UNIVERSITÄT TRIER

IAFPA 2026

Book of Abstracts
July 2026

Phonetics Lab
TRIER UNIVERSITY

TRIER
IAFPA
2026

Welcome to IAFPA 2026, welcome to Trier!

Dear colleagues and friends,

It is our great pleasure to welcome you to the **34th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)** at Trier University.

Trier, Germany's oldest city and a former imperial capital of the Roman Empire, is home to one of the country's youngest universities. Although the original University of Trier was founded in 1473, it was closed by Napoleon in 1798 and re-established only in 1970. Today, around 12,000 students and 2,000 staff study and work across six faculties, with a particular strength in the Humanities and Social Sciences. We are delighted to welcome you to our university and hope you will also have the opportunity to discover the rich history and unique atmosphere of the city during your stay.

Hosting IAFPA in Trier is particularly meaningful because forensic phonetics has a long and distinguished tradition at our university. Since Jens-Peter Köster established the Department of Phonetics in 1985, Trier has become an internationally recognized center for forensic phonetics. Together with Herbert Masthoff, Angelika Braun, the late Hermann Künzel, and many colleagues and graduates over the years, the department has made lasting contributions to both research and practice. Forensic phonetics remains an integral part of our teaching and research, with students engaging in authentic casework and producing numerous Master's and doctoral theses in the field. At the same time, the department's research has expanded considerably. Current work focuses on prosody, regional speech variation, and second language acquisition, while continuing to build bridges between basic phonetic research and its forensic applications. We are proud that this tradition continues today and that Trier remains closely connected to the international forensic phonetics community. Our connection with IAFPA goes back more than three decades. Trier hosted the Association's second conference in 1993 – the first to be held outside York – and welcomed the community again in 2010. We are therefore especially honored to host the conference for a third time, making Trier, alongside York, one of only two venues to have welcomed the IAFPA Annual Conference on three occasions.

Both IAFPA and forensic phonetics have grown tremendously since that first meeting in Trier. The Association has expanded into a truly global community, annual conferences have been held on four continents, forensic phonetics has become a well-established topic at the International Congress of Phonetic Sciences, and the field continues to thrive through an ever-growing body of innovative research. We are delighted that this year's conference reflects this continued development.

For **this year's scientific program**, we received submissions from researchers across the globe, reflecting both the international character of our community and the breadth of current research in forensic phonetics and acoustics. The program covers a remarkable range of topics, from inhalation patterns, phonetic variation, and voice quality to emerging challenges such as AI-generated speech and deepfake detection. Altogether, the conference features 21 oral presentations and 29 poster presentations, including several Work in Progress contributions.

This year's **Special Session**, Expertise in Context, focuses on the implications of judicial systems for practical forensic casework. It highlights the role of the (voice) expert as a key participant in criminal proceedings and explores the interplay between experts, judges, and defense lawyers across different judicial systems. By placing forensic expertise within its broader legal context, the session offers an important perspective on the practical application of our research.

We are especially pleased to see a strong participation of **early-career researchers**, whose enthusiasm and innovative work promise a bright future for our field. Their contributions are an essential part of this conference, and we look forward to recognizing outstanding work through the Early Career Researcher Awards.

This conference would not have been possible without the support of many people and institutions. We would like to express our sincere gratitude to Trier University for hosting the event, to the IAFPA President and Executive Committee for their continued support and guidance, and to the local organizing team, whose effort, dedication, and enthusiasm have made this conference possible.

In the spirit of the department's long-standing commitment to forensic phonetics and international collaboration, we warmly welcome you to Trier. We hope you enjoy the scientific program, engage in stimulating discussions, establish new collaborations, renew old friendships, and take the opportunity to experience the unique history and hospitality of our city.

We wish you an inspiring, enjoyable, and memorable conference.

For the local team

Angelika Braun and Katharina Zahner-Ritter

Phonetics Lab
TRIER UNIVERSITY

Committees

Scientific Committee

Anil Alexander
Angelika Braun
Ruth Huntley Bahr
Paul Foulkes
Peter French
Andrea Fröhlich
Lauren Harrington
Vincent Hughes
Michael Jessen
Finnian Kelly
Kirsty McDougall
Richard Rhodes
Dominic Watt

Local Organizing Committee

Angelika Braun
Katharina Zahner-Ritter
Sophia Fünfgeld
Zarah Kampschulte
Peter Knopp
Jacek Kudera
Elizabet Gogolashvili
Gabriele Jacobs

Student Assistants

Andreas Frank
Melissa Hildebrand
Lotta Hilliger
Philipp Hutmacher
Johanna Lentz

Some practical information

Venue and conference website

The conference will take place at

Trier University, Campus II (Room K101, Kapelle)
Behringstraße 21
54296 Trier

from 26–29 July 2026.

