuous data: applied to forensic MDMA comparison. *Law, Probability and Risk*, 14, 243–266.
- Drygajlo, A., Meuwly, D. & Alexander, A. (2003). Statistical methods and Bayesian interpretation of evidence in forensic automatic speaker recognition. In *Proceedings of EUROSPEECH 2003*, Geneva, 689–692.
- Leegwater, A. J., Vergeer, P., Alberink, I., van der Ham, L. V., van de Wetering, J., El Harchaoui, R., Bosma, W., Ypma, R. J. F. & Sjerps, M. J. (2024). From data to a validated score-based LR system: A practitioner's guide. *Forensic Science International*, 357, 111994.
- Meuwly, D. & Drygajlo, A. (2001). Forensic speaker recognition based on a Bayesian framework and Gaussian Mixture Modelling (GMM). In *Proceedings of ODYSSEY 2001*, Crete, 145–150.
- Meuwly, D., Ramos, D. & Haraksim, R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276, 142–153.
- Morrison, G. S., Enzinger, E., Ramos, D., González-Rodríguez, J. & Lozano-Díez, A. (2021). Statistical models in forensic voice comparison. In D. Banks, K. Kafadar, D. H. Kaye & M. Tackett (Eds.), *Handbook of forensic statistics* (pp. 451–497). CRC Press.
- Neumann, C., Hendricks, J. & Ausdemore, M. (2021). Statistical support for conclusions in fingerprint examinations. In D. Banks, K. Kafadar, D. H. Kaye & M. Tackett (Eds.), *Handbook of forensic statistics* (pp. 277–324). CRC Press.
- Ommen, D. M. & Saunders, C. P. (2018). Building a unified statistical framework for the forensic identification of source problems. *Law, Probability and Risk*, 17, 179–197.
- Rose, P. (2006). The intrinsic forensic discriminatory power of diphthongs. In *Proceedings of the 11<sup>th</sup> Australasian International Conference on Speech Science and Technology*, Auckland, 64–69.
- Vergeer, P. (2023). From specific-source feature-based to common-source score-based likelihood-ratio systems: ranking the stars. *Law, Probability and Risk*, 22, mgad005.

# Language proficiency as a useful index in predicting individual performance of trilingual speakers in cross-language forensic voice comparison using an automatic speaker recognition system

*Grace W. L. Cao<sup>1</sup>, Vincent Hughes<sup>2</sup>, Justin J. H. Lo<sup>3</sup> and Peggy Mok<sup>4</sup>*

<sup>1</sup>*School of Languages, Cultures and Linguistics, University College Dublin, Ireland*  
grace.cao@ucd.ie

<sup>2</sup>*Department of Language and Linguistic Science, University of York, UK.*  
vincent.hughes@york.ac.uk

<sup>3</sup>*Department of Linguistics and English Language, Lancaster University, UK.*  
j.h.lo@lancaster.ac.uk

<sup>4</sup>*Department of Linguistics and Modern Languages, The Chinese University of Hong Kong, Hong Kong SAR, China.*  
peggy Mok@cuhk.edu.hk

Previous studies show that automatic speaker recognition (ASR) systems can perform very well in cross-language forensic voice comparison (FVC) when there is a mismatch between known and questioned samples (Lo, 2021; Nuttall, Harrison and Hughes, 2023). While previous studies have looked at overall system performance, they have not considered aspects related to individual speakers. The current study fills this gap and explores the effect of language proficiency on cross-language forensic voice comparison performance using an ASR system.

Fifty-one female L1 Cantonese-L2 English-L3 Mandarin trilingual speakers (28 young and 23 older participants) participated in three mock police interviews. Participants' proficiency in English and Mandarin was rated by Hong Kong Cantonese-L1 listeners due to the lack of standard proficiency scores for the older group. The young participants scored 6.8 and 6.7 (out of 10) in English and Mandarin, respectively, while the older participants scored 5.3 in English and 3.8 in Mandarin. For each participant, two samples of 30-second speech were extracted from their interviews in three languages. The commercial x-vector ASR system, Phonexia Voice Inspector (v.4.0.0), was used for testing and in total 306 samples were used as input for the ASR system. A total of 204 same-speaker (SS) and 10200 different-speaker (DS) comparison scores were generated for each language pair, which were exported and calibrated using cross-validated logistic regression (Brümmer, et al., 2006). Individual speaker performance was evaluated using SS log-likelihood ratios (SS-LLRs) and DS log-likelihood ratios (DS-LLRs). To test the effect of language proficiency, a series of linear models were run using mean SS-LLRs and DS-LLRs by speaker as dependent variables, and proficiency ratings, age and their interaction as fixed effects.

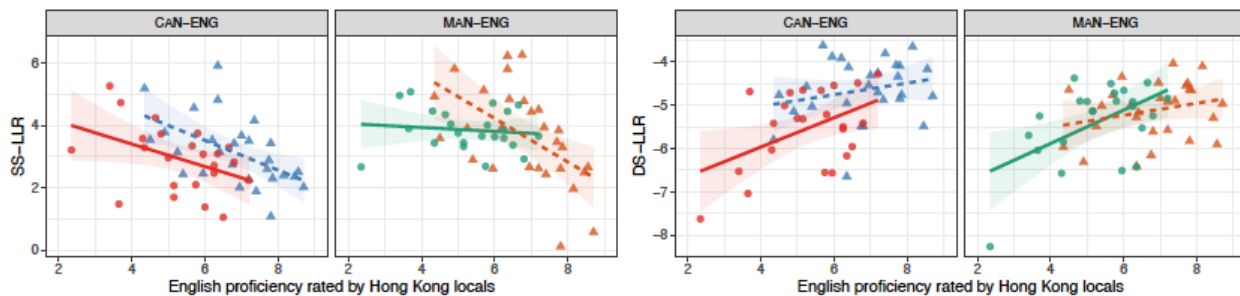
Results suggest that English proficiency was significant for both SS-LLRs and DS-LLRs in the Cantonese-English mismatch condition, but no significant effect was found for Mandarin proficiency. As shown in Table 1 and Figure 1, when comparing speakers' Cantonese with English, those with lower English proficiency tend to have a higher SS-LLR and a lower DS-LLR, indicating a stronger discriminatory performance in the ASR system. In contrast, speakers with higher English proficiency are likely to have relatively poorer discriminatory performance. One explanation for the significant effect of English proficiency is that participants who have high English proficiency might be more likely to develop two separate articulatory settings in their L1 and L2, whereas speakers with lower English proficiency might use similar articulatory settings, resulting in more shared features between their L1 and L2. Another significant finding is that participants with a larger L2-L3 proficiency difference were more difficult to be distinguished from other speakers in the Mandarin-English comparisons, particularly among the older speakers.

There are two implications. First, different types of individual bilingualism might affect speakers' discriminatory performance, such as the balanced Cantonese-English speakers (i.e., speakers with

high English proficiency) are more likely to be a challenge for FVC using an ASR system. Second, practitioners should also consider the types of societal bilingualism where the population lies as it may affect how “typical” a bilingual is in the reference population. (Word count: 500 words)

	CAN vs ENG		MAN vs ENG	
	$\beta$	$p$	$\beta$	$p$
<b>SS-LLR</b>				
Age	1.54	0.272	<b>4.24</b>	<b>0.009</b>
Proficiency	<b>-0.36</b>	<b>0.033</b>	-0.06	0.740
Age $\times$ Proficiency	-0.12	0.606	<b>-0.638</b>	<b>0.016</b>
<b>DS-LLR</b>				
Age	1.76	0.113	1.37	0.190
Proficiency	<b>0.34</b>	<b>0.012</b>	<b>0.39</b>	<b>0.003</b>
Age $\times$ Proficiency	-0.20	0.254	-0.25	0.144

**Table 1.** Summary of linear models for English proficiency ratings by Hong Kong locals (significant effects in bold). Reference level for age is elderly.



**Figure 1.** Scatterplot of SS-LLRs (left panels) and DS-LLRs (right panels) in mismatched conditions against English proficiency rated by Hong Kong locals (elderly: red/green circles; young: blue/orange triangles) with best-fit lines (elderly: solid; young: dashed).

## References

- Brümmer, N., Burget, L., Černocký, J., Glembek, O., Grézl, F., Karafiát, M., van Leeuwen, D. A., Matejka, P., Schwarz, P., & Strasheim, A. (2007). Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7), 2072–2084.
- Lo, J. J. H. (2021). *Issues of bilingualism in likelihood ratio-based forensic voice comparison* (Doctoral dissertation, University of York, UK).
- Nuttall, B., Harrison, P., & Hughes, V. (2023). Automatic speaker recognition performance with matched and mismatched female bilingual speech data. *Proceedings of INTERSPEECH 2023*, 601–605. Dublin, Ireland.

# The effect of a language mismatch: Strength-of-evidence in multilingual forensic speaker comparison

Meike de Boer<sup>1</sup> and Willemijn Heeren<sup>1</sup>

<sup>1</sup>Leiden University Centre for Linguistics, Leiden University, the Netherlands  
{m.m.de.boer, w.f.l.heeren}@hum.leidenuniv.nl

A significant portion of the world’s population speaks multiple languages. This is reflected in forensic casework, where speech evidence in a case can be in more than one language (cf. Van der Vloed et al., 2014). This raises concerns about the impact of language mismatches in forensic speaker comparisons (FSCs) with multilingual evidential recordings. IAFPA’s Code of Practice (IAFPA, 2020) seemingly advises against multilingual FSCs, by recommending to “exercise particular caution”. However, any mismatch between the questioned and reference material is expected to deteriorate speaker comparisons at least to some extent, and other mismatch situations (e.g. recording device, speech situation) are not explicitly mentioned. Although we agree that FSC can be expected to be affected given a language mismatch between speech samples (see also Cao et al., 2024; Lo, 2021; Nuttall et al., 2023), performance may still be high enough to meaningfully contribute to FSC. As the size of the degradation is likely to depend on the language combination investigated, in this study we ran a Likelihood Ratio (LR) analysis examining the effect of a language mismatch with speakers of L1 Dutch and L2 English.

## Method

Data were taken from prior work investigating the language dependency of filled pause vowels (De Boer & Heeren, 2020), the bilabial nasal /m/ (De Boer & Heeren, 2023), and the voiceless sibilant /s/ (De Boer & Heeren, 2024). The features included are presented in Table 1. Several monolingual and cross-linguistic LR systems were built, which differed in the language combination and the precise set of tokens included. All LR analyses were performed twice, once including all features, and once including only those features that are language independent according to the prior work.

<i>Segment</i>	<i>Type of features considered</i>	<i>All features</i>	<i>Language-independent features</i>
uh, um	vowel formants	F1, F2, F3	F3
/m/	nasal formants	N1, N2, N3	N1, N2, N3
/s/	spectral moments	CoG, SD	SD

**Table 1.** Overview of the segments and features included in the analysis.

Speech recordings from 47 speakers were taken from D-LUCEA (Orr & Quené, 2017). Speakers were compared to themselves, splitting the recordings in half, and to all other speakers, giving 47 same-speaker and 1,081 different-speaker comparisons. LRs were computed using a MATLAB script (Morrison, 2011) which uses a leave-one-out implementation of Aitken and Lucy (2004)’s method. The background population consisted of all speakers from the dataset who were not involved in the comparison at hand (cf. Morrison, 2011). ELUB boundaries were applied to limit LR values, taking into account the size of the dataset (Vergeer et al., 2016). System performance was evaluated through (a) LLRs, (b) Equal Error Rate (EER), and (c) the log LR cost function ( $C_{LR}$ ; Brümmer & du Preez, 2006), using the R package *sretools* (Van Leeuwen, 2008).

## Findings

The results for each language combination (i.e. system) are presented in Table 2. As expected, systems in which there is a language mismatch (systems 3 to 6), perform worse than monolingual systems (systems 1 and 2). Monolingual systems with language-independent features only showed slightly lower performance compared to all-feature systems. For cross-language systems, however, performance improved with language-independent features only. Results show that cross-language knowledge of individual segments could help to increase FSC performance.

System	Languages		All features				Language-independent features			
	Com- parison	Back- ground	Median LLR		System performance		Median LLR		System performance	
			SS	DS	$c_{llr}$	EER	SS	DS	$c_{llr}$	EER
1	L1–L1	L1	0.85	−0.95	0.63	10.74	0.85	−0.80	0.69	14.71
2	L2–L2	L2	0.85	−0.80	0.67	11.68	0.70	−0.65	0.79	18.90
3	L1–L2	L1	0.70	−0.50	0.84	23.11	0.55	−0.65	0.80	18.93
4	L1–L2	L2	0.10	−0.05	0.97	23.60	0.55	−0.65	0.80	19.37
5	L2–L1	L2	0.55	−0.50	0.88	26.54	0.55	−0.50	0.84	21.31
6	L2–L1	L1	0.70	−0.50	0.83	22.31	0.55	−0.65	0.81	20.35

**Table 2.** Overview of the median log Likelihood Ratios (LLRs) for Same-Speaker (SS) and Different-Speaker (DS) comparisons, the LLR cost function ( $c_{llr}$ ), and Equal Error Rate (EER) for each system. Calculations were done including all features or including language-independent features only.

## References

- Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 53(1), 109-122.
- Cao, G. W., Hughes, V., Wang, B. X., & Mok, P. (2024, November). Cross-language forensic voice comparison of Hong Kong trilingual speakers using filled pauses and an automatic speaker recognition system. In *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)* (pp. 279-283). IEEE.
- de Boer, M. M., & Heeren, W. F. (2020). Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English. *Journal of the Acoustical Society of America*, 148(6), 3612-3622.
- de Boer, M. M., & Heeren, W. F. (2023). The language dependency of /m/ in native Dutch and non-native English. *The Journal of the Acoustical Society of America*, 154(4), 2168-2176.
- de Boer, M. M., & Heeren, W. F. L. (2024). Language Dependency of /s/ Production: Native Dutch Versus Non-Native English. *Language and Speech*, 68(1), 87-99.
- Brümmer, N. & du Preez, J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language*, 20(2-3): 230-275.
- IAFPA (2020). Code of practice. International Association for Forensic Phonetics and Acoustics, <https://www.iafpa.net/about/code-of-practice/>
- Lo, J. H. (2021). *Issues of bilingualism in likelihood ratio-based forensic voice comparison*. PhD thesis, University of York.
- Morrison, G.S. (2011). A comparison of procedures for the calculation of forensic likelihood ratios from acoustic-phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model-universal background model (GMM-UBM). *Speech Communication*, 53: 242-256.
- Nuttall, B., Harrison, P., & Hughes, V. (2023). Automatic Speaker Recognition performance with matched and mismatched female bilingual speech data. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 601-605).
- Orr, R., & Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project data. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 177–190). Ubiquity Press.
- van Leeuwen, D. A. (2008). *SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations*. Downloaded from <http://sretools.googlepages.com/>
- van der Vloed, D. L., Bouten, J. S., & van Leeuwen, D. A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. In *Proceedings of the Odyssey Speaker and Language Recognition Workshop 2014*, June 16–19, Joensuu, Finland, pp. 6–13.
- Vergeer, P., van Es, A., de Jongh, A., Alberink, I., Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: When to stop extrapolating? *Sci Justice*. 56(6): 482-491.

## Forensic Analysis of the Singing Voice

Ana Flavia Zuim<sup>1</sup>, Dominic Watt<sup>2</sup>, Geddy Warner<sup>1</sup>

<sup>1</sup>Steinhardt School, MPAP - New York University

afz1@nyu.edu, geddywarner@gmail.com

<sup>2</sup>J P French International, York, UK/Zurich, Switzerland

Speaker discrimination in sung speech presents unique challenges compared to ordinary speech, particularly in distinguishing speakers using instrumental analysis methods, as previous studies have shown (June, 2024; Taylor, 2024; Loni & Subbaraman, 2015). The role of voice quality in forensic speaker comparison is well-documented as a key factor in forensic analysis and speaker differentiation (Nolan, 2007). Singing has been used in threatening messages, to disguise voices, and in cases where chanting or singing incites hatred or threats. For example, UK legislation made sectarian chanting or singing at football games an offense in 2012 if it incites hatred based on religious or cultural affiliations. Cases like a caller singing racist lyrics to “Go West” highlight the potential benefits of research distinguishing speakers.

This study investigates whether formants, specifically F3, can reliably distinguish speakers in singing, even in the absence of ordinary speech for comparison. We hypothesize that F3 serves as a reliable marker of vocal identity, independent of pitch accuracy or vocal registration, with less variability than F1 and F2. To test this hypothesis, we analyzed the acoustic properties of five singers performing three phrases—“Here I go,” “Here I am,” and “I am here”—in two keys: G4 to C5 and Bb4 to Eb5. The singers repeated the first pitch twice for the first two syllables before shifting to a higher pitch in a perfect fourth interval. The dataset includes formant frequencies (F1–F3), fundamental frequencies (F0), and their standard deviations across both keys, providing insights into the consistency and variability of these parameters.

Mixed effects model and Tukey post hoc analysis revealed significant inter-singer variability in F3, with statistically significant differences in F3 values across participants at  $\alpha = .95$ . Intra-singer analysis showed most individuals maintained relatively stable formant frequencies, with coefficient of variation values ranging from 1.9% to 5.6%, suggesting consistent F3 values. In contrast, F1 and F2 exhibited higher variation, with coefficient of variation values ranging from 9.1% to 18.4% for F1 and 11.1% to 18% for F2. The study also explored how the *passaggio*—a shift commonly known as the transition between ‘chest voice’ and ‘head voice’—affects vocal acoustics (Titze, 1988). Analysis showed significant shifts in formant frequencies across the *passaggio* with key changes for F1 ( $p = 0.018$ ) and F3 ( $p = 0.0138$ ), but no significant effect for F2 ( $p = 0.711$ ).

These findings have practical implications for forensic voice identification, demonstrating that distinguishing individuals by their singing voice can improve speaker discrimination methods. The study shows that formants, especially F3, are reliable markers for vocal identity, providing a foundation for future research in forensic contexts. This analysis could help identify individuals involved in sectarian chanting, vocal disguise, or threats, contributing to the fight against hate speech. Additionally, the results suggest that singing may be less variable than speech due to the larger mouth opening required for singing, leading to more stable formant frequencies. Future studies with larger sample sizes are needed to explore this hypothesis and its implications for forensic voice analysis.

### References

- L. June, L. Cirelli, and M. Eitel, “Who is singing? Voice recognition from spoken versus sung speech,” *J. Acoust. Soc. Am. Express Lett.*, vol. 4, no. 6, p. 065203, 2024. [Online]. Available: <https://doi.org/10.1121/10.0026385>
- K. Taylor, A. Gully, and H. Daffern, “Familiar and unfamiliar speaker identification in speech and singing,” In *Proc. Interspeech*, 2024. doi: 10.21437/Interspeech.2024-1763

- Loni, D. Y., & Subbaraman, S. (2015). Singing voice identification using harmonic spectral envelope. *Proceedings of the 2015 International Conference on Information Processing (ICIP)*, 2015. <https://doi.org/10.1109/INFOP.2015.7489362>
- F. Nolan, "Voice quality and forensic speaker identification," 2007. <https://hrcak.srce.hr/file/256311>
- Titze, Ingo R. "A Framework for the Study of Vocal Registers." *Journal of Voice*, vol. 2, no. 3, 1988, pp. 183-194. [https://doi.org/10.1016/S0892-1997\(88\)80075-4](https://doi.org/10.1016/S0892-1997(88)80075-4).

# Long-term formant measurement in casework: The more formants, the better?

Anja Moos, Michael Jessen, Katharina Klug, and Almut Braun

Department of Text, Speech and Audio, Bundeskriminalamt, Germany

{Anja.Moos|Michael.Jessen|Katharina.Klug|Almut.Braun}@bka.bund.de

**Introduction and motivation:** When dealing with forensic voice comparison, vowel formants are among the most frequently analysed phonetic characteristics (Gold & French 2011). Besides vowel-specific analysis, there has been research on long-term formant measurements (LTF), looking at different aspects of LTF for speaker discriminating purposes, starting with Nolan & Grigoras (2005) and followed up, for example, by Moos (2010). Gold et al. (2013), Asadi et al. (2018), Hughes et al. (2018) and Chan & Wang (2024) are examples of LTF research using automated formant tracking. Our casework, however, makes use of manual correction of formant tracking.

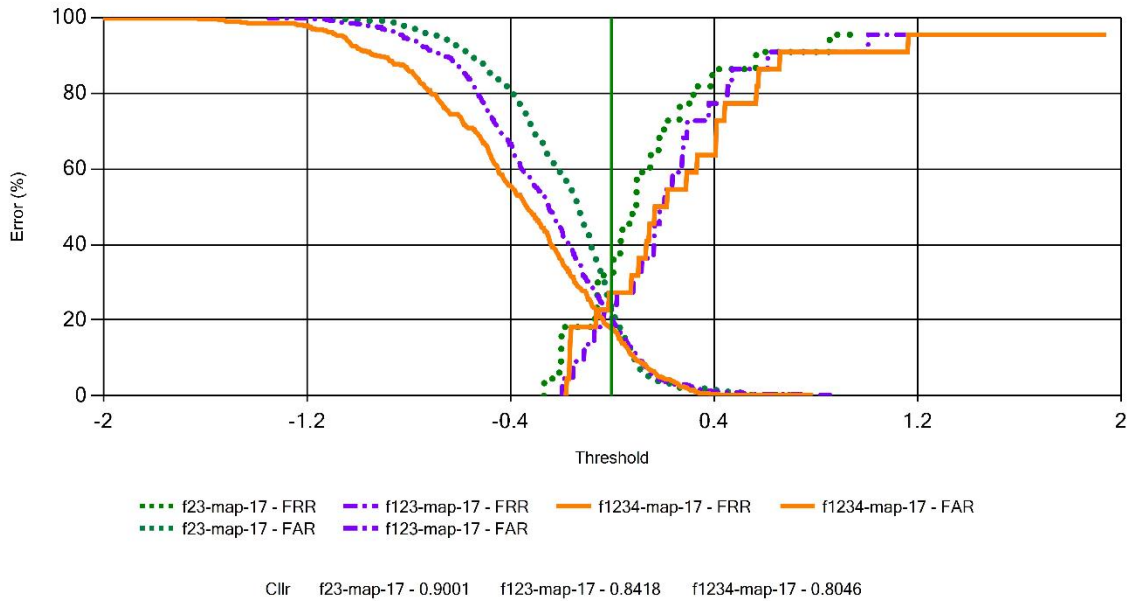
With a rising occurrence of voice messages in case work, limitations of the telephone bandwidth play less of a role for the analysis. That means, F1 and F4 could also be used for discriminating purposes, all of which have been found to be useful for speaker discrimination in better than telephone quality recordings (e.g. Cavalcanti et al., 2024, Cao & Dellwo, 2019). Our research investigates whether the additional measurement of LTF1 and especially LTF4 increases the speaker discriminatory power and whether the manual correction process of formant tracking is worthwhile to increase the validity of the data.

**Methodology:** Only voice messages from authentic casework were used for this investigation based on male voices speaking German. In order to (a) compare the speaker discriminatory power when different numbers of formants are used, and to (b) examine the usefulness of manually corrected data in opposition to non-corrected formant tracking, we conducted semi-automatic speaker recognition using the GMM-UBM approach (Gaussian Mixture Model – Universal Background Model) with MAP (maximum a posteriori) adaptation and 17 Gaussians using *Vocalise Legacy*. See Jessen (2021) for an explanation of the GMM-UBM approach used here as well.

**Results:** Results with a dataset of 22 speakers with two recordings per speaker and 25 speakers with single recordings for the UBM suggest that it is best to (a) include all formants F1-F4 to increase speaker discriminatory power and to (b) manually correct the formant tracking. Table 1 lists the EER's and Cllr's of manually corrected and uncorrected formant data. EER and Cllr are best when all four formants are used and manually corrected. The Tippett plot in Figure 1 visualizes the findings for answering research question (a). For casework, this means there is no need to refrain from using F1 and F4 when transmission range is better than telephone quality. Also it is worth investing in the manual work of phonetic experts and correct the automated formant tracking for discrimination purposes even though the benefit of manual correction seems to lose power when more formants are included.

		$F1+F2+F3+F4$	$F1+F2+F3$	$F2+F3$
EER (%)	corrected	18.3	20.0	26.9
	uncorrected	19.1	22.8	30.6
Cllr	corrected	0.57	0.61	0.78
	uncorrected	0.60	0.65	0.82

**Table 1.** Equal error rates (EER) and log-likelihood ratio cost (Cllr) for different combinations of formants (multivariate models), each for corrected and uncorrected LTF data. Cllr calculated after applying cross validation logistic regression calibration.



**Figure 1.** Tippett plot showing the performance of manually corrected LTF data, displaying same-speaker comparisons (rising to the right) and different-speaker comparisons (falling to the right). The plot and the Cllr values shown are before logistic regression calibration; GMM-UBM alone has reasonable calibration capabilities when applied to the LTF data.

## References

- Asadi, H., Nourbakhsh, M., Sasani, F. & Dellwo, V. (2018). Examining long-term formant frequency as a forensic cue for speaker identification: An experiment on Persian. In *Proc. of First International Conference on Laboratory Phonetics and Phonology*, Tehran, Iran, 21-28.
- Cao, H. & Dellwo, V. (2019). The role of the first five formants in three vowels of Mandarin for forensic voice analysis. In *Proc. of ICPHS 2019*, Melbourne, Australia, 617-621.
- Cavalcanti, J.C., Eriksson, A., Barbosa, P.A. & Madureira, S. (2024). Revisiting the speaker discriminatory power of vowel formant frequencies under a likelihood ratio-based paradigm: The case of mismatches speaking styles. *PLoS ONE* 19(12): e0311363.
- Chan, R.K.W. & Wang, B.X. (2024). Do long-term acoustic-phonetic features and mel-frequency cepstral coefficients provide complementary speaker-specific information for forensic voice comparison? *Forensic Science International*, 363, 112199.
- Gold, F. & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18, 293-307
- Gold, E., French, P. & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. In *Proc. Mtgs. Acoust.* 2013; 19 (1): 060041.
- Hughes, V, Harrison, P, Foulkes, P., French, P., Kavanagh, C. & San Segundo, E. (2018). The individual and the system: assessing the stability of the output of a semi-automatic forensic speaker recognition system. In *Proceedings of INTERSPEECH 2018*, Hyderabad, India, 227–230.
- Jessen, M. (2021). MAP adaptation characteristics in forensic long-term-formant analysis. In *Proc. Interspeech 2021*, 411-415.
- Nolan, F. & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *International Journal of Speech, Language and the Law*, 12(2), 143-173.
- Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech, *The Phonetician* vol. 101/102, 7–24.
- VOCALISE software. <https://oxfordwaveresearch.com/products/vocalise/>

# Between-speaker variability in information flow rate through temporal changes of spectral composition in speech signals

Lei He<sup>1</sup> Bruce Xiao Wang<sup>2</sup>

<sup>1</sup>*Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China*  
helei@fudan.edu.cn

<sup>2</sup>*Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong SAR, China*  
brucex.wang@polyu.edu.hk

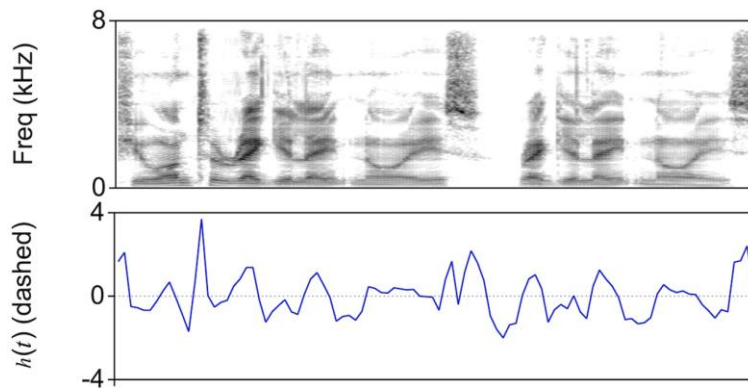
How fast an individual speaks is an acknowledged phenomenon that contains rich speaker-specific information and has been measured in terms of speaking rate or articulation rate calculated as the *number of syllables ÷ total utterance duration with(out) pauses* (see Gold 2014 for a comprehensive review in the context of forensic phonetics). The major drawbacks of this method are (i) annotations/segmentations are needed to count the number of syllables in each utterance, adding to the time cost of such analyses, and (ii) the syllable lengths (and the segment lengths therein) are not uniform, adding to the inherent imprecision issue of this type of measurements. **Here, we follow a new way to objectively characterize the rate of changes in the spectral compositions of a speech signal (He 2025) and examine how well this method can reveal between-speaker variability.**

The information in speech is delivered via temporal changes of spectral compositions; a prolonged [ə:::] with little spectral variability over time carries little information, as a counterexample (He 2025). Such spectral changes evolve temporally because of dynamic source-and-filter activities. To summarize the shape of the spectrum taken at each frame (shape = Hann, length = 25ms, overlap = ¾), each spectral slice was treated as a probability distribution whereby its entropy was calculated. The entropy increases as the spectral slice more closely resembles a uniform distribution and decreases as it approaches a spike. The concatenated entropy values from successive frames provide an estimate of spectral shape evolution, serving as a proxy for the information flow (See Fig. 1 for an illustration) (He 2025). To obtain comprehensive knowledge of the rate of information changes, the entropy time series was Fourier transformed to obtain its spectrum. From this spectrum, the centroid, spread, skewness and kurtosis (similar to the method reported in He 2022) were calculated as summary statistics for different aspects of the information changing rates. We investigated speaker differences from these four variables.

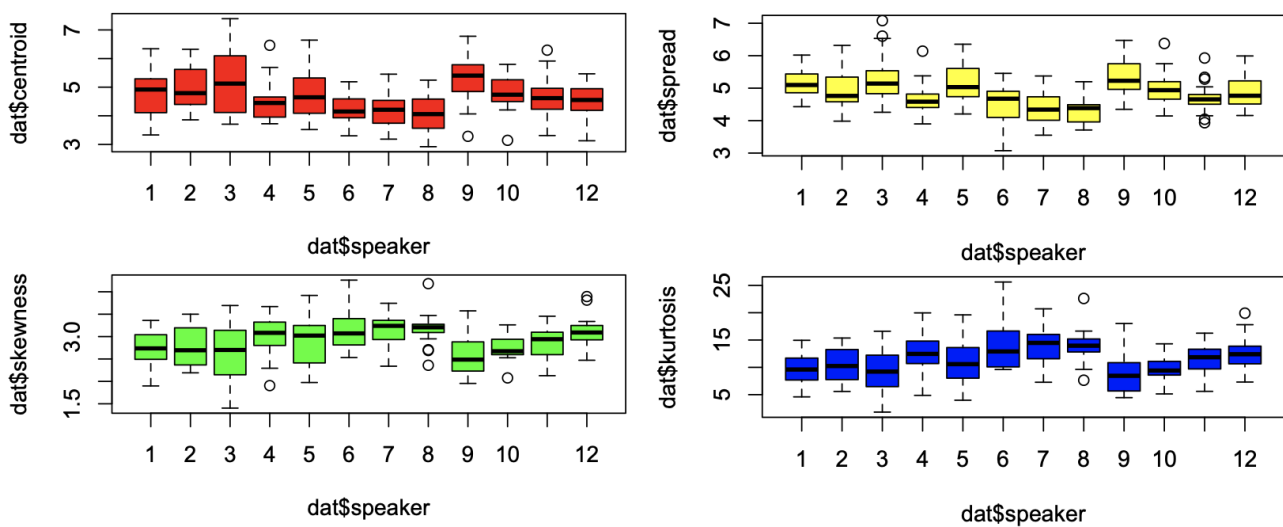
Till now, we have data from 12 native speakers of Mandarin (each produced 20 utterances). Using linear mixed-effects models, all four variables showed significant *speaker* effect (see Fig. 2 and 3). We are keeping increasing our sample sizes both in terms of the number of speakers and the amount of speech materials produced by each speaker. By the time of the conference, we will have a decent amount of data to report more on the discriminatory power of these variables.

## References

- Gold, E. (2014). Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters [Doctoral dissertation], University of York.
- He, L. (2025). Mouth rhythm as a “packaging mechanism” of information in speech: A proof of concept. *J. Acoust. Soc. Am.*, 157(3), 1612–1617.
- He, L. (2022). Characterizing first and second language rhythm in English using spectral coherence between temporal envelope and mouth opening-closing movements. *J. Acoust. Soc. Am.*, 152(1), 567–579.



**Figure 1.** Illustration on the spectral entropy time series (blue curve, centered around zero) as an approach to reveal the changes of spectral composition in speech. This curve is used as a proxy to show the rate of information changes in speech.