On the conference website (www.iafpa2026.uni-trier.de), you will find all practical information for your stay, including directions to Trier and the conference venue, accommodation, campus maps, room locations, and nearby lunch options.

Publication of Book of Abstracts

We are pleased to announce that the Book of Abstracts (BoA) for IAFPA 2026 has now been published in our **Open Journal System (OJS):**

https://iafpa-ojs.uni-trier.de/index.php/iafpa/en/index

The Book of Abstracts at OJS includes all regular oral presentations and posters. To make them easy to find regardless of the conference schedule, the contributions are listed in alphabetical order by author. Together, they reflect the broad range of topics covered at this year's conference, spanning everything from breathing patterns to the identification of deepfakes. All abstracts are freely accessible online and have been assigned individual Digital Object Identifiers (DOIs), ensuring that they are permanently available, citable, and easy to reference. We hope that publishing the BoA in this format will increase the visibility of your research and facilitate proper citation of your work.

Abstracts of Work-in-Progress papers (WiP) will be made available via the conference website and in this analogue booklet.

Code of conduct

By participating in the events related to the 34th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA) at Trier University, you agree to participate in the creation of a space free from any form of discrimination or violence. Please consider the detailed Code of Conduct for IAFPA 2026, published on our conference website. Contact persons during the IAFPA 2026 at Trier University are: JProf. Dr. Katharina Zahner-Ritter: k.zahner-ritter@uni-trier.de and Prof. Dr. Angelika Braun: brauna@uni-trier.de.

Emergency information

In case of an emergency in Germany, please call **112** (ambulance and fire brigade) or **110** (police). If you require assistance during the conference or encounter any urgent problems, please contact the conference organizers at +49 651 201 3399, +49 (0)171 3683196, or iafpa2026@uni-trier.de. We are happy to help whenever possible.

IAFPA 2026 | Trier University

Campus II (K101, Kapelle) | Behringstraße 21, 54296 Trier

July 26-29, 2026

Program

(as of July 19, 2026)

Sunday, July 26		
16:00	Scientific Working Group on Forensic Speaker Recognition (OSAC) *Larry Kincaid*	HZ 202
16:00 – 19:00	Student Reception	K101
17:00 – 19:00	Welcome Reception	K101

Monday, July 27 (K101)		
09:00 – 09:30	**Opening Session** • JProf. Dr. Katharina Zahner-Ritter – Local organizers • Prof Dr. Eva Martha Eckkrammer – President of Trier University • Prof. Dr. Christian Soffel – Vice Dean of Faculty II, Trier University • Dr. Tina Cambier-Langeveld – President of the IAFPA • Dr. Richard Rhodes – Chair of the IAFPA • Prof Dr. Angelika Braun – Local organizers	
09:30 – 10:30	**Oral Session I** – Chair: Ruth Huntley Bahr	
09:30 – 09:50	A question for forensic phoneticians from privacy officers: when are speech recordings identifiable?	*Tina Cambier-Langeveld*
09:50 – 10:10	Insights from the ENFSI FSAAWG Collaborative Exercise on Audio Enhancement	*Andrea Fröhlich, Anna Bartle, Sarah Lim, Dagmar Boss, Olof Lundgren Källmark, Sunčica Kuzmić*
10:10 – 10:30	Experience and results from a comprehensive audio forensics test – the MdCE 2025 project	*Attila Fejes*
10:30 – 11:00	**Coffee Break**	
11:00 – 12:00	**Oral Session II** – Chair: Sandra Ferrari Disner	
11:00 – 11:20	Voice familiarity across speaking styles: Effects on speech intelligibility and voice recognition	*Alessandra Metzger, Elisa Pellegrino, and Volker Dellwo**
11:20 – 11:40	Forensic Approaches to Voice Perception: Acoustic Variation Across Gender-Diverse Vowel Spaces	*Katelyn Taylor and Kirsty McDougall**

11:40 – 12:00	Perceived voice distinctiveness predicts speaker discrimination but is not described by acoustic voice spaces	*Alessandro De Luca and Volker Dellwo**
12:00 – 13:00	**Oral Session III** – Chair: Gea de Jong-Lendle	
12:00 – 12:20	The impact of accent imitation on voice identity perception	*Emilia Fletcher and Justin J. H. Lo*
12:20 – 12:40	Automatic identification of dialects in the western part of Germany	*Vatho Chem and Reinhold Greisbach*
12:40 – 13:00	The role of temporal information in automatic speech and voice recognition	*Erdem Baha Topbas, Srikanth Madikeri, Elisa Pellegrino and Volker Dellwo**
13:00 – 14:00	**Lunch Break**	
14:00 – 15:00	**Poster Session I** (Foyer in Building K) – Chair: Sophia Fünfgeld	
15:00 – 15:30	**Coffee Break (Poster Session I continued)**	
15:30 – 16:30	**Oral Session IV** – Chair: Michael Jessen	
15:30 – 15:50	Empirical Validation of Synthetic Speech as a Forensic Proxy for Mismatch and Recording-Condition Research	*Peter Milne, Colleen Driscoll and Emily Lawrie-Munro*
15:50 – 16:10	The inadequacy of statistically-calculated likelihood ratios in forensic voice comparison: the Italian case	*Sonia Cenceschi, Chiara Meluzzi and Francesco Roberto Dani*
16:10 – 16:30	Evaluation and Application of Joint Speech Enhancement and Automatic Speaker Recognition in Forensic Science	*Wei Li, Wenlin Huang, Weili Jiang and Xin Chen*
16:45	Meeting for VOCALISE/Fauxdio users	HZ 203

Tuesday, July 28 (K101)		
09:00 – 11:00	**Special Session: Expertise in Context** Chair: Angelika Braun	
09:00 – 09:15	The Expert in German Criminal Trials	*Dr. Matthias Quarch*
09:15 – 09:30	The Role of the Expert Witness in Criminal Proceedings in the Continental European Legal System	*Prof. Dr. Jan Bockemühl*
09:30 – 09:45	Expert Testimony in the U.S.	*Prof Dr. Ruth Huntley Bahr*
09:45 – 10:00	Expert Testimony across Jurisdictions	*Prof. Dr. Angelika Braun*
10:00 – 11:00	Panel Discussion	
11:00 – 11:30	**Coffee Break**	
11:30 – 12:30	**Oral Session V** – Chair: Volker Dellwo	
11:30 – 11:50	Um, Do These Uhs Matter? – An Investigation into the Influence of Filled Pauses on the Perceptions of the Interviewee	*Maya Ely Joshi, James Tompkinson, and Lauren Harrington**
11:50 – 12:10	Final F0 Fall in Forensic Speaker Comparison: The Role of Fillers	*Leia Weber, Gea de Jong-Lendle, and Mathias Scharinger**
12:10 – 12:30	Forensic phonetics before 1980: Some examples	*Angelika Braun*
12:30 – 13:30	**Lunch Break**	
13:30 – 15:00	**Poster Session II** (Foyer in Building K) – Chair: Peter Knopp	
15:00	IAFPA AGM	K101
19:00	Conference Dinner	Restaurant Weinwirtschaft Friedrich-Wilhelm, Weberbach 75, 54290 Trier

Wednesday, July 29 (K101)		
9:00 – 10:00	Roundtable: Checking Procedures	*Bryony Nuttall, Katherine Earnshaw and Richard Rhodes*
10:00 – 10:30	**Coffee Break**	
10.30 – 11:30	**Oral Session VI** – Chair: Katharina Zahner-Ritter	
10:30 – 10:50	Towards explainable linguistic phonetic methods to authenticate allegedly deepfake audio	*Ben Gibb-Reid, Jessica Wormald, and Vincent Hughes*
10:50 – 11:10	Perceiving speaker identity in authentic vs voice-converted speech: An eyetracking study	*Justin J. H. Lo*
11:10 – 11:30	Dialect, voice quality and perceived speaker identity in AI voice conversion	*Hope McVean**
11:30 – 12:30	**Oral Session VII** – Chair: David van der Vloed	
11:30 – 11:50	Assessing automatic, acoustic-phonetic, and auditory speaker comparison in the era of open-source deepfakes	*Linda Gerlach, Amaia Sagasti and Anil Alexander*
11:50 – 12:10	Forensic Automatic Speaker Recognition: reflecting on technological progress and casework applications	*Finnian Kelly and Anil Alexander*
12:10 – 12:30	Beyond 'Real' or 'Fake': The challenge of benign AI-based processing in forensic deepfake detection	*Amaia Sagasti, Finnian Kelly and Anil Alexander*
12:30	**Closing Session** • Announcement of IAFPA Student Award winners • Announcement of venue for IAFPA 2027 • Announcement of OWR Award winner	
13:00 – 16:00	SWG on FSR (OSAC) – *Larry Kinkeid*	HZ 202