**Figure 2.** Boxplots illustrating speaker variations on centroid (red), spread (yellow), skewness (green), and kurtosis (blue).

```

> mod1 = lmer(centroid ~ speaker + (1|sentence), data = dat, REML = F)
> mod1R = lmer(centroid ~ 1 + (1|sentence), data = dat, REML = F)
> anova(mod1, mod1R)
Data: dat
Models:
mod1R: centroid ~ 1 + (1 | sentence)
mod1: centroid ~ speaker + (1 | sentence)
      npar  AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
mod1R   3 522.72 533.16 -258.36  516.72
mod1    14 436.44 485.17 -204.22  408.44 108.28 11 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> mod2 = lmer(spread ~ speaker + (1|sentence), data = dat, REML = F)
> mod2R = lmer(spread ~ 1 + (1|sentence), data = dat, REML = F)
> anova(mod2, mod2R)
Data: dat
Models:
mod2R: spread ~ 1 + (1 | sentence)
mod2: spread ~ speaker + (1 | sentence)
      npar  AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
mod2R   3 428.78 439.22 -211.39  422.78
mod2    14 330.27 379.00 -151.14  302.27 120.51 11 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> mod3 = lmer(skewness ~ speaker + (1|sentence), data = dat, REML = F)
> mod3R = lmer(skewness ~ 1 + (1|sentence), data = dat, REML = F)
> anova(mod3, mod3R)
Data: dat
Models:
mod3R: skewness ~ 1 + (1 | sentence)
mod3: skewness ~ speaker + (1 | sentence)
      npar  AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
mod3R   3 277.74 288.18 -135.868  271.74
mod3    14 221.42 270.15 -96.711  193.42 78.315 11 3.121e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
> mod4 = lmer(kurtosis ~ speaker + (1|sentence), data = dat, REML = F)
> mod4R = lmer(kurtosis ~ 1 + (1|sentence), data = dat, REML = F)
> anova(mod4, mod4R)
Data: dat
Models:
mod4R: kurtosis ~ 1 + (1 | sentence)
mod4: kurtosis ~ speaker + (1 | sentence)
      npar  AIC    BIC logLik deviance  Chisq Df Pr(>Chisq)
mod4R   3 1295 1305.4 -644.48  1289
mod4    14 1234 1282.7 -602.98  1206 82.989 11 3.888e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

**Figure 3.** R output: model comparison results from linear mixed-effects models (full models vs. speaker-reduced models) revealing that the models containing *speaker* effects have better fit.

## Assessing the suitability of f0 estimators with respect to recording condition and voice quality

*Katharina Klug and Markus Niermann*

*Department of Text, Speech and Audio, Bundeskriminalamt, Germany*

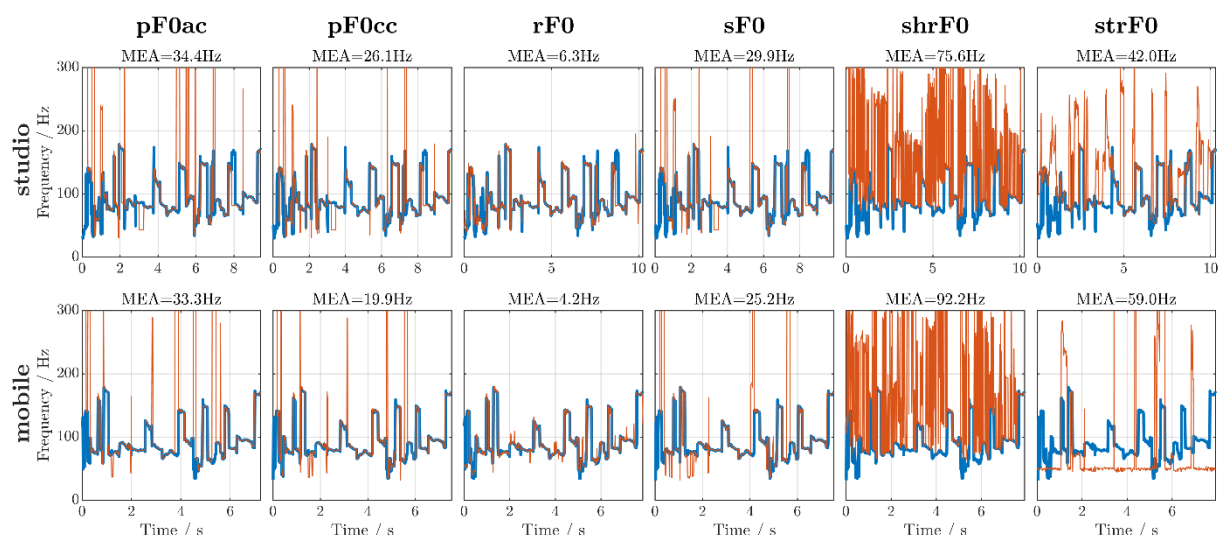
{Katharina.Klug|Markus.Niermann}@bka.bund.de

This exploratory study investigates the influence of recording condition and voice quality (VQ) on the performance of fundamental frequency (f0) estimators. Both degraded recording condition and non-modal VQ are hypothesised to negatively affect the validity of f0 measurements. In forensic casework usually the default f0 estimator of the preferred software is used instead of critically questioning its suitability for the material at hand. Same applies for VQ studies. However, valid automatic f0 estimation reduces the time-consuming manual f0 correction process in casework. Also, valid f0 estimates are required to obtain valid spectral slope measurements in VQ studies, as voice analysis programs such as VoiceSauce (Shue et al. 2011) locate harmonics based on f0. The study therefore investigates the robustness of f0 estimators using controlled productions of sustained cardinal vowels phonated by one male and one female speaker under two recording conditions (*studio* and *mobile phone*) in *modal*, *breathy* and *creaky* VQ.

Six f0 estimators were tested that are commonly used in our field: *Praat Autocorrelation* (pF0ac) (Boersma, 1993), *Praat Cross-Correlation* (pF0cc) (Boersma and Weenick, 1992-2025), *REAPER* (rF0) (Talkin, 2015), *Snack* (sF0) (Talkin, 1995), *Subharmonic-to-harmonic ratio* (shrF0) (Sun, 2000; Sun, 2002), and *STRAIGHT* (strF0) (Kawahara et al., 2008; Kawahara et al., 2012). VoiceSauce (Shue et al., 2011) was used to automatically conduct the f0 measurements of the six tested f0 estimators.

To assess the performance of the f0 estimators, the f0 ground truth was manually determined in both recording conditions by measuring the length of each pitch period to derive f0 ( $f_0 = 1/T_0$ ). The manually corrected f0 path was interpolated with the data points of the f0 estimators. The error measure *Mean Error Absolute* (MEA) is used to indicate the difference between the ground truth f0 and the automatically extracted f0 of each tested f0 estimator in Hz. MEA specifies the mean distance between these two frequencies without weighting the error.

The results show that five out of six f0 estimators perform well for modal voice under studio conditions. However, attention should be given to automatic f0 tracking when the voice of interest deviates from modal VQ. The GSM mobile phone network did not deteriorate the performances of f0 estimation in general, but only for some f0 estimators on individual non-modal VQs. Overall, the *Snack* estimator and the two *Praat* estimators proved to be all-rounders for most of the voice qualities and recording conditions tested. Furthermore, we found that *REAPER* performs well for vowels containing *aperiodic creak*, where f0 is particularly difficult to detect (see Fig. 1) and that the *STRAIGHT* f0 estimator should be avoided for mobile-filtered recordings. The results allow an informed choice of f0 estimators for the f0 analysis in casework and the VQ analysis in voice studies.



**Figure 1.** Performance of the individual f0 estimators for the FEMALE speaker for CREAKY cardinal vowels under STUDIO and MOBILE recording condition. The blue line shows the f0 ground truth, the orange line represents the performance of the tested f0 estimators. The name of the f0 estimators and the error metric MEA (mean error absolute) are indicated above each plot. Abbreviations are based on those used in VoiceSauce: PRAAT Autocorrelation (pF0ac), PRAAT Cross-Correlation (pF0cc), REAPER (rF0), Snack (sF0), Subharmonic-to-Harmonic-Ratio (shrF0), and STRAIGHT (strF0).

## References

- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *IFA Proceedings, Amsterdam*, Vol: 17, 97–110.
- Boersma, P., & Weenick, D. (1992-2025). Praat manual. [www.fon.hum.uva.nl/praat/manual/Sound\\_\\_To\\_Pitch\\_\\_raw\\_cc\\_\\_\\_\\_.html](http://www.fon.hum.uva.nl/praat/manual/Sound__To_Pitch__raw_cc____.html)
- Kawahara, H., Morise, M., Nisimura, R., & Irino, T. (2012). Deviation measure of waveform symmetry and its application to high-speed and temporally-fine F0 extraction for vocal sound texture manipulation. In *Proceedings of Interspeech, USA*, 386–389.
- Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., & Banno, H. (2008). Tandem- STRAIGHT: a temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation. In *Proceedings of IEEE ICASSP, USA*, 3933–3936.
- Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. In *Proceedings of ICPhS, Hong Kong*, 1846–1849.
- Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio, In *IEEE ICASSP, USA*, Vol: 1, 333–336.
- Sun, X. (2000). A pitch determination algorithm based on subharmonic-to-harmonic ratio. In *Proceedings of ICSLP, China*, Vol: 4, 676–679.
- Talkin, D. (2015). Reaper: Robust epoch and pitch estimator [computer program]. [github.com/google/REAPER](https://github.com/google/REAPER)
- Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT). In W. B. Kleijn & K. K. Paliwal (Eds.), *Speech Coding and Synthesis* (pp. 497–518). Elsevier Science B.V.

# Properties of ENF interference in audio deepfakes: an exploratory study using ElevenLabs

Molly Schutten and Amelia Gully<sup>1,2</sup>

<sup>1</sup> Department of Language and Linguistic Science, University of York, UK

<sup>2</sup> Oxford Wave Research, Oxford, UK

m911ys@gmail.com, amelia.gully@york.ac.uk

## Introduction

The use of electrical network frequency (ENF) interference for audio authentication is well established (e.g. Grigoras, 2007). With the increasing prevalence of audio deepfakes or ‘spoofed’ speech (Wang et al., 2024), it is of interest to assess existing authentication techniques, like ENF analysis, for their potential in identifying synthetic speech.

This exploratory study uses voice conversion spoofs (where both the utterance content and speaker-specific information are derived from audio files, rather than written text) to investigate how ENF interference is affected by the spoofing process.

## Materials and methods

Spontaneous and read speech were recorded for three speakers (2F, 1M) in a quiet room, using both a close microphone (Rode SmartLav+) and a desk-mounted microphone (Shure SM58). Recordings took place in the UK, so the expected frequency of ENF interference is around 50Hz. Some ENF interference was detected in the original recordings. Recordings were also notch filtered at 50, 100 and 150 Hz, and assessed to ensure no ENF interference was present, providing ENF-free control recordings. Finally, ENF interference obtained from archive recordings was artificially added to the control recordings. As such there were three audio conditions, referred to as ‘no ENF’, ‘recorded ENF’, and ‘archive ENF’.

For the present study, one female voice (F2) was always used as the *target* file (the voice that the spoofing system would attempt to mimic), and the remaining male (M1) and female (F1) voices were used as *source* files (the source of the spoofed utterance content). Voice conversion spoofs were created for both source speakers using every combination of speech task and microphone configuration. Several voice conversion systems were tested; results reported here are for the Eleven Labs ‘Instant Voice Clone’ speech-to-speech system (ElevenLabs, 2024) using 50-second audio clips taken from the start of each recording.

## Results

The main finding of this study, illustrated in Figure 1, is that when ENF interference is present in the *target* audio file, distinctive ‘ENF-like’ content also appears in the spoof; this is true whether ENF is present in the *source* file or not. However, this ENF-like interference does not show the narrow bandwidth of real ENF content, and would not be suitable for a typical ENF analysis process. Where ENF interference is present in the *source* file but not the *target* file, it does not appear in the spoof. This general behaviour was not affected by the source speaker sex, speech tasks, or microphone configurations tested in this study.

Several additional observations were made. In some cases, the target audio file contained archive ENF with only a 100Hz harmonic present, and no 50Hz fundamental, but the resulting spoofs contained ENF-like content around 50Hz as well as 100Hz; this is assumed to be due to co-occurrence of these frequencies in the training data.

Frequency (Hz)

Frequency (Hz)

**Figure 1.** With a 100Hz ENF harmonic present in the target audio file (left), ‘ENF-like’ frequency content appears in the spoof (right). No ENF was present in the source file for this example. Top row: spectrograms of 50s audio files; bottom row: average spectra over same regions.

Several other examples showed that the amplitude of the ENF-like component was often highly variable, dropping in and out more abruptly than would be expected with genuine ENF interference. Future work is planned to explore the outcomes with different voice conversion systems, and to quantify the behaviour of the ENF-like interference observed here.

## Conclusion

This exploratory study has shown that ENF interference is not directly reproduced by spoofing systems; instead a distinctive ‘ENF-like’ trace is visible in the spoofer where ENF was present in the target audio file. The distinctive properties of this trace make it a promising candidate for further study, as one tool in the toolbox of audio deepfake detection.

## References

- ElevenLabs (2024). ElevenLabs Instant Voice Clone [computer software]. Eleven Multilingual V2 model; accessed August 2024. <https://elevenlabs.io/>.
- Grigoras, C. (2007). Applications of ENF criterion in forensic audio, video, computer and telecommunication analysis. *Forensic Science International*, 167(2-3), 136-145.
- Wang, X. et al. (2024). ASVspoofer 5: Crowdsourced Speech Data, Deepfakes, and Adversarial Attacks at Scale. In *Proceedings of The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024)*, Kos, Greece, 1-8.

# Empirical study on the application of forensic audio authentication evidence in Chinese courts (2016-2020)

Honglin Cao, Danyang Li

Key Laboratory of Evidence Science (China University of Political Science and Law), Ministry of Education, China.

caohonglin@cupl.edu.cn | li\_danyang0910@sina.com

## Introduction

Since the *Provision on the Online Issuance of Judgment Documents by People's Courts* passed by the Supreme People's Court of China in 2013, over 153 million judgment documents (JDs) have been publicly accessible online, providing an unprecedented resource for big data analysis in forensic studies. In the field of forensic phonetics (FPs), Cao and Zhang (2020) analyzed 244 FP-related JDs in 2017 (one year) to investigate the characteristics and existing problems of the FP evidence application in Chinese courts. To explore long-term trends in FP development, this study extends the analysis to a five-year period (2016–2020) and focuses specifically on forensic audio authentication (FAA).

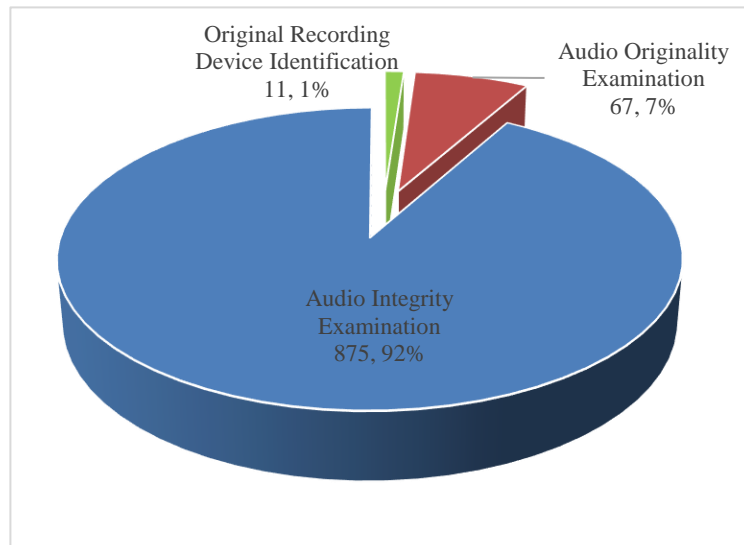
## Method

A total of 1,451 FP-related JDs were retrieved from *China Judgments Online* using seven keyword combinations related to FP. Of these, 497 valid JDs involving FAA were identified. Key variables analyzed included: case volume, case nature, cause of action, forensic institutions, recording types, recording devices, authentication tasks, expert opinions, expert testimony, and re-examination outcomes.

## Results

- (1) **Case volume:** The dataset included 497 JDs with 953 authentication examinations. Annual trends exhibited a “decline-rise-decline” pattern, peaking in 2018 (114 cases). Geographically, Guangdong (55 cases), Fujian (46), and Beijing (44) dominated, while western provinces had minimal caseloads.
- (2) **Case type:** Civil cases accounted for 82% (predominantly contract disputes, 74.5%), with private lending (46%) and sales contracts (22%) most common. Criminal cases accounted for 17%, with drug-related offenses and fraud (47.6%) as leading categories. Administrative cases represented less than 1%.
- (3) **Forensic institution:** Seventy-three forensic institutions across 24 provinces were identified. Private entities comprised 49%, followed by university-affiliated labs (29%) and government-affiliated institutes (22%; police, procuratorate, etc.). The Institute of Forensic Science in Shanghai served 11 provinces, while the Southwest University of Political Science and Law Forensic Center exhibited the broadest geographic coverage (13 provinces).
- (4) **Recording type:** Call recordings (52.9%) were most common, followed by face-to-face recordings (30%) and WeChat voice messages (12.3%). Mobile phones (67.9%) dominated as recording devices, while digital voice recorders saw a 50% usage decline over five years.
- (5) **Expert opinion:** In audio integrity examinations, 65% of conclusions stated “no evidence of editing”, and 26% confirmed “unedited”. For originality examinations, 56% of the opinions concluded “no signs of non-original recording”. Eleven cases involved a disputed task “original recording device identification”. The relationships among these concepts (authenticity, integrity, originality, original device) reflect the conflicts between the two guidelines in FAA in China, the Ministry of Justice guideline vs the Ministry of Public Security guideline.
- (6) **Expert testimony:** Expert witness appearance rate was very low (2.4%, 12/497 cases). Cross-examination focused on procedural validity (35%), evidence originality (30%), and method of examination (20%), etc.
- (7) **Re-examination:** Only three re-examination cases were documented, all yielding conflicting conclusions with initial reports. For example, a fraud case in 2017 illustrated contradictory

opinions: the initial conclusion of a private institution in Tianjin “confirmed edited” vs. the conclusion of the re-examination of a public institution in Beijing changed to “inconclusive”. All re-examinations arose from parties’ dissatisfaction with original conclusions.



**Figure 1.** A pie chart for different types of the forensic audio authentication evidence

## References

Cao, H. & Zhang, X. (2020). An Empirical Study on the Present Status of the Application of Evidence of Forensic Phonetics in Courts of China. *Journal of Chinese Phonetics.*,1:90-104. (in Chinese)

# Towards an interpretation framework for forensic audio deepfake detection

*Finnian Kelly<sup>1</sup>, Anil Alexander<sup>1</sup>, Anna Bartle<sup>2</sup>, Colleen Driscoll<sup>3</sup>  
and Peter Milne<sup>3</sup>*

<sup>1</sup>*Oxford Wave Research, Oxford, UK*

<sup>2</sup>*Forensic Services, Metropolitan Police, UK*

<sup>3</sup>*Royal Canadian Mounted Police, Canada*

{finnian|anil}@oxfordwaveresearch.com, anna.bartle@met.police.uk,  
{Colleen.Kavanagh|Peter.Milne}@rcmp-grc.gc.ca

Deepfake audio detection systems are designed to analyse an input speech sample and produce a detection score, which can be used to inform a decision about whether the speech sample in question is *real* (has been produced by a human speaker) or *deepfake* (has been generated using a text-to-speech or voice conversion model of a specific speaker’s voice).

A key question facing the use of such systems in a forensic context is how to reliably interpret the resulting detection score in a way that is suitable to inform legal decision-making. A natural solution is to apply a likelihood ratio (LR) framework to convert the detection score into an interpretable format.

A deepfake LR could be defined as a measure of the relative strength of support for two competing hypotheses: the likelihood of the evidence (i.e. the detection score) if  $H_{DF}$  (the hypothesis that the sample is deepfake) is true, divided by the likelihood of the evidence if  $H_R$  (the hypothesis that the sample is real) is true. Evaluating such an LR requires a representative set of real and fake samples, based on how these hypotheses are defined. Drawing on the LR framework as commonly applied in speaker recognition (Drygajlo, 2015), we could consider both specific-source and common-source approaches to defining the hypotheses:

- A *common-source* approach is agnostic to the specific identity of the speaker in the questioned sample, with potential hypotheses  $H_{DF}$ : “the speech in the questioned sample is deepfake” and  $H_R$ : “the speech in the questioned sample is human”, requiring a representative set of deepfake and real samples.
- A *specific-source* approach poses hypotheses specific to the speaker in the questioned sample, with potential hypotheses  $H_{DF}$ : “the speech from the speaker in the questioned sample is deepfake” vs “the speech from the speaker in the questioned sample is human”, requiring a representative set of deepfake and real samples from the specific speaker in the questioned sample.

If there are multiple real samples available for the speaker in the questioned sample, it may be possible to generate sufficient deepfake samples to adopt a specific-source approach; however, the common-source approach is likely to be the most practical option.

With either approach, the selection of representative real and deepfake samples is of central importance to the LR. To inform this selection, it is important first to understand factors affecting deepfake detection performance, including:

- Technical: recording device, noise and compression, duration.
- Speaker: sex/gender, age, language, accent.
- Algorithmic: the specific deepfake generation method.

To begin the process of developing an interpretation framework for audio deepfake detection, we will explore the application of different hypotheses and identify some of the key factors to consider in the selection of representative data. We will demonstrate some possibilities through examples with controlled and in-the-wild data, using FAUXDIO deepfake detection software.

We will also discuss the practitioners' perspective on interpreting the output of a deepfake detector, which raises questions about handling unknowns like the deepfake generation algorithm, and in a speaker comparison case involving a questioned deepfake, whether to apply a two-stage process of deepfake detection followed by speaker comparison, or to combine the two in tandem.

## References

Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen J., and T. Niemi (2015), Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition, *Frankfurt: Verlag für Polizeiwissenschaft*

## **FAUXDIO: An audio deepfake detector for law enforcement and forensics**

*Anil Alexander<sup>1</sup>, Linda Gerlach<sup>1</sup>, Thomas Coy<sup>1</sup>, Oscar Forth<sup>1</sup>,  
Liam Lonergan<sup>2</sup> and Finnian Kelly<sup>1</sup>*

*<sup>1</sup>Oxford Wave Research, Oxford, UK*

*{anil|linda|tom.coy|oscar|finnian}@oxfordwaveresearch.com*

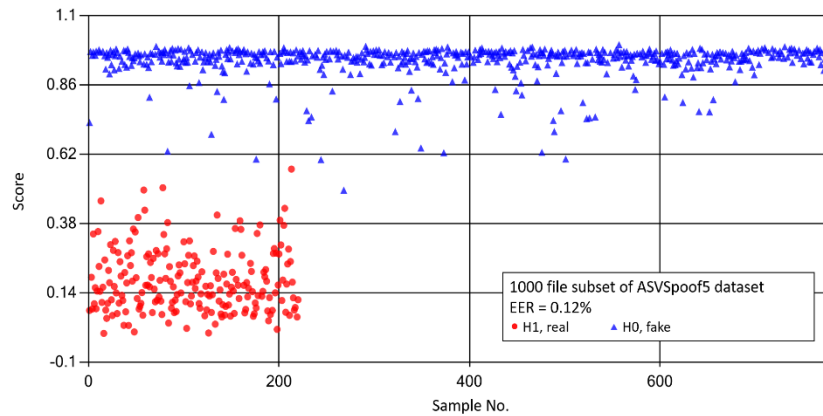
*<sup>2</sup>Phonetics and Speech Laboratory, Trinity College Dublin, Dublin, Ireland.*

*llonerga@tcd.ie*

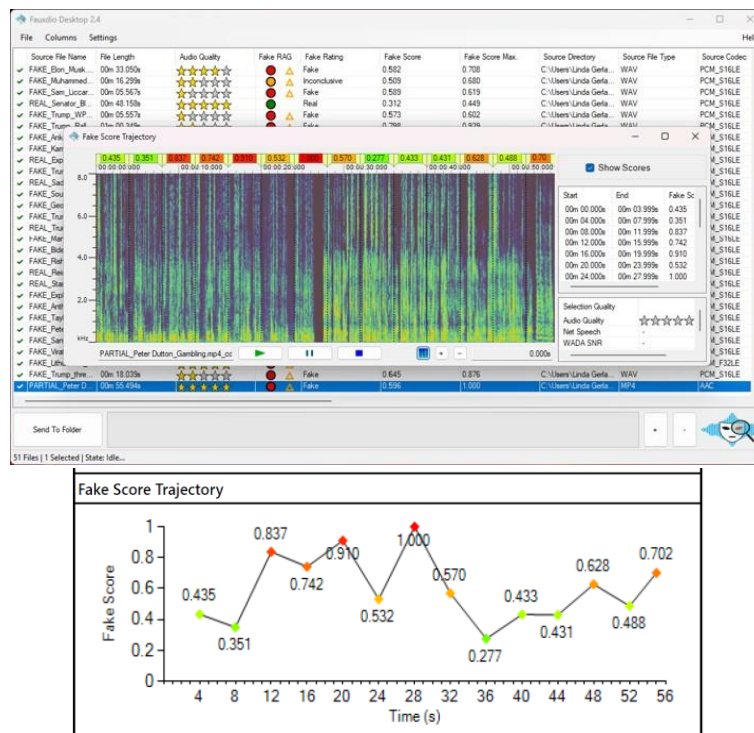
Synthetic speech generation has now reached a level of sophistication allowing natural and realistic speech, capable of easily fooling the human ear, to be generated with relative ease either through text-to-speech synthesis or voice conversion. Linguistic and phonetic cues that could previously be relied upon as reliable indicators of fake content can now be masked to a certain extent using voice conversion with a suitable donor voice. Synthetic speech targeting a specific individual (often referred to as audio deepfakes) can now play a part in forensic cases including extortion and blackmail, false implication of innocent individuals, fraud and political disinformation. Development of countermeasures and tools to detect and defeat deepfakes has become the focus of governments, commercial institutions, and academia around the world. As part of the 2024 UK Home Office Deepfake Detection Challenge (Shanks, 2024), we developed an audio deepfake detection solution called FAUXDIO, which is capable of ingesting an audio or video file and automatically outputting an indication of whether the speech contained within the file is potentially real or fake. FAUXDIO relies on a DNN (deep neural network) detection model trained with many examples of real and fake speech and has a configurable decision threshold. The detection model can be run fully offline (e.g., on-premise) within a Windows desktop application for audio extraction (from video), conversion, and playback. FAUXDIO Desktop version, aimed at forensic users, enables user control via selectable detection models and configurable RAG (Red Amber Green) decision thresholds. The same detection models can also be run within FAUXDIO Web, which is a collaborative tool for multiple analysts to leverage insights obtained from samples provided by many users.

We calibrated the system using publicly available deepfake data including real deepfakes ‘in the wild’, as well as our internal test sets to provide meaningful RAG ratings. The tool further allows fine-tuning of decision thresholds and selection of detection models for specific use cases. We also developed a partial fake detection functionality which would highlight to the user potential partial fake regions in an otherwise largely real file. FAUXDIO Web provides an overall fake rating and score for an input file, along with a transcription of the speech, which is colour-coded to indicate which (if any) regions may be fake. Furthermore, the deepfake detector was connected with our MADCAT audio fingerprinting tool (Alexander et al. 2015) to recognise previously-seen fakes, thereby leveraging the ‘wisdom of the crowd’.

To demonstrate the detection performance on controlled data, we tested a subset of 1000 samples from the ASVspoof5 dataset (Wang et al. 2025) (780 fake samples, involving 9 different fake speech generation algorithms, and 220 real samples), resulting in an EER of 0.12%. Figure 1 shows a scatter plot of the data with fakes as blue triangles and reals as red circles. The FAUXDIO audio deepfake detection tool (see Figure 2) aced the 2024 Deepfake Challenge and has demonstrated similar strong performance in subsequent 2025 Home Office office user-trials and benchmarking.



**Figure 1.** Scatter plot of FAUXDIO results based on ASVspoof5 data.



**Figure 2.** FAUXDIO Desktop interface with fake score segments highlighted in red for a partially fake audio file (above) and its fake score trajectory from the generated report (below).

## References

- Accelerated Capability Environment (ACE; 2025, February 05). Innovating to detect deepfakes and protect the public. *UK government case study*. <https://www.gov.uk/government/case-studies/innovating-to-detect-deepfakes-and-protect-the-public>
- Alexander, A., Forth, O., Atreya, A. (2015). Audio fingerprinting to detect illegal content in digital media files. In *Proc. International Association for Forensic Phonetics and Acoustics (IAFPA) conference 2015*. Leiden, The Netherlands.
- Shanks, K. (2024, July 30). Innovative solutions unveiled at the Deepfake Detection Challenge Showcase. *Accelerated Capability Environment (ACE) blog*. <https://ace.blog.gov.uk/2024/07/30/innovative-solutions-unveiled-at-the-deepfake-detection-challenge-showcase/>, retrieved on 28.03.2025.
- Wang, X. et al. (2025), ASVspoof 5: Design, Collection and Validation of Resources for Spoofing, Deepfake, and Adversarial Attack Detection Using Crowdsourced Speech, *arXiv preprint*. <https://arxiv.org/abs/2502.08857>

# What can voice quality features do in detecting deepfake speech in forensic scenarios

Kang Jintao<sup>1</sup>, Jin Tian<sup>2</sup>, and Peng Cheng<sup>1</sup>

<sup>1</sup>*Institute of Forensic Science, Ministry of Public Security, China*

<sup>2</sup>*Criminal Investigation Department, Jiangsu Provincial Department of Public Security, China.*

kangjintao@cifs.gov.cn

In recent years, generative artificial intelligence has made remarkable advancements in the field of speech generation technology (Kim, Kong & Son, 2021), posing serious threats to social security (CNN, 2024). Various deepfake speech detection algorithms (Jung, et al. 2022; Tak, Patino, et al. 2021; Tak, Todisco, et al. 2021) have been proposed and achieved excellent results on their respective test sets. However, in practical applications such as forensic scenarios, these models often fail due to various reasons (Yan, Zhao & Wang, 2024). Meanwhile, there is a critical need for the explainability in forensic methods, which current models lack but traditional acoustic-phonetic features can provide. This study explores the effectiveness of voice quality features in detecting deepfake speech and uses feature perturbation to assess their importance.

## Data and Methods

**Bona fide speech:** 50 recordings from RASC863 (Li and Wang, 2003) were split into 10-second segments, resulting in 1324 segments, from which 1000 were randomly selected.

**Fake speech:** Zero-shot voice cloning from CosyVoice 2 (Du, et al. 2024), GPT-SoVITS-V3 (GPT-SoVITS, 2025) and Spark-TTS (Wang, Jiang, et al. 2025) generated 50 recordings per tool using RASC863 texts. After splitting 150 recordings into 10-second segments, 1000 segments were randomly selected.

**Features:** Pitch (mean and standard deviation), jitter, shimmer and Harmonic-to-Noise Ratio (HNR) were extracted using Parselmouth (Jadoul, Thompson & De Boer, 2018) and preprocessed with Scikit-learn.

**Classifier:** A Pytorch-based LSTM network (Warren, et al. 2025) was constructed, with architecture shown in Table 1.

<i>Layers</i>	<i>Input</i>	<i>Output</i>	<i>Parameters</i>
LSTM1	(B, T, 1)	(B, T, 100)	40,400
BatchNorm1d	(B, 100, T)	(B, 100, T)	200
LSTM2	(B, T, 100)	(B, T, 50)	30,200
BatchNorm1d	(B, 50, T)	(B, 50, T)	100
Linear1	(B, 50)	(B, 50)	2,550
Linear2	(B, 50)	(B, 1)	51
Sigmoid	(B, 1)	(B, 1)	0
<b>Sum:</b>			73501

**Table 1.** Network architecture of the classifier

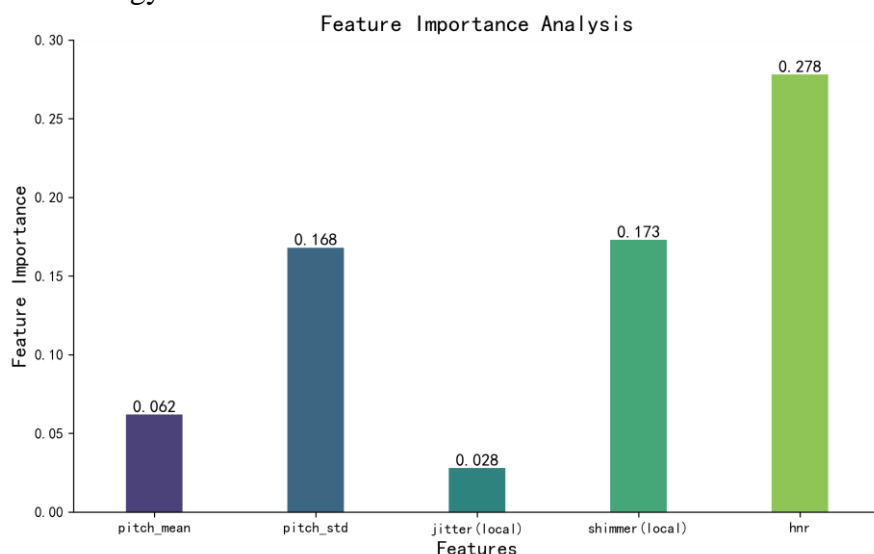
## Results

Only the network architectures of RawNet2 (Tak, Patino, et al. 2021) and AASIST2 (Tak, Todisco, et al. 2022) were adopted in this study for using F-bank. Table 2 shows the results. Though not as good as AASIST2, the voice quality model did show its potential in detecting deepfake speech.

<i>Models</i>	<i>Accuracy</i>	<i>F<sub>1</sub></i>	<i>EER</i>
VoiceQuality+LSTM	0.942	0.933	0.031
F-bank+RawNet2	0.936	0.932	0.035
F-bank+AASIST2	0.958	0.954	0.019

**Table 2.** Performance of three classifiers

Feature Importance Analysis via perturbation method (Brocki and Chung, 2023) in Figure 1 revealed critical voice quality features, which could facilitate forensic explainability and guide new speech authentication methodology.



**Figure 1.** Feature importance analysis of the VQ+LSTM model

## . References

- Brocki, L., & Chung, N. C. (2023). Feature perturbation augmentation for reliable evaluation of importance estimators in neural networks. *Pattern Recognition Letters*, 176, 131-139.
- CNN. (2024). Finance worker pays out \$25 million after video call with deepfake 'chief financial officer'. (2024, February 4). *CNN*. <https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html> (Accessed: 2025, March 7)
- Du, Z., Wang, Y., Chen, Q., et al. (2024). Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arxiv preprint arxiv:2412.10117*.
- GPT-SoVITS project. (2025, March 2). *GitHub*. <https://github.com/RVC-Boss/GPT-SoVITS> (Accessed: 2025, March 7)
- Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A python interface to praat. *Journal of Phonetics*, 71, 1-15.
- Kim, J., Kong, J., & Son, J. (2021). Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. *International Conference on Machine Learning*. PMLR, 5530-5540.
- Li, A. J., Wang, T. Q., et al. (2003). RASC863 - Voice Corpus for Speaker Recognition. In *Proceedings of the Seventh National Conference on Human-Machine Voice Communication*.
- Tak, H., Patino, J., Todisco, M., et al. (2021). End-to-end anti-spoofing with rawnet2. *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6369-6373.
- Tak, H., Todisco, M., Wang, X., et al. (2022). Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arxiv preprint arxiv:2202.12233*.
- Wang, X., Jiang, M., Ma, Z., et al. (2025). Spark-TTS: An Efficient LLM-Based Text-to-Speech Model with Single-Stream Decoupled Speech Tokens. *arxiv preprint arxiv:2503.01710*.
- Warren, K., Olszewski, D., Layton, S., et al. (2025). Pitch Imperfect: Detecting Audio Deepfakes Through Acoustic Prosodic Analysis. *arxiv preprint arxiv:2502.14726*.
- Yan, Z., Zhao, Y., & Wang, H. (2024). Voicewukong: Benchmarking deepfake voice detection. *arxiv preprint arxiv:2409.06348*.

# How Do Expert Witnesses Survive the Hot Seat? Discourse Strategies of Expert Testimony in Cross-Examination

Yuyao Yang

Department of Environment, Education and Development, University of Manchester, UK  
17832551567@163.com

The adversarial system, characterized by structured opposition between legal parties, places significant weight on the cross-examination phase (Doak et al., 2021). During cross-examination, both sides often exert verbal pressure on the expert witnesses, challenging their professional conclusions, questioning their analytical methods, and emphasizing potential biases to undermine the credibility of the testimony (Lazer, 2021). As a vital source of evidence, expert testimony impacts the case outcome (Griffin, 2023), so it is crucial for expert witnesses to employ suitable discourse strategies in response to different case contexts. Hedging (Vass, 2017; Ward, 2015) and syntactically complex constructions (Shuy, 2011) are two widely applied discourse techniques for expert witness. However existing studies have predominantly focused on describing these strategies separately, with limited examination of their application in response to distinct cross-examination types. Moreover, comparative analyses of these discourse features across civil and criminal cases remain scarce.

This study fills these gaps by exploring how expert witnesses utilize hedging and syntactically complex constructions to counter different types of cross-examination, namely fact-elucidation, method-interrogation, and credit-undermining (Cairns, 2016). In addition, in order to compare the adaptive adjustment of expert witnesses to language strategies under different standards of proof and reveal the law systematic impact of legal procedure differences on expert discourse patterns, this study selects representative adversarial trial cases from the Anglo-American legal system, covering both criminal and civil cases. Furthermore, this study uses corpus analysis to examine the frequency and distribution patterns of hedging and syntactically complex constructions in expert testimony, combining critical discourse analysis (CDA) to delve into power dynamics in the courtroom as well as the construction of ideology.

In detail, this study will focus on three research questions. Firstly, How do expert witnesses employ hedging and syntactically complex constructions to maintain testimonial credibility when facing different types of cross-examination in adversarial trials? Secondly, how do expert witnesses adapt their discourse strategies to meet the different standards of evidence in both criminal and civil cases? Thirdly, How do expert witnesses' discourse strategies affect jurors' perceptions of credibility and influence their decision-making in adversarial trials? Theoretically, this research contributes to forensic linguistics by providing a new perspective and reference for the understanding expert testimony and legal discourse features. Practically, it offers empirical insights for expert witnesses to optimize their discourse strategies, thereby enhancing the credibility and acceptance of their testimony in adversarial proceedings.

## References

- Cairns, D. J. (1999). *Advocacy and the making of the adversarial criminal trial 1800–1865*. Oxford University Press.
- Doak, J., Jackson, J., Saunders, C., Wright, D., Gomez Farinas, B., & Durdiyeva, S. (2021). *Cross-examination in criminal trials towards a revolution in best practice?*
- Griffin, L. K. (2023). *False Accuracy in Criminal Trials: The Limits and Costs of Cross-Examination*. *Tex. L. Rev.*, 102, 1011.
- Lazer, S. (2021). *The principle of orality: An analysis of the principles governing the prevalence of direct oral testimony in the English adversarial trial system and the impact of reforms to reduce its status* [University of Huddersfield].
- Shuy, R. W. (2011). *The language of perjury cases*. Oxford University Press.

Vass, H. (2017). Lexical verb hedging in legal discourse: The case of law journal articles and Supreme Court majority and dissenting opinions. *English for Specific Purposes*, 48, 17-31.

Ward, H. V. (2015). *A comparative analysis of hedging in a corpus of two written legal discourse genres*. Universidad Politécnica de Madrid.

# Teaching practices in Forensic Language Analysis

James Tompkinson<sup>1</sup>

<sup>1</sup>*Department of Language and Linguistic Science, University of York, UK*  
james.tompkinson@york.ac.uk

As the field of Forensic Language Analysis (hereafter FLA)<sup>1</sup> has expanded over recent decades, increasing amounts of research has been devoted to what might be described as ‘the state of the field’ (Gold and French, 2011; 2019; Morrison et al., 2016; Elstein and Kredens, 2023; Clarke and Kredens, 2018). However, relatively little scholarly attention has focused on the teaching of FLA within university settings. This paper aims to provide a greater understanding of current FLA teaching practices in relation to topics which have either been researched in wider pedagogical literature, or have parallels with existing research relating to practice and evidence provision in FLA. These issues are 1) the provision of content warnings and the protection of mental health and wellbeing, 2) the use of real and relevant forensic cases in teaching, and 3) the development of teaching guidelines and standards for FLA.

An online survey was designed and distributed to people who either currently teach FLA in higher education or have taught it in the past. Thirty participants took part in the research, recruited using convenience sampling. In the survey, participants were asked to state the extent to which they agreed with a series of statements using a five-point scale. The first six statements related to the provision of content warnings and the protection of mental health for both staff and students. This was followed by seven statements relating to teaching materials and the use of data from real forensic cases in the classroom. Finally, there were five statements relating to ethics and the provision of standards and guidelines for teaching FLA. After indicating their level of agreement with each set of statements, participants could provide further opinions in a free-text response.