List of Posters

Poster Session I | Monday, July 27, 14:00-15:30 | Foyer in Building K

PI-1	Individual Variation in the Production of Austrian German Rhotics	*Nathalie Elsässer, Hendrik Behrens-Zemek and Eva Reinisch***
PI-2	The Effect of Speech Rate on Dynamic Acoustic Features of Mandarin Triphthongs	*Kai Gao, Weili Jiang, Xin Chen, Wenlin Huang, Wei Li*
PI-3W	The Social Salience of tapped /r/ [ɾ] in Dublin Accents and Linguistic Attitudes Towards it in a Forensic Context	*Elin Cat Stüer and Grace Wenling Cao***
PI-4W	/s/ and /ʃ/ Variation Across Gender-Diverse Speakers: Implications for Forensic Casework Involving Minority Speaker Populations	*Katelyn Taylor and Kirsty McDougall***
PI-5W	Phoneme Inventory Effects on Speaker Specificity in /n/: A Cross-Linguistic Study	*Annie Baker and Eleanor Chodroff **
PI-6W	Fundamentally challenging: The role of f0 in cross-language speaker discrimination	*Leah Bradshaw, Julien Plante-Hebert and Elizabeth Allyn Smith*
PI-7W	One Speaker, Multiple Dialects: Insights from the Language Familiarity Effect on Forensic Speaker Recognition	*Sophie Möller, Radek Skarnitzl, Volker Dellwo, Elisa Pellegrino and Tyler Perrachione***
PI-8	Grammatical Gender in Bidialectal Speakers of Swiss and Standard German	*Niels O. Schiller*
PI-9	On the speaker-discriminatory potential of breath pauses in speech	*Sascha Schäfer and Jürgen Trouvain*
PI-10W	The role of voice quality in voice memorability: an experiment	*Alice Paver and Kirsty McDougall **
PI-11W	Transcription in noise: investigating how errors produced by lay listeners vary across increasing noise levels	*Lauren Harrington*

PI-12W	Vowel Distribution and the Stability of Long-Term Formants in German	*Philipp Hutmacher, Angelika Braun and Katharina Zahner-Ritter***
PI-13W	Signs of Partner Proximity in Formant Measurements: A First Look at the Relevance of Entrainment on Voice Identification Analysis	*Ellen Grand and Georgina Brown***
PI-14	Can listeners learn to estimate speaker age?	*Johanna Lentz, Katharina Zahner-Ritter and Angelika Braun***
PI-15	The Construction of a Voice Parade – Stimulus playback and identification accuracy	*Pauline Winkler, Gea de Jong-Lendle, Ulrike Stühler, Mathias Scharinger***