Results indicated that participants were generally in favour of the use of content warnings, that they believed students and staff should be provided with wellbeing support, that the FLA syllabus should refer to authentic cases, and that teaching materials should not be adapted to remove references to serious crimes. Participants were also generally in favour of more discussion of issues relating to teaching at academic conferences, and indicated they had a clear understanding of the key issues around responsible teaching practice in FLA. Participants also responded with cautious positivity to the idea of guidelines for FLA teaching, with some explaining concerns regarding the potentially restrictive nature of any guidelines or standards. These results indicate that the calls made by Carlyle-Davis (2022) and Mullen et al. (2024) for a greater degree of discussion, research, standardisation and guidance in forensic science teaching may be merited for FLA.

The goal of this paper is not to promote a particular model for good practice or directly suggest standards and guidelines for teaching, as this would require a great deal of further consultation and collaboration. The aim of this paper is simply to provide an illustration of the current state of the field with respect to pedagogical issues, and provide a springboard for a potentially long-overdue discussion of teaching practices in FLA.

## References

Carlyle-Davis, F. (2022). Do we need a forensic science teaching network?. *Science & Justice*, 62(6), 827-829.

---

<sup>1</sup> Throughout this paper, I use the term “Forensic Language Analysis” as a label for any discipline which involves the analysis of language for forensic or legal purposes. This includes Forensic Phonetics / Forensic Speech Science, Forensic Linguistics, and Language and the Law.

- Clarke, I., & Kredens, K. J. (2018). I consider myself to be a service provider: Discursive identity construction of the forensic linguistic expert. *International Journal of speech, Language and the Law*, 25(1), 79-107.
- Elstein, S., & Kredens, K. (2023). Occupational stress in forensic linguistic practice. *Journal of Applied Linguistics & Professional Practice*, 17(1), 50-72.
- Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language & the Law*, 18(2), 293-307.
- Gold, E., & French, P. (2019). International practices in forensic speaker comparisons: second survey. *International Journal of Speech, Language and the Law*, 26(1), 1-20.
- Morrison, G. S., Sahito, F. H., Jardine, G., Djokic, D., Clavet, S., Berghs, S., & Dorny, C. G. (2016). INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic science international*, 263, 92-100.
- Mullen, C., Gallacher-Graham, S., Hammond, K., Myles, H., and Tidy, H. (2024). Sensitive Subjects in Forensic Science Education. *Chartered Society of Forensic Science conference*, Leeds.

# Thank you for watching: automatically evaluating transcriptions for hallucinations and missing meaning

*Jadd Virji and Finnian Kelly*

*Oxford Wave Research, Oxford, UK*

{jadd|finnian}@oxfordwaveresearch.com

The performance of automatic speech recognition (ASR) systems has advanced rapidly, making them useful for investigation and triage of speech samples in forensic contexts. Such conditions, however, present challenges for ASR (Loakes, 2022). First, words can be mistranscribed or omitted altogether. Second, ‘hallucinations’—fabricated words transcribed in the absence of intelligible speech—can appear in transcriptions (Koenecke et al., 2024; Barański et al., 2025). Using Whisper large-v3, a multilingual ASR model (Radford et al., 2023), we find that hallucinations primarily occur when transcribing audio files that are overly noisy or contain significant non-speech (Barański et al., 2025). This may occur as the model attempts to transcribe such unintelligible audio as speech. Hallucinations can take the form of bonafide phrases in the transcription repeated incorrectly, or of words transcribed that are unrelated to the audio. To evaluate transcriptions in the presence of these issues, we separate ASR errors into two different types: simple mistranscriptions, affecting how a transcription can convey the meaning of the source speech, and hallucinations.

The ‘gold standard’ evaluation is human judgement of a transcription against the ground truth, which is expensive in terms of time and cost. The standard word error rate (WER) between a transcription and the ground truth is simple to calculate, but is inadequate to determine semantic similarity. For example, the incorrect insertion of the word “not” in a transcription would result in a minute increase in WER, but reverse the meaning of the transcribed sentence. This motivates an approach in the middle-ground between WER and human judgements.

We propose a new methodology using large language models (LLMs), such as OpenAI’s GPT-4 (Achiam et al., 2023), with a carefully designed prompt, to evaluate ASR transcriptions. We use few-shot in-context learning—providing the LLM examples of human-rated transcriptions, as well as descriptors of each quality level (Table 1)—to induce accurate results (Dong et al., 2022; Sahoo et al., 2024). Scores are on a Likert-style scale from 1 to 7 of the semantic similarity between the candidate transcription and the ground truth (Joshi et al., 2015). Mitigating against LLMs’ indeterminism, we average scores over several runs (Klishevich et al., 2025). A hallucination score is similarly obtained by combining information-theoretic and LLM-generated metrics. We use separate scores for semantic conveyance and hallucination. Our initial pilot studies indicate that these measures correlate with human judgements.

A useful application of these LLM metrics is to assess potential improvements to the ASR pipeline; in Table 2 we show an example of a transcription of pilot communications evaluated before and after a signal conditioning process (voice activity detection and noise removal). Conditioning the audio greatly reduces the number of hallucinations in the transcription, reflected in the score increasing from 1 (worst possible) to 7 (best possible).

Therefore, LLMs can be used to rapidly assess the quality of ASR-produced transcriptions against ground truths, in terms of semantic content and the presence of hallucinations. Further, these automated metrics correlate with human judgements, and are thus a valuable tool for supporting the development of improved ASR approaches for forensics.

<i>Leve</i>	<i>Excerpt of description</i>
<i>l</i>	
1 (worst)	none or almost none of the meaning of the ground truth transcription— terrible
2	a fairly minimal amount of the meaning of the ground truth transcription
3	a little bit of the meaning of the ground truth transcription
4	some of the meaning of the ground truth transcription—about half of it
5	most of the meaning of the ground truth transcription
6	almost all of the meaning of the ground truth transcription
7 (best)	identical or almost identical meaning as the ground truth transcription— perfect or almost perfect

**Table 1.** Excerpts of descriptors given to the LLM when determining semantic similarity between a candidate transcription and the ground truth.

<i>Conditioning</i>	<i>Transcription</i>	<i>Semantic score</i>	<i>Hallucination score</i>
Default	“Red eye one one, jeep nine one boom. You guys can go ahead and depart the area. Copy that, we’ll descend low, we’ll see you guys in about an hour. Copy, thanks man. <i>We have Squallow today. Super cool, as it was in the past. Please get to Sister tatsächlich. Over the radio as fast as you can. But I’ll let つてる do it later phải. I’ll call from Little Baby. Awesome. Alright, time Chicken. Thanks for watching!</i> ”	5	1
Conditioned	“I-11, jeep 911 boom, that’s the go ahead and depart the area. Copy that. We’ll descend low. We’ll see you guys in about an hour. Copy. Thanks, man.”	5	7

**Table 2.** An example of transcription scores before and after signal conditioning.

## References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Barański, M., Jasiński, J., Bartolewska, J., Kacprzak, S., Witkowski, M., & Kowalczyk, K. (2025). Investigation of Whisper ASR hallucinations induced by non-speech audio. *arXiv preprint arXiv:2501.11378*.
- Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., ... & Sui, Z. (2022). A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.
- Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: explored and explained. *British Journal of Applied Science & Technology*, 7(4), 396.
- Klishevich, E., Denisov-Blanch, Y., Obstbaum, S., Ciobanu, I., & Kosinski, M. (2025). Measuring determinism in large language models for software code review. *arXiv preprint arXiv:2502.20747*.
- Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., & Sloane, M. (2024). Careless Whisper: speech-to-text hallucination harms. *In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1672-1681).
- Loakes, D. (2022). Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7, 803452.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *In International Conference on Machine Learning* (pp. 28492-28518). PMLR.
- Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: techniques and applications. *arXiv preprint arXiv:2402.07927*.

# Discovery and retrieval of speakers from large unlabelled datasets using scalable clustering

*Thomas Coy, Finnian Kelly and Anil Alexander*

Oxford Wave Research Ltd, Oxford, United Kingdom

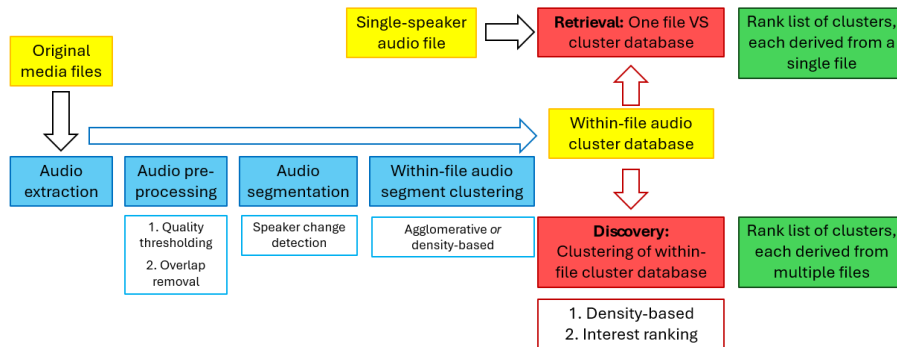
{finnian|tom.coy|anil}@oxfordwaveresearch.com

In digital forensic investigations containing large quantities of multimedia files it may be necessary to rapidly triage hundreds or thousands of unlabelled files, to either find previously unknown speakers of interest or to search for specific speakers. Without any indication as to how many speakers are in each file, who those speakers are, or whether speakers reappear elsewhere in the unlabelled data, this could become exceedingly challenging for both manual or computational methods. We propose a two-tier clustering-based approach to address these issues. The first stage is to run within-file clustering to create a cluster database. This then enables:

1. Retrieval - compare a single-speaker reference sample against the database
2. Discovery - second-level clustering to group within-file clusters that contain the same speaker

Naive clustering involves comparing every identified cluster in each file against all clusters in the database. This has  $O(n^2)$  complexity that scales poorly as the number of files increases, making clustering with thousands of files computationally prohibitive. Here we propose a scalable clustering that sidesteps this issue.

Pre-processing includes the removal of audio of inadequate quality and regions where speakers are overlapping. The remaining audio is then segmented and x-vectors extracted, before running agglomerative hierarchical clustering to estimate how many speakers are in a file and where they are speaking, creating a within-file cluster database.



**Figure 1.** A flowchart showing the sequence of clustering processes.

For retrieval, VOCALISE (Kelly et al, 2019) is used to compare a reference sample against the database. For discovery, a density-based clustering algorithm is used to find groups of within-file clusters belonging to the same speaker. Each discovery group undergoes an all versus all VOCALISE comparison to generate comparison metrics to rank groups by match strength (i.e. a high mean score suggests a more coherent group).

To evaluate performance, a curated subset of 386 files from VoxCeleb1 (Nagrani et al, 2017) was used. Files contain speech from at least two speakers, only one of which is labelled, and speakers appear in more than one video. There is diversity in recording conditions and speaker population.

Discovery elicited a total of 407 groups containing clusters from 2 or more different files. Of these, 26 groups were determined to be pure (i.e. containing a single, labelled speaker). Pure groups covered 22 of the 129 unique labelled speakers in the curated dataset. Thus, there were several labelled speakers who appeared in more than 2 pure groups. Of the remaining ‘impure’ groups, some were manually inspected and discovered to contain the same unlabelled speaker across files. For example, the fourth, blue-highlighted group in Figure 2 appears to contain results from Speakers 68 and 79. This group actually consists of speech from a mutual interviewer who appears in both files. As such the algorithm found links that were unknown to us.

By deploying clustering algorithms, rather than non-scalable, traditional NvN comparisons, thousands of completely unlabelled multimedia files can be rapidly analysed.

File Name	Audio Quality	Group Mean	No. of Original Files	Ind. Mean
speaker029_F_USA_sample001_cluster0002.wav	4	61.61	2	61.61
speaker029_F_USA_sample003_cluster0003.wav	4	61.61	2	61.61
speaker057_F_USA_sample003_cluster0005.wav	3	52.68	2	52.68
speaker068_F_USA_sample004_cluster0002.wav	3	52.68	2	52.68
speaker010_M_USA_sample002_cluster0001.wav	5	41.92	2	41.92
speaker010_M_USA_sample004_cluster0001.wav	5	41.92	2	41.92
speaker068_F_USA_sample002_cluster0002.wav	4	36.64	2	36.64
speaker079_M_USA_sample001_cluster0005.wav	5	36.64	2	36.64
speaker057_F_USA_sample001_cluster0004.wav	4	31.49	4	21.72
speaker057_F_USA_sample002_cluster0004.wav	4	31.49	4	30.08
speaker057_F_USA_sample004_cluster0004.wav	3	31.49	4	36.64
speaker057_F_USA_sample006_cluster0004.wav	3	31.49	4	37.53
speaker002_M_IND_sample001cluster0004.wav	3	28.27	2	28.27
speaker002_M_IND_sample002cluster0003.wav	5	28.27	2	28.27
speaker056_M_USA_sample001_cluster0002.wav	2	27.87	4	38.34
speaker056_M_USA_sample003_cluster0004.wav	2	27.87	4	24.26
speaker056_M_USA_sample004_cluster0004.wav	2	27.87	4	22.83
speaker056_M_USA_sample005_cluster0003.wav	4	27.87	4	26.05
speaker083_F_USA_sample001_cluster0002.wav	5	25.29	3	39.26
speaker083_F_USA_sample002_cluster0004.wav	4	25.29	3	30.81
speaker083_F_USA_sample003_cluster0002.wav	3	25.29	3	5.79
speaker034_F_MEX_sample002_cluster0001.wav	3	24.47	4	-7.90
speaker034_F_MEX_sample003_cluster0004.wav	4	24.47	4	36.86
speaker034_F_MEX_sample005_cluster0002.wav	4	24.47	4	33.79
speaker034_F_MEX_sample006_cluster0003.wav	4	24.47	4	35.15

**Figure 2.** Discovery results showing groups of within-file clusters that contain the same speaker.

## References

- Kelly, F., Forth, O., Kent, S., Gerlach, L. & Alexander, A. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proc. AES International Conference 2019*, Paper 27.
- Nagrani, A., Chung, J. S. & Zisserman, A. (2017). VoxCeleb: a large-scale speaker identification dataset. *INTERSPEECH*, (2017).
- Ruch, H., Fröhlich, A. & Lim, S. (2023). Clustering a large number of unknown voices. *Proc. International Association for Forensic Phonetics and Acoustics (IAFPA) Conference, Zurich, Switzerland, 23-24.*

## LiRI – voxplorer: an interactive dashboard to extract, visualise, and interact with feature-rich large speech corpora.

*Alessandro De Luca<sup>1,2</sup>, Srikanth Madikeri<sup>2</sup>, and Volker Dellwo<sup>2</sup>*

<sup>1</sup>*Linguistic Research Infrastructure, University of Zurich, Zurich, Switzerland*

<sup>2</sup>*Department of Computational Linguistics, University of Zurich, Switzerland.*

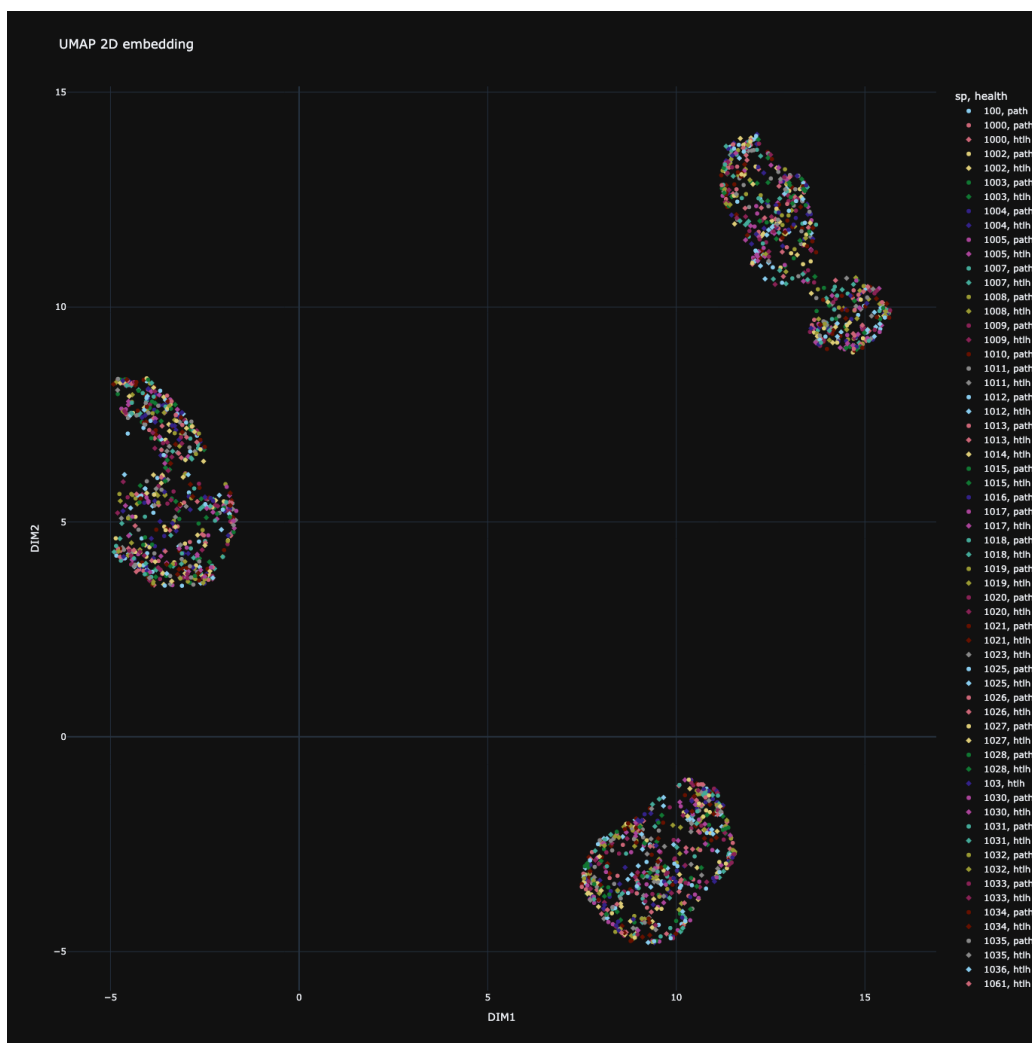
{alessandro.deluca|srikanth.madikeriraghunathan|volker.dellwo}@uzh.ch

Speech corpora are increasing in size and complexity, with more speakers and longer recordings, while advancements in computational power enable extraction of high-dimensional feature sets. However, these developments also introduce challenges for phonetics and speech sciences research, as contemporary datasets require powerful tools for exploration, visualisation and interpretation—tools that demand advanced programming skills and familiarity with dimensionality reduction algorithms.

There are a great deal of existing tools to analyse and extract features from speech samples; Praat (Boersma & Weenink, 1992–2022) and VoiceSauce (Shue, 2010) are two very well known examples. Online there are also several visualisation toolsets or interfaces, although it is rarer to find one that integrates dimensionality reduction. The LiRI – voxplorer dashboard aims to allow all speech scientists and students to have access to an integrated interface incorporating interactive visualisation of high-dimensional data through dimensionality reduction, including PCA (Pearson, 1901), UMAP (McInnes & Healy, 2018), MDS (Borg & Groenen, 1997), and t-SNE (van der Maaten & Hinton, 2008). It is built in Python and is freely available and open-source (GNU GPL v3.0). Users can upload a data table containing previously extracted speech features and categorical variables (metavariables) which are specified by the user after data upload. Alternatively, users can also upload a set of recordings and choose to extract either MFCCs or DNN speaker embedding features (metavariables are specified in the filename). Once the data is uploaded, users can run one of the dimensionality reduction methods implemented in *voxplorer* and visualise the embeddings directly in the dashboard. The platform enables data labelling using the user-specified categorical variables, which can be used for colour coding and shaping points in the visualisation and supports filtering and selection capabilities. Both the original features table and the embedded features table can be downloaded, as well as the currently selected subset of the data. *Voxplorer* will be available both online as well as a standalone tool that can be locally installed.

An important upcoming feature is “the recogniser”, which will integrate speaker verification functionalities for researchers. This mode will include state-of-the-art open-source speaker verification DNN pre-trained models alongside legacy models such as UBM-GMM (both pre-trained and user-trainable UBM). The inclusion of legacy models is intended to address research questions where accuracy is not pivotal and the extremely high accuracy of modern DNN approaches might mask the magnitude of more subtle effects.

The LiRI – voxplorer dashboard is our contribution to the field of voice communication sciences, offering an accessible platform for all researchers for data interaction and experimental speaker verification. With this tool we aim to empower voice researchers with access to state-of-the-art exploration, visualisation, and analytical tools that do not require extensive programming expertise. The project can be found at the following repository: <https://github.com/liri-uzh/voxplorer>. We envision the LiRI – voxplorer as the tool that can enable a cohesive analysis of complex feature sets and even shape the future researchers’ vision on voice communication sciences.



**Figure 1.** Example LiRI–voxplore visualisation results. UMAP 2–dimensional projection of 13 MFCCs (synthetic data simulating MFCCs extraction from Saarbrücken Voice Database (Putzer & Barry)).

## References

- Boersma P. & Weenink, D. (1992–2022). Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 23 January 2022 from <https://www.praat.org>.
- Borg, I. & Groenen, P. (1997). Modern multidimensional scaling – Theory and Applications. Springer Series in Statistics.
- McInnes, L. & Healy, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e–prints 1802.03426.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. In *Philosophical Magazine*, Vol: 2 (11), 559–572, doi:10.1080/14786440109462720.
- Putzer, M., & Barry, W.. Saarbrücken Voice Database. Institute of Phonetics, University of Saarland. <http://www.stimmdatenbank.coli.uni-saarland.de/>
- Shue, Y.–L. (2010). The voice source in speech production: Data, analysis and models. UCLA dissertation.
- van der Maaten, L. J. P. & Hinton, G. E. (2008). Visualizing high–dimensional data using t–SNE. In *Journal of Machine Learning Research*, Vol: 9, 2579–2605.

# Variability in the performance of automatic speaker recognition systems across modelling approaches

Lauren Harrington<sup>1</sup>, Vincent Hughes<sup>1</sup>, Philip Harrison<sup>1</sup>, Paul Foulkes<sup>1</sup>, Jessica Wormald<sup>1</sup>, Finnian Kelly<sup>2</sup>, and David van der Vloed<sup>3</sup>

<sup>1</sup>Department of Language and Linguistic Science, University of York, UK

{firstname.lastname}@york.ac.uk

<sup>2</sup>Oxford Wave Research, UK

finnian@oxfordwaveresearch.com

<sup>3</sup>Netherlands Forensic Institute, Netherlands

d.vandervloed@nfi.nl

Over the last three decades, there have been a series of generational changes to speaker modelling approaches used in automatic speaker recognition (ASR) systems. Improvements in performance are generally reported from one generation to the next using system-level metrics, such as Equal Error Rate (EER), which can mask performance variability as a function of speaker or other factors. However, in forensic contexts, it is crucial to understand speaker-specific variability. In this study, we use individual speaker-level metrics to evaluate the performance of four speaker modelling approaches for ASR.

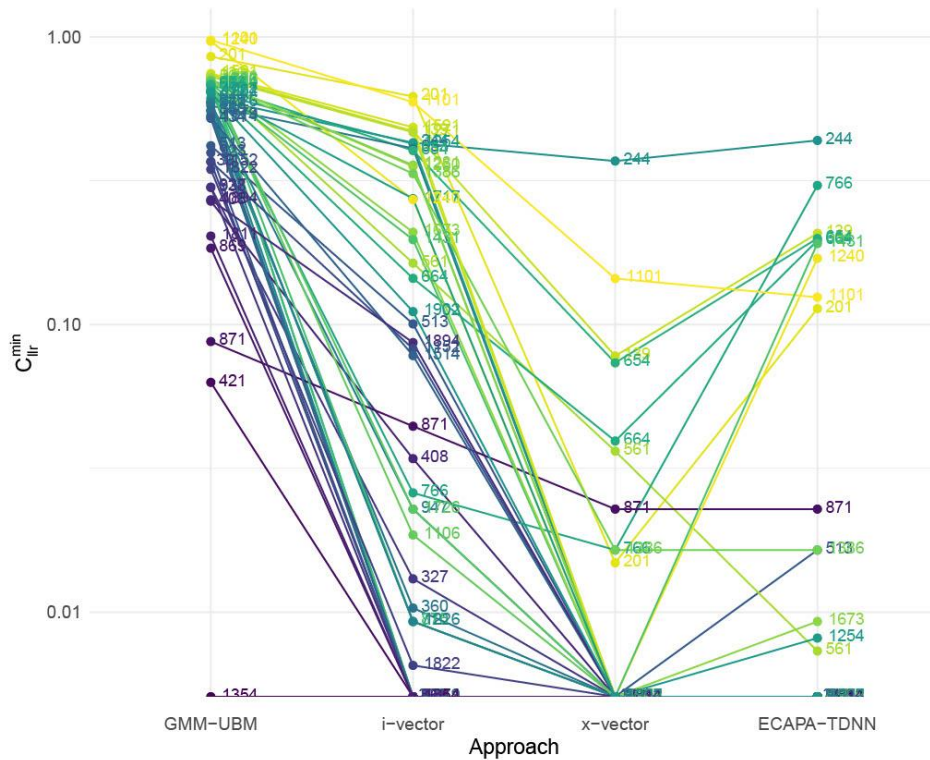
We used a set of forensically realistic recordings from the GBR-ENG corpus provided by the UK Government, which contains spontaneous conversational recordings of speakers with considerable variability in age and regional and social backgrounds. The data in this study is a subset (also used in Hughes et al., 2024) comprising mobile telephone recordings of 98 male speakers of British English, divided into a test set (48 speakers, 3-7 files each) and a calibration set (50 speakers, 2 files each). Testing was carried out using VOCALISE 2021 (version 3.0.0.1746; Kelly et al., 2019) and four speaker modelling approaches: (1) Gaussian Mixture Modelling with Universal Background Model (**GMM-UBM**) and Maximum A Posteriori (MAP) adaptation, (2) **i-vector** with dimension reduction via Linear Discriminant Analysis (LDA) and scoring with a pre-trained Probabilistic LDA (PLDA) model, (3) **x-vector** with dimension reduction via LDA and scoring with a pre-trained PLDA model, and (4) Emphasized Channel Attention, Propagation and Aggregation within a Time-Delay Neural Network (**ECAPA-TDNN**) with Cosine Distance scoring. Same-speaker (SS) and different-speaker (DS) scores were computed for each set. The calibration scores were used to train a logistic regression model and the coefficients were applied to the test scores to produce calibrated  $\log_{10}$  likelihood ratios (LLRs).

On a system- (Table 1) and individual speaker-level (Figure 1), we observe improvements from GMM-UBM to i-vector to x-vector but not for ECAPA-TDNN, which is outperformed by the x-vector system. Strong positive correlations are found between the speaker scores for each generation with its successor, and the rankings of speakers within a system are strongly positively correlated across GMM-UBM and i-vector, and x-vector and ECAPA-TDNN. Perfect separation of SS and DS scores (i.e. a  $C_{it}^{\min}$  of 0) is achieved for 38 of the 48 speakers using the x-vector system. We investigated the comparisons involving the 10 speakers for whom  $C_{it}^{\min}$  exceeded 0 (ranging from 0.015 to 0.37) and found that all speakers had mean average SS LLRs at the lower end of the SS LLR distribution for the x-vector system. Closer inspection of the SS LLRs demonstrated that 8 of the 10 speakers had a “problem” file, whereby comparisons involving that file resulted in low or negative LLRs but comparisons not involving that file resulted in high positive LLRs.

Approach	EER (%)	$C_{llr}$	$C_{llr}^{\min}$	$C_{llr}^{\text{cal}}$
GMM-UBM	44.5	0.97	0.92	0.05
i-vector	23.5	0.67	0.58	0.09
x-vector	3.0	0.13	0.10	0.03
ECAPA-TDNN	7.0	0.27	0.21	0.06

**Table 1.** Overall performance of the four modelling approaches. The metrics include Equal Error Rate (EER), Log Likelihood Ratio Cost Function ( $C_{llr}$ ) and its two components,  $C_{llr}^{\min}$  and  $C_{llr}^{\text{cal}}$ .

Preliminary auditory and acoustic analysis of the “problem” files revealed observable differences between the “problem” file and other files for each speaker, related to a range of technical, speaker and stylistic factors. The remaining 2 speakers did not have a “problem” file and the  $C_{llr}^{\min}$  above 0 was not caused by SS comparisons; rather, the overlap in SS and DS LLRs was mostly the result of high positive LLRs for multiple DS comparisons with one other speaker. Preliminary assessment of these files indicated more similarity in speaker-related factors (e.g. voice quality) rather than technical (e.g. distance from microphone, background noise) factors. Thus, the cause of the non-optimal performance for the 10 speakers is likely a complex interaction between technical, speaker-related and stylistic (e.g. increased vocal effort, pitch variability) factors. The findings clearly demonstrate that system-level performance metrics mask a wealth of detail about the behaviour of ASR systems. Understanding individual speaker and file behaviour will ultimately allow us to predict what types of behaviour are more likely to influence system performance, which in turn will assist analysts using ASR systems in forensic voice comparison cases.



**Figure 1.** By-speaker  $C_{llr}^{\min}$  across four different speaker-modelling approaches. Note that the y-axis is  $\log_{10}$  scaled to provide better resolution at the lower end of the scale (particularly for the x-vector and ECAPA-TDNN approaches).

## References

- Hughes, V., Xu, C., Foulkes, P., Harrison, P., Welch, P., Wormald, J., Kelly, F. & van der Vloed, D. (2024, April). Exploring individual speaker behaviour within a forensic automatic speaker recognition system. *Proceedings of Odyssey: The Speaker and Language Workshop*. Quebec.
- Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S., & Alexander, A. (2019). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic\_eval\_01). *Speech Communication*, 112, 30-36.

# Multiple Enrollments and Neural Back-End Modeling for Automatic Speaker Verification

Aron Paulsson<sup>1,5</sup>, Torbjörn Onshage<sup>1,5</sup>, Greta Öhlund Wistbacka<sup>2,5</sup>,  
Susanna Whitling<sup>3,5</sup>, and Andreas Jakobsson<sup>4,5</sup>

<sup>1</sup>*Faculty of Engineering (LTH), Lund University, Sweden.*

{aron.paulsson|torbjorn.onshage}@voiceprint.se

<sup>2</sup>*Department of Public Health and Caring Sciences, Uppsala University, Sweden.*

greta.ohlund.wistbacka@uu.se

<sup>3</sup>*Department of Logopedics, Phoniatrics and Audiology, Lunds University, Sweden.*

susanna.whitling@med.lu.se

<sup>4</sup>*Div. Of Mathematical Statistics, Lund University, Sweden.*

andreas.jakobsson@matstat.lu.se

<sup>5</sup>*Voice Print Sweden AB, Sweden.*

The rapid development of artificial intelligence offers intriguing possibilities in speech forensics, such as the development of automatic speaker verification (ASV) tools able to not only determine the similarity between voices but also focus the attention on the key aspects of the voices for such a comparison. The topic is experiencing a rapid development, but there are still many outstanding issues in need of further investigation.

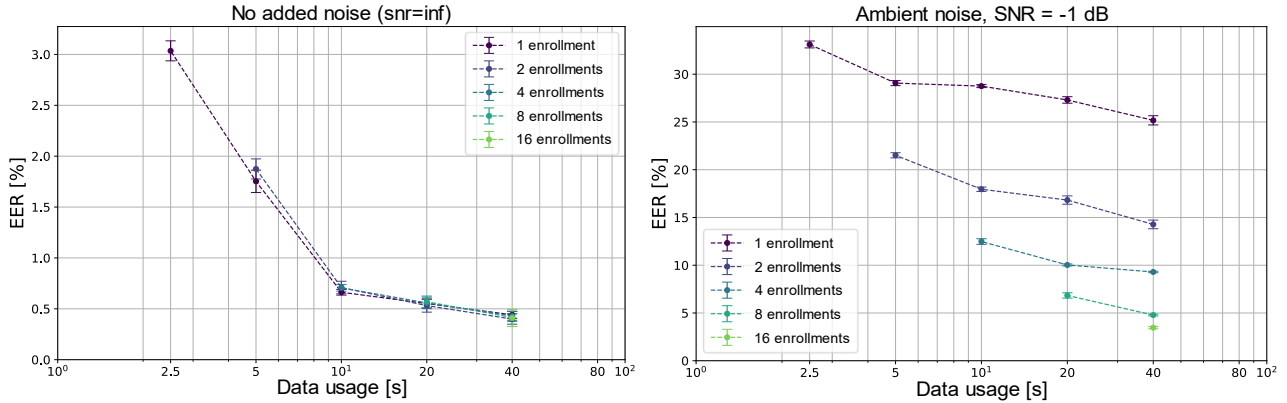
Modern ASV combines front-end feature extraction, using state-of-the-art methods based on Deep Neural Networks (DNNs), with back-end modeling used to differentiate between speakers based on the extracted features. Examples of the current development include the works by Yakovlev *et al.* (2024), Brümmer *et al.* (2022), and Silnova *et al.* (2023). Yakovlev *et al.* (2024) utilized reshaping to combine one-dimensional (1D) time delay neural network (TDNN) structures with 2D convolutional techniques for front-end feature extraction. They also made use of the additive angular margin (AAM) loss function, which is seeing more and more widespread use. On the other hand, Brümmer *et al.* (2022) introduced the Probabilistic Spherical Discriminant Analysis (PSDA) back-end model a variant of the commonly used Probabilistic Linear Discriminant Analysis (PLDA) better suited for combined use with AAM trained front-end feature extractors.

In this work, we build on their contributions, examining how sub-dividing longer speech sequences into multiple shorter utterances processed in parallel by the back-end model affects the ASV reliability. This yields a set of x-vectors, one for each shorter enrollment utterance, which are then sent to the front-end model. Our study evaluates the performance of the off-the-shelf ReDimNet-b6 back-end presented in Yakovlev *et al.* (2024), combined with various front-end models under various noise conditions and enrollment configurations.

Currently, most widely adopted front-end models, e.g., cosine similarity, PSDA, and Neural PLDA, are limited to only comparing two x-vectors. That is, they are designed only for the singular enrollment setting and cannot make full use of a set of x-vectors obtained through sub-divisions. In this work, we introduce a novel front-end model, termed an Adaptive Neural PLDA, that can be used to compare two sets of x-vectors of any size without having to be retrained. This allows for increased flexibility and performance across all multiple enrollment scenarios.

The front-end models were trained using a combined dataset with both English and Swedish data while keeping the weights of the back-end model fixed. The combined models were then evaluated on a purely Swedish dataset. In particular, the influence of the number of sub-divisions and enrollment lengths are investigated. The analysis demonstrates that splitting longer speech recordings into multiple shorter enrollments significantly improves performance in noisy environments, as illustrated in Figure 1, likely due to, firstly, the increased robustness as a result of the subsequent combination

of multiple x-vectors, and, secondly, to the better fit with the shorter training utterances used to build the ReDimNet back-end models. In particular, the observed increase to robustness is of notable interest in forensic applications, where recordings are often subject to varying levels of background noise. Our study underscores the critical role of data partitioning and back-end modeling in optimizing ASV systems for real-world applications, providing insights into improving robustness and generalizability across languages and acoustic conditions.



**Figure 1.** The effects of increasing the combined data usage on the Equivalent Error Rate (EER) with different amounts of sub-divisions, i.e., different levels of multiple enrollments, are shown for undisturbed speech (left), and speech with ‘ambient’ background noise (right).

## References

- Brümmer, N., Swart, A., Mosner, L., Silnova, A., Plhot, O., Stafylakis, T., & Burget, L. (2022). Probabilistic Spherical Discriminant Analysis: An Alternative to PLDA for length-normalized embeddings. *Interspeech 2022*. (pp.1446-1450). DOI: 10.21437/Interspeech.2022-731
- Silnova, A., Brümmer, N., Swart, A., & Burget, L. (2023). Toroidal probabilistic spherical discriminant analysis. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (pp. 1–5). DOI: 10.1109/ICASSP49357.2023.10095580.
- Yakovlev, I., Makarov, R., Balykin, A., Malov, P., Okhotnikov, A., & Torgashov, N. (2024). Reshape Dimensions Network for Speaker Recognition. *Interspeech 2024*. (pp. 3235–3239). DOI: 10.21437/Interspeech.2024-2116

## Engaging with Government and Police stakeholders regarding the use of speech technology in UK investigative interviewing

*Jessica Wormald, Lauren Harrington, James Tompkinson and Eloísa Monteoliva García*

*Department of Language and Linguistic Science, University of York, UK*  
 {firstname.lastname}@york.ac.uk

Safely facilitating the implementation of AI-technologies has become a key priority for the UK government (DSIT, 2025). Police forces at local and national levels are exploring ways to implement speech technologies across policing, focussing on the use of speech recognition and translation software (e.g. Muir & O’Connell 2025). However, there is a disparity between the expectations of law enforcement and current technological capabilities (Muir and O’Connell 2025). Examples of this include the belief that AI will facilitate the automatic transcription of emergency calls and simultaneously search police databases for names and addresses, and the idea that AI is capable of producing accurate real-time translation of multilingual spoken interactions between officers and members of the public.

One area which is likely to be affected by advances in AI technology is the police interview. Interviews are complex conversational contexts and can have significant investigative and evidential value. A great deal of linguistic research has been carried out in this area from a variety of perspectives, including interpreter-mediated interviews (Böser 2013; Nakane 2014; Monteoliva-García 2020), transcription of interviews (Haworth 2018; Richardson et al. 2023; Tompkinson et al. 2023), and the incorporation of automatic speech recognition into investigative interviewing (Harrington, 2023). Haworth (2018) highlighted an increasing reliance on transcripts of interviews rather than audio recordings, despite the two being fundamentally different modalities. Monteoliva-García (2020) showed that the ‘stand-by’ interpreting regime could shape the “distribution of interactional power” and how interpreters’ role shifts from facilitating communication to monitoring communication and preventing miscommunication. Harrington (2023) highlighted the potential for automatic systems to be used to reduce time constraints in the production of interview transcripts, but also warned of the dangers of using automatic systems in a way which is not transparent, systematic and linguistically-informed.

The existing linguistic research around investigative interviewing is directly pertinent to current priorities within government and policing and the implementation of speech technologies. There is a significant opportunity to support and facilitate the safe uptake of appropriate technologies by those of us with expertise in this area. To this end, we are hosting a one-day knowledge exchange workshop to bring together key stakeholders in policing to discuss and engage with current research on investigative interviewing, interpreted interviews, and human- and machine-led transcription. The workshop will provide an opportunity to present ongoing research and facilitate discussions about best practice and the potential integration of new technologies in these areas.

In this talk, we will present the outcomes from the workshop, focussing on how linguistic research can inform police practice. One envisioned outcome is the development of continuing professional development courses that can be hosted by FoSS - a new project at the University of York looking to bridge the gap between research and practice. These courses could be offered to police officers needing to upskill or increase awareness of the risks and benefits of these technologies. At this critical time in policing, expertise in forensic speech analysis can support the safe implementation of technologies ensuring appropriate and robust use by capable and trained individuals.

## References

- Böser, U. (2013). "So tell me what happened!": Interpreting the free recall segment of the investigative interview. *Translation and Interpreting Studies. The Journal of the American Translation and Interpreting Studies Association*, 8(1), 112-136.
- Department for Science, Innovation & Technology (2025). *AI Opportunities Action Plan*. Available online: <https://www.gov.uk/government/publications/ai-opportunities-action-plan/ai-opportunities-action-plan#contents> (accessed 27th March 2025)
- Harrington, L. (2023). Incorporating automatic speech recognition methods into the transcription of police-suspect interviews: factors affecting automatic performance. *Frontiers in Communication*, 8. <https://doi.org/10.3389/fcomm.2023.1165233>
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *Int. J. Evid. Proof* 22, 428–450.
- Haworth, K., Tompkinson, J., Richardson, E., Deamer, F., & Hamann, M. (2023). "For the Record": applying linguistics to improve evidential consistency in police investigative interview records. *Frontiers in Communication*, 8. <https://doi.org/10.3389/fcomm.2023.1178516>