Poster Session II | Tuesday, July 28, 13:30-15:00 | Foyer in Building K

PII-1	From Absence to Construction: The Rise of *Fonoaudiologia Forense* in Forensic Speaker Comparison in Brazil and Paraguay	*Pablo Vourliotis Alvarenga, Ana Paula Sanches and Samantha Moraes*
PII-2	It's the Evidence That Matters. The Forensics of LOID	*Andreea Dragu and Mihaela Stan*
PII-3	A survey of forensic practitioners on speaker attribution methods	*Tallulah Buckley, Kirsty McDougall and Richard Rhodes***
PII-4W	Audio Enhancement and Transcription Interlaboratory Comparison Exercises	*Katherine Earnshaw, Bryony Nuttall and Richard Rhodes*
PII-5	DIVERSE: a corpus of cisgender female speakers for phonetic and forensic research	*Alice Paver, Chloe Patman, Kirsty McDougall, Paul Foulkes and Jay Paulson***
PII-6	F0 corpus for case-adapted reference populations	*Almut Braun, Michael Jessen, Anja Moos and Isolde Wagner*
PII-7	Effect of combining supralaryngeal with laryngeal information in a forensic telephone-based dataset	*Michael Jessen and Canan Konrat*
PII-8	Choosing an appropriate reference population can mitigate limitations in automatic cross-language forensic speaker comparison	*Kristina Tomić, Dominic Watt, Valeriia Vyshnevetska, Volker Dellwo and Peter French*
PII-9W	AI-ccent judgements: Accent stereotypes persist with AI-generated voices	*Jun Liu, Leah Bradshaw and Eleanor Chodroff***
PII-10W	Developing a forensic methodology for deepfake audio detection	*Anna Bartle, Matthew Boyle, Ben Gibb-Reid, Ronan Given, Amelia Gully, Philip Harrison, Vincent Hughes, Lucy Morville, Luca Resti, Andrew Robinson and Jessica Wormald*
PII-11W	Can Spoofed Voices Compromise an ASR system?	*Julia Roegner, Simon Werner and Angelika Braun***
PII-12	Comparing humans and neural networks in speech deepfake detection	*Manuel Otero-Gonzalez, Aurora López-Jareño, Eugenia San Segundo and Joaquin Gonzalez-Rodriguez***
PII-13	Calibration of ASV Systems and Robustness Thereof	*Torbjörn Onshage, Aron Paulsson, Greta Öhlund Wistbacka, Susanna Whitling and Andreas Jakobsson***
PII-14	The Risk of Data Contamination in Forensic Voice Comparison: a Perception Experiment using Voice-mixed Speech Samples	*Scarlett Vanroose-Peer***

Legend:

* = eligible for IAFPA Student Award – Best oral presentation;

** = eligible for IAFPA Student Award – Best poster

W = Work-in-Progress Poster (WiP)

Oral Sessions I-VII

A question for forensic phoneticians from privacy officers: when are speech recordings identifiable?

Tina Cambier-Langeveld
Netherlands Forensic Institute, Immigration and Naturalisation Service, and Leiden University, the Netherlands
Gm.cambier.langeveld@ind.nl

Classic forms of speech data collection involve explicit consent from speakers, which is the most well-known legal basis for processing personal data. More recently, however, researchers use data scraped from platforms like YouTube or data recorded from broadcast media, in response to an increasing appetite for more challenging real-world settings (Leschanowsky et al., 2025). The (re)usability of speech data is a pressing issue, since access to representative data is key for the development of speech technology. Furthermore, access to forensically realistic audio materials (ideally: casework data) is invaluable for validation and teaching (Brown, Ross & Kirchhübel, 2021).

Whether it is actually legal to collect and use certain speech data is a complex matter, and depends on the purpose as well as the jurisdiction. Aside from copyright and intellectual property issues, I will focus on the topic of personal data protection as laid down in the EU General Data Protection Regulation (GDPR, European Parliament and Council, 2016). A central question for the applicability of the GDPR is whether we are dealing with information that can be related to a natural person.

It is clear that speech data is *usually* considered to be personal data. Speech data *can* also fall within the scope of sensitive data[1], since it *can* reveal identity, ethnic origin, political opinions, religious or philosophical beliefs, health conditions and sexual orientation (Nautsch et al. 2019a,b). Yet, even if speech data is likely to qualify as sensitive data, this still leaves open the possibility of (a comprehensive list of) conditions under which speech data is *not* (or no longer) sensitive or even personal data. The IAFPA community would appear well suited to explain the difficulties of identifying an unknown speaker in an open identification task.

A first question that we can shed some light on is: when can we say that it is practically impossible to trace a speech recording back to its speaker, i.e., that a recording is effectively anonymized? It's relatively straightforward to understand what kind of content and metadata should be removed, but then the question remains if the speaker could still be identified through voice recognition. Privacy officers and those wanting to anonymize speech data struggle with this question, as they usually have no expertise in the area of voice recognition.

A second question for us to consider is: when does speech data meet the GDPR's definition of biometric data? A close look at the legal definition of biometric data is required before we can begin to answer this question. An outstanding matter here is whether speaker embeddings, often part of DNN-based speech technology applications, should always be regarded as biometric data.

This talk will present an outline of the relevant legal definitions, before offering an initial answer to the questions posed above. Given that the European Data Protection Board (EDPB) has issued extensive guidance on the processing of photos and videos carrying face images, we will also examine the parallels between face and voice recognition.