- Monteoliva-García, E. (2020). The collaborative and selective nature of interpreting in police interviews with stand-by interpreting. *Interpreting*, 22(2), 262-287.
- Muir, Rick & O'Connell, Felicity (2025). *Policing and Artificial Intelligence*. The Police Foundation. Available online: <https://www.police-foundation.org.uk/publication/policing-and-artificial-intelligence/>
- Nakane, I. (2014). *Interpreter-mediated police interviews: A discourse-pragmatic approach*. Springer.
- Richardson, E., Hamann, M., Tompkinson, J., Haworth, K., & Deamer, F. (2023). Understanding the role of transcription in evidential consistency of police interview records in England and Wales. *Language in Society*, 54(1), 135–166. <https://doi.org/10.1017/S004740452300060X>
- Tompkinson, J., Haworth, K., Deamer, F., & Richardson, E. (2023). Perceptual instability in police interview records: Examining the effect of pauses and modality on people's perceptions of an interviewee. *International Journal of Speech, Language and the Law*, 30(1), 22–51.

## Assessing the ability of deepfake voice clones to produce accent and context-specific phonological features

*Ben Gibb-Reid<sup>1</sup>, Vincent Hughes<sup>1,2</sup>, and Jessica Wormald<sup>1,2</sup>*

<sup>1</sup>*Department of Language and Linguistic Science, University of York, York, UK*  
 {ben.gibb-reid|vincent.hughes|jessica.wormald}@york.ac.uk

<sup>2</sup>*Forensic Speech Services (FoSS), York, UK.*

Considerations of accent variation are lacking in audio deepfake detection methods (Almutairi & Elgibreen, 2022). Previous studies have assessed the ability of voice cloning to recreate specific phonetic features (e.g. Khanjani et al., 2023) and attempted to use phonetic knowledge to aid in human deepfake detection (Kirchhübel & Brown, 2022; Lee et al., 2023). However, there is still no established methodology for deepfake detection in the forensic context: i.e. a comparison between a verified authentic sample and a recording that is allegedly synthesized. This leads to the question: how can we use phonological knowledge and phonetic observations of real speech to assess the ability of deepfake audio techniques to accurately reproduce a speakers' voice? The present study investigates phonological and contextual variation to assess not only whether a deepfake system can successfully replicate speaker-specific phoneme realisations, but whether it can replicate the phoneme in new contexts.

As part of this IAFPA-funded study, we will record the following accent groups: southern standard British (SSBE), Yorkshire, Scottish, and Hong Kong English (HKE), specifically investigating the PRICE vowel and post-vocalic coda /ɪ/. We will be informed by participants' own pronunciations and predictions from previous literature (Beal, 2008, p.133; Foulkes, 1997, p.74; Setter et al., 2010, pp.23-27; Stuart-Smith, 2008). For SSB and Yorkshire English, PRICE is predicted to be [ɹɪ] and [ɑ:] respectively. In urban Scottish English, PRICE varies in quality depending on whether it occurs before a voiceless/voiced consonant. In HKE, PRICE is expected to be more monophthongal pre-consonantally. SSB and Yorkshire Englishes lack coda /ɪ/, unless an orthographic <r> is in a 'linking' context (e.g. *later* on). In HKE, sometimes /ɪ/-colouring occurs on vowels preceding /ɪ/, but this depends on individual speaker alignments between British/US Englishes. Scottish English is expected to always produce coda /ɪ/.

The *Elevenlabs* text-to-speech voice cloning system is selected to create deepfake audio speech samples as it is widely-used and accessible to non-specialists – making it a potential method for nefarious use. Recordings of participants will be carefully edited to only contain the given variables in certain contexts. For example, a Yorkshire sample may provide a pre-consonantal /ɪ/ context (e.g. *park*), but no linking /ɪ/ contexts (e.g. *car alarm*). This means the deepfake software would not be provided with knowledge of linking /ɪ/ contexts for this speaker. The system will be asked to produce a voice clone of each sample reading the North Wind and the Sun passage (which provides phonological variables across many contexts).

The study will test the ability of deepfake audio to recreate both provided and withheld contextual occurrences of the segments. To accurately reproduce variables that are not offered in the input, the deepfake system must rely on the knowledge of phonological variation provided by its own underlying model. The research will provide insight into how deepfake systems recreate the accents of speakers, and how auditory-acoustic forensic voice comparison methodologies can be applied to human deepfake detection.

## References

- Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5).
- Beal, J. (2008). English dialects in the North of England: Phonology. In B. Kortmann, C. Upton, & E. W. Schneider (Eds.), *A handbook of varieties of English* (Vol. 1, pp. 122–144). Mouton de Gruyter.
- Foulkes, P. (1997). English [r]-sandhi—A sociolinguistic perspective. *Histoire Épistémologie Langage*, 73–96.
- Khanjani, Z., Davis, L., Tuz, A., Nwosu, K., Mallinson, C., & Janeja, V. P. (2023). Learning to listen and listening to learn: Spoofed audio detection through linguistic data augmentation. *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, 01–06. <https://doi.org/10.1109/ISI58743.2023.10297267>
- Kirchhübel, C., & Brown, G. (2022). Spoofed speech from the perspective of a forensic phonetician. *Proceedings of Interspeech 2022*, 1308–1312. [https://www.isca-archive.org/interspeech\\_2022/kirchhubel22\\_interspeech.html](https://www.isca-archive.org/interspeech_2022/kirchhubel22_interspeech.html)
- Lee, D. D., McDougall, K., Kelly, F., & Alexander, A. (2023). PASS (Phonetic Assessment of Spoofed Speech): Towards a human-expert-based framework for spoofed speech detection. *IAFPA 2023–31st Conference of the International Association of Forensic Phonetics and Acoustics*, 31.
- Setter, J., Wong, C. S. P., & Chan, B. H. S. (2010). *Hong Kong English*. Edinburgh University Press.
- Stuart-Smith, J. (2008). Scottish English: Phonology. In B. Kortmann, C. Upton, & E. W. Schneider (Eds.), *A handbook of varieties of English* (Vol. 1, pp. 48–70). Mouton de Gruyter.

# Perception of deception through prosodic information in 911 emergency calls.

*Julien Plante-Hébert<sup>1</sup> and Lucie Ménard<sup>1</sup>*

<sup>1</sup>*Department of Linguistics, Université du Québec à Montréal, Montréal, Canada  
plante-hebert.julien@uqam.ca, menard.lucie@uqam.ca*

## Introduction

The detection of deception in speech has long been of interest to scientists and investigators. However, very little studies have been conducted on deception in contexts with actual stakes. In emergency settings such as 911 calls, dispatchers are occasionally suspicious regarding the manner the caller speaks, but what generates such suspicion and whether it is based on reliable on accurate markers remains unknown. The present study aims to determine the reliability of such perceptions of deception in speech and what guides them using actual 911 emergency calls.

## Methods

A total of 43 adult participants were recruited for the present experiment. All participants reported a normal hearing and were native Quebec French speakers.

The stimuli used were created using anonymized 911 emergency calls recordings obtained through a partnership with Sureté du Québec (SQ) during a previous phase of the study (Laforest, Rioux-Turcotte et St-Yves, 2020). SQ indicated whether each call was truthful or deceitful in regard of their investigation. A total of 58 calls were selected for the present experiment on the basis of the quality of the recording, the intelligibility of the speech, the perceived sex and native variety of French (Quebec French) of the caller.

All speech from the 911 dispatcher and long silences were removed as well as short answers such as “yes” or “no” and non-linguistic vocalizations. Thirty-second samples were kept for each call, with the onset of these samples at the moment the caller starts explaining the reason of their call (see (Laforest et al., 2020). Finally, the samples were low-pass filtered at a 350 Hz cutoff frequency in order to delexicalize them while preserving as much prosodic information as possible (Boucher, Gilbert et Rossier-Bisaillon, 2018).

Participants were asked to listen to each sample and to tell if the speaker was being truthful or deceitful. In a second phase, 21 of the 43 participants were also asked to rate each sample in regard of their speech rate, pitch, pitch variability and emotivity on continuous scales.

## Results

The results regarding the accuracy of the participants, with an average of 48% (SD = 6.24%) of correct answers, demonstrate the difficulty of the task. Generalized mixed models were used to analyze the participants’ ratings of speech rate, pitch, pitch variability and emotivity (fixed factors) as predictors of their judgements of deception and truthfulness and, in a different GMM, as predictors of the callers’ actual truthfulness/deception of the callers. Theses GMM showed that participants significantly relied on speech rate and emotivity to decipher deception, whereas pitch and pitch variability were the only two significant predictors in this regard.

## Discussion

Overall, our experiment showed that human participants were mostly unable to distinguish truthful emergency calls from deceitful ones by relying only on prosodic information. This could be explained by the speech parameters they use in comparison with the parameters established to significantly predict deceitfulness. In a forensic perspective, such observations should serve to invite caution when

assessing the truthfulness of a speech in an emergency setting as well as to guide further research on deception in speech in such conditions.

## References

- Boucher, V. J., Gilbert, A. C. et Rossier-Bisaillon, A. (2018). The structural effects of modality on the rise of symbolic language: A rebuttal of evolutionary accounts and a laboratory demonstration. *Frontiers in Psychology, 9*, 2300. doi: 10.3389/fpsyg.2018.02300
- Laforest, M., Rioux-Turcotte, J. et St-Yves, M. (2020). Détecter l'appelant dissimulateur au service d'urgence 9-1-1: Une analyse discursive et interactionnelle de la tromperie. *Criminologie, 53*, 193-218. doi: 10.7202/1074193ar

# The grandparent scam – a perceptual study

*Sara-Sophie Schedel and Gea de Jong-Lendle*

*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*

s.schedel@cranium.de | dejong@staff.uni-marburg.de

## Introduction

“Hello grandma, guess who this is”. The typical start of a telephone call using highly realistic synthetic voices based on just a short selection of real speech of “the grandchild”: criminals have found a new way to deceive victims. Elderly people are made to believe their grandchild is in trouble and in urgent need of money. The German Federal Network Agency registered over 11.000 complaints related to this scam in only the first half of 2024. The increasing accessibility of affordable speech-software increases the risk of voice-based deception, especially affecting vulnerable populations like elderly people or children (e.g. sexual online grooming). The question arises – how real is this threat? Is it possible for an average criminal with limited IT-skills and finances to commit such a crime?

This study investigates listeners’ ability to differentiate between natural and synthetic voices. The experimental design was loosely based on the structure of a typical grandparent scam scenario. Two listener groups were tested: one familiar with the speakers and a non-familiar group. A survey of commercially available text-to-speech systems was conducted, that are inexpensive, require no advanced programming skills and produce acceptable speaker imitations. ElevenLabs (2025: 20€/month) and Speechify (2024: 25€/month) were selected and their performance subsequently compared. Four acoustic conditions were compared: studio-quality, telephone transmission, background noise, and a combination of telephone and noise. The voices were synthesized based on 60 seconds of recorded speech from the same speakers whose natural voices were also presented. Thus, the natural and synthetic voices were directly comparable. Four acoustic conditions were compared: studio-quality, telephone transmission, background noise, and a combination of telephone and noise. These conditions were chosen to simulate realistic scam environments. However, recognition scores reported here are averaged across all four conditions.

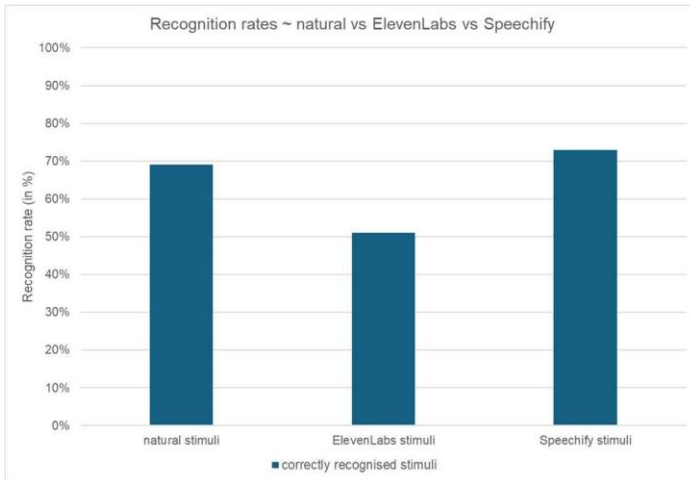
## Research on the perception of synthetic voices

Previous studies have shown that synthetic voices are difficult to detect when listeners are not expecting them (Prudký et al., 2023). Roswandowitz et al. (2024) found that deepfake voices activate different brain regions than natural voices, suggesting higher processing effort. Müller et al. (2022) demonstrated that both humans and algorithms struggle to detect synthetic speech in realistic scenarios, particularly with TTS-based stimuli. Mai et al. (2023) confirmed that even under ideal conditions, human recognition performance remains low.

## Methodology

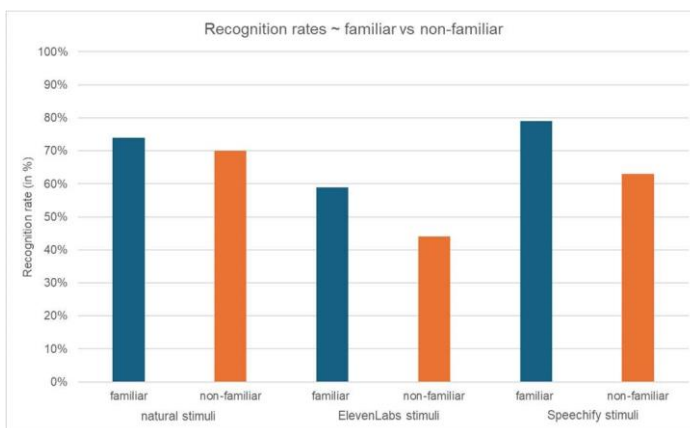
Fifty-eight native German speakers participated in the study. The familiar group (n = 26, Mean age = 44, range: 16-85) consisted of people who knew the speakers personally; the non-familiar group (n = 32, Mean age = 25, range: 17-44) had no prior exposure. Participants listened to 36 randomised voice samples (18 per speaker) presented online (PsychoPy/Pavlova). One third were natural recordings, one third synthetic recordings generated by ElevenLabs and one third by Speechify. Each voice was modified in four acoustic conditions. In a forced-choice task, listeners judged whether each voice was natural or synthetic.

## Results and conclusions



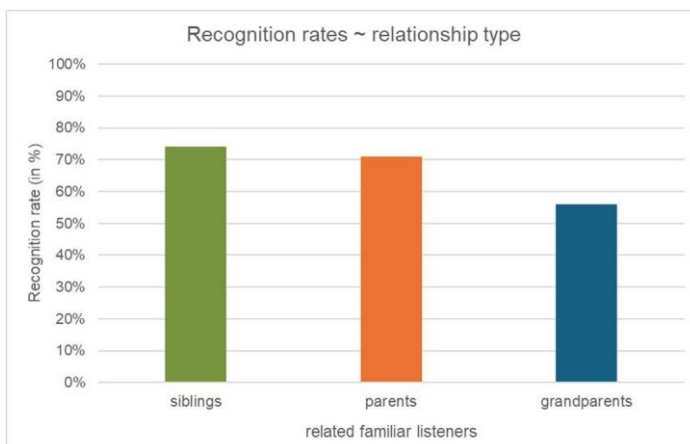
Participants were better at recognizing natural voices (69%) than synthetic voices overall (62%). The fake stimuli from Elevenlabs were detected less often (51%) than from Speechify (73%). This difference suggests that ElevenLabs currently produces more perceptually convincing output than Speechify. Preliminary statistical analysis suggests that difference is significant.

**Figure 1.** Recognition rates of all participants (N = 58) comparing natural, synthetic stimuli by Elevenlabs and by Speechify, derived from 60sec. of speech.



Familiar listeners outperformed the nonfamiliar group in all three conditions (natural versus Elevenlabs versus Speechify).

**Figure 2.** Recognition rates comparing familiar listeners (N = 15, Mean age = 27) with non-familiar listeners (N = 15, M age = 27) for natural synthetic stimuli by Elevenlabs and by Speechify, derived from 60sec. of speech.



Recognition rates comparing different types of relationships show the best performance for siblings (74%), followed by parents (71%). Grandparents achieved the lowest recognition rate at 56%. This result is especially relevant, as grandparents are the most likely targets of such scams.

**Figure 3.** Recognition rates among related listeners (siblings: n = 6, parents: n = 4, grandparents: n = 7)

Overall, the findings show that nowadays affordable speech synthesis systems exist, that can generate convincing voices requiring minimal input and IT-skills.

## References

- Müller, N. M., Pizzi, K., & Williams, J. (2022). Human perception of audio deepfakes. In Proceedings of the ACM Conference on Computer and Communications Security (pp. 85–91).
- Roswadowitz, C., Keitel, A., Lisker, M., & Obleser, J. (2024). The neural basis of voice deepfake detection. Manuscript submitted for publication.
- Prudký, L., Svoboda, M., & Krčál, O. (2023). The illusion of authenticity: Human detection of speech deepfakes in spontaneous communication. *Journal of Cyberpsychology*, 17(2), 112–125.
- Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLOS ONE*, 18(8), e0288477.
- Bundesnetzagentur (2024). Rufnummernmissbrauch: Schwerpunkt Einzeltrick. Retrieved from <https://www.bundesnetzagentur.de>

## Developing a collection of mock police interviews for use in research and teaching

*James Tompkinson<sup>1</sup>, Lotte Eijk<sup>2</sup>, Sarah Knight<sup>3</sup> and Eloisa Monteoliva-Garcia<sup>1</sup>*

<sup>1</sup>*Department of Language and Linguistic Science, University of York, UK*  
 {james.tompkinson|eloisa.monteolivagarcia}@york.ac.uk

<sup>2</sup>*Department of Psychology, University of York, UK.*  
 lotte.eijk@york.ac.uk

<sup>2</sup>*Department of Psychology, Newcastle University, UK.*  
 Sarah.Knight@newcastle.ac.uk

In the UK, police interviews are the most common method of gaining information from both suspects and witnesses. Police interview recordings are also the most common form of known sample used in forensic speaker comparison cases. Evidence gathered via this routine process is done so entirely through spoken language, with the dialogue between interviewers and interviewees recorded and preserved for use throughout the judicial process. It is therefore important to understand how judgements made by listeners about interviewees can shape decisions that might be made in the legal process. We know that features of the human voice can influence personality judgements (Dixon et al., 2004; Paver et al., 2025), and that such judgements can be problematic in high-stakes scenarios such as police interviews (Deamer et al., 2022; Tompkinson et al., 2023). However, most linguistic research in this area suffers from the problem that there is no established database of real UK police interviews. This means that researchers wishing to work on police interviews must either engage directly with UK police forces (see, for example, Haworth et al., 2023), work with the minimal amount of data available in the public domain, or use simulated police interview data from projects such as WYRED (Gold, 2020) and DyViS (Nolan et al., 2006). Furthermore, research in this area has not yet explored speakers who do not have ‘typical’ voices.

In this poster, we present a new, small database of mock police interview recordings collected as part of the YorVoice-funded “Perceptions of Atypical Voices” project, and explore the different ways in which the database can be used for future research. To create the database, we recruited participants to take part in a recording task. Participants were compensated for their time and participation. Recording sessions took place in a sound-attenuating booth in the Department of Psychology at the University of York between July 2024 and April 2025. There were three tasks within each session. First, participants read a standardized, phonetically balanced passage, followed by the sentence “hello my name is Sam”. Each participant then took part in a mock police interview task, using a slightly adapted format to the interviews in DyViS and WYRED. We recruited people from three different categories for the database, as follows:

1. **L1 Spanish speakers.** These speakers took part in two recording sessions, one in English and one in Spanish using an interpreter to translate the interviewer’s questions and the interviewee’s answers.
2. **Speakers with a stammer.** These participants self-identified as having a stammer or stutter.
3. **Control speakers.** These participants did not fall into either of the categories in (1) or (2).

The database contains 58 mock interview recordings, which will be made available to researchers through the UK Data Service ReShare platform. We anticipate that the resource will be a useful tool for people looking for open access police interview-style data for research purposes, particularly those looking for controlled data for voice perception experiments, transcription tasks, and research on cross-language intra-speaker variation.

## References

- Deamer, F., Richardson, E., Basu, N., & Haworth, K. (2022). For the Record: Exploring variability in interpretations of police investigative interviews. *Language and Law/Linguagem e Direito*, 9(1), 25-46.
- Dixon, J. A., & Mahoney, B. (2004). The effect of accent evaluation and evidence on a suspect's perceived guilt and criminality. *The Journal of social psychology*, 144(1), 63-73.
- Gold, E. (2020). *WYRED - West Yorkshire Regional English Database 2016-2019*. [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854354
- Haworth, K., Tompkinson, J., Richardson, E., Deamer, F., & Hamann, M. (2023). "For the Record": applying linguistics to improve evidential consistency in police investigative interview records. *Frontiers in Communication*, 8, 1178516.
- Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2006). 'Introducing DyViS: a dynamic study of British English for forensic purposes'. Paper presented at *the International Association for Forensic Phonetics and Acoustics Annual Conference*, Gothenburg, 23-26 July 2006.
- Paver, A., Wright, D., Braber, N., & Pautz, N. (2025). Stereotyped accent judgements in forensic contexts: listener perceptions of social traits and types of behaviour. *Frontiers in Communication*, 9, 1462013.
- Tompkinson, J., Haworth, K., Deamer, F., & Richardson, E. (2023). Perceptual instability in police interview records. *The International Journal of Speech, Language and the Law*, 30(1), 22-51.

# Spectral characteristics of sibilant fricative /s/ in voice disguise via age modification

*Payam Ghaffarvand-Mokari*

*Department of Linguistics, University of Eastern Finland, Joensuu, Finland*

*payam.ghaffarvand.mokari@uef.fi*

Despite significant progress in automatic speaker verification (ASV) systems, it remains a challenge to recognize speakers who intentionally alter their voice to conceal their identity. Intentional voice change—referred to as voice disguise—is sometimes employed in situations such as blackmail calls or robberies, where the offender intends to remain anonymous. While state-of-the-art deep learning models and architectures have significantly advanced speaker identification and verification tasks, their inherent complexity often renders the decision-making processes opaque, functioning much like a black box. Hence, a fine-grained analysis of speech sounds can help identify which acoustic properties are most robust or vulnerable to this type of intentional voice change. Fricatives are particularly important sounds, as they contribute to the overall quality and identity of a person's speech. Following previous studies on the acoustics of voice disguise (González Hautamäki et al., 2017; 2019), this study examines the spectral and temporal properties of the fricative /s/ in speech produced using the speaker's modal voice, compared to speech produced when the speaker imitates the voice of a child or an elderly person.

A total of 15,992 /s/ tokens were analyzed after labeling and boundary adjustments using data from the Age-related Voice Disguise corpus (AVOID; González Hautamäki et al., 2018). Eleven Finnish sentences produced by 58 speakers in 'modal', 'child', and 'old' conditions from the corpus were used. For this study, spectral center of gravity (COG), root mean square (RMS) amplitude, and duration of the tokens were measured. The COG was estimated in a frequency range between 1 and 11 kHz and was calculated using the multitaper method (Percival and Walden, 1993; Blacklock, 2004) with seven data tapers. Normalized durations of the /s/ sounds were calculated to account for the potential influence of speech rate on sibilant duration. Linear mixed-effects models were fitted to assess the effects of voice condition on spectral and temporal features of /s/ sounds.

The results showed intensity, center of gravity, and duration of the fricative /s/ varied significantly across speaking conditions ('modal', 'child', 'old'), with notable gender differences. Female speakers consistently reduced vocal intensity when imitating child and elderly voices, while male speakers showed a significant decrease only in the 'child' condition and more variability in the 'old' condition. At the individual level, 90% of females and 69% of males reduced intensity in the 'child' condition. Spectral center of gravity (COG) values were higher for females than for males across all conditions. In the 'child' condition, COG increased relative to 'modal' for both genders, while in the 'old' condition, COG decreased, more noticeably in females. Regarding duration, /s/ sounds were significantly longer in both 'child' and 'old' conditions, with 95.8% of speakers increasing duration in the 'old' condition. Gender did not significantly affect duration changes, suggesting a general tendency across speakers to lengthen /s/ sounds in voice disguise, particularly when imitating an elderly voice.

These findings highlight that specific acoustic features of fricatives—such as intensity, COG, and duration—are relatively systematically affected by voice disguise, with female speakers showing more consistent modifications.

## Acknowledgement

This work was supported by a grant from the Kone Foundation awarded to the author.

## References

- Blacklock, O. S. (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods* (Doctoral dissertation). University of Southampton, Southampton, UK.
- Percival, D. B., and Walden, A. T. (1993). *Spectral Analysis for Physical Applications: Multitaper and Conventional Univariate Techniques* (Cambridge University Press, Cambridge), pp. 1–583.
- González Hautamäki, R., Sahidullah, M., Hautamäki, V., & Kinnunen, T. (2017). Acoustical and perceptual study of voice disguise by age modification in speaker verification. *Speech Communication, 95*, 1–15.
- González Hautamäki, R., Sahidullah, M., Hautamäki, V., Bentz, M., Werner, S., and Kinnunen, T. (2018). Corpus of age-related voice disguise (AVOID), [Dataset]. <http://urn.fi/urn:nbn:fi:lb-2018060621>
- González Hautamäki, R., Hautamäki, V., & Kinnunen, T. (2019). On the limits of automatic speaker verification: Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. *The Journal of the Acoustical Society of America, 146*(1), 693–704.

# Validating the auditory-acoustic phonetic method for forensic speaker comparison

Vincent Hughes<sup>1</sup>, Lauren Harrington<sup>1</sup>, Philip Harrison<sup>1</sup>, Finnian Kelly<sup>2</sup>, David van der Vloed<sup>3</sup>, and Richard Rhodes<sup>4</sup>

<sup>1</sup>*Department of Language and Linguistic Science, University of York, UK.*  
 {vincent.hughes|lauren.harrington|philip.harrison}@york.ac.uk

<sup>2</sup>*Oxford Wave Research, UK.*  
 finnian@oxfordwaveresearch.com

<sup>3</sup>*Netherlands Forensic Institute, Netherlands.*  
 d.van.der.vloed@nfi.nl

<sup>4</sup>*The Forensic Voice Centre, UK.*  
 richard.rhodes@forensicvoicecentre.com

Forensic sciences must demonstrate the validity of methods used in evidential comparisons; this is reflected in the requirements of regulation and international standards (FSR 2025). Yet, there has never been a large-scale validation exercise of the auditory-acoustic phonetic method for speaker comparison using realistic casework procedures and materials (for debate, see Kirchhübel et al. 2023, Morrison 2023).

In this paper, we provide an update on a validation exercise for auditory-acoustic phonetic speaker comparison being run at the University of York. The exercise involves testing the method as an end-to-end process, including a peer-review stage. It contains a series of 40 comparisons from a forensically-realistic database from the UK Government and is being undertaken by two experienced forensic analysts. This paper will also consider the wider landscape of validation and regulation, the specific challenges for speaker comparison, and how we addressed those challenges in our validation exercise:

- (1) **Separation of analyst and method:** It can be argued that since speaker comparison analysis and interpretation are human-based, the analyst and the method are inseparable. However, this undermines speaker comparison as a forensic *science*, implying a) no consistent methodology across analysts and cases, b) that competence is innate, or learned through professional or everyday experience, and that this intuition is given a key role in the analysis, and c) that divergence in conclusions across analysts is to be expected. We argue that there is considerable methodological consistency across cases and in our exercise, we provide a prescriptive methodological framework based on the 2021 ENFSI Best Practice Manual. Peer review also helps mitigate the risks of human-based interpretation. We therefore consider peer review a non-negotiable part of the auditory-acoustic phonetic method, and hence include it in the validation exercise.
- (2) **Replicating real casework methods:** Speaker comparison is time- and labour-intensive, with 1-to-1 comparisons taking 10 to 15 hours or more. The time required to undertake large numbers of comparisons for validation is therefore a considerable barrier. We are in a fortunate position that we have resources to run the exercise through a funded project at the University of York. In addition, we reduced the time our analysts spend on comparisons by removing the preparatory phase of the process. We also provided our analysts with ancillary

materials to streamline the analysis, such as orthographic transcripts and premade reporting forms.

- (3) **Choosing appropriate files:** A criterion in Kirchhübel et al.'s (2023) test was that it should not be 'too easy' or 'too hard'. The wider issue is that for a validation exercise to be meaningful, the comparisons must be representative of casework conditions. We focused on conditions that are typical of casework, rather than extremes (e.g. very short or noisy samples), i.e. police interview-like recordings and non-contemporaneous mobile phone recordings with reasonable duration (> 45s). We did not make decisions about comparisons we deemed 'too easy' or 'too hard', as this requires subjective decisions which are unrelated to the accurate application of the method, but did limit the scope to within-accent comparisons.

The exercise is currently on-going, and once complete, the results will be shared with the community as a basis for accreditation.

## References

- ENFSI (2021) *Best Practice Manual for the Methodology of Forensic Speaker Comparison*. ENFSI-BPM-FSC-01. Version 01.
- Forensic Science Regulator (2025) Draft Code of Practice 2025 (Version 2).  
<https://www.gov.uk/government/publications/forensic-science-regulator-code-of-practice>
- Kirchhübel, C., Brown, G. and Foulkes, P. (2023) What does method validation look like for forensic voice comparison by a human expert? *Science and Justice*, 63, 251—257.
- Morrison, G. S. (2023) A single test pair does not a method validation make: A response to Kirchhübel et al. (2023). *Science and Justice*, 63, 327—329.

# Towards a Transparent and Interpretable Strategy for Spoofed Speech Detection

Carolina Lins Machado<sup>1</sup>, Xin Wang<sup>2</sup>, and Junichi Yamagishi<sup>2</sup>

<sup>1</sup>*Netherlands Forensic Institute, The Hague, The Netherlands.*

c.machado@nfi.nl

<sup>2</sup>*National Institute of Informatics, Tokyo, Japan.*

{wangxin|jyamagis}@nii.ac.jp

The quick spread of artificially-generated (henceforth spoofed) speech applied for malicious purposes poses unprecedented challenges for forensic investigators and legal systems (Gambin et al., 2024; Verdoliva, 2020). The "black box" character of many systems used in the detection of spoof speech poses a problem in forensic contexts where the interpretability of the conclusions drawn by such detection methods are crucial. (Mitchell, 2010; Hall et al., 2022). Therefore, in order to ensure a fair justice outcome, emerging regulations governing the use of complex systems in the detection of artificially-generated media require that the decisions made by these system be understandable and justified to all parties involved in the process (Hall et al., 2022; Siegel et al., 2024). This suggests that interpretability and transparency are necessary for a method to be forensically valid. This work aims to address this need by examining how acoustic-phonetic features and explainable machine learning approaches may provide clarity on the process of spoofed speech detection. Moreover, this exploratory study attempts to (i) understand how acoustic-phonetic features perform in various spoofing types and (ii) provide a baseline against which future state-of-the-art attacks can be compared by displaying how these features and their discriminative performance change in various voice spoofing attack types.

## Method

From the datasets ASVspoof 2015 (Wu et al., 2015), 2019 (Wang et al., 2020), 2021 (Yamagishi et al., 2021), and 5 (Wang et al., 2024), two binary classification experiments were conducted to classify speech samples as human or spoofed. For each sample, white-box<sup>1</sup> acoustic-phonetic features were extracted using Praat (Boersma & Weenink, 2024). These features were grouped as global, when measures were taken over a whole sentence, and local, when measures were taken at/around places of intensity maxima in a syllable (see Table 1). In the first experiment, classifications were performed using decision trees (full and pruned), since they enable the visualization of the features being considered and their influence on the decision outcome. In the second experiment, an AutoML (Automated Machine Learning) pipeline using Lazy Predict<sup>2</sup> was employed to assess the performance of these features in more complex, albeit less transparent, algorithms. Finally, we used the database Deepfake-Eval-2024 (Chandra et al., 2025), to assess how generalizable these features are with "in-the-wild" spoofed samples.

## Results

Preliminary results on the ASVspoof 2015 database revealed that the overall classification performance was much worse compared to state-of-the-art opaque features and methods. Our best model had an EER of 31.7% on seen attacks (samples present in the training phase), and of 45.8% on unseen attacks. Moreover, the results of the AutoML pipeline revealed an interplay between algorithms and features in terms of generalizability (i.e., seen and unseen attacks). On seen attacks a LightGBM was the best performing method (Balanced Acc = 74%; F1 Score = 80%), and on unseen attacks a Nearest Centroid Classifier (Balanced Acc = 60%; F1 Score = 60%). Thus far, our results suggest that an explainable approach may not be competitive with state-of-the-art "black box" methods in terms of accuracy. Nonetheless, it provides a significant advantage over opaque approaches when explainability and transparency are necessary, since interpretable features are able

<sup>1</sup> The term "white-box" in this paper refers to features whose explanations stem from well-known physiological aspects of speech production.

<sup>2</sup> <https://github.com/shankarpandala/lazypredict>

to illustrate the differences between spoofed and natural speech, and transparent models can reveal how classification decisions were made.

Feature	Local Measurement	Feature	Global Measurement
Formants	F1; F2; F3	Harmonic-to-noise ratio	Mean
Spectral tilt	H1-H2; H1-A1; H1-H2; H1-A3 A1-A2; A1-A3; A2-A3	Peaks-per-second	Standard Deviation
Jitter	Local Absolute Relative average perturbation Difference of difference of periods Five-point period perturbation quotient	Intensity slopes Signal periodicity	Mean Standard Deviation 2kHz-4 kHz 4 kHz-6 kHz 6 kHz-8 kHz
Shimmer	Local Three-point amplitude perturbation quotient Five-point amplitude perturbation quotient Average absolute difference	F0 wiggleness F0 spaciousness F0 slopes Spectral flatness Spectral centroid	Mean Standard Deviation

**Table 1.** Overview of local and global features implemented. Local features were extracted at vocalic centers while global features encompass a whole sentence. Features without measurement signal that the feature name and measurement are the same.

## References

- Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.08) [Computer software]. <http://www.praat.org/>
- Chandra, N. A., Murtfeldt, R., Qiu, L., Karmakar, A., Lee, H., Tanumihardja, E., Farhat, K., Caffee, B., Paik, S., Lee, C., Choi, J., Kim, A., & Etzioni, O. (2025). *Deepfake-Eval-2024: A Multi-Modal In-the-Wild Benchmark of Deepfakes Circulated in 2024*.
- Gambín, Á. F., Yazidi, A., Vasilakos, A. V., Haugerud, H., & Djenouri, Y. (2024). Deepfakes: Current and future trends. *Artificial Intelligence Review*, 57(3).
- Hall, S. W., Sakzad, A., & Choo, K.-K. R. (2022). Explainable artificial intelligence for digital forensics. *WIREs Forensic Science*, 4(2), e1434.
- Mitchell, F. (2014). The use of Artificial Intelligence in digital forensics: An introduction. *Digital Evidence and Electronic Signature Law Review*, 7(0). <https://doi.org/10.14296/deeslr.v7i0.1922>