[1] 'Sensitive data' is used as short-hand term for what the GDPR calls 'special categories of personal data'. This includes, but is not limited to, 'biometric data for the purpose of uniquely identifying a natural person' (GDPR Art. 9 sub 1). Sensitive data merits specific protection (i.e. stricter processing requirements), since their processing could create significant risks to fundamental rights and freedoms (GDPR Recital 51).

References

Brown, G. Ross, S. & Kirchhübel, C. (2021). Voicing concerns: The balance between data protection principles and research developments in forensic speech science. *Science & Justice*, 61, 311-318. https://doi.org/10.1016/j.scijus.2021.05.006

European Data Protection Board (2020). Guidelines 3/2019 on processing of personal data through video devices, version 2.0, adopted on 29 January 2020. https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201903_video_devices_en_0.pdf

European Data Protection Board (2023). Guidelines 05/2022 on the use of facial recognition technology in the area of law enforcement, version 2.0, adopted on 26 April 2023. https://www.edpb.europa.eu/system/files/2023-05/edpb_guidelines_202304_frtlawenforcement_v2_en.pdf

European Data Protection Board (2024). Opinion 11/2024 on the use of facial recognition to streamline airport passengers' flow (compatibility with Articles 5(1)(e) and (f), 25 and 32 GDPR, version 1.1, adopted on 23 May 2024. https://www.edpb.europa.eu/system/files/2024-05/edpb_opinion_202411_facialrecognitionairports_en.pdf

European Data Protection Board (2024). Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models, adopted on 17 December 2024. https://www.edpb.europa.eu/system/files/2024-12/edpb_opinion_202428_ai-models_en.pdf

European Parliament and Council (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A02016R0679-20160504

Leschanowsky, A., Rusti, C., Quinlan, C., Pnacek, M., Gorce, L., & Hutiri, W. (2025). A Data Perspective on Ethical Challenges in Voice Biometrics Research. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(1), 118-131. https://doi.org/10.1109/TBIOM.2024.3446846

Nautsch, A., Jasserand, C., Kindt, E., Todisco, M., Trancoso, I. & Evans, N. (2019a). The GDPR & speech data: reflections of legal and technology communities, first steps towards a common understanding. *Proceedings of Interspeech 2019*, 3695-3699. https://doi.org/10.21437/Interspeech.2019-2647

Nautsch, A., Jimenez, A., Treiber, A., Kolberg, J., Jasserand, C., Kindt, E., Delgado, H., Todisco, M., Hmani, M.A., Mtibaa, A., Abdelraheem, M. A., Abad, A., Teixeira, F., Matrouf, D., Gomez-Barrero, M., Petrovska-Delacrétaz, D., Chollet, G., Evans, N., Schneider, T., Bonastre, J.-F., Raj, B., Trancoso, I. & Busch, C. (2019b). Preserving privacy in speaker and speech characterisation. *Computer Speech and Language*, 58, 441-480. https://doi.org/10.1016/j.csl.2019.06.001

Insights from the ENFSI FSAAWG Collaborative Exercise on Audio Enhancement

Andrea Fröhlich[1], Anna Bartle[2], Sarah Lim[1], Dagmar Boss, Olof Lundgren Källmark[3], Sunčica Kuzmić[4]

[1]*Speech & Audio Group, Zurich Forensic Science Institute, Zurich, Switzerland*
andrea.froehlich@for-zh.ch
sarah.lim@for-zh.ch
[2]*Audio Laboratory, Digital Forensics, Metropolitan Police, UK*
anna.bartle@met.police.uk
[3]*Speech & Audio Analysis, National Forensic Centre, Sweden*
olof.lundgren-kallmark@polisen.se
[4]*Forensic Science Centre "Ivan Vučetić", Ministry of the Interior, Croatia*
skuzmic@mup.hr

The increasing availability of recording devices in public and private spaces has direct implications for law enforcement, where the number of recordings requiring processing is steadily increasing (Brayne et al. 2018). In these contexts, recordings from CCTV cameras, for example, are typically of degraded quality; however, they may document highly relevant case-related information. Accordingly, law enforcement agencies often request audio specialists to enhance the recording quality with the goal of better understanding what is being said.