- Siegel, D., Kraetzer, C., Seidlitz, S., & Dittmann, J. (2024). Media Forensic Considerations of the Usage of Artificial Intelligence Using the Example of DeepFake Detection. *Journal of Imaging*, 10(2).
- Verdoliva, L. (2020). Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5), 910–932.
- Wang, X., Delgado, H., Tak, H., Jung, J., Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T. H., Evans, N., Lee, K. A., & Yamagishi, J. (2024). ASVspooF 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspooF 2024)*, 1–8.
- Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., Juvola, L., Alku, P., Peng, Y.-H., Hwang, H.-T., Tsao, Y., Wang, H.-M., Maguer,

- S. L., Becker, M., Henderson, F., ... Ling, Z.-H. (2020). ASVspooF 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64, 101114.
- Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilçi, C., Sahidullah, M., & Sizov, A. (2015). ASVspooF 2015: The first automatic speaker verification spoofing and countermeasures challenge. *Interspeech 2015*, 2037–2041.
- Yamagishi, J., Wang, X., Todisco, M., Sahidullah, M., Patino, J., Nautsch, A., Liu, X., Lee, K. A., Kinnunen, T., Evans, N., & Delgado, H. (2021). ASVspooF 2021: Accelerating progress in spoofed and deepfake speech detection. *2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge*, 47–54.

## Introducing the International Network for Forensic Transcription: initial plans and overall aims

*Lauren Harrington*  
*University of York, UK*  
 lauren.harrington@york.ac.uk

Forensic transcription is an under-researched area of forensic speech science, despite being relatively frequently conducted by practicing forensic speech experts. A survey on forensic transcription practices carried out by Harrington and Rhodes (2025) demonstrated that there is considerable variability in the methods employed by forensic practitioners, and that there are many issues with and questions surrounding current approaches. The survey and its follow-up IAFFPA workshop initiated discussions amongst practitioners and brought transcription practices to the attention of the forensic speech science community.

The *International Network for Forensic Transcription* is a new initiative that will be officially launched in 2026. The network will promote the continuation of discussions centred around forensic transcription methods amongst practitioners and will provide a forum for knowledge exchange between those practicing and researching forensic transcription. The main aims of the network are:

1. To **bridge the gap between academia and practice**, fostering ongoing discussion and collaboration between researchers and practitioners, and ensuring that academic research is guided by practitioners' experiences and casework requirements.
2. To encourage **more research on forensic transcription** and provide a forum for knowledge exchange where academics can receive feedback on their research directly from practitioners.
3. To work towards **developing guidelines or a framework for practice** using a standardised, validated methodology which will address regulatory requirements of international standards such as ISO/IEC 17025.

Forensic practitioners and academics interested in or actively conducting research on forensic transcription will be invited to join the *International Network for Forensic Transcription* to create a community dedicated to sharing current knowledge and furthering our understanding of forensic transcription and best practices. The network will aid the dissemination of new research and organise knowledge exchange events amongst its members, while also focusing on ensuring and enhancing quality of transcripts and methods. Several practical initiatives are under consideration, including:

- Organising events centred around sharing best practices amongst forensic transcribers
- Generating and/or sharing materials for proficiency and validation testing
- Facilitating interlaboratory comparisons and sharing validation/verification data
- Producing guidance documents and standard operating procedures for forensic transcribers

At this stage, these initiatives remain in the planning phase and are open to development. Feedback, particularly from practitioners, on which activities would be most valuable or impactful is warmly welcomed.

### References

Harrington, L. & Rhodes, R. (2025). Survey on forensic transcription practices. *International Journal of Speech, Language and the Law*, 31(2), pp. 236-266.

# Voice quality across a speaker's languages: investigating the case of L1 Persian – L2 English

Janan Shalpush<sup>1</sup>, Willemijn Heeren<sup>1</sup>, and Niels O. Schiller<sup>1,2</sup>

<sup>1</sup>*Department of Linguistics, Leiden University, Leiden, The Netherlands*

<sup>2</sup>*Department of Linguistics and Translation, City University of Hong Kong, Hong Kong SAR*

j.shalpush@hum.leidenuniv.nl |

w.f.l.heeren@hum.leidenuniv.nl | Niels.Schiller@cityu.edu.hk

This study investigates whether a shift from a listener's first language (L1) to their second language (L2), or vice versa, modulates their ability to accurately identify speakers' voices. We followed the method of Orena et al. (2019) on L1 English - L2 French while studying the ability of monolingual Persian speakers and unbalanced Persian-English bilingual speakers to distinguish between distinct voices after a language shift. Previous research has demonstrated that voice recognition relies on both language-dependent and language-independent cues. For instance, Winters et al. (2008) showed that listeners can successfully identify speakers across languages, underscoring the role of stable, cross-linguistic voice characteristics. Conversely, Orena et al. (2019) highlighted that bilinguals' identification accuracy is influenced by language dominance and switch direction, indicating the involvement of language-specific information.

## Method

Four unbalanced, female Persian-English bilingual speakers (L1 Persian; L2 English) recorded 20 matched sentences in each language. Sentences were controlled for syllable counts and processed in Praat (Boersma & Weenink, 2023). Adobe Podcasts was used to reduce acoustic background differences while preserving phonetic integrity.

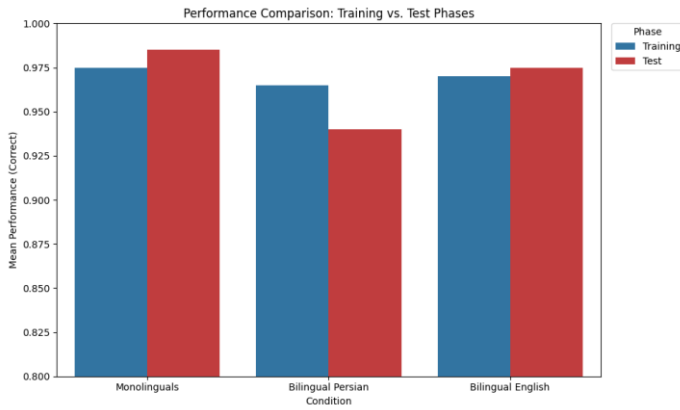
Participants included 26 Persian monolinguals and 52 unbalanced Persian-English bilinguals, randomly assigned to one of two switch conditions: Persian-trained group (trained in L1, tested in L2) and English-trained group (trained in L2, tested in L1). The experiment, implemented in PsychoPy (Peirce *et al.*, 2019) and adapted from Orena et al. (2019), comprised four phases: familiarization, training, accuracy check after training, and testing after language switch. Participants heard sentences from four speakers and identified the speaker via keyboard. They completed 80 training trials (20 per speaker) with feedback, requiring at least 85% accuracy to proceed to the final test. In the test phase, no feedback was given.

## Results

Preliminary results showed high overall accuracy: monolingual baseline (98.5%, L1 trained – L2 tested), English-trained bilinguals (97.1%, L1 tested), and Persian-trained bilinguals (94.6%, L2 tested) (see Figure 1). To assess performance, we used a linear mixed-effects model (*lmer()*, *lme4* package; Bates *et al.*, 2015) in R (R Core Team, 2024), with Group, Phase, and their interaction as predictors (monolinguals excluded due to absence of a language switch). No main effects of Group or Phase were found, but the Group  $\times$  Phase interaction was significant ( $\beta = -0.0266$ ,  $p = .0011$ ), indicating a reduction in accuracy from training to test in the Persian-trained group. This suggests that L1-to-L2 switching may disrupt voice recognition more than L2-to-L1.

To further examine the high overall accuracy, we acoustically analyzed the stimuli for speaker-specific cues, focusing on mean pitch (Hz), duration (s), jitter, and shimmer (see Table 1 for participant-level statistics). Language had no significant effect on pitch,  $F(1, 3.00) = 1.058$ ,  $p = .379$ ; duration,  $F(1, 31.39) = 2.018$ ,  $p = .165$ ; or jitter and shimmer, indicating stable voice quality across conditions. Although some speaker-level differences were observed (e.g., P4 showed a wider pitch range, P1 a lower overall pitch), these were not statistically supported.

In sum, L1 to L2 switching reduced recognition accuracy, while speaker-specific acoustic variation requires further investigation.



**Figure 1.** Participants' performance during the training and test phases across different experimental conditions.

Speaker	Language	Mean Pitch (Hz)	SD Pitch (Hz)	Mean Duration (s)	SD Duration (s)	Mean Jitter	SD Jitter	Mean Shimmer	SD Shimmer
P1	English	165.6	11.5	3.135	0.416	0.021	0.0046	0.0665	0.0139
P1	Persian	171	10.5	3.028	0.440	0.0216	0.0049	0.0654	0.0155
P2	English	235.7	10.1	2.856	0.396	0.0241	0.0049	0.0657	0.0111
P2	Persian	227.8	7.9	2.818	0.441	0.0213	0.0033	0.0657	0.0083
P3	English	205.5	6	3.287	0.660	0.0155	0.0035	0.0506	0.0083
P3	Persian	205.8	7.7	3.227	0.590	0.0149	0.003	0.0539	0.0084
P4	English	240.3	14.3	3.269	0.674	0.0175	0.0028	0.0539	0.0085
P4	Persian	212.3	16.5	3.002	0.438	0.0199	0.0027	0.0588	0.009

**Table 1.** Acoustic measures by participant and language.

## References

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Boersma, P., & Weenink, D. (2023). *Praat: Doing phonetics by computer* (Version 6.3.00) [Computer software]. University of Amsterdam.
- Orena, A. J., Theodore, R. M., & Polka, L. (2019). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, 185, 1–9.
- Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1), 195–203.
- R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.x) [Computer software]. R Foundation for Statistical Computing.
- Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, 123(6), 4524–4538.

# Distinctiveness vs. Deception: Assessing the success of deepfake technologies for perceptually distinctive voices

*Leah Bradshaw*<sup>1,2</sup>

<sup>1</sup>*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*

*leah.bradshaw@uzh.ch*

<sup>2</sup>*JP French International, York, UK*

Audio deepfakes (or *voice clones*) are artificially generated speech signals created using machine learning techniques, such as neural text-to-speech (TTS) and voice conversion models, to mimic the voice and speaking style of a real individual. Recent years have seen substantial improvements to both the quality and accessibility of deepfake technologies, making it increasingly feasible to create speech samples that sound convincingly like specific individuals. As such, they have been the subject of substantial discussion in the forensic community with researchers and practitioners alike sharing large concerns surrounding their potential misuse in criminal contexts. Indeed, forensics practitioners must now not only consider the authenticity of speech evidence but also the possibility of synthetic interference. As these technologies continue to evolve, understanding their capabilities, limitations and detectability becomes crucial for maintaining the integrity of forensic speaker analysis and safeguarding against audio-based deception.

Our ability to detect deepfake voices, both using automatic technologies or some type of phonetic analysis, has dominated discussions amongst the forensic community in recent years. For example, studies have assessed the performance of both expert and lay-listers for detecting deepfake voices<sup>1,2</sup>. In addition, the role of the specific voice, with reference to the accent, in detecting deepfakes has been considered<sup>3</sup>. Despite this, studies are yet to address the role that perceptually distinctive voices may play in both detecting and developing audio deepfakes. Perceptually distinctive voices can be characterised by numerous acoustic features which create a complex voice profile that leads to some voices simply standing out more than others. For example, fundamental frequency (f0), timbre and prosody have all been shown to influence the distinctiveness of a voice (e.g.,<sup>4,5</sup>). In addition, distinctive voices are frequently highlighted in forensic speech science, showing both positive and negative attributes. For one, voice distinctiveness plays a large role in voice memory, with listeners showing a greater ability to recall highly distinctive voices<sup>6,7</sup>. Comparatively, highly distinctive voices pose challenges for selecting suitable foils when preparing a voice parade.

Given that such voices often contain atypical acoustic and prosodic features, such as unique pitch patterns, timbre, or articulation, it is possible that these characteristics may also pose specific challenges for generative models, as a result of underrepresentation in training data. Moreover, because distinctive voices are more salient and memorable, even subtle imperfections in a synthetic clone may be more easily detected by listeners. This prompts further investigation into whether the multi-dimensional nature of distinctive voice profiles makes them inherently more resistant to convincing deepfake replication.

The overarching aim of this research project is to develop a detailed understanding of the relationship between voice distinctiveness and the quality of an audio deepfake. To better understand this, I will analyse the detectability of voice clones of perceptually distinctive voices generated by a selection of the most popular voice cloning systems from various companies. During this project, I will conduct a series of experiments designed to explore both human and machine performance for detecting deepfake voices to assess if highly distinctive voices are indeed more easily detected. This presentation will outline the overall goals and plans for the project and demonstrate the overall contribution of the research for the field of forensic phonetics.

## References

1. Kirchhübel, C. & Brown, G. Spoofed speech from the perspective of a forensic phonetician. in *Interspeech 2022* 1308–1312 (ISCA, 2022). doi:10.21437/Interspeech.2022-661.
2. Terblanche, C., Harrison, P. & Gully, A. J. Human Spoofing Detection Performance on Degraded Speech. in *Interspeech 2021* 1738–1742 (ISCA, 2021). doi:10.21437/Interspeech.2021-1225.
3. Lee, D. D., McDougall, K., Kelly, F. & Alexander, A. Exploration of contemporary commercially available voice clone generators and detectors. in *IAFPA 2024* (2024).
4. Nolan, F., McDougall, K. & Hudson, T. Some Acoustic Correlates of Perceived (dis) Similarity Between Same-Accent Voices. in *ICPhS 2011* 1506–1509(2011).
5. McDougall, K., Paver, A. & Nolan, F. Voice Distinctiveness: An Investigation of the Role of Speakers' Position in a Population with Respect to F0. in *ICPhS 2023* 3790–3794 (2023).
6. Orchard, T. L. & Yarmey, A. D. The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology* **9**, 249–260 (1995).
7. Yarmey, A. D. Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society* **31**, 421–428 (1991).

## **Analysing the accuracy of the public's perception on characteristics of synthetic voices and how the concept of "expertise" affects the accuracy of AI identification.**

*Emily Verry, Ben Gibb-Reid and Amelia Gully*

*Department of Language and Linguistic Science, University of York, York, UK*

{ev652|ben.gibb-reid|amelia.gully}@york.ac.uk

Given that the focus of research on synthetic voices over the past few years has been centred around detection of spoofing in Automatic Speaker Verification systems, there is surprisingly little knowledge about how the public perceive these voices to sound, and about different types of listeners' ability to detect synthetic speech. Current papers on this topic cover auditory-acoustic phonetic approaches to spoofed speech (Brown and Kirchhubel, 2022; Lee et al., 2022), whilst others find that the public are generally unsuccessful in identifying synthetic speech (Gully et al., 2021). An experiment exploring different groups of listeners' accuracy of perception and identification with different levels of expertise could highlight the type of knowledge required for casework, in a world where it is becoming increasingly plausible that spoofed speech will begin to appear in forensic casework scenarios.

Given the lack of current research on synthetic voice, and that the existing papers have focused on the auditory-acoustic aspects of synthetic speech, this project proposes two hypotheses:

*1a. Public perceptions of characteristics of synthetic speech are inaccurate.*

*1b. Linguists will be more successful in AI identification than computer-scientists, who will be more successful than lay listeners.*

Part a) of this study will involve qualitative research about perceptions of synthetically produced voices. This will be done in the style of a research form which will be posted on forums and emailed out to students and staff of the particular experimental groups, (linguistic/computer science expertise). Whilst responses are being collected for this part of the study, an auditory and acoustic analysis will be conducted on a range of synthetic voices to see how their features align with the public's perceptions of them. Existing literature will also be used. This will be the foundation of determining how accurate the public's perception of synthetic voice characteristics are compared to an expert analysis. Due to the general public's inexperience in describing voices, the questions have been designed as open-ended; responses may be used as direct quotes as evidence to provide a more general overview of the public's perception of synthetic voices.

Part b) will present synthetic voices to the participants fitting into the groups mentioned above. Subjects will have the task of describing the voice and identifying whether it is synthetic or not. Using percentage accuracy rates, this should help identify which level of expertise is best for identifying synthetic speech. The two parts of this experiment will harmonise to give a nice picture of how listeners' confidence in identifying synthetic voices compares to their actual ability to identify one in a real-life scenario, and whether previous knowledge and experience with synthetic speech improves detection accuracy. The research will provide insight into how the public's perception of synthetic speech can be improved, and whether, in terms of language based crimes involving synthetic speech, methodologies need to be adapted to conform to what will lead to the most successful rate of synthetic speech identification.

## References

- Brown, G., & Kirchhubel, C. (2022). Spoofed speech from the perspective of a forensic phonetician. In *Proceedings of INTERSPEECH*, Incheon, Korea, 2022, pp. 1308–1312.  
<https://doi.org/10.21437/Interspeech.2022-661>
- Gully, A., Harrison, P., & Terblanche, C. (2021). Human spoofing detection performance on degraded speech. In *Proceedings of INTERSPEECH*, Brno, Czechia, 2021, pp. 1738–1742.  
<https://doi.org/10.21437/Interspeech.2021-1225>
- Lee, D. D., McDougall, K., Kelly, F., & Alexander, A. (2023). PASS (Phonetic Assessment of Spoofed Speech): Towards a human-expert-based framework for spoofed speech detection. *IAFPA 2023–31st Conference of the International Association of Forensic Phonetics and Acoustics*, 31.
- Nadja Schinkel-Bielefeld, Netaya Lotze, Frederik Nagel; Audio quality evaluation by experienced and inexperienced listeners. *Proc. Mtgs. Acoust.* 2 June 2013; 19 (1): 060016.  
<https://doi.org/10.1121/1.4799190>

# Sound Judgments: ASR Accuracy and Trust in ROTI Transcripts

*Lorimer Pepper, Paul Foulkes, and Lauren Harrington*  
*Department of Language and Linguistic Science, University of York, UK*

{lp1377|paul.foulkes|lauren.harrington}@york.ac.uk

Work in Progress poster

Eligible for Best Student Paper Award: Yes

## Background

The UK government and policing bodies are increasingly exploring speech-to-text technology for policing practices. Transcripts of police-suspect interviews (Records of Taped Interviews; ROTIs) are often produced as summaries rather than verbatim records; although full transcripts would better serve high-stakes contexts, time constraints limit human transcription. Automatic Speech Recognition (ASR) offers a potential solution by generating near-verbatim text efficiently, and recent developments show growing interest among police forces. However, because transcribers must still review AI output, it is important to understand their attitudes toward ASR - overreliance or improper use could allow errors to slip through unchecked. Although the College of Policing (2024) has outlined AI integration opportunities and risks, Haworth (2018) highlighted that existing human practices suffer from consistency and accuracy issues that can distort evidence. Because jurors and other lay decision-makers frequently depend on written transcripts without hearing the original audio, even minor errors or omissions could bias perceptions of credibility and guilt (Fraser & Stevenson, 2014).

## Methods

This paper describes a Master's project that aims to update our understanding of AI adoption in UK policing and probe how legal professionals and laypeople judge the reliability and trustworthiness of ASR- versus human-generated police transcripts. This will be investigated over two studies: a Freedom of Information (FOI) request and an online experiment.

In February 2022, an FOI request sent by researchers at Aston University to all 43 territorial police forces in England and Wales revealed no force was using ASR for ROTI transcription, though three reported plans to adopt it (Tompkinson et al., 2022). UK forces operate independently, so practices vary widely. To assess how ROTI procedures and the adoption of ASR has progressed since then, a follow-up FOI will be issued to the same 43 forces as part of the research project. The request will include all 10 questions of the 2022 FOI, but the main interest centres on the question below:

- “Does [Force Name] currently use, or plan to use, automatic transcription systems for producing ROTI (Record of Taped Interview) transcripts? If so, please specify the system(s) under evaluation or in use.”

Comparing responses from 2022 with 2025 responses will highlight any shifts in the adoption of automatic systems and situate the perception experiment within real-world policing trends.

The transcription evaluation experiment aims to examine how well participants detect transcript errors and how source labels (Human vs. ASR) affect trust. Simulated multi-speaker police interviews (Q&A format) will be used and a manually verified ground truth transcript will be produced. Audio will then be degraded with realistic background noise. An ASR transcript will be generated using CrisperWhisper (Radford et al., 2022), and a human transcript will be produced by a layperson with no linguistic training, much like ROTI clerks in the UK. Each participant receives the same ~2-minute audio recording and a transcript under one of four conditions:

1. ASR output, accurately labeled as ASR-generated
2. Human transcript, accurately labeled as human-generated
3. ASR output, mislabeled as human-generated
4. Human transcript, mislabeled as ASR-generated

Each participant will read their assigned transcript, rate trustworthiness on a Likert scale, and be asked to highlight any specific errors they notice. Flagged errors will be compared to the ground truth to calculate accuracy (true positives ÷ total errors) and over-flagging (false positives ÷ total flagged). We will examine whether source labelling affects error-detection rates and trust ratings. This design will evaluate both ASR versus human error profiles and the influence of labelling bias in a realistic police-interview context.

## Results

These results will inform our understanding of how human transcribers perceive ASR-generated transcripts - whether they trust machine output or scrutinise it more closely. Understanding these perceptions could lay the groundwork for developing targeted policy safeguards, bespoke training for transcription practitioners (Haworth et al. 2023), and evidence-based procurement criteria (Harrington, 2024) - so that, if ASR tools are ever adopted for ROTI transcription, they preserve the integrity of spoken evidence in court.

## References

- College of Policing. (2024). Covenant for using artificial intelligence (AI) in policing. <https://science.police.uk/delivery/resources/covenant-for-using-artificial-intelligence-ai-in-policing/>
- Essex Police. (2022, February). Record of Taped Interview (ROTI) and Record of Video Interview (ROVI) transcripts and transcribers [Freedom of Information response]. FOI 1039068/22
- Harrington, L. (2024). Towards improving transcripts of audio recordings in the criminal justice system (Unpublished doctoral dissertation, University of York, Department of Language and Linguistic Science). [Embargoed until 20 June 2025].
- Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *International Journal of Evidence & Proof*, 22(4), 428-450. <https://doi.org/10.1177/1365712718798656>
- Haworth, K., Tompkinson, J., Richardson, E., Deamer, F., & Hamann, M. (2023). "For the Record": Applying linguistics to improve evidential consistency in police investigative interview records. *Frontiers in Communication*, 8, Article 1178516. <https://doi.org/10.3389/fcomm.2023.1178516>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2022). Whisper: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356
- Tompkinson et al. (2022, June). For the Record: Assessing force-level variation in the transcription of police-suspect interviews in England and Wales [Conference presentation]. Spoken Interaction in Legal Contexts Symposium, Aston University, Birmingham, UK.

## A new resource for Cantonese Forensic Voice Comparison

*Bruce X. Wang<sup>1</sup>, Shuming Huang<sup>1</sup>, and Lei He<sup>2</sup>*

<sup>1</sup>*Department of English and Communication, The Hong Kong Polytechnic University, Hong Kong SAR, China*

{brucex.wang;shuming.huang}@polyu.edu.hk

<sup>2</sup>*Institute of Modern Languages and Linguistics, Fudan University, Shanghai, China.*

helei@fudan.edu.cn

A central challenge in forensic voice comparison (FVC) lies in identifying speech variables that have the strongest speaker-discriminatory potential. Effective features must exhibit low within-speaker variability and high between-speaker variability (Nolan, 1997; Rose, 2002). Empirical testing of such variables requires speech samples with known ground-truth comparisons, typically drawn from existing corpora. While forensic-grade corpora exist for English (Gold et al., 2018; Morrison et al., 2015; Nolan et al., 2009; Watt et al., 2018), French (Ajili et al., 2016), Spanish (Segundo et al., 2013), and Mandarin (Zhang, unpublished), no comparable resource is publicly available for Cantonese. Prior Cantonese FVC studies have relied on small cohorts (e.g., ~20 speakers: Cao et al., 2024; Rose & Wang, 2016), falling short of the 60–90 speakers recommended for robust validation (Hughes, 2017). This limitation likely reflects the absence of a dedicated Cantonese forensic speech corpus, hindering methodological rigor in the field.

As such, the current work-in-progress project addresses this gap by providing speech data from 102 male Hong Kong Cantonese speakers (mean age: 23, median: 22). Building on forensically oriented protocols (Gold et al., 2018; Morrison et al., 2012; Nolan et al., 2009), the corpus includes spontaneous non-contemporaneous recordings for each speaker, captured across two sessions separated by at least a 3-week interval. Each session features distinct speaking styles, including a mock police interview and a conversation with an accomplice—tasks adapted from the English DyViS corpus (Nolan et al., 2009) but localised with Hong Kong street and shop names. Additionally, a quasi-map task (Pang & Rose, 2012) based on the Hong Kong Mass Transit Railway (HKMTR) map elicits controlled, natural speech for likelihood ratio-based FVC. Typical questions involve asking participants how to get from station A to station B or how many stations between A and B. The choice of the station names was carefully selected aiming to cover all possible Cantonese vowels and consonants. A small talk task from Wormald (2016), using topic cards (e.g., hobbies, family), further diversifies speech styles. Table 1 summarises the tasks and their average net speech durations per session.

Recording sessions were conducted in a soundproof room at The Hong Kong Polytechnic University using a Shure SM58 microphone positioned approximately 30 cm from participants. Audio was captured at a 44.1 kHz sampling rate with 16-bit depth in Audacity (Audacity Team, 2021) and saved in WAV format. All participants were compensated and provided written informed consent under a protocol approved by the Human Subjects Ethics Sub-Committee at The Hong Kong Polytechnic University (Ref: HSEARS20231117002).

Automatically transcribed texts are currently undergoing manual correction, to be followed by forced alignment and additional manual refinement. The finalised corpus will be made publicly available for non-commercial research purposes in Cantonese forensic voice comparison and general Cantonese phonetic studies.

<i>Session 1</i>		<i>Session 2</i>	
Tasks	Net speech (min.)	Tasks	Net speech (min.)
Mock police interview	6.4	Conversation with an accomplice	6
HKMTR	4	HKMTR	4.2
Small talk	3.2	Small talk	3.2

**Table 1.** Tasks and average net speech duration in Sessions 1 and 2. **Mock police interview & Conversation:** adapted from the English DyViS corpus (Nolan et al., 2009) but localised with Hong Kong street and shop names. **Hong Kong Mass Transit Railway (HKMTR):** a quasi-map task (adapted from Pang & Rose, 2012) based on the HKMTR map to elicit controlled, natural speech. **Small talk:** Topic cards (e.g., hobbies, family) prompting spontaneous, free-form speech (adapted from Wormald, 2016).

*This work was supported by the Hong Kong Polytechnic University Start-up Fund for Research Assistant Professors under the Strategic Hiring Scheme P0049447 [1-BDUM].*

## References

- Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016). FABIOLÉ, a Speech Database for Forensic Speaker Comparison. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 726–733). European Language Resources Association (ELRA). <https://aclanthology.org/L16-1115>
- Audacity Team. (2021). *Audacity* [Computer software]. <https://audacityteam.org/>
- Cao, G. W., Hughes, V., Wang, B. X., & Mok, P. (2024). Cross-Language Forensic Voice Comparison of Hong Kong Trilingual Speakers using Filled Pauses and an Automatic Speaker Recognition System. *2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 279–283. <https://doi.org/10.1109/ISCSLP63861.2024.10800518>
- Gold, E., Ross, S., & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the Generalizability of Reference Populations for Forensic Speaker Comparison Casework. *Interspeech 2018*, 2748–2752. <https://doi.org/10.21437/Interspeech.2018-65>
- Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: How many speakers are enough? *Speech Communication*, 94, 15–29. <https://doi.org/10.1016/j.specom.2017.08.005>
- Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *The Speaker and Language Recognition Workshop*, 62–77.
- Morrison, G. S., Zhang, C., Enzinger, E., Ochoa, F., Bleach, D., Johnson, M., Folkes, B. K., De Souza, S., Cummins, N., & Chow, D. (2015). *Forensic database of voice recordings of 500+ Australian English speakers*. <https://opus.bibliothek.uni-augsburg.de/opus4/frontdoor/index/index/docId/67850>
- Nolan, F. J. (1997). Speaker recognition and forensic phonetics. In *The Handbook of Phonetic Sciences* (In Hardcastle, W. J. and Laver, J. (eds.), pp. 744–767). Oxford: Blackwell.
- Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*, 16(1), 31–57. <https://doi.org/10.1558/ijssl.v16i1.31>
- Pang, J., & Rose, P. (2012). Likelihood Ratio-Based Forensic Voice Comparison with the Cantonese Diphthong /ei/ F-Pattern. *Proceedings of the 14th Australasian International Conference on Speech Science and Technology*, 205–208.
- Rose, P. (2002). *Forensic Speaker Identification*. Taylor & Francis.
- Rose, P., & Wang, X. (2016). Cantonese forensic voice comparison with higher-level features: Likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. 326–333. <https://doi.org/10.21437/Odyssey.2016-47>
- Segundo, E. S., Alves, H., & Trinidad, M. F. (2013). CIVIL Corpus: Voice Quality for Speaker Forensic Comparison. *Procedia - Social and Behavioral Sciences*, 95, 587–593. <https://doi.org/10.1016/j.sbspro.2013.10.686>
- Watt, D., Llamas, C., French, P., Braun, A., & Robertson, D. (2018). A new corpus of Northern Englishes: Building the TUULS database for sociolinguistic and forensic research on phonetic variation in the Northeast of England. *The 8th Northern Englishes Workshop*, 21. [https://blogs.ncl.ac.uk/northernenglishes8/files/2018/03/NEW\\_8\\_Book\\_of\\_Abstracts\\_final.pdf](https://blogs.ncl.ac.uk/northernenglishes8/files/2018/03/NEW_8_Book_of_Abstracts_final.pdf)
- Wormald, J. (2016). *Regional Variation in Panjabi-English*. [PhD Thesis]. University of York.

# Sent to Coventry: A Study of Accent Perception in British Midlands Varieties

*Erin Broadhurst<sup>1</sup>, Paul Foulkes<sup>2</sup>, and Ben Gibb-Reid<sup>2</sup>*

<sup>1</sup>*MSc Forensic Speech Science, Language and Linguistic Sciences, University of York*  
rrb550@york.ac.uk

<sup>2</sup>*Project Supervisors, Language and Linguistic Sciences, University of York*

## Background

Midlands varieties of British English have been underrepresented in the study of regional variation. The city of Coventry, in the rare occasion of its mention, is often included in descriptions of West Midlands English alongside Birmingham, Wolverhampton, Solihull and other areas (Asprey and Hickey, 2015). However, Asprey and Hickey themselves stated that “Coventry...is underresearched in linguistic terms and rarely features in any linguistic discussion of the West Midlands” (2015:395). This lack of dedicated research has resulted in Coventry being considered a West Midlands variety with very little empirical backing, perhaps due to its inclusion in the former administrative county of the West Midlands. However, local rhetoric suggests that this classification should not be taken at face value. For example, Reddit user Dolly\_Wobbles stated that “as a migrant to the city [of Coventry] there’s definitely an accent. It’s got a West Midlands twang but a lot softer. Like a semi northern Derbyshire type thing with a slight touch of Brum.” (Dolly\_Wobbles, 2023). This suggests that Coventry English represented a perceptual middle ground between West and East Midlands varieties, supported by the anecdotal inability of locals to identify clear shibboleths of the variety relative to the more stereotypical accents of the West Midlands.

Within the field of forensic speech science, inappropriate categorisation of regional varieties could be highly problematic. Valid conclusions depend on the accurate estimation of typicality compared to a well-defined relevant population. The lack of empirical research into Coventry English, combined with apparently conflicting accounts of its nature between the literature and the locals, prohibits clear definition of relevant population.

## Aims

In order to gain further understanding of the perceptual boundaries between midlands varieties, the following research questions are proposed:

- 1) Do listeners perceive clear boundaries between varieties of Midlands English?
- 2) Does this perception vary as a function of listener variety?

## Procedure

These questions will be addressed through an accent sorting task based on the methodology used in Nije, Lavan and McGettigan’s (2023) study of the effect of accent familiarity on voice identification in Derry English. Listeners will form four experimental groups based on their regional variety: Birmingham, Coventry, Leicester, Non-Midlands. They will be presented with audio clips of speakers from 3 experimental midlands accent groups (Birmingham, Coventry, Leicester) and tasked with forming clusters according to perceived regional variety. Listeners will not be instructed as to how many clusters they should form.

## Implications

Results from this task will allow insight into several aspects of accent perception for midlands varieties:

- 1) Do participants demonstrate accurate perception *between* midlands varieties, such that each cluster contains only speakers of one variety?

- 2) Do participants demonstrate accurate perception *within* midlands varieties, such that a single cluster is formed per variety?
- 3) Do between-variety perception and within-variety perception vary as a function of the listener's own variety?

In this way, not only can accent discrimination be assessed, but the status of Coventry as a well-defined, independently perceived variety can too be investigated. As such, the nature of Coventry English relative to West and East Midlands varieties can be better understood.

## References

- Asprey, E., & Hickey, R. (2015). The West Midlands. *Researching Northern English*, 393-416.
- Njie, S., Lavan, N., & McGettigan, C. (2023). Talker and accent familiarity yield advantages for voice identity perception: A voice sorting study. *Memory & Cognition*, 51(1), 175-187.
- "Dolly\_Wobbles" (2023). *Is there a "Coventry Accent"?* [Online forum post]. Reddit. [Is there a "Coventry accent"? : r/coventry](https://www.reddit.com/r/coventry)

# Evaluating state-of-the-art generators and detectors of audio deepfakes

Daniel Denian Lee<sup>1</sup>, Linda Gerlach<sup>2</sup>, and Kirsty McDougall<sup>1</sup>

<sup>1</sup>Phonetics Laboratory, University of Cambridge, UK.

{ddl26|kem37}@cam.ac.uk

<sup>2</sup>Oxford Wave Research, Oxford, UK.

linda@oxfordwaveresearch.com

With significant progress in deepfake technology in recent years (Masood et al., 2023), *voice cloning* (in the form of *text-to-speech*, TTS; but also *voice conversion*) has shown remarkable naturalness in mimicking human speech. Detectors of audio deepfakes require regular updating to ensure they offer robust protection against presentation attacks (Almutairi & Elgibreen, 2022). To that end, an ongoing investigation into some currently available generators and detectors is presented, with the main research question being:

- How generalisable and robust is the performance of state-of-the-art detectors?

## Fake speech dataset

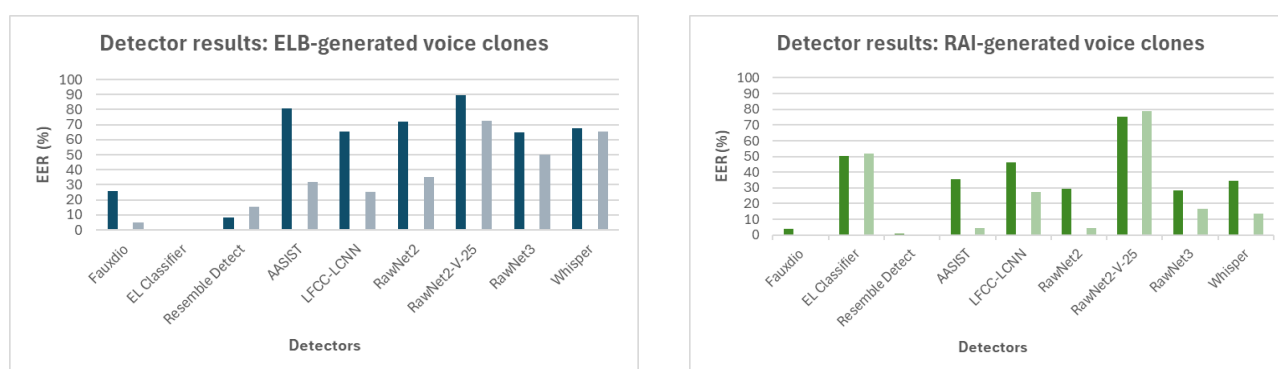
A growing dataset of fake speech—developed through TTS voice cloning—is introduced in this paper. This dataset can facilitate interdisciplinary collaboration on fake speech research, such as large-scale testing of audio deepfake detectors. In this ongoing work on generative fake speech, two commercial providers are presented in this study: namely, *ElevenLabs* (ELB) and *Resemble AI* (RAI) for their basic and ‘professional’ (PRO) tier TTS services. Five speakers with diverse accents in English were recorded in a sound-treated room of the Phonetics Laboratory, Cambridge, henceforth designated as the ‘LAB’ group of speakers, and four speakers’ bona fide speech data were retrieved from the CSTR VCTK Corpus (Yamagishi et al., 2019), henceforth the ‘VCTK’ group. Speaker 2 (Singapore English) had his voice cloned using both basic and ‘PRO’ tier TTS services through ELB and RAI. Basic voice cloning only needs a few seconds of speech as training data; PRO cloning requires substantially more speech training material. Thus, an approximately 30-minute recording of Speaker 2 reading aloud a semi-formal podcast transcript was used. The resultant PRO TTS voice clones show considerable improvements in naturalness compared to their basic counterparts (Lee et al., 2023). Future investigators are invited to enquire about utilising this growing dataset if their projects require thousands of TTS (and in the future, voice-converted) voice clones in a variety of accents and languages.

## Evaluating the detectors

To better understand the realistic levels of protection that current detectors provide against state-of-the-art voice clones, nine detector systems were examined, as shown in Table 1. ELB and RAI, which also provide TTS services, have in-house fake speech detectors—*ElevenLabs Classifier* (EL Classifier) and *Resemble Detect*. Thus, it is interesting to investigate how well they detect their own voice clones and other-generated voice clones. *Fauxdio* (Alexander et al., 2025), another commercial audio deepfake detector, was also examined. Finally, through the *Deepfake-O-Meter* (DFoM) interface (UB Media Forensic Lab, 2025), six open-source detectors were evaluated (see 4.1–4.6 in Table 1). Overall, variation is observed in the robustness between the nine detectors, with Resemble Detect and Fauxdio being the most reliable in correctly discriminating between bona fide speech and TTS voice clones from the different generators, achieving between 0% to 25.83% equal error rate (EER; Figure 1A and 1B), whereas the six DFoM models demonstrate limited effectiveness in classifying state-of-the-art fake speech. EL Classifier, while effective in detecting ELB-generated voice clones (Figure 1A), it is ineffective against non-ELB voice clones (Figure 1B).

#	Provider	Detector
1	ElevenLabs	EL Classifier
2	Resemble AI	Resemble Detect
3	Oxford Wave Research	Fauxdio
4	UB Media Forensics Lab	Deepfake-O-Meter
4.1		AASIST (2021)
4.2		LFCC-LCNN (2021)
4.3		RawNet2 (2021)
4.4		RawNet2-Vocoder (2023)
4.5		RawNet3 (2023)
4.6		Whisper (2023)

**Table 1.** An overview of the audio deepfake detectors tested.



**Figure 1A** (Left) Overview of EER values for ELB-generated basic voice clones, separated by LAB (blue) and VCTK (light blue) speakers. **Figure 1B** (Right) Similar overview for RAI-generated basic voice clones, separated by LAB (green) and VCTK (light green) speakers.

## References

- Alexander, A., Gerlach, L., Coy, T., Forth, O., Lonergan, L., & Kelly, F. (2025). FAUXDIO: An audio deepfake detector for law enforcement and forensics. 33rd International Association for Forensic Phonetics and Acoustics (IAFPA) Conference, Den Haag, The Netherlands.
- Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms*, 15(5), 155.
- Lee, D. D., McDougall, K., Kelly, F., & Alexander, A. (2023). PASS (Phonetic Assessment of Spoofed Speech): Towards a human-expert-based framework for spoofed speech detection. 31st International Association for Forensic Phonetics and Acoustics (IAFPA) Conference, Zurich, Switzerland.
- Masood, M., Nawaz, M., Malik, K. M., Javed, A., Irtaza, A., & Malik, H. (2023). Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward. *Applied Intelligence*, 53(4), 3974–4026.
- UB Media Forensic Lab. (2025). Deepfake-O-Meter [Computer software]. [https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/home\\_login](https://zinc.cse.buffalo.edu/ubmdfl/deep-o-meter/home_login)
- Yamagishi, J., Veaux, C., & MacDonald, K. (2019). CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (Version 0.92) [Sound].

# Reference Population Effects on Automatic Speaker Recognition Performance

*Sophie Möller*<sup>1</sup>, *Andrea Fröhlich*<sup>2</sup>, *Sarah Lim*<sup>2</sup>, *Adrian Leemann*<sup>3</sup>  
and *Gea de Jong-Lendle*<sup>1</sup>

<sup>1</sup>*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*  
{moeller9|dejong}@students|staff.uni-marburg.de

<sup>2</sup>*Zurich Forensic Science Institute, Switzerland*  
{sarah.lim|andrea.froehlich}@for-zh.ch

<sup>3</sup>*Institute of Germanic Languages and Literatures, University of Bern, Switzerland*  
adrian.leemann@unibe.ch

## Introduction

Morrison et al. (2021) and Drygajlo et al. (2015), both recommendation papers, state that forensic ASR evaluations should be sufficiently representative of casework conditions, including appropriate reference populations. However, recent studies have shown that ASR systems can be relatively robust against language mismatch (Watt et al., 2020; van der Vloed, 2024). In practice, suitably matched corpora are often unavailable—especially in Switzerland, where dialect diversity and limited data pose particular challenges. Unlike in Germany, where dialects have declined in favour of a shared standard variety, Swiss dialects remain relatively robust and widely used even in formal contexts (Ruch et al., 2023). They differ substantially in phonetic, phonological, and lexical dimensions (Leemann et al., 2024).

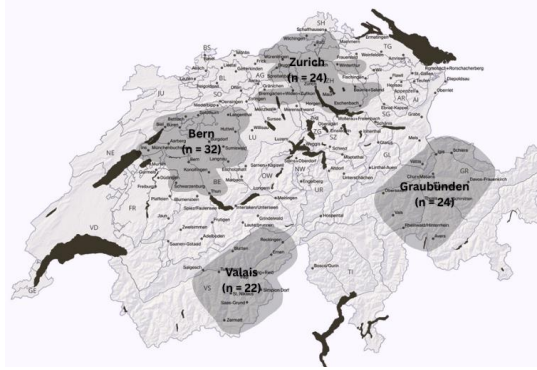
## Objectives

Current research increasingly investigates how mismatched reference data may affect ASR performance and evidential strength. Van der Vloed (2020; 2024) emphasises the need for systematic evaluations across varying conditions. While Watt et al. (2020) suggest that ASR systems may be robust to accent and even language mismatch due to holistic feature modelling, they also note that few studies have examined this and conclude that accent mismatch may not affect performance substantially but can impact evidential strength. Using Swiss corpora recorded under realistic conditions, this study investigates whether a standard-language corpus could serve as a reference population or whether dialect-matched corpora improve ASR performance.

## Methodology

Testing was conducted using VOCALISE 2021 and Bio-Metrics 2024, combining quantitative evaluation via equal Error Rate (EER) and Log-Likelihood Ratio cost ( $C_{lr}$ ) metrics with a qualitative analysis of how different reference databases affected likelihood ratios (LRs). The dataset consists of 102 male speakers from the SDATS corpus (Leemann et al., 2020), aged 17–43, from four Swiss dialect regions: Zurich ( $n = 24$ ), Bern ( $n = 32$ ), Valais ( $n = 22$ ), and Graubünden ( $n = 24$ ). For each speaker, three recordings were selected, representing different speech modes and dialectality levels: spontaneous speech (1 min.), read speech in standard German (25 sec.) and in dialect (2 min.), acknowledging that the regional groupings used here are simplified and that dialect boundaries are fluid, influenced by age, locality, education, and social meaning (Leemann et al., 2024). Due to COVID-19, data were collected via a supervised remote app procedure, resulting in field-like recordings with varied background noise. Therefore, a dedicated noise classification scheme based

on Jessen (2012) was developed for annotation. This approach enabled retrospective diagnostics of potential confounding factors, though further validation is needed.



**Figure 1.** The four dialect regions examined: Zurich (n = 24), Bern (n = 32), Valais (n = 22), and Graubünden (n = 24).

### Preliminary observations

Low EER (0,3-1,1%, convex hull) and  $C_{llr}$  (0,02-0,09) values after cross-validation were observed, indicating strong baseline performances. Lower values indicating higher performance were observed, however, when reference data did not include standard-language samples. These preliminary findings suggest that the ASR system performed well on dialectal speech data, in line with previous studies (Watt et al., 2020). Ongoing work focuses on 1-vs.-1-comparisons and a diagnostic analysis of other factors that may play a role.

### References

- Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., & Niemi, T. (2015). *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Verlag für Polizeiwissenschaft Frankfurt.
- Jessen, M. (2012). *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. LINCOM Europa.
- Leemann, A., Steiner, C., Studerus, M., Oberholzer, L., Jeszenszky, P., Tomaschek, F., & Kistler, S. (2024). *Dialäktatlas: 1950 bis heute*. vdf Hochschulverlag AG. <https://doi.org/10.3218/4184-2>
- Leemann, A., Studerus, M., Messerli, J., Jeszenszky, P., & Steiner, C. (2020). *SDATS Corpus – Swiss German Dialects Across Time and Space*. <https://doi.org/10.17605/OSF.IO/S9Z4Q>
- Morrison, G. S., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., Ypma, R. J. F., Zhang, C., Anonymous, A., & Anonymous, B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice: Journal of the Forensic Science Society*, 61(3), 299–309. <https://doi.org/10.1016/j.scijus.2021.02.002>
- Ruch, H., Fröhlich, A., & Lim, S. (2023). Grosse sprachliche Vielfalt auf kleinem Raum: Chancen und Herausforderungen für die forensische Phonetik in der Schweiz. *Kriminalistik* (4), 236–244. [https://www-wiso-net-de.ezproxy.ub.uni-marburg.de/document/KRIM\\_\\_69de2f72bb7f4241a1d17455f373dcee6eeafcf1](https://www-wiso-net-de.ezproxy.ub.uni-marburg.de/document/KRIM__69de2f72bb7f4241a1d17455f373dcee6eeafcf1)
- van der Vloed, D. (2024). Interchangeability of Calibration Audio Datasets for Forensic Automatic Speaker Recognition. In *2024 12th International Workshop on Biometrics and Forensics (IWBF)* (pp. 1–6). IEEE. <https://doi.org/10.1109/iwbf62628.2024.10593938>
- Watt, D., Harrison, P., Hughes, V., French, P., Llamas, C., Braun, A., & Robertson, D. (2020). Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system. *The International Journal of Speech, Language and the Law*, 27(1), 1–34. <https://doi.org/10.1558/ijsl.41466>

## Casework Conundrums

Megan Thomas<sup>1</sup>, Katherine Earnshaw<sup>1,2</sup>, Bryony Nuttall<sup>1,2</sup> and Richard Rhodes<sup>1,2</sup>

<sup>1</sup>The Forensic Voice Centre, York, UK

<sup>2</sup>Department of Language and Linguistic Science, University of York, York, UK  
megan.thomas@forensicvoicecentre.com

We would like to use the poster session to facilitate a conversation between practitioners within the IAFPA community regarding a range of practical issues in casework that we frequently encounter at the Forensic Voice Centre; primarily in order to discuss, develop and share solutions. We also welcome practical issues from other practitioners and we will use the poster session to discuss these as well (these can be provided in advance). There are five main areas that we would like to explore:

### 1) Combining voice notes

Increasingly, voice notes (e.g., from WhatsApp, Signal) are submitted as questioned recordings in voice comparison casework. These recordings are often short and do not contain sufficient speech for a full analysis when treated as individual recordings. In some cases, it may be possible to produce a suitable questioned sample by extracting and combining the speech from multiple voice notes, assuming that the voice notes are part of one conversation and are (semi-)continuous. In cases where we have grouped voice notes, we include a caveat in the report to explain that they have been grouped based on the information provided by our instructing party, and if additional information is provided to refute this assumption, we would need to re-evaluate the findings. We are interested in discussing: *what approaches should we use when dealing with voice notes? How continuous is 'continuous'?*

### 2) Unconfirmed-identity reference material

We typically request all available reference material for a known individual in a voice comparison case as it allows us to best assess within-speaker variation. Consequently, we are often provided with supplementary unconfirmed-identity reference material such as bank or insurance calls or non-emergency police calls in which the speaker confirms personal information, but their identity may not be visually confirmed as it would be in an interview. These are sometimes the best source of reference if confirmed-identity reference recordings - e.g. from police interviews - contain limited speech. At the Forensic Voice Centre we separate the unconfirmed-identity recordings from the confirmed-identity recordings, producing different conclusions for each type of reference material. This ensures that, if the identity of the person in the unconfirmed recording(s) is later contested, we can refer the court to the comparison using the confirmed sample(s) only. We are interested to discuss: *how do other practitioners deal with different statuses of reference material? And whether they group reference recordings in a similar or different way? Have practitioners ever received information partway through a case that contests the identity of a speaker in a reference recording?*

### 3) Suspected synthetic speech

We are starting to receive enquiries relating to whether speech is real or generated through the use of Artificial Intelligence (AI), i.e., it is claimed or suspected that the recordings contain 'deepfaked', 'synthesised' or 'spoofed' speech. Our current position is that if there is a suggestion or suspicion that AI interference has occurred, this would be outside our area of expertise. *But what happens if we have reason to believe that a recording submitted in a different type of case is unnatural, are we qualified to comment on this? What would a statement to this effect look like?* Currently there are few experts in the United Kingdom who could work on the assessment of deepfakes; *how, as an industry, should we deal with this? What would qualify an expert to be able to do this assessment? What items do they need, and what information/case circumstance do they need to know? Will we eventually need full authentication of every recording we work on? How do different jurisdictions deal with this issue in legal proceedings?*

#### 4) Localised attributions in long transcripts

Most forensic transcription and attribution work requires the detailed analysis of short sections of speech/dialogue in a recording. Attribution of speakers in short sections of recordings made on the same device can be relatively straightforward, and there may be many aspects that assist us, for example, speech context, distance, visual information, *etc.* However, this isn't always the case, particularly where we are instructed to transcribe much a longer series of events, recorded on multiple device types (e.g., CCTV recordings, mobile phone recordings, telephone calls, *etc.*) in different locations; these factors can make attribution of speech more difficult. *How can we attribute speakers in different localised events in the clearest way for the trier of fact? Is it helpful to have local attributions such as Male X or Male Y, which may or may not be the same person as Male 1/Male 2, or is this confusing? What is the clearest way we can convey these ideas to a jury?*

#### 5) Peer-review procedures

All analyses undertaken by the Forensic Voice Centre undergo a full checking procedure in line with the [UK] FSR Code of Practice. This ensures that appropriate procedures have been followed from start to finish, and reduces the risk of errors in the analyses and final reports. Our current checking procedure in each voice comparison involves a second suitably qualified analyst reaching their own independent (semi-blind) conclusion and then confirming that all work carried out by the primary analyst is correct, appropriate, fully documented, compliant to our policies and procedures, and consistent with the contents of the report or statement. This is a time-consuming process that relies on multiple staff members being kept from potentially biasing case information, something that may be more difficult for smaller teams or sole providers. *What do other organisation's checking procedures involve? What level of independence is required? How do checking procedures need to change to meet the requirements of the new FSR codes?*

#### References

[UK] Forensic Science Regulator. (2023). Forensic science activities: Statutory Code of Practice (version 1 - March 2023; version 2 consultation draft also available from February 2024) - [URL: <https://www.gov.uk/government/publications/statutory-code-of-practice-for-forensic-science-activities>]

# Visualising latent representations: An interactive approach to improving the linguistic interpretability of MFCC-based phoneme models

*Samantha Williams*

*Department of Language and Linguistic Science, University of York, York, UK*  
*samantha.ej.williams@gmail.com*

## Introduction

Building models of speech often requires dimension reduction due to the high dimensionality of input features such as MFCCs. However, it can be challenging to determine which speech information is being captured during this process, often requiring extensive additional testing (e.g., Hughes et al., 2023) or post hoc explanations.

In this work, the Variational Autoencoder plus Spectrogram Reconstruction (VAE+SR) method is proposed as a linguistically interpretable approach to phoneme modelling for forensic and sociophonetic applications. The VAE neural network architecture (Kingma and Welling, 2014) reconstructs the data directly from the latent representations leading to less reliance on post hoc explanations of the model. While the addition of spectrogram reconstruction and an interactive interface improves the linguistic interpretability, making it easier to communicate which speech features are being captured by the model and therefore used in any downstream tasks.

## Methods

The VAE+SR method as a proof-of-concept was tested on a combination of unsupervised individual L2 English phoneme models and a supervised multi-vowel model. The models were trained on segmental phoneme data labelled based on the target General American phoneme, from 425 speakers representing 11 L2 varieties of English (Weinberger, 2015). To support exploration and analysis, an interactive application was built enabling real-time spectrogram reconstruction from model outputs.

## Evaluation

A domain-specific evaluation framework based on guidelines from explainable AI (Phillips et al., 2021) and guidance for forensic speech science practitioners (Rhodes and Cambier-Langeveld, in press) was used to evaluate the interpretability of the VAE+SR method in the context of a forensic phonetician as the initial ‘end-user’ of the application:

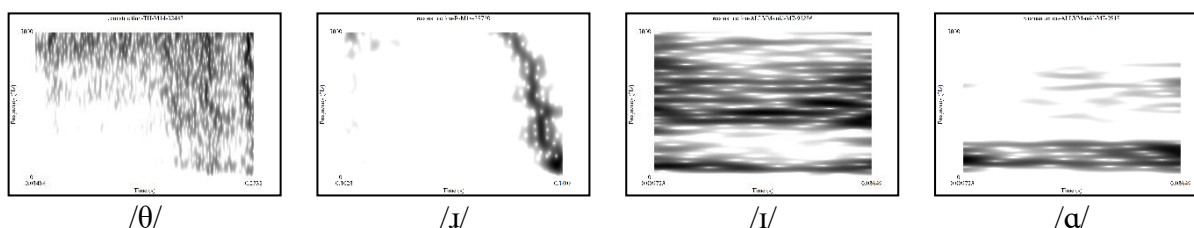
- (i) The model provides an *explanation* in its output
- (ii) The output is *meaningful* to the end-user
- (iii) The output *accurately* reflects the information used by the model
- (iv) The model captures *relevant* information for forensic speech analysis

Criteria (i)-(iii) address the interpretability, while (iv) is added to ensure the potential for use in real-world applications. A final criterion regarding understanding the limitations of the method would be evaluated at a later stage in model development.

## Results

The proposed VAE+SR method meets (i-iii) by providing a phonetically meaningful output in the form of a reconstructed audio file and/or spectrogram. Specifically, (ii) and (iii) are met as the expected spectral features for each phoneme model are visible in the corresponding reconstructions (Figure 1) and there is a clear relationship between the reconstruction and its position within the

model. However, (iv) was only partially met. The positioning of data in the multi-vowel model, which captured broad formant information, was consistent with the vowel space distribution from manual formant measurements, but some relevant acoustic information was lost in favour of more salient patterns in the data.



**Figure 1.** Example reconstructions of /θ/, /ɪ/, /ɪ/, and /ɑ/ phonemes. Note, the /θ/ and /ɪ/ models include zero padding resulting in the reconstruction containing a period of ‘silence’ before the phoneme. The /θ/ example reconstruction shows some high frequency noise in this period, while the /ɪ/ reconstruction does not. The y-axis shows frequency (Hz), up to 5 kHz, while time (s) is shown along the x-axis.

While improvements to model design such as optimising hyperparameters and features could improve the method’s ability to capture more detailed phonetic information, this work presents a step towards developing more linguistically interpretable methods to assist the forensic expert in making data-driven decisions that can be explained and supported by existing research in linguistics and phonetics.

## References

- Hughes, V., Wormald, J., Foulkes, P., Harrison, P., Kelly, F., Vloed, D. van der, Welch, P., & Xu, C. (2023, August 20). Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics. *INTERSPEECH 2023*. INTERSPEECH 2023. <https://doi.org/10.21437/interspeech.2023-443>
- Kingma, D. P., & Welling, M. (2014). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations, ICLR*. <https://doi.org/10.48550/arXiv.1312.6114>
- Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., & Przybocki, M. A. (2021). Four principles of explainable artificial intelligence (Vol. 2019). National Institute of Standards and Technology (U.S.). <https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8312.pdf>
- Rhodes, R., & Cambier-Langeveld, T. (in press). Guidance for Practitioners. In Oxford University Press: *Handbook of Forensic Phonetics*.
- Weinberger, S. (2015). *Speech Accent Archive*. Speech Accent Archive. <http://accent.gmu.edu/index.php>

# Forensic phonetic analysis of spoofed European Portuguese speech

Lina Almeida<sup>1,2</sup>, Amelia Gully<sup>2</sup>, and Paul Foulkes<sup>2</sup>

<sup>1</sup> Forensic Speech Solutions, Lisbon, PT

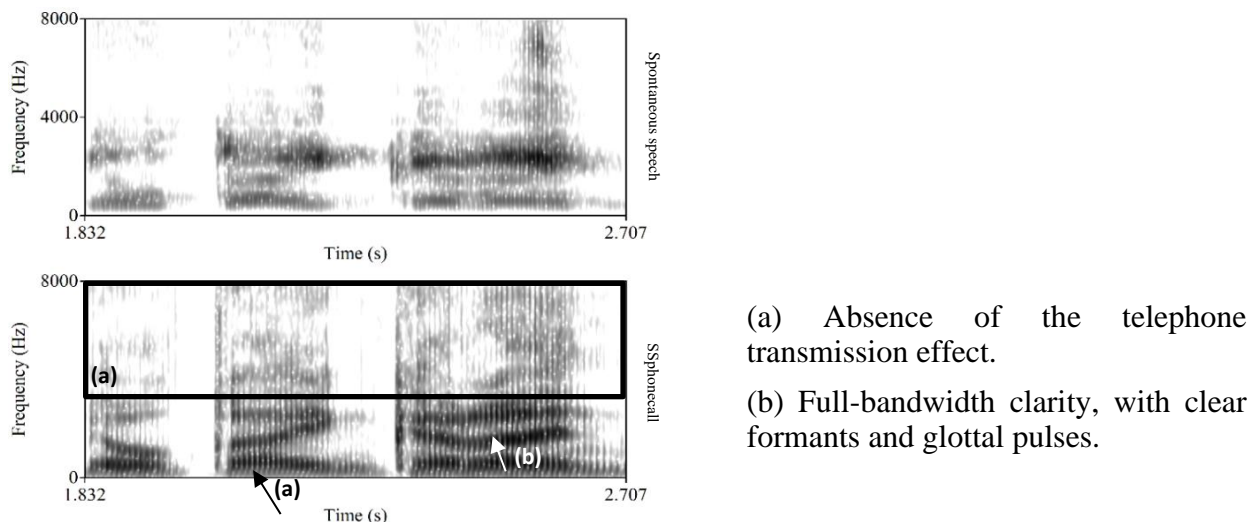
linalmeida@gmail.com

<sup>2</sup> Department of Language and Linguistic Science, University of York, UK

{amelia.gully|paul.foulkes}@york.ac.uk

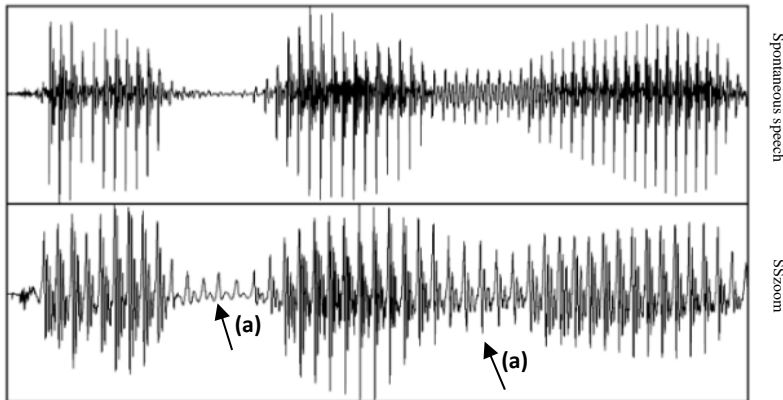
The increasing similarity between spoofed speech samples and human voices poses a significant challenge for systems tasked with accurately detecting and distinguishing between human and synthetic speech. Recent studies have raised concerns about the applicability of speech technology across different languages and the optimal features for speaker comparison in the context of spoofed speech. Automatic spoofing detection systems, such as those submitted to ASVspoof 2024 (Wang et al., 2024), have demonstrated increasing accuracy in detecting spoofed speech. However, it remains unclear how these systems determine whether a speech sample is genuine, and their performance in challenging audio conditions has yet to be fully validated. Kirchhübel and Brown (2022) were among the first to examine the acoustic characteristics of spoofed speech, assembling a dataset of 300 English read speech samples from ASVspoof 2015 and 2019. Lee et al. (2023) introduced the Phonetic Assessment of Spoofed Speech (‘PASS’) framework, designed to assist human experts in detecting spoofed speech. While both studies aimed to identify distinguishing characteristics of spoofed speech, Lee et al. focused on a structured forensic assessment framework, whereas Kirchhübel and Brown provided a phonetic analysis of various spoofing techniques. In this study, we expand the focus to include voice conversion spoofed samples and use European Portuguese (EP) due to its relative underrepresentation in the speech training data of synthesis systems, as well as its typologically unusual phonetic features, such as nasal vowels.

We recorded spontaneous conversations via phone, Zoom, and WhatsApp involving five male Portuguese speakers aged 28 to 54. These recordings were then converted using a voice conversion model (Li et al., 2022). The recordings were analysed phonetically and acoustically to identify any distinctive features. The spoofed speech was compared to the spontaneous conversations, taking into account features identified in previous studies and new features that are particular to features of EP. Our analysis revealed a number of distinctive features in spoofed speech by voice conversion. These include the absence of the telephone effect (Künzel, 2001, Byrne and Foulkes, 2004) (marked **a** in Figure 1), and the presence of consistent, clear and strong glottal pulses and formant marks in spoofed speech samples (**b** in Figure 1).



**Figure 1.** Comparison of spontaneous and spoofed speech in *phonecall* channel condition.

The so-called *telephone effect* refers to the narrow bandpass of approximately 300-3500 Hz introduced by telecommunication systems. This filtering results in the attenuation of formant frequencies—particularly F1 and F3—which are crucial for vowel identification and speaker characterization. As a consequence, natural speech transmitted via telephone often exhibits reduced spectral detail and less clearly defined formant structures. In contrast, our spoofed samples consistently lacked these distortions, displaying energy well above 3.5 kHz, unnaturally clear and stable formant tracks, and pronounced glottal pulses.



(a) Wider, sharper, and more regular periodic waveform amplitude characteristics in spoofed speech.

**Figure 2.** Comparison of waveforms in spontaneous and spoofed speech in Zoom condition.

Differences were also observed in amplitude peaks at the onset of /t/ and in the greater energy spread across the frequency range during nasal vowels. Additionally, spoofed speech in the WhatsApp and Zoom channel conditions exhibited wider, sharper, and more regular periodic waveform amplitude characteristics, as illustrated in Figure 2 (marked a).

This study underscores the crucial role of auditory and acoustic analysis in forensic speech science, offering valuable insights for future speaker comparisons and investigations involving synthetic speech. By identifying and characterizing these distinctions, this work contributes to improving the reliability of forensic assessments and the overall security of speech-based technologies.

## References

- Byrne, C. & Foulkes, P. (2004) The mobile phone effect on vowel formants. *International Journal of Speech, Language and the Law* 11:83-102.
- Kirchhübel, C., & Brown G. (2022) Spoofed speech from the perspective of a forensic phonetician. *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, Korea, 18-22.
- Künzel, H.J. (2001) Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic Linguistics: the International Journal of Speech, Language and the Law* 8(1):80-99.
- Lee, D., McDougall, K., Kelly, F., & Alexander, A. (2023) *PASS (Phonetic Assessment of Spoofed Speech): Towards a human-expert-based framework for spoofed speech detection*. Paper presented at the 31st IAFPA Conference, Zürich. Abstract available at: <https://iafpa2023.uzh.ch/dam/jcr:519171ea-58b0-4ff4-8333-91664dc44a54/BoA-IAFPA23.pdf> (pp. 31-32).
- Li, J., Tu, W., and Xiao, L. (2022) FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion, arXiv, <https://arxiv.org/abs/2210.15418>
- Liu, X., Wang, X., Sahidullah, M., Patino, J., Delgado, H., Kinnunen, T., Todisco, M., Yamagishi, J., Evans, N., Nautsch, A., & Lee, K. (2023) ASVspooF 2021: Towards Spoofed and Deepfake Speech Detection in the Wild. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31:2507-2522. doi: [10.1109/TASLP.2023.3285283](https://doi.org/10.1109/TASLP.2023.3285283).
- Wang, X., Delgado, H., Tak, H., Jung, J.-w., Shim, H.-j., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T.H., Evans, N., Lee, K.A., Yamagishi, J. (2024) ASVspooF 5: crowdsourced speech data, deepfakes, and adversarial attacks at scale. Proc. The Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspooF 2024), 1-8, doi: [10.21437/ASVspooF.2024-1](https://doi.org/10.21437/ASVspooF.2024-1)

# The more the better: assessing formant-based features for speaker differentiation with random forest--a pilot study

Kang Jintao<sup>1</sup>, Gao Kai<sup>1</sup>, Huang Wenlin<sup>1</sup>

<sup>1</sup>*Institute of Forensic Science, Ministry of Public Security, China*  
 {kangjintao|gaokai|huangwenlin}@cifs.gov.cn

## Introduction

The demand for interpretability and explainability of forensic sciences has made formant-based features (Nolan, 1983) one of the most popular acoustic parameters in forensic speaker identification (Gold and French, 2011; ENFSI, 2021). However, within the auditory-phonetic-acoustic methodology framework, there is no unified way to quantify these feature values. The present study proposes a method using random forests as classifiers to assess the values of acoustic features and uses Mandarin monophthongs as the subjects to evaluate and compare the values of their formant features.

## Data and Methods

This study selected 100 male speakers from RASC863 (Li and Wang, 2003) and RASC863-G2 as subjects. Each recording consists of a spontaneous monologue on a pre-designated topic and lasts about 4 minutes.

The Montreal Forced Aligner (McAuliffe, Socolof, et al. 2017) was used to automatically annotate these 100 recordings. For each monophthong, the segment duration was divided into 15 equal parts (for the 15-point results) and 5 equal parts (5-point results) to compare the performance of different representative data volumes. The first four formant frequencies were extracted and averaged at each part using a Praat script (Boersma and Weenink, 2001) with the following settings: 25ms analysis window, pre-emphasis from 25 Hz, Burg algorithm, maximum formant frequency of 4000 Hz, and 4 formants.

Besides, for each vowel, the means of the first 4 formants were calculated using the 15-point data. Their trajectories were fitted with Legendre polynomials (McDougall, 2006; Kang, Li, et al. 2022) and the first 4 coefficients were kept representing dynamic features of these formants.

The random forest algorithm (Breiman, 2001) was chosen as the classifier for two reasons: (1) It performs well in many classification tasks in theory and practice. (2) It offers a good feature selection indicator, which is very useful in showing the relative importance of each feature in this study. Its parameters were set as: n\_estimators: 100, criterion: Gini, min\_samples\_split: 5. The rest were kept as the default values.

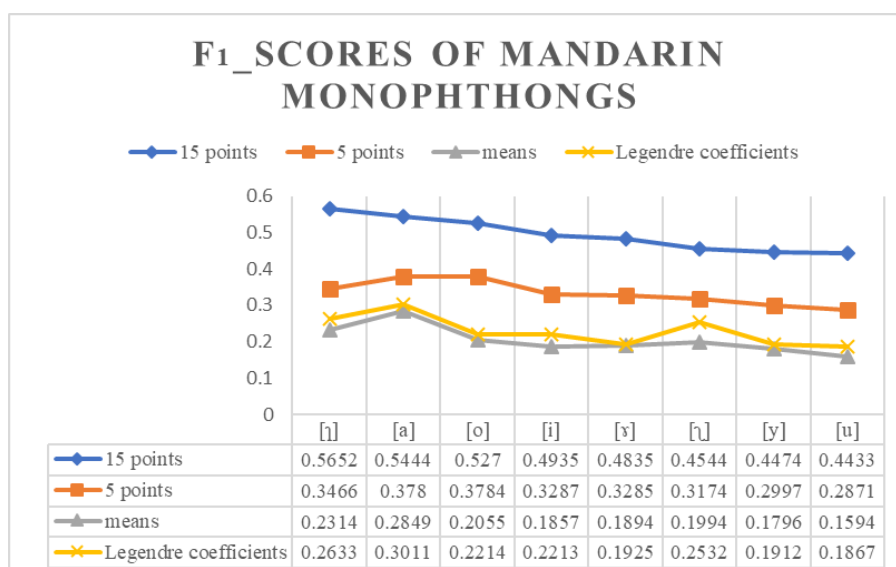
F<sub>1</sub> score was chosen as the metric for its good balance between precision and recall. Its formula is:

$$F_1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} = \frac{TP}{TP + \frac{FP + FN}{2}}$$

where TP, FP and FN are the numbers of true positive, false positive and false negative of the classifier respectively.

## Results

Figure 1 shows the F<sub>1</sub> scores of the eight vowels from four kinds of input (15 points, 5 points, means and Legendre coefficients). As shown in Figure 1, the higher the data dimensionality (60 dimensions for 15 points and 4 dimensions for means), the better the classifier's performance. For specific vowels ([ɿ], [a] and [o]), they exhibit stronger speaker differentiation abilities compared to other monophthongs. The formant bandwidth data also exhibits the same trend, although its performance is not as good as that of formant frequencies. Due to space limitations, their results are not listed here. As for feature importance, we can see from Table 1 that F4 and F3 have generally higher importance values than F2 and F1, which may indicate that the higher formants carry more speaker-specific information.



**Figure 2.** The F1 scores of the random forest classifiers for 4 kinds of input

Monophthongs	F1	F2	F3	F4
[a]	<b>0.1696</b>	0.1647	0.1581	<b>0.1743</b>
[ɿ]	0.1633	0.1599	<b>0.1738</b>	0.1697
[i]	0.1625	<b>0.1703</b>	0.1645	<b>0.1694</b>
[o]	0.1664	0.1573	<b>0.1757</b>	0.1674
[u]	0.1623	0.1606	<b>0.1731</b>	<b>0.1707</b>
[y]	<b>0.1704</b>	0.1657	0.1607	<b>0.1700</b>
[ɥ]	0.1598	0.1624	0.1719	<b>0.1726</b>
[ɨ]	0.1587	0.1580	<b>0.1702</b>	<b>0.1797</b>

**Table 3.** The feature importance values of formant frequencies (based on impurity)

## Discussion

This method has proven effective for assessing feature values within a limited scale in a specific corpus. Next, we are planning to experiment with Mandarin diphthongs and triphthongs in other corpora.

## References

- Boersma, P., & Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *GLOT International*, 5, 341-347.
- Breiman, L. (2001). Random Forests. *Machine Learning*, 45, 5-32.
- ENFSI. (2021). Best Practice Manual for the Methodology of Forensic Speaker Comparison (Version 01).
- Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18(2), 293-307.
- Kang, J. T., Li, J. Y., & Li, A. J. (2022). Formant Dynamics of Chinese Compound Vowels with Implication for Forensic Speaker Identification. In *Proceedings of Odyssey 2022* (pp. 396-401).
- Li, A. J., Wang, T. Q., et al. (2003). RASC863 - Voice Corpus for Speaker Recognition. In *Proceedings of the Seventh National Conference on Human-Machine Voice Communication*.
- McAuliffe, M., Socolof, M., et al. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498-502).
- McDougall, K. (2006). Dynamic Features of Speech and the Characterisation of Speakers: Towards a New Approach Using Formant Frequencies. *International Journal of Speech, Language and the Law*, 13(1), 89-126.
- Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.

# Phonological Familiarity and Voice Discrimination: A Study on Quranic Reciters without Arabic Comprehension

*Sadia Azad<sup>1</sup>, Elisa Pellegrino<sup>1</sup>, Eleanor Chodroff<sup>1</sup>, Volker Dellwo<sup>1</sup>*

<sup>1</sup>*Department of Computational Linguistics and Phonetics, University of Zurich,*

*+41 44 634 11 11, Zurich, Switzerland*

*sadia.azad@uzh.ch*

*eleanor.chodroff@uzh.ch | elisa.pellegrino@uzh.ch | volker.dellwo@uzh.ch*