Depending on the case and context, audio enhancement may pursue different goals, such as improving listenability, intelligibility, or both, for example to facilitate transcription by an interpreter or playback in court (Zajlic, 2021). Overall, there has been sparse research into audio enhancement practices and what might constitute an effective enhancement; prevalent assumptions are that metrics such as improving SNR can be used to gauge the effectiveness of audio enhancement. While metrics can give insight, the real-life case circumstances as well as the client listener are of crucial importance when attempting to improve intelligibility or listenability of a recording. This circumstance has received less attention, with the notable exception of studies in which enhancements were evaluated before and after processing (Fraser 2018, Hu & Loizou 2007). Given the many open questions surrounding the subjective role of the listener in audio enhancement, we initiated a project that ultimately resulted in a collaborative exercise (CE) on audio enhancement and the subjective evaluation of participants' enhancement results.

Methods

With the goal of closely mirroring recording qualities from casework practice, two audio recordings with severely degraded quality were generated; one recording was produced in a stairway, resulting in high reverberation, while the other was recorded outside on a noisy street.

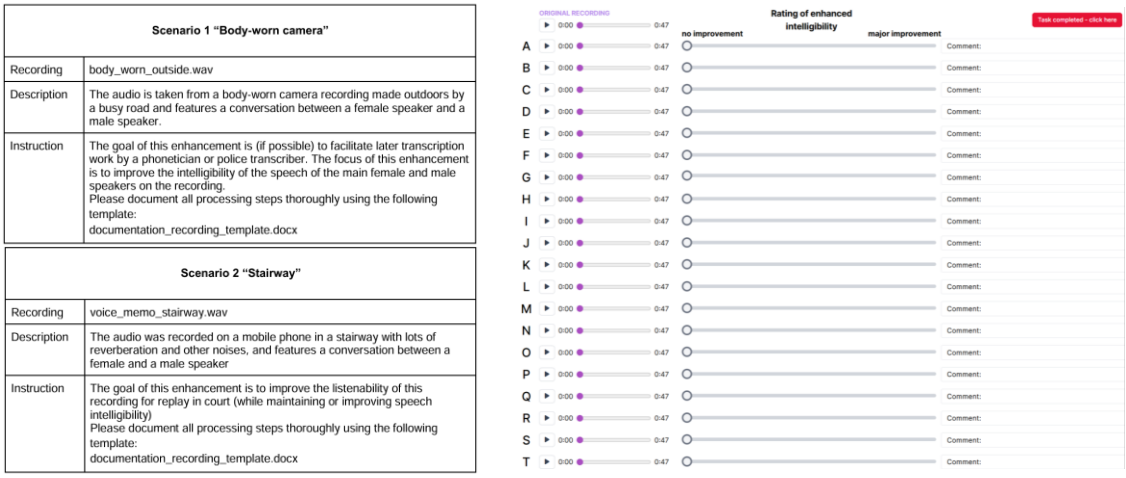
In the first phase, participating laboratories were instructed to enhance the two recordings with two scenarios in mind, see detailed instructions in Figure 1 (left). Alongside submitting enhancements,

the participating labs were asked to provide detailed documentation of the enhancement steps taken as well as answer a short questionnaire.

In a second step, all submitted enhanced versions were anonymized, including the removal of metadata, and were then included in a rating task implemented in Gorilla.sc (Anwyl-Irvine et al. 2020), see illustration below, Figure 1 (right). The recordings were evaluated by an internal rater group consisting of laboratories that participated in Phase 1 (18 of the 20 participating laboratories submitted internal ratings), as well as by an additional external rater group (n = 26) comprising individuals from law enforcement with diverse professional backgrounds and varying levels of experience in audio enhancement.

Results

Twenty laboratories submitted enhanced versions of the two recordings. Overall, perceived improvements in listenability were greater than those in intelligibility. Ratings varied considerably for each enhancement. Additionally, substantial variability was observed in both the software employed and the enhancement procedures applied. The talk provides insights from a meta-analysis comparing the enhancement steps across all participating laboratories and presents results from the rating task.