**Introduction.** Previous studies have demonstrated that listeners are better at discriminating voices in languages they are familiar with, an effect known as the language familiarity effect (LFE; Goggin et al., 1991; Perrachione et al., 2011). It has been debated whether the LFE is rooted in semantic or phonological knowledge of the language. To highlight the role of phonological knowledge, prior research has often employed techniques such as time-reversed speech (Fleming et al., 2014). Most studies on voice discrimination focus on native speakers or those who understand the language—but what about those who only hear and repeat it without comprehension? Our study bridges a crucial gap in our understanding of how phonological knowledge without language comprehension shapes our ability to recognize voices by investigating Quranic reciters against native and non-Arabic speakers. The native Arabic speakers serve as a baseline group to establish standard performance in Arabic speaker discrimination. The reciters are the primary experimental group, testing whether phonological familiarity alone enhances voice recognition. The non-Arabic speakers serve as a control group with no phonological familiarity with Arabic, allowing the isolation of phonological exposure effects.

**Research question.** To what extent does phonological knowledge of the language alone facilitate speaker discrimination performance, but when the presented speech is natural and the listener has almost no comprehension (i.e., for Quranic reciters without Arabic comprehension)?

**Methodology.** The experiment involved native Arabic speakers from Saudi Arabia ( $n = 9$ ; data collection ongoing), Reciters (Huffaz-e-Quran,  $n = 30$ ) without Arabic Comprehension, and non-Arabic speakers (Non-Reciters,  $n = 30$ ). The reciters and non-Arabic speakers were from Pakistan. The study used an Arabic speech corpus comprising 20 Arabic words, spoken by 50 native male Arabic speakers (Alalshekmubarak & Smith, 2014). Audio files were processed at 44,100 Hz, 16-bit stereo (converted to mono), standardized to 70 dB for uniform loudness. Participants performed a speaker discrimination task on Gorilla Experiment Builder and listened to 80 word pairs (40 same-speaker, 40 different-speaker pairs), which were randomized to prevent order effects. Fixed-duration pauses were inserted between word pairs to minimize carryover effects in perception. Each participant was presented with a unique randomized sequence, and participants were not allowed to replay audio clips to prevent learning effects.

**Results and Discussion.** In terms of accuracy, native speakers performed at 84% accuracy, reciters at 74% and non-Arabic speakers at 48%. A logistic mixed effects model was implemented assessing accuracy based on participant group. In line with expectations, native speakers performed significantly better than reciters or non-Arabic speakers (reciters:  $\beta = -0.53$ ; non-Arabic:  $\beta = -1.77$ , each  $p < 0.001$ ). Critically, however, reciters also performed significantly better than non-Arabic speakers ( $\beta = -1.14$ ,  $p < 0.001$ ). In a future analysis, we aim to perform a signal detection theory analysis. Overall, these findings suggest that phonological familiarity alone contributes strongly to voice discrimination abilities, even in the absence of language comprehension. It challenges models that tie the language familiarity effect strictly to comprehension and also supports a phonologically grounded account of voice recognition.

## References

- Alalshekmubarak, A., & Smith, L. S. (2014). On improving the classification capability of reservoir computing for Arabic speech recognition. In S. Wermter, C. Weber, W. Duch, T. Honkela, P. Koprinkova-Hristova, S. Magg, G. Palm, & A. E. P. Villa (Eds.), *Artificial neural networks and machine learning—ICANN 2014: 24<sup>th</sup> International Conference on Artificial Neural Networks (Lecture Notes in Computer Science, Vol. 8681, pp. 225–232)*. Springer.
- Fleming, D. E., Giordano, B. L., Caldara, R., & Belin, P. (2014). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, 111(38), 13795–13798.
- Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, 19(5), 448–456.
- Köster, O., & Schiller, N. O. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, 4(1), 18-28.
- Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, 333(6042), 595–598.

# An exploration of the other-accent effect in Quebec and Hexagonal French.

*Julien Plante-Hébert<sup>1</sup> and Pamela Bautista-Boivin<sup>1</sup>*

<sup>1</sup> *Department of Linguistics, Université du Québec à Montréal, Montréal, Canada*  
 plante-hebert.julien@uqam.ca, bautista\_boivin.pamela@courrier.uqam.ca

## Introduction

Speaker recognition is sometimes necessary in a forensic context where visual means of identification are unavailable. It has been shown that the human ability to recognize and identify voices is impaired when the spoken language is not the native language of the listeners (Goggin et al., 1991; Perrachione, 2018; Philippon et al., 2007). Such an effect has also been reported, but not as consistently, for different varieties of a same language (Kerstholt et al., 2006; Stevenage et al., 2012; Yu et al., 2021). The objectives of the present study are to attest this other-accent effect (OAE) in different French varieties (Quebec French, QF and hexagonal French, HF) and to determine whether such an OAE is asymmetric in consideration of the more dominant of both varieties (HF) (Perrachione et al., 2010; Stevenage et al., 2012). Finally, we investigated the effect of the duration of the stays of HF speakers living in Quebec on their scores.

## Methods

Thirty-four adult participants were recruited for the present experiment. Eighteen were native QF speakers and 16 HF native speakers. HF participants all lived in Montreal at the time of testing and were divided into 2 groups according to the duration of their stay in Quebec: less than 5 years (n= 7) and 5 years and more (n= 9).

Two voice line-ups (one in FQ and one in FH) were created, each containing 1 target voice and 5 foils recorded by female volunteers. All volunteers were recorded reading a short neutral text easily understandable in both French varieties.

The volunteers also recorded a sample by reading a second and slightly longer text to be used as training material. The training recording of only one speaker for each French variety was selected for the training phase based on its perceived neutrality in terms of accent and pronunciation.

Participants were asked to listen to the training recording of a given French variety (randomized) before listening to the entire line-up of the corresponding accent. They then had to indicate which voice was that of the training sample. Participants were also asked to report their confidence level regarding each answer. The same procedure then followed with stimuli of the other variety.

## Results

Generalized mixed models were used to investigate the OAE on speaker recognition. Statistics showed no significant effect when comparing the results of participants when listening to the same or a different accent. Moreover, statistics also showed no significant asymmetry in relation to the dominant French variety. Finally, even though statistical analyses showed no significant effect of the duration of stay in Quebec for native French participants, a trend was observed, suggesting an increase in correct recognition for the QF voice with longer stays, while a decrease was observed for the HF voice was observed.

## Discussion

Our data suggests that the OAE is not to be systematically assumed and should be considered with precautions, especially when speaker recognition is used in a forensic or legal framework. The results

regarding the duration of the stay for HF participants, while not statistically significant, underline the importance of further research focusing specifically on the duration of stay effect on the OAE.

## References

- Goggin, J. P., Thompson, C. P., Strube, G. & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & Cognition*, 19(5), 448-458. <https://doi.org/10.3758/bf03199567>
- Kerstholt, J. H., Jansen, N. J. M., Amelsvoort, A. G. V. & Broeders, A. P. A. (2006). Earwitnesses: effects of accent, retention and telephone. *Applied Cognitive Psychology*, 20(2), 187-197. <https://doi.org/10.1002/acp.1175>
- Perrachione, T. K., Chiao, J. Y. & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition*, 114(1), 42-55. <https://doi.org/10.1016/j.cognition.2009.08.012>
- Perrachione, T.K. (2018). Speaker recognition across languages. In S. Frühholz & P. Belin (Eds.). *The Oxford Handbook of Voice Perception*. Oxford University Press. <https://open.bu.edu/handle/2144/23877>
- Philippon, A. C., Cherryman, J., Bull, R. & Vrij, A. (2007). Earwitness identification performance: the effect of language, target, deliberate strategies and indirect measures. *Applied Cognitive Psychology*, 21(4), 539-550. <https://doi.org/10.1002/acp.1296>
- Stevenage, S. V., Clarke, G. & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology*, 24(6), 647-653. <https://doi.org/10.1080/20445911.2012.675321>
- Yu, M. E., Schertz, J. & Johnson, E. K. (2021). The other accent effect in talker recognition: Now you see it, now you don't. *Cognitive Science*, 45(6), e12986. <https://doi.org/10.1111/cogs.12986>

# The effect of sample size on the evaluation of speaker discriminatory power of diphthong /ei/

Honglin Cao<sup>1</sup>, Xuehui Li<sup>2</sup>, Danlin Wang<sup>3</sup>

<sup>1</sup>Key Laboratory of Evidence Science (China University of Political Science and Law), China.

<sup>2</sup>Xinxing Development Group Co., Ltd. Beijing, China

<sup>3</sup>Department of Forensic Science, Beijing Police College, Beijing, China.

caohonglin@cupl.edu.cn

## Introduction

Formant frequency of vocalic sounds is a critical acoustic parameter in Forensic Voice Comparison (FVC) employing the auditory-acoustic phonetic methodology. Linear discriminant analysis (LDA) is widely used to quantify speaker discriminatory power in vowel formant analysis (Fairclough et al., 2023). However, a persistent methodological challenge lies in determining the minimum sample size (MSS) required to ensure robust and replicable classification rates (CRs). Previous studies investigating MSS for likelihood ratio (LR)-based systems suggested thresholds of 20 speakers (Hughes 2017) or 30 speakers (Hughes & Foulkes 2014) were needed to obtain relatively stable LR data. However, crucially, LR and LDA represent distinct statistical frameworks—open-set versus closed-set classifier—limiting the applicability of LR-derived MSS thresholds to LDA-based approaches. This study aims to fill this gap by empirically establishing MSS thresholds for LDA applications in FVC.

## Method

In this study, 660 male Mandarin speakers (aged 18-22 years) from the Beijing Police College were involved. They were asked to read a short passage (consisting 256 Chinese characters) 5 times, using a dynamic vocal microphone in a sound-attenuated room. The syllable “/pei/, 北” in the word “Beijing” appeared 5 times in the passage. For each speaker, all 25 repetitions (5×5) of the target diphthong /ei/ in “/pei/, 北” were analyzed. WaveSurfer was used for F1-F4 extraction. The formant contours were time-normalized into 10 intervals and fitted with a cubic polynomial. The coefficients of the polynomial were used as predictors in the LDA. Matlab was utilized to conduct LDA cross-validation via the “leave-one-out” method.

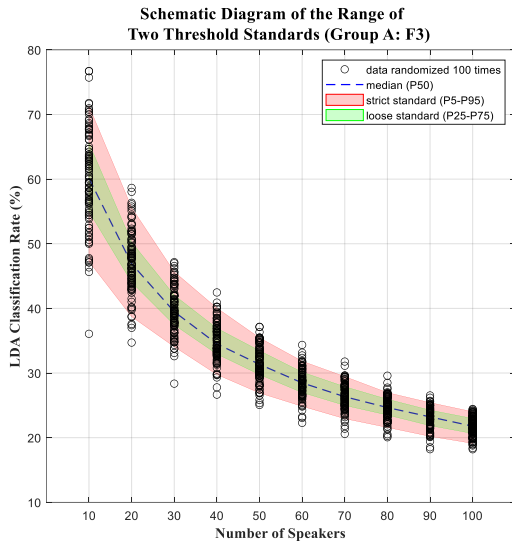
To verify the reliability and repeatability of the results, the 660 speakers were randomly divided into two independent groups (A/B, n=330 each). For each group, 33 subsets were generated with incremental sample sizes (ranging from 10 to 330 speakers, with 10-speaker increments). To mitigate order effects, the original order of speakers in each group was randomized 100 times. For each formant parameter in each group, the CRs were calculated under different sample-size conditions, resulting in a total of 3,300 calculations (33 subsets × 100 times).

Normality tests revealed that the CRs of some formant parameters did not follow a normal distribution, and outlier tests indicated that outliers were present in 43.8% of the measurements. Therefore, a specific percentile range was used to define the valid data interval, which is immune to the influence of outliers or extreme values. Two threshold standards, the “loose standard” and the “strict standard”, were established based on different percentile ranges. The “loose standard” is defined as the interquartile range (IQR) that encompasses 50% of the observed CRs. This range lies between the 25th percentile (P25) and the 75th percentile (P75) of the 100 CRs obtained from 100 random repeated calculations (see the green-shaded area in Figure 1). The red-shaded area in Figure 1 represents the “strict standard,” which includes 90% of the observed values between the 5th (P5) and 95th (P95) percentiles.

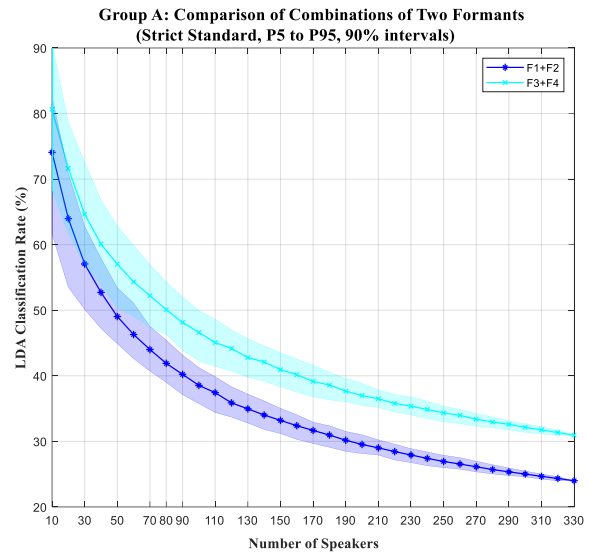
A pairwise comparison method was employed to determine the MSS. Under the two threshold standards, the overall distributions of the CRs of two formant parameters across the 33 subsets were compared. The MSS was defined as the smallest sample size at which the non-overlapping CR ranges (shaded area) stabilized across incremental subsets. Figure 2 depicts the comparison of the CRs of the F1 + F2 combination and the F3 + F4 combination using the strict standard range, indicating that the MSS is 80.

**Result**

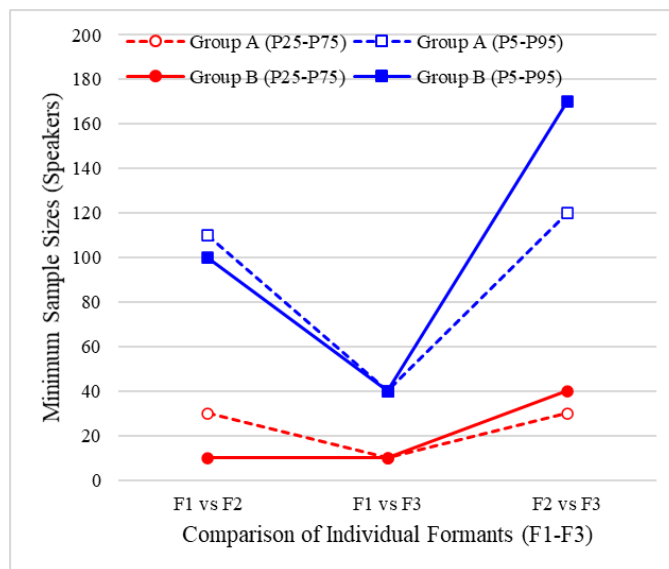
The results showed that when F4 was not taken into account, at least 40 speakers were needed to stably reflect the relative values of speaker discriminatory power among the individual formants F1 - F3 (see Figure 3). When analyzing combinations of F1 - F3 or considering F4, the MSS increased significantly, but it was difficult to provide an exact number.



**Figure 1.** A schematic diagram of the range of the loose and strict threshold intervals



**Figure 2.** Comparison of combinations of two formants under the strict standard in Group A



**Figure 3.** MSS when comparing individual formants (F1-F3)

**References**

Fairclough, L., Brown, G. & Kirchhübel, C. (2023). Reviewing the performance of formants for Forensic Voice Comparison: a meta-analysis of forensic speech science research. *Proceedings of the 20th ICPhS*, Prague, Czech.

Hughes, V. & Foulkes, P. (2014). Variability in analyst decisions during the computation of numerical likelihood ratios. *International Journal of Speech Language and the Law*, 21(2): 279-315.

Hughes, V. (2017). Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough? *Speech Communication*, 94: 15-29.

# New Conclusion Framework for the Forensic Speaker Recognition methodology in the Hungarian Audio Forensics

*Attila Fejes*

*Special Service for National Security, Institute for Expert Services, Budapest, Hungary.*

fejes.attila@nbsz.gov.hu

How the partial results of the Forensic Speaker Recognition (FSR) analysis are summarised is crucial for the correctness of the final statement in the expert report (Drygajlo et al., 2015). In the Hungarian FSR methodology, the results of the methodological elements were not scored separately in the past, but the results of the acoustic-phonetic (ACPA) and auditory-phonetic (AUPA) analysis were combined and merged with the Automatic Speaker Recognition (ASR) measurements using a matrix. In order to avoid cognitive bias (Czebe & Kovács, 2016) and to increase objectivity, a new framework was created to describe in more detail the methods of analysis and measurement on the audio recordings and the conclusions drawn by the expert, while the expert remains in control of the identification process.

In the new conclusion framework, the results of the three methodological elements (ACPA, AUPA, ASR) of the assessment are evaluated separately on a eleven-point scale, ranging from minus five to plus five points. In the case of the ASR (Drygajlo & Haraksim, 2017), the arithmetic mean of the LR values (Craig, 2010) (Meester & Slooten (2021) of the systems gives the result. If the data produced by the systems differ significantly, a detailed justification by the expert is required in the expert report. Under the analysis the minus five represents the difference between the speakers, while the plus five represents their similarity with the highest probability. The three scores are added together and is determined the verbal expert statement.

To create the framework, the FSR analysis under a blind test were made on speech samples from 136 female and the same number of male speakers, recorded in realistic forensic-like test recordings (Kelly et al, 2019) (Morrison & Enzinger 2016). The samples were compared using FSR methods according to a defined procedure and conclusions were drawn from their numerical and verbal results. It has been established that even when using new software in AUPA and ACPA analyses, it is not possible to make conclusions that are independent of the expert in contrast to the ASR measurements, which are independent of the expert when using the same test and reference audio data. However, ASR results are independent of the expert. During ACPA analyses using the OTExpert software (<https://ot-contact.com/en/>) were pitch measurements, comparisons of spectrographic images were performed, LTA and LPC curve analyses were applied, and the values and characteristics of formants were measured.

Measurements, empirical methods and logical methods were used to determine the principles for the development of the rating scale. For the acoustic-phonetic analysis, was separated the audio forensics expert dependent procedures from the semi-automatic measurements (pitch analyses), and for the biometric measurements, and was analysed the probability distribution of the identification system scores using five different systems (Batvox 3.1, Batvox 4.1, Nuance Forensics 12.2.0, Phonexia 3.16, VOCALISE2021) and a Hungarian ASR pilot software (PyForVoice) was developed Budapest University of Technology and Economics (Sztaho & Fejes, 2021).

## References

- Craig, A. (2010). Essential Mathematics and Statistics for Forensic Science. John Wiley & Sons, Oxford.
- Czebe, A., Kovacs, G. (2016). How Cognitive Infocommunications Play a Critical Role in Shaping the Future of Forensic Sciences: Defining Forensic Cognitive Infocommunications. Proceedings of 7th IEEE Conference on Cognitive Infocommunications (COGINFOCOM) IEEE Hungary Section., (pp. 283-287.).

- Drygajlo, A., Haraksim, A. (2017). Biometric Evidence in Forensic Automatic Speaker Recognition, *Handbook of Biometrics for Forensic Science*, Springer International Publishing, Cham. (pp. 221-229.).
- Drygajlo, A., Jessen, M., Gfroerer, S., Wagner, I., Vermeulen, J., Niemi, T. (2015). Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition. European Network of Forensic Science Institute Forensic Speech and Audio Analysis Working Group (pp. 42-46.).
- Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S., Alexander, A. (2019). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic\_eval\_01). *Speech Communication.*, Vol: 112, 30-36.
- Meester, R., Slooten, K. (2021). *Probability and Forensic Evidence*, Cambridge University Press. (pp. 30-35.).
- Morrison, G., Enzinger, E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic\_eval\_01) – Introduction, *Speech Communication*, Vol: 85, 2016. (pp. 119-126.).
- Sztaho, D., Fejes, A. (2023). Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings, *Journal of Forensic Sciences* Vol: 68. (pp. 871-883.).

# Speech discernment using signal analysis technology

*Terese Anderson<sup>1</sup> and Grandon Goertz<sup>2</sup>*

<sup>1</sup> *Office of Research Safety, University of Chicago, Chicago, Illinois*  
b13@uchicago.edu

<sup>2</sup> *Educational Linguistics, University of New Mexico, Albuquerque, New Mexico*  
sordfish@unm.edu

## Introduction

This paper reports on the progress made in speech analysis technology that compares speakers by the lag of speech cross-correlations. This research uses the mathematical measurements of the speech sound from the speaker as recorded by the microphone.

## Research method

Speech is represented by numerical listing of the frequencies that a microphone responds to (Quatieri, 2002). Individual speech sounds may be analyzed using *signal similarities technology*, (SST) and employing the Matlab Signal Processing toolbox (The MathWorks, 2022).

The Matlab program determines both the frequencies of the speech sample and the frequency of the speech pitch. The pitch frequency value is used to create a sine wave which acts as a benchmark. The pitch represents the fundamental frequency of the person's voice (Gold & Morgan, 2000, Hardcastle, Laver, & Gibbon, 2012).

This research compares the speech frequency waveform to its benchmark sine wave waveform, to produce detailed graphic depictions. Cross-correlating the speech frequency waveform to its pitch waveform acts as a type of autocorrelation (Mansali, et al. 2022; McLoughlin, 2016). The cross-correlation values are plotted and the plots show how much and in which direction the signal lags, which is the difference between the speech signal and its sine wave. (Bourke, 1996, Quatieri, 2002).

## Procedure

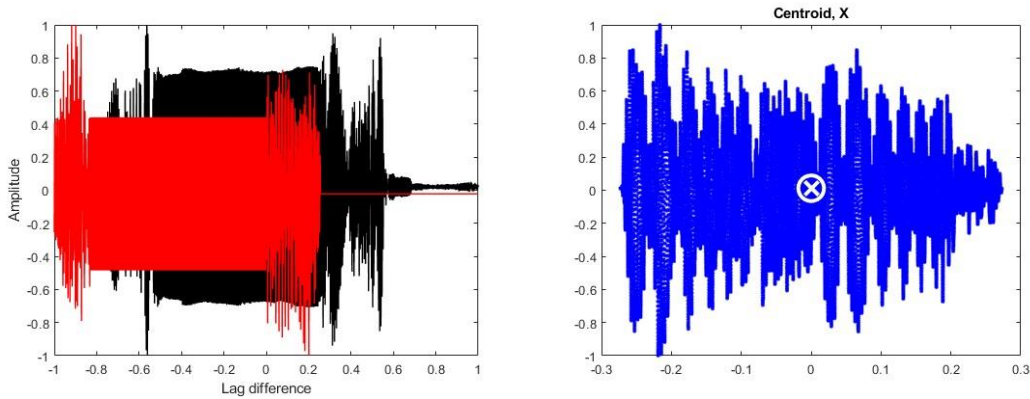
Computations were made using whole-word samples that were obtained from six English speakers who read word lists in a sound booth and were recorded as .wav files. Each participant produced 61 single-syllable words with vowels representing the vowel space corners, diphthongs, and central vowels. Each recorded vowel list was then cloned, producing a cloned list of each speaker.

The individual sounds and their clones were processed using our modified version of the Matlab 2022b Signal Processing Toolbox. The Toolbox calculates signal similarities and compares the frequency content of two signals. This experiment shows the cross-correlation between the speech signal as recognized by the microphone and the sine wave that represents the speech signal's fundamental pitch. We compared naturally spoken word to their pitch and to their respective clones. We also compared natural speech to clones of the same words.

Signal cross-correlation similarities are represented in lag, which is the plot of one signal has advanced or retarded compared to a standard, in this case the pitch. The difference in the location of the plots was shown by calculating the geometric mean of the wave.

## Findings

Uniqueness in each speech signal is observed by speech lag, compared to the sine value of each speech sample's pitch. Cloned speech samples show that they lag further in the negative direction, compared to the un-cloned speech sample. For comparison purposes, a k-means clustering function was used to determine the optimal centroid of each signal. The distance between the centroids was used as a metric for lag differential determination, and the results are shown in Chart 1. It appears that each speaker has distinctive lag values and clone lag values. When comparisons are made between speakers saying the same word, lag differences are noted.



Plot 1 (left). The word 'bite'. Natural speech shown in black and cloned speech in red illustrating lag shift. Plot 2 (right). Example of a female speaker saying 'buy', and centroid marked by x.

	Natural	Clone	Distance
	speech lag	speech	between centroids
		lag	
bad	-0.02802048		0.027019198
clone		-0.001	
ban	0.001077483		0.018042886
clone		-0.01697	
bed	0.010758773		0.032536891
clone		0.043296	
bee	-0.01015767		0.015374166
clone		0.005216	
ben	-0.00507966		0.0026029
clone		-0.00248	
bird	0.001552911		0.000860365
clone		0.002413	
bite	0.020309866		0.041940045
clone		-0.02163	
boat	0.006872466		0.014101735
clone		0.020974	
book	-0.00560907		0.014668924
clone		-0.02028	

Chart 1. A sample of the data sheet showing lag values for a male speaker.

## References

- Bourke, P. (1996). Cross correlation. *Cross Correlation”, Auto Correlation—2D Pattern Identification*, 596.
- Gold, B. & Morgan, N. (2000). *Speech and Audio Signal Processing*. New York, John Wiley and Sons.
- Hardcastle, W. J., Laver, J., & Gibbon, F. E. (Eds.). (2012). *The handbook of phonetic sciences*. New York, John Wiley & Sons.
- Measure Signal Similarities. Signal Processing Toolbox. (2025). Matlab.  
<https://www.mathworks.com/help/signal/ug/measuring-signal-similarities.html>.
- Mansali, M., Ramos, D., S. Kadiri, & P. Alku. (2022). *Introduction to Speech Processing*, 2nd Edition. URL:  
<https://speechprocessingbook.aalto.fi>. DOI: 10.5281/zenodo.6821775.
- McLoughlin, I. (2016). *Speech and Audio Processing: A Matlab®-based Approach*. Cambridge, Cambridge University Press.
- The MathWorks Inc. (2022). MATLAB version: 9.13.0 (R2022b), Natick, Massachusetts: The MathWorks Inc.
- Quatieri, T. (2002). *Discrete-Time Signal Processing*. Upper Saddle River, NJ, Prentice Hall

# Authorship analysis on speech data as a counter to AI voice cloning

Anneke Visser

*Institute for Forensic Linguistics, Aston University, Birmingham, UK*  
aviss24@aston.ac.uk

Speaker comparison is very well established in the field of Forensic Speech Science and the advancements of automatic systems have been particularly impressive in recent years (Kelly et al., 2019). A considerable problem facing the application of these approaches, however, is the ever-increasing popularity and quality of deepfakes/voice cloning. An acoustic/phonetic approach may no longer be viable when the audio is fake. Recent attempts to combat this trend have focused on the detection of deepfakes, with real success (Shanks, 2024). But these methods only address the detection of fake audio, not the identification of the person responsible. We, therefore, need an approach to speaker comparison that is resistant to the changes that the deepfake process can make to a voice. As the linguistic content of the speech is not altered by this process, authorship analysis methods could be a solution to this problem. Hence, this study investigates the speaker discriminatory power of authorship analysis techniques on transcriptions of speech data. Though there have been past explorations of this approach (Doddington, 2001; Kredens, 2002), it has only recently seen a resurgence in interest (Brown et al., 2024; Sergidou et al., 2023; Tompkinson & Nini, 2024). A novel aspect of this research is that it will be contextualised within the process of creating text-to-speech deepfake audio. This means that, unlike prior investigations, both speech and written text will be included in the analysis.

We propose to collect experimental data which includes multiple instances of spontaneous speech data and one instance of written data that specifically emulates a text-to-speech process. The aim is to have data from around 100 speakers. The speech data will then be transcribed using various approaches, both manual and automatic. This will enable scrutiny of the impact of the transcription process and its conventions on the outcomes of the authorship analysis. For example, authorship analysis has previously been shown to be very successful at using punctuation to distinguish between authors (Grieve, 2007). This does not work for speech because punctuation is not inherent to the spoken medium and any punctuation that is present in a transcript will have been imposed by a transcriber. Finally, an word n-gram based authorship analysis approach will be used across the transcripts for all the speakers. The success of classification based on these features will be determined and the linguistic underpinnings of any features that prove particularly speaker-discriminatory will be investigated. Our findings will then be considered in relation to existing theories of idiolect and the linguistic individual. In the event that authorship analysis is successful when applied to speech data, attribution of a deepfake recording to a potential ‘author’ by this method should be achievable.

## References

- Brown, G., Nini, A., & Kirchhübel, C. (2024). *Likelihood ratio-based authorship verification methods applied to forensic voice comparison tasks*. 5th European Conference of the IAFL: Applying Linguistics, Improving Justice., Birmingham. <https://doi.org/10.5281/zenodo.12688496>
- Doddington, G. (2001). Speaker recognition based on idiolectal differences between speakers. *7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2521–2524. <https://doi.org/10.21437/Eurospeech.2001-417>
- Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, 22(3), 251–270. <https://doi.org/10.1093/lc/fqm020>
- Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). *Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors*. Audio Engineering Society International Conference on Audio Forensics, Porto, Portugal.

- Kredens, K. (2002). Towards a corpus-based methodology of forensic authorship attribution: A comparative study of two idiolects. In P. Lang & B. Lewandowska-Tomaszczyk (Eds.), *Practical Applications in Language Corpora*.
- Sergidou, E.-K., Scheijen, N., Leegwater, J., Cambier-Langeveld, T., & Bosma, W. (2023). Frequent-words analysis for forensic speaker comparison. *Speech Communication*, 150, 1–8.  
<https://doi.org/10.1016/j.specom.2023.03.010>
- Shanks, K. (2024, July 30). Innovative solutions unveiled at the Deepfake Detection Challenge Showcase. *Accelerated Capability Environment*. <https://ace.blog.gov.uk/2024/07/30/innovative-solutions-unveiled-at-the-deepfake-detection-challenge-showcase/>
- Tompkinson, J., & Nini, A. (2024). *Evaluating the usefulness of embedding phonetic representations into an authorship analysis-based framework for the comparison of spoken data*. 5th European Conference of the IAFLL: Applying Linguistics, Improving Justice., Birmingham.

# A Pilot Acoustic Study of Mandarin /ʃ/ in Southern Min Speakers: Implications for Forensic Speaker Comparison

Jinjin Lin<sup>1</sup>, Paul Foulkes<sup>1,2</sup>, and Vincent Hughes<sup>1,2</sup>

<sup>1</sup>*Department of Language and Linguistic Science, University of York, York, UK*  
bdh541@york.ac.uk

<sup>2</sup>*Forensic Speech Services, University of York, York, UK*

This pilot acoustic study investigates the realization of the Mandarin sibilant /ʃ/ by native Southern Min speakers. This phoneme is contrastive with /s/ in Standard Mandarin but they are often reported to merge in regional varieties. The aim of this study is to examine whether speakers with a Southern Min background maintain an acoustic distinction between /s/ and /ʃ/, using spectral analysis as the primary method, and to examine the value of these segments as speaker discriminants. Furthermore, this work aims to contribute to the small body of forensic phonetic work on languages of China, and specifically Southern Min (e.g., Chang & Shih, 2012; Goh et al., 2024; Svantesson, 1986). Twenty male native speakers of Southern Min, aged 18 to 24, participated in the study. Each speaker took part in two different recordings, allowing for within-speaker comparison across separate sessions. Each participant read a script of a mock fraud conversation twice, with each session lasting approximately 5 minutes. The recordings were made using Xiaomi 10 smartphones via regular phone calls. The sample rate was 44.1 k Hz and all recordings were saved in MP3 format. Due to the phone's built-in noise reduction features, the resulting spectrograms showed some patchiness, although the fricative segments remained analyzable. The target words were "身" (/ʃən/) and "是" (/ʃɿ/), both containing the retroflex /ʃ/. These words were selected because they occurred frequently across the scripted conversations, making it easier to identify and extract them reliably for analysis.

They are picked up from the mock fraud recordings. These words were elicited in controlled environments and embedded in carrier sentences. The acoustic analysis focused on the fricative segments of the two target words. Each fricative was segmented using Praat, and spectral moments—including center of gravity, standard deviation, skewness, and kurtosis—were measured using a script (Rentz, 2017) to assess their acoustic properties.

For "是", the mean COG was 5994 Hz in Session 1 and 6053 Hz in Session 2. For "身", the mean COG was 6071 Hz and 6111 Hz. Other spectral parameters showed similar patterns: standard deviation values ranged from 4329 to 4503 Hz, skewness from 1.88 to 1.94, and kurtosis from 4.17 to 4.56 across sessions. No statistically significant differences were found between sessions for either word ("是":  $p = .516$ ; "身":  $p = .479$ ), indicating overall acoustic stability. These results suggest that most participants did not produce a consistent acoustic contrast between /s/ and /ʃ/, and that their fricative realizations remained stable across sessions. The substantial overlap in spectral properties supports the interpretation that the contrast may be weakened or neutralized in the Mandarin of Southern Min speakers. However, between-speaker variability was observed in the degree of spectral overlap between /s/ and /ʃ/. While some speakers produced clearly distinct fricative categories, others showed substantial overlap, suggesting that individual differences may affect the extent of category separation. This variability may be useful for assessing speaker discriminatory potential in forensic or sociophonetic contexts. Likelihood ratio-based testing is currently underway to further assess the strength of acoustic separation between categories at both group and individual levels.

## References

- Chang, Y. H., & Shih, C. (2012). Using map tasks to investigate the effect of contrastive focus on the Mandarin alveolar-retroflex contrast. *Proceedings of Speech Prosody 2012*.
- Chang, Y. H. S., & Shih, C. (2015). Place contrast enhancement: The case of the alveolar and retroflex sibilant production in two dialects of Mandarin. *Journal of Phonetics*, 50, 52-66.
- Goh, H. L., Woon, F. T., Moisk, S. R., & Styles, S. J. (2024). Contrastive alveolar/retroflex phonemes in Singapore Mandarin bilinguals: Comprehension rates for articulations in different accents, and acoustic analysis of productions. *Language and Speech*, 67(4), 924-944.
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America*, 108(3), 1252–1263. <https://doi.org/10.1121/1.1288413>
- Keith, E., & Kinoshita, Y. (2024). Sub-band parametric cepstral distance measurement of voiceless alveolar fricative segments as a tool for identifying speaker-characteristic information robust to emotional variation. *The International Journal of Speech, Language and the Law*, 31(2), 267-290.
- Künzel, H. J. (2001). Beware of the “telephone effect”: The influence of telephone transmission on the measurement of formant frequencies. *International Journal of Speech, Language and the Law*, 8(1), 80–99. <https://doi.org/10.1558/sll.2001.8.1.80>
- Rentz, B. (2017, January 14). Praat script for measuring spectral moments and duration [Praat script]. [https://github.com/rentzb/praat-scripts/blob/master/spectral\\_moments.praat](https://github.com/rentzb/praat-scripts/blob/master/spectral_moments.praat)
- Stevens, K. N. (1998). *Acoustic Phonetics*. MIT Press.
- Stuart-Smith, J., Sonderegger, M., & Turk, A. (2003). The interaction of speech production and perception in socially-indexed variation: Evidence from Glasgow English. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1851–1854.
- Svantesson, J. O. (1986). Acoustic analysis of Chinese fricatives and affricates. *Working papers/Lund University, Department of Linguistics and Phonetics*, 25.
- Turk, A. E., Nakai, S., & Sugahara, M. (2006). Acoustic segment durations in prosodic research: A practical guide. In S.-A. Jun (Ed.), *Prosodic Typology: The Phonology of Intonation and Phrasing* (pp. 1–28). Oxford University Press.

# Accent Copycats: How Accurately Can Non-Linguists Mimic a Known Voice?

*Amy Cope, Ben Gibb-Reid*

<sup>1</sup>*Department of Language and Linguistic Science, University of York*  
ac2796@york.ac.uk

<sup>2</sup>*Department of Language and Linguistic Science, University of York*  
ben.gibb-reid@york.ac.uk

Mimicry is the act of impersonating another individual's voice- be that for the purpose of comedy, storytelling or for fraudulent intention. This study is concerned with mimicry from a forensic phonetic position, specifically investigating how accurately non-linguists can mimic a known voice. This is assessed via a lay-listener perceptual identification task and via acoustic comparisons between real and mimicked voices.

The focus of this study is to assess to what degree mimicry in forensic contexts can lead to difficulty in forensic voice comparisons. Multiple production studies involving professional imitators indicate that the acoustic proximity of the imitation to target was minimal, and that formant frequencies were by far the least replicable feature (compared to F0 and speech rate, Eriksson & Wretling, 1997; Zetterholm, 2003). In terms of perception of natural voices, Foulkes and Barron (2000) found that, even with speakers who are highly familiar to each other, voice recognition is not highly accurate. Rather, distinctiveness in pitch/regional accent was a large determining factor. My study is tailored to be more applicable to forensic contexts, in that the participants are untrained in mimicry, and recordings were made across phone bandwidth using voicemail messages.

The two speakers were both male close friends, aged 20-21, with no prior linguistic teaching. The first, (referred to as 'Luke') has a West Midlands accent; the other (referred to as 'Andy') has a Manchester accent with elements of Adoptive RP (Wells, 1982). Each participant was recorded doing a map task in their natural speaking voice with the interviewer. Then, Andy was provided with the recording of Luke's map task and vice-versa. Further recordings were made of each participants' mimic attempt. To assess the accuracy of the mimicry, an auditory acoustic analysis was undertaken of the recordings. F1 and F2 measures were extracted from the midpoint of vowels FLEECE, GOOSE, FOOT and BATH. Mean F0 was also measured. A perception study was also undertaken with seven familiar listeners (five friends as well as Luke and Andy themselves). A Qualtrics survey asked these listeners to respond to recordings of real and mimicked speech from both Luke and Andy - assessing whether the recording was either Luke's natural voice, Andy's natural voice, Luke mimicking Andy or Andy mimicking Luke.

The results largely agreed with that of Eriksson and Wretling (1997), as the formants showed some difference between the mimicker's real and mimicking voice, but very little accuracy towards the target. Perception results revealed that a small number of participants were fooled by the mimic. Similar to Foulkes and Barron's (2000) results, Luke managed to misidentify his own voice. The results indicate that lay listeners are generally not accurate in identifying even highly familiar voices.

## References

- Eriksson, A., & Pär Wretling. (1997). How flexible is the human voice? a case study of mimicry.
- Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *Forensic Linguistics*, 7(2), 180–198.
- Wells. (1982). *Accents of English. / 2, The British isles*. Cambridge University Press.
- Zetterholm, E. (2003). *Voice Imitation: A Phonetic Study of Perceptual Illusions and Acoustic Success* [PhD Thesis]

# Exploring the interpretability of deep speaker-representations from a phonetic perspective

Guangmou Deng<sup>1</sup>

<sup>1</sup> *Department of Language and Linguistic Science, University of York, UK*  
lsm577@york.ac.uk

Deep Neural Network (DNN)-based architectures (e.g., x-vector/r-vector PLDA baseline) have become increasingly prominent as the state-of-the-art in forensic automatic speaker recognition (FASR) systems, outperforming previous techniques such as i-vectors. DNN-based systems use embeddings from the layer within the DNN as the speaker model. These embeddings are compact representations of a speaker’s speech generated from front-end acoustic features (e.g., MFCC or Fbank), which are then used for probabilistic interpretation. However, FASR systems are often criticized for their limited transparency and interpretability, especially in forensic contexts where methodological clarity is crucial for the accessibility of evidence (H. Wang & Zhang, 2015). A key interpretability challenge lies in understanding what phonetic information is actually encoded in speaker embeddings. Both long-term acoustic features (LTAFs) and DNN embeddings are holistic speaker-specific representations derived from variable-length and continuous speech. Recent studies (Xu et al., 2023) suggest that mismatches in LTAFs can influence DNN-based FASR outputs. This offers a promising analytical paradigm: by mapping the relationship between LTAFs and speaker embedding, we may uncover the phonetic dimensions that underlie DNN speaker representations.

This project aims to better understand the interpretability of DNN speaker embeddings from a phonetic perspective. By modelling the relationship between LTAFs and speaker embeddings, the project seeks to provide insight into what types of phonetic information are encoded in DNN pipelines used in state-of-the-art FASR systems. The broader goal is to establish an analytical framework that may help profile FASR systems under different conditions and contribute to forensic practice. For example, by supporting the pre-screening of ASR in cases where behavioral or technical mismatches are likely to lead to unreliable results, or by identifying acoustic-phonetic features that are poorly encoded and may be candidates for the further system fusion. In addition, the project considers how different DNN architectures and training datasets influence the nature of speaker representations, with a view to better understanding the variability across DNN pipelines.

The study will use the *Standard Chinese: 68 Female Speakers* dataset (Zhang & Morrison, 2011) as test data, which consists of Mandarin recordings from 68 female speakers across multiple sessions and speaking styles. Long-term acoustic features, including formant frequencies (F1-F3), bandwidths (B1-B3), fundamental frequency (F0), and voice quality measures (e.g., jitter, shimmer, HNR), will be extracted. Speaker embeddings will be obtained from several pre-trained DNN models implemented in the *Wespeaker* toolkit (S. Wang et al., 2024), covering different architectures and training datasets (see **Table 1**).

Baseline architecture	Training datasets	Embedding length
ResNet34	<i>CNCeleb</i>	256
ResNet34	<i>VoxCeleb</i>	256
ResNet152	<i>VoxCeleb</i>	256
ResNet221	<i>VoxCeleb</i>	256
ResNet293	<i>VoxCeleb</i>	256
ECAPA-TDNN (512 channels)	<i>VoxCeleb</i>	192
ECAPA-TDNN (1024 channels)	<i>VoxCeleb</i>	192

**Table 1.** DNN pipelines for comparison

Inspired by Huckvale (2025), we plan to apply Principal Component Analysis (PCA) to retaining the top  $N$  principal components that explain  $\geq 90\%$  of the embedding variance. The PCA components will then be used as predictors in regression analyses, with the corresponding LTAFs (mean and standard deviation statistics) of each recording as dependent variables. The predictive performance of the regression models will be interpreted as an approximate indicator of how much the pipeline encodes the respective phonetic feature, thereby revealing the encoding patterns of the DNN pipeline.

## References

- Huckvale, M. (2025, January 23). Understanding dimensions of speaker variation found in large corpora.
- Wang, H., & Zhang, C. (2015). Forensic Automatic Speaker Recognition Based on Likelihood Ratio Using Acoustic-phonetic Features Measured Automatically. *Journal of Forensic Science and Medicine*, 1(2), 119.
- Wang, S., Chen, Z., Han, B., Wang, H., Liang, C., Zhang, B., Xiang, X., Ding, W., Rohdin, J., Silnova, A., Qian, Y., & Li, H. (2024). Advancing speaker embedding learning: Wespeaker toolkit for research and production. *Speech Communication*, 162, 103104.
- Xu, C., Foulkes, P., Harrison, P., Hughes, V., Welch, P., Wormald, J., Kelly, F., & Vloed, D. van der. (2023, July 9). Impact of mismatches in long-term acoustic features on different-speaker ASR scores. 31st International Association for Forensic Phonetics and Acoustics Conference, Zurich, Switzerland.
- Zhang, C., & Morrison, G. (2011). Forensic database of audio recordings of 68 female speakers of Standard Chinese [Dataset]. <http://databases.forensic-voice-comparison.net/>

# DSER: Dialog-Structure-Aware Metric for Speaker Diarization Evaluation in Forensics

Tim Fries<sup>1</sup>, David Grünert<sup>1,2</sup>, Alexandre de Spindler<sup>1</sup>, and Volker Dellwo<sup>2</sup>

<sup>1</sup>ZHAW Zurich University of Applied Sciences, Switzerland

<sup>2</sup>Department of Computational Linguistics, University of Zurich, Switzerland

david.gruenert@zhaw.ch, volker.dellwo@uzh.ch

Automatic speaker diarization can be used to segment and classify recordings in forensic investigations. In (Grünert, 2023) we showed that relevant parts of a recording can be found by analyzing communication structures extracted by speaker diarization systems. However, today’s evaluation metrics such as the diarization error rate (DER) (Fiscus, 2006) and the Jaccard error rate (JER) (Ryant, 2019) are not adequate for such applications. In (Grünert, 2023) we have shown that incorrectly recognized communication structures often have little effect on the error rate. More recent metrics such as the segment-level error rate (SER), the balanced error rate (BER) (Liu, 2022), and the communication diarization error rate (CDER) (Cheng, 2022) overcome some of these limitations. However, we show in this work that recognizing incorrect communication structures is still an issue and we present our ideas for a new, dialog-structure-aware metric (DSER).

## Prime-Based Encoding of Speaker Activity

At the core of DSER is a symbolic encoding of speaker activity using prime numbers. Each speaker is assigned a unique prime number. For every frame in time, the set of active speakers is represented as the product of their corresponding primes. This encoding is compact and particularly effective in representing overlapping speech. Formally, for a speaker  $S_i$  assigned prime  $P(S_i)$ , and a set of active speakers  $S_t$  at time  $t$ , the encoded value  $e_t$  is given by:

$$e_t = \prod_{s_i \in S_t} P(s_i)$$

## Overlap-aware Levenshtein Distance for Structure Matching

To compare speaker structures between reference and hypothesis, the DSER employs a modified version of the Levenshtein distance (Levenshtein, 1966). This metric is commonly used in string alignment tasks and calculates the minimal number of operations (insertions, deletions and substitutions) required to transform one sequence into another. As in the original Levenshtein distance, insertions and deletions are assigned a fixed cost of 1. To evaluate substitutions, the DSER uses a Jaccard-based cost function. For every substitution between the ground truth  $e^{ref}$  and the diarization result  $e^{hyp}$ , the prime values are factorized into sets of active speakers, denoted as  $f^{ref}$  and  $f^{hyp}$ . The speaker mismatch is quantified using the Jaccard distance:

$$\text{cost}(e^{ref}, e^{hyp}) = 1 - \frac{|f^{ref} \cap f^{hyp}|}{|f^{ref} \cup f^{hyp}|}$$

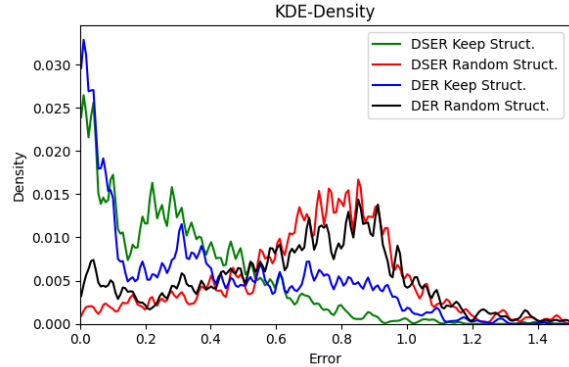
The new, overlap-aware Levenshtein Distance with Jaccard-based substitution cost can recursively be defined as  $L(E^{ref}, E^{hyp})$  with  $E^{ref} = [e_1^{ref}, \dots, e_n^{ref}]$  and  $E^{hyp} = [e_1^{hyp}, \dots, e_n^{hyp}]$ . To obtain the final DSER, the score is normalized to the length of  $E^{ref}$ .

$$L(E^{ref}, E^{hyp}) = \begin{cases} \text{len}(E^{ref}) & \text{if } \text{len}(E^{hyp}) = 0, \\ \text{len}(E^{hyp}) & \text{if } \text{len}(E^{ref}) = 0, \\ L(\text{tail}(E^{ref}), \text{tail}(E^{hyp})) & \text{if } \text{head}(E^{ref}) = \text{head}(E^{hyp}), \\ \min \begin{cases} L(\text{tail}(E^{ref}), E^{hyp}) + 1 \\ L(E^{ref}, \text{tail}(E^{hyp})) + 1 \\ L(\text{tail}(E^{ref}), \text{tail}(E^{hyp})) + \text{cost}(\text{head}(E^{ref}), \text{head}(E^{hyp})) \end{cases} & \text{otherwise} \end{cases}$$

## Metric Evaluation

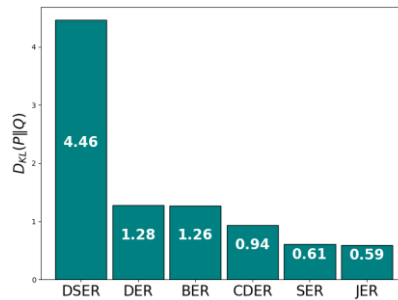
To evaluate the structural sensitivity of the DSER metric, we simulate realistic diarization errors and assess whether the dialog structure is preserved. From the VoxConvers dataset (Chung, 2022) we generated 2761 RTTM files with diarization errors, from which 1480 had similar communication structures and 1281 had random structures. Figure 1 shows the average scores across all runs in every regime and the KDE density estimates.

Metric	Keep Str.	Random Str.
DSER	0.289	0.714
DER	0.377	0.671
JER	0.412	0.671
CDER	0.568	1.729
SER	0.467	0.749
BER	0.486	0.813



**Figure 1.** Left: Average mean values for structure-preserving versus random manipulation. Right: KDE density estimates under the Keep Structure and Random Structure for DSER and DER.

Finally, we quantified how cleanly every metric separates the two regimes by calculating the Kullback-Leibler divergence between their score distributions (Figure 2). The DSER achieved the highest KL scores ( $\approx 4.5$ ), above DER and JER (at  $\approx 1.3$ ), well above SER and BER (round 0.6) and CDER ( $\approx 0.9$ ). This confirms that DSER delivers the strongest distinction between preserved and disrupted dialog structures.



**Figure 2.** KL-Divergence between Keep Structure and Random Structure.

Our evaluation shows that DSER performs better in separating correct from incorrect dialog structures than all other error rates. DSER achieves an ideal balance: it is sensitive enough to detect meaningful structural changes, yet robust against minor timing shifts.

## References

- Cheng G, et al (2022). Dataset, evaluation metric and baselines. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP), pages 488–492. IEEE.
- Chung, Joon Son, et al. (2020). Spot the conversation: speaker diarisation in the wild, INTERSPEECH, <https://www.robots.ox.ac.uk/~vgg/data/voxconverse/>
- Fiscus, J. G. et al. (2006). The rich transcription 2006 spring meeting recognition evaluation. Proceedings of International Workshop on Machine Learning and Multimodal Interaction, pp. 309–322.
- Grünert D. et al (2023). Speaker Diarization Systems in the Context of Forensic Audio Analysis. The 31rd Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)
- Levenshtein, V. I. (1966), Binary Codes Capable of Correcting Deletions, Insertions and Reversals. Soviet Physics Doklady, Vol. 10, p.707. Bibcode 1966SPhD...10..707L
- Ryant, N. et al. (2019). The second DIHARD diarization challenge. Proceedings of the Annual Conference of the International Speech Communication Association, pp. 978–982.
- Tao Liu et al. (2022). Ber: Balanced error rate for speaker diarization. arXiv preprint arXiv:2211.04304.

# Interaction of linguistic contrast and speaker specificity: an investigation into F3 and vowel rounding

Annie Baker<sup>1</sup>, Eleanor Chodroff<sup>1</sup>

<sup>1</sup>*Department of Computational Linguistics, University of Zurich, Switzerland*  
 annie.baker@uzh.ch eleanor.chodroff@uzh.ch

## Introduction

F3 has been reported as a relatively speaker-specific acoustic feature, relevant for forensic applications such as voice comparison (McDougall, 2004; Bosch, 2003; Gold & French, 2011), with the additional benefit of typically falling within the frequency range of telephone call recordings (Jessen, 2008). F3 is affected by coarticulation from neighbouring segments, particularly those with a velar place of articulation, as well as, crucially for this study, vowel rounding.

The aim of this analysis is to identify possible presence and degree of speaker specificity of F3, both within and across languages, focusing on the potential effect of a language-specific linguistic contrast (rounded/unrounded vowel distinction) in which F3 plays an important role. With an initial focus on high back vowels, three languages with differing vowel inventories were selected: Turkish (/u, u/), Ukrainian (/u/), and Japanese (/u/). Independent comparisons of Ukrainian and Turkish (rounded vowel /u/), and Japanese and Turkish (unrounded vowel /u/) were carried out, in addition to within-language comparisons of F2 and F3. The predicted outcome was that languages without a vowel rounding contrast would have more ‘freedom’ in their use of F3 and therefore show higher levels of speaker specificity.

## Methods

The data originates from the VoxCommunis corpus (Zhang et al., 2025) consisting of read speech from the validated section of the Mozilla Common Voice crowdsourced speech corpus (Ardila et al., 2020), as well as TextGrids containing time-aligned phonetic transcriptions generated using the Montreal Forced Aligner (McAuliffe et al., 2017). For each language, F2 and F3 values were extracted from speech samples of 200 speakers using Praat (Boersma & Weenink, 2024). Formant frequencies were originally measured in Hertz and subsequently transformed to the BARK scale. With the intention of avoiding potential influence of coarticulation, values were extracted from the vowel midpoint and vowel tokens with adjacent velar segments were discarded. Speaker specificity was operationalised as Kullback-Leibler divergence (KL), which quantifies the dissimilarity of two distributions (in this case two individual speaker distributions), taking into account both the mean and the standard deviation (Kullback & Leibler, 1951). A KL value of zero indicates that the two distributions compared are identical, with higher KL values interpreted here as being more speaker specific.

## Results

Mixed effects models predicting KL divergence included fixed effects of formant, language, and their interaction, along with a random by-speaker intercept. For /u/, F3 was found to be significantly more speaker specific than F2 for Ukrainian ( $p < .001$ ,  $\beta = -1.15$ ) while for Turkish F2 was more speaker specific; however, this difference did not reach statistical significance ( $p = .77$ ,  $\beta = 0.016$ ). For the between-language F3 comparison, the KL divergence values were higher for Ukrainian than for Turkish. This difference was statistically significant ( $p < .001$ ,  $\beta = 0.70$ ). For /u/, F3 KL divergence values were found to be significantly higher than F2 for both Turkish ( $p < .001$ ) and Japanese ( $p < .001$ ). Regarding the between-language F3 comparison, the KL divergence values were higher for Japanese than for Turkish, reaching statistical significance ( $p < .001$ ). These results conform to the prediction that in languages without a vowel rounding contrast, F3 shows higher degrees of speaker specificity.

## References

- Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). *Common Voice: A massively-multilingual speech corpus*. arXiv. <https://arxiv.org/abs/1912.06670>
- Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer* (Version 6.4.04) [Computer software]. <http://www.praat.org/>
- Bosch, J. C. (2003, August). Acoustic study of the vowel formant frequencies and F0: a contribution to Catalan forensic phonetics. In *Proceedings of the 15th International Congress of Phonetic Sciences. Barcelona, Spain: Universitat Autònoma de Barcelona* (pp. 687-90).
- Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *International Journal of Speech, Language and the Law*, 18(2), 293–307. <https://doi.org/10.1558/ijssl.v18i2.293>
- Jessen, M. (2008). Forensic phonetics. *Language and Linguistics Compass*, 2(4), 671–711. <https://doi.org/10.1111/j.1749-818X.2008.00066.x>
- Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1), 79–86. <https://doi.org/10.1214/aoms/1177729694>
- Mathur, S., & Vyas, J. (2016). Acoustic analysis for comparison and identification of normal and disguised speech of individuals. *Journal of Forensic Science & Criminology*, 4(4), 403.
- McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017* (pp. 498–502). <https://doi.org/10.21437/Interspeech.2017-1386>
- McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aɪ/. *International Journal of Speech, Language and the Law*, 11(1), 103–130. <https://doi.org/10.1558/sll.2004.11.1.103>
- Zhang, M., Ahn, E., Baker, A., & Chodroff, E. (2025). *VoxCommunis Corpus*. OSF. <https://doi.org/10.17605/OSF.IO/T957V>

# The influence of bilingualism on voice quality in cross-language voice comparison

D.J. de Graaff<sup>1</sup>

<sup>1</sup>*Faculty of Humanities, Leiden University, Leiden, The Netherlands*  
d.j.de.graaff@umail.leidenuniv.nl

Bilingualism and cross-language voice comparisons are becoming more prevalent in forensic cases (Lo, 2021). This means that research on forensic voice comparison should adopt a multilingual perspective and requires more knowledge on the influence of language on acoustic parameters (de Boer & Heeren, 2023). An international survey on forensic voice comparison practices among experts shows that voice quality (VQ) is mentioned most often as most useful for discriminating speakers (Gold & French, 2011). Lo (2021) mentions that studies on the discriminatory potential of VQ parameters are limited. The studies that are done show that VQ parameters are robust in forensic voice comparisons (Hughes et al., 2019; Lo, 2021). However, studies that compare VQ parameters across languages show contradictory results (Zhu et al., 2023).

The current study, focused on the effect of bilingualism on the stability of VQ parameters. The aim of the study is to contribute to the knowledge about the influence of language and bilingualism on the stability of VQ parameters in (forensic) cross-language voice comparisons. Based on previous research the VQ parameters investigated are F0, jitter, shimmer, mean spectral energy and spectral tilt. The language combination in the current research is Dutch and English. The research question is: How does L1 Dutch and L2 English bilingualism influence the stability of voice quality parameters in cross-language voice comparison?

The expectations are that F0 is a stable VQ parameter and that jitter, shimmer, mean spectral energy (MSE) and spectral tilt (ST) are unstable VQ parameters compared across Dutch and English voice comparisons.

For this research, audio recordings from 35 speakers from the database of the D-LUCEA Accent Project are used (Orr & Quené, 2017). The D-LUCEA Accent Project collected data from speakers of L1 Dutch and L2 English to study the convergence of accents (Orr & Quené, 2017).

The acoustic analysis was done in Praat (Boersma & Weenink, 2024). The samples that were used include the schwa-like vowel from filled pauses “uh” and “um” that last at least 160 milliseconds. The statistical analysis was performed using R Statistical Software (R Core Team, 2024). The effect of bilingualism on VQ was tested by performing linear mixed effect models in R using the *lme4* package (Bates, Maechler, & Bolker, 2012).

The results show that ST is influenced by language and F0, jitter, shimmer and MSE are not influenced by language when compared across Dutch and English. The stability of F0, jitter, shimmer and MSE could be explained by first language transfer to the second language as certain phonatory settings might be difficult to acquire by Dutch-English bilingual speakers who have not been exposed to an English-speaking environment for a long time (Lo, 2021; Laver, 1987). The instability of ST across languages could be attributed to ST being more directly influenced by subtle phonatory and resonance shifts than MSE and other VQ parameters (Hanson, 1997; Kreiman et al., 2021). Further research with other language combinations is necessary to better understand the influence of bilingualism and language on VQ parameters in cross-language voice comparisons.

## References

- Bates, D.M., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using Eigen and S4 classes*. R package version 0.999999-0.
- Boersma, P., & Weenink, D. (2024). Praat: doing phonetics by computer (version 6.4.25) [Computer program]. Retrieved December 15th, 2024, from <https://www.fon.hum.uva.nl/praat/>

- de Boer, M. M., & Heeren, W. F. L. (2023). The language dependency of /m/ in native Dutch and non-native English. *The Journal of the Acoustical Society of America*, 154(4), 2168–2176. <https://doi.org/10.1121/10.0021288>
- Gold, E. & French, P. (2011). International practices in forensic speaker comparison. *The international journal of speech, language and the law*, 18(2), 293-307. <https://doi.org/10.1558/ijssl.v18i2.293>
- Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, 101(1), 466–481. <https://doi.org/10.1121/1.417991>
- Hughes, V., Cardoso, A., Harrison, P., Foulkes, P., French, P., & Gully, A. J. (2019). Forensic voice comparison using long-term acoustic measures of laryngeal voice quality. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia 2019 (pp. 1455–1459). Australasian Speech Science & Technology Association Inc.
- Kreiman, J., Gerratt, B. R., & Ito, M. (2021). The relative contribution of source and filter characteristics to voice quality. *The Journal of the Acoustical Society of America*, 149(5), 3438–3450. <https://doi.org/10.1121/10.0004757>
- Laver, J. (1987). *Individual features in voice quality* [PhD thesis, University of Edinburgh]. <http://hdl.handle.net/1842/6732>
- Lo, J. H. (2021). *Issues of bilingualism likelihood ratio-based forensic voice comparison* [PhD thesis, University of York]. [https://etheses.whiterose.ac.uk/30007/1/Lo\\_JJH\\_Thesis\\_Final.pdf](https://etheses.whiterose.ac.uk/30007/1/Lo_JJH_Thesis_Final.pdf)
- Orr, R. & Quené, H. (2017). D-LUCEA: Curation of the UCU Accent Project Data. In: Odijk, J., & van Hessen, A. (eds.) *CLARIN in the Low Countries*, 181-193. Ubiquity Press. <https://doi.org/10.5334/bb1.15>
- R Core Team (2024). "R: A language and environment for statistical computing.". <https://www.R-project.org/> (Last viewed 18 June 2025).
- Zhu, S., Chong, S., Chen, Y., Wang, T., & Ng, M. L. (2022). Effect of Language on Voice Quality: An Acoustic Study of Bilingual Speakers of Mandarin Chinese and English. *Folia Phoniatrica et Logopaedica*, 74(6), 421–430. <https://doi.org/10.1159/000525649>

# **CANDORspeech: A large-scale corpus of phonetically annotated conversational speech from dyadic online conversations with human quality control**

*Valeriia Vyshnevetska*<sup>1,2</sup>, *Alessandro De Luca*<sup>1,2</sup>, *Nadine Lavan*<sup>3</sup>, *Carolyn McGettigan*<sup>4</sup>, *Gus Cooney*<sup>5</sup>, *Andrew Reece*<sup>6</sup>, and *Volker Dellwo*<sup>2</sup>

<sup>1</sup>*Linguistic Research Infrastructure, University of Zurich, Zurich, Switzerland*

<sup>2</sup>*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*

<sup>3</sup>*Centre for Brain and Behaviour, School of Biological and Behavioral Sciences, Queen Mary University of London, London, United Kingdom*

<sup>4</sup>*Division of Psychology and Language Sciences, University College London, London, United Kingdom*

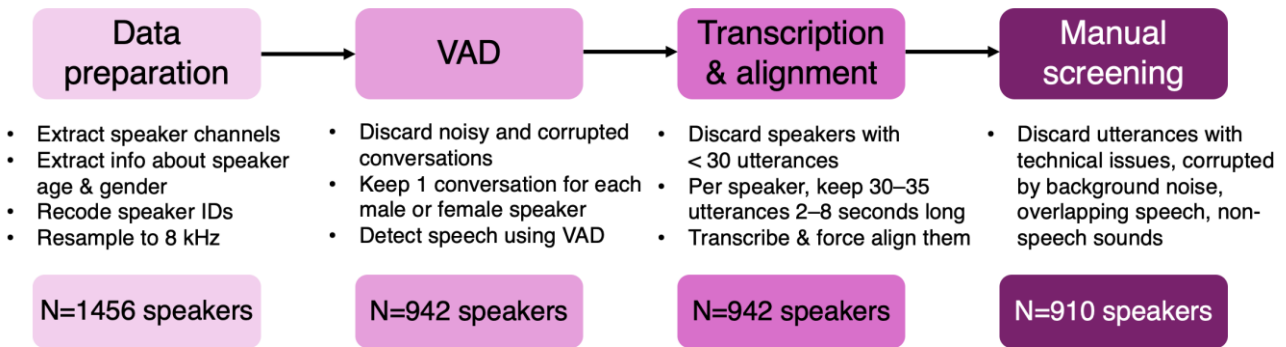
<sup>5</sup>*The Wharton School, University of Pennsylvania, Philadelphia, United States*

<sup>6</sup>*BetterUp Inc., Austin, Texas, United States*

{valeriia.vyshnevetska|alessandro.deluca|volker.dellwo}@uzh.ch,  
n.lavan@qmul.ac.uk, c.mcgettigan@ucl.ac.uk, guscooney@gmail.com,  
andrew.reece@betterup.com

Existing speech corpora for phonetic analysis are typically recorded in controlled environments, have restricted sample sizes (in terms of voices and materials recorded), and frequently focus on read speech, making them less suitable for studying naturalistic speech and voice phenomena (Lin et al., 2024; Nagrani et al., 2017; Nolan et al., 2009; Panayotov et al., 2015). The CANDOR corpus was recently introduced as a large multimodal dataset of naturalistic conversations containing audio and video from 1656 dyadic conversations in English along with their transcripts (Reece et al., 2023). CANDOR includes a large and diverse sample of speakers from the United States (N=1456) who spanned a broad range of gender, educational, ethnic, and generational identities. The multimodal nature of CANDOR ensured more naturalistic interactions compared to telephone-based conversational corpora (Godfrey & Holliman, 1993; Sadjadi, 2021), and CANDOR is further enriched by extensive metadata, including pre- and post-conversation surveys about participants experiences and personality traits, as well as a wealth of pre-extracted linguistic, acoustic, visual, and behavioral features. Since most speakers engaged in multiple conversations with different interlocutors, the corpus also enables robust analysis of within-speaker variability across sessions.

Here, we present CANDORspeech: a transcribed, force aligned and manually screened audio version of the original CANDOR corpus. We outline the processing pipeline (Figure 1) and the resulting CANDORspeech corpus with around 30 utterances from more than 900 speakers, with potential for expansion to include additional speech material and visual modality data. CANDORspeech offers a new unique environment to study speech and voice phenomena in online dyadic conversations that became increasingly popular and in many cases the norm for dyadic interactions in professional as well as in private contexts. The humanly screened automatic transcription allows the study of phonetic phenomena of this particular speech environment at large quantities. The database is also suitable for understanding voice variability within- and between-speakers of a representative random sample of speakers of the varied US population.



**Figure 1.** Construction pipeline of the CANDORspeech corpus.

## References

- Godfrey, J. J., & Holliman, E. (1993). *Switchboard-1 Release 2* (p. 14610176 KB) [Dataset]. Linguistic Data Consortium. <https://doi.org/10.35111/SW3H-RW02>
- Lin, Y., Cheng, M., Zhang, F., Gao, Y., Zhang, S., & Li, M. (2024). *VoxBlink2: A 100K+ speaker recognition corpus and the open-set speaker-identification benchmark* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2407.11510>
- Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. *Interspeech 2017*, 2616–2620. <https://doi.org/10.21437/Interspeech.2017-950>
- Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *The International Journal of Speech, Language and the Law*, 16(1), 31–57. <https://doi.org/10.1558/ijssl.v16i1.31>
- Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. <https://doi.org/10.1109/ICASSP.2015.7178964>
- Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13), eadf3197. <https://doi.org/10.1126/sciadv.adf3197>
- Sadjadi, S. O. (2021). NIST SRE CTS *Superset: A large-scale dataset for telephony speaker recognition* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2108.07118>

## Speaker characteristics of categorical data on filler particles in German

*Shunichi Ishihara<sup>1</sup>, Michael Jessen<sup>2</sup> and Beeke Muhlack<sup>3</sup>*

<sup>1</sup>*Speech and Language Laboratory, The Australian National University*

*shunichi.ishihara@anu.edu.au*

<sup>2</sup>*Department of Text, Speech and Audio, Bundeskriminalamt, Germany*

*michael.jessen@bka.bund.de*

<sup>3</sup>*Sachgebiet Phonetik, Bayerisches Landeskriminalamt, Germany*

*beeke.muhlack@polizei.bayern.de*

Categorical forensic evidence in the form of count data has been a neglected topic in forensic voice comparison. An initial research design for how such type of data might be processed in order to obtain likelihood ratios (LRs) was presented by Aitken and Gold (2013), based on data of click sounds as filler particles. Extensive research on this type of categorical evidence has been conducted in forensic authorship analysis (Ishihara 2023 including further references). Carne (2023) transferred the methodology developed in authorship analysis to forensic voice comparison by focussing on the speaker discrimination characteristics of filler particles in English. In this contribution we want to extend upon this research by, among other aspects, considering more types of filler particles.

This study investigates five German filler particles (FPs) – vocalic ‘uh’, vocalic-nasal ‘um’, nasal ‘hm’, glottal ‘gl’ and tongue clicks ‘cl’ using recordings from 100 speakers (two recordings each, one in normal, one in Lombard condition) in the Pool2010 database. For ‘uh’, ‘um’ and ‘hm’ four pause-context variants were also considered: +FP+ (FP within speech), +FP- (FP before silent pause), -FP+ (FP after silent pause) and -FP- (FP in isolation, i.e. surrounded by silent pauses). See Muhlack et al. (2023) for description of the data and the typology of FPs used here; see also Tschäpe et al. (2005) for previous work on some FPs in this database and Braun et al. (2023) for other recent forensically-guided work on a broad range of disfluencies in German.

Two FP sets were evaluated for forensic voice comparison:

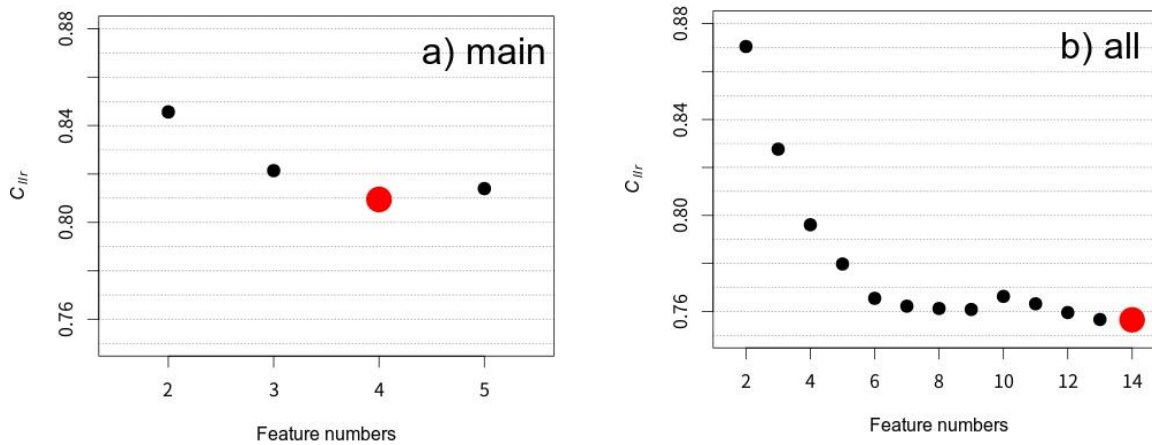
- Main FPs: the five base FPs ['uh', 'um', 'hm', 'gl', 'cl']
- All FPs: the full set including pause-context variants ['+uh-', '-uh-', '+uh+', '-uh+', '+um-', '-um-', '+um+', '-um+', '+hm-', '-hm-', '+hm+', '-hm+', 'gl', 'cl']

FP occurrences were tallied per recording to form count-based feature vectors. Speakers were randomly divided into three batches (34, 33, 33) to serve, in turn, as background, test, and calibration sets in a six-fold cross-validation. Scores for same-speaker (SS) and different-speaker (DS) comparisons were generated via a Dirichlet–multinomial model (Bolck and Stamouli 2017; Ishihara 2023) with hyperparameters estimated from the background data. Test-set scores were converted into LRs using logistic-regression calibration trained on the calibration-set scores.

All combinations of the Main and All FP sets were assessed. System performance was assessed by the log-likelihood-ratio cost (Cllr) and the resultant LRs were visualised with Tippett plots. Because recordings were captured under both normal and Lombard speech conditions, the effect of matched versus mismatched conditions on DS comparisons was also examined.

Finally, results will be compared with prior English-language findings (Carne 2023), and directions for extending this work will be discussed.

To illustrate some of the results so far, Figure 1 shows, separately for Main FPs (a) and All FPs (b) the best Cllr value for each possible combination of features within each number of features. For example, for three features in the Main FPs set, there are 10 combinations of features; for six features in the All FPs set, there are 3003 combinations.



**Figure 1.** Best Cllr values plotted as a function of the number of features combined. Panel (a) shows results for Main FPs, and Panel (b) for All FPs. Large red circles = Best Cllr values of all.

Overall, Figure 1 shows that All FPs outperform Main FPs when multiple features are combined. Although using all features yields the highest performance, Panel (b) demonstrates that a nearly equivalent result can be obtained with just six features, which are:

+uh-, -uh-, -uh+, +um-, -um-, -um+.

The combination of four features in Panel (a) that yielded the best result is:

uh, um, gl, cl.

## References

- Aitken, C. & Gold, E. (2013). Evidence evaluation for discrete data. *Forensic Science International*, 230, 147–155.
- Bolck, A. & Stamouli, A. (2017). Likelihood ratios for categorical evidence: Comparison of LR models applied to gunshot residue data. *Law, Probability and Risk*, 16, 71–90.
- Braun, A., Elsässer, N. & Willems, L. (2023). Disfluencies revisited – Are they speaker-specific? *Languages*, 8, 155.
- Carne, M. (2023). Evaluating discrete forensic voice evidence: A preliminary investigation based on filled pause occurrence. In *Proceedings of the 20<sup>th</sup> International Congress of Phonetic Sciences (ICPhS)*, Prague, 3795–3799.
- Ishihara, S. (2023). Weight of authorship evidence with multiple categories of stylometric features: A multinomial-based discrete model. *Science & Justice*, 63, 181–199.
- Muhlack, B., Trouvain, J. & Jessen, M. (2023). Distributional and acoustic characteristics of filler particles in German with consideration of forensic-phonetic aspects. *Languages*, 8, 100.
- Tschäpe, N., Trouvain, J., Bauer D. & Jessen, M. (2005). Idiosyncratic patterns of filled pauses. Paper presented at the IAFPA Conference 2005, Marrakesh.

## Conference Organization

The 33rd International Association for Forensic Phonetics and Acoustics (IAFPA) Conference is held in The Hague, The Netherlands, July 20-23, 2025.

### The conference is organized by

- Leiden University Centre for Linguistics (LUCL)
- Netherlands Forensic Institute (NFI)
- Immigration and Naturalisation Service (IND)

### Organizing committee

Tina Cambier-Langeveld, Arjan van Dijke, Willemijn Heeren, Mirjam de Jonge, Sjef van Lier, Carolina Lins Machado, Gerard Tolsma, David van der Vloed

*With the support of six student assistants from Leiden University and VU Amsterdam.*

### Scientific committee

Anil Alexander, Ruth Bahr, Anna Bartle, Georgina Brown, Colleen Driscoll, Paul Foulkes, Helen Fraser, Andrea Fröhlich, Philip Harrison, Vincent Hughes, Michael Jessen, Gea de Jong-Lendl, Finnian Kelly, Elisa Pellegrino, Richard Rhodes, Radek Skarnitzl, Dominic Watt

### Venue

The conference will take place July 20-23 2025 at the The Hague Campus of Leiden University (Turfmarkt 99, 2511 DP The Hague).

### Contact information

E-mail: [iafpa2025@hum.leidenuniv.nl](mailto:iafpa2025@hum.leidenuniv.nl)

Conference website: <https://www.universiteitleiden.nl/en/events/2025/07/iafpa>

### Acknowledgements

The conference organization gratefully acknowledges sponsoring of the conference by the following institutions:

- Leiden University Centre for Linguistics (LUCL)
- Netherlands Forensic Institute (NFI)
- Immigration and Naturalisation Service (IND)

We thank Irina and Margot of LUCL, Leiden University.

Cover photo by [Alireza Parpaei](#) on [Unsplash](#)