Scenario 1 "Body-worn camera"	
Recording	body_worn_outside.wav
Description	The audio is taken from a body-worn camera recording made outdoors by a busy road and features a conversation between a female speaker and a male speaker.
Instruction	The goal of this enhancement is (if possible) to facilitate later transcription work by a phonetician or police transcriber. The focus of this enhancement is to improve the intelligibility of the speech of the main female and male speakers on the recording. Please document all processing steps thoroughly using the following template: documentation_recording_template.docx

Scenario 2 "Stairway"	
Recording	voice_memo_stairway.wav
Description	The audio was recorded on a mobile phone in a stairway with lots of reverberation and other noises, and features a conversation between a female and a male speaker
Instruction	The goal of this enhancement is to improve the listenability of this recording for replay in court (while maintaining or improving speech intelligibility) Please document all processing steps thoroughly using the following template: documentation_recording_template.docx

Figure 1. Left: Instructions provided to participants for enhancement tasks. Right: Experimental interface displaying the twenty enhanced versions of the original recordings. The degree of subjective enhancement was rated on a scale ranging from no improvement to major improvement.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior research methods*, *52*(1), 388-407.

Brayne, S. (2018). The criminal law and law enforcement implications of big data. *Annual Review of Law and Social Science*, *14*(1), 293-308.

Fraser, H. (2020). 'Enhancing' forensic audio: what if all that really gets enhanced is the credibility of a misleading transcript? *Australian Journal of Forensic Sciences*, *52*(4), 465-476.

Hu, Y., & Loizou, P. C. (2007). A comparative intelligibility study of single-microphone noise reduction algorithms. *The Journal of the Acoustical Society of America*, *122*(3), 1777-1786.

Zjalic, J. (2020). *Digital audio forensics fundamentals: from capture to courtroom*. Focal Press.

Experience and results from a comprehensive audio forensics test – the MdCE 2025 project

Attila Fejes
Special Service for National Security, Széchenyi István University Doctoral School of Law and Political Sciences
fejes.attila@nbsz.gov.hu, fejes.attila@sze.hu

Multidisciplinary collaboration exercises (MdCE) (Zampa et. al., 2024) are an important part of forensic science, as cooperation among experts is essential in investigations involving multiple disciplines. During the examinations, the experts must not compromise traces of other disciplines. The European Network of Forensic Science Institutes (ENFSI) has been organizing cooperation exercises for several years, and this project is the first in Audio Forensics. The test materials for the audio examinations in the exercise consisted of audio recordings of Hungarian speakers, which were analyzed by experts from 16 countries on behalf of 23 laboratories. The tasks included authenticity analysis, speaker comparison (Drygajlo et. al., 2015), and audio enhancement (Fraser, 2020). Accreditation according to the ISO/IEC 17025 standard (Pádár et. al., 2015) is widespread in expert institutes. This includes detailed description, testing, and validation of each testing method, preparation of authentic documents, and regular proficiency tests and collaborative exercises. Collaborative Exercises (CE) and Proficiency Tests (PT) ensure that the performance of participating testing laboratories is measured and documented.

The speaker comparison task in the exercise may rely on Forensic Semiautomatic Speaker Recognition (FSASR), Forensic Automatic Speaker Recognition (FASR) and/or other suitable comparison methods (auditory or acoustic). Regardless of the chosen approach, the CERTAIN-FORS WP7 multilingual reference population database (Fejes et. al., 2024) can be utilized. Results should be expressed as a Likelihood Ratio (LR) and/or with a verbal statement. Guidance on interpreting LR values can be found in the ENFSI guideline for evaluative reporting. The authenticity analysis task involves the examination of an unknown audio file and ENF-reference data (Rodríguez et. al, 2010).

Within the authenticity analysis task majority of laboratories correctly identified the tampered point on the audio. This may be due to the effect of speech recorded in an unknown language on the experts or the features of the unknown audio; nevertheless, it would be worthwhile to investigate the detailed cause of this phenomenon in the future. The results of authenticity test were evaluated in accordance with the ENFSI Guideline for Evaluative Reporting in Forensic Science. More than half of the experts used voice enhancement to perform the speaker comparison task, however, a significant number of laboratories did not consider it effective before speaker recognition. Those who considered noise reduction important did so primarily for FASR. Based on the speaker comparison's results, the non-native language posed a challenge in several cases; however, overall, it hindered speaker comparison only in a few instances. Typically, if a participant was unable to use one methodology, they could still complete the test using the other. Several laboratories reported the different types of speech recorded on the two audio recordings—read versus spontaneous—as a problem; however, many participants reported that this corresponded to the type of data available in the reference population. Here, too, it was evident that for some participants, the poor quality of the unknown recording prevented them from performing speaker recognition, whereas many participants performed audio enhancement and subsequently identified the speakers.

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 [online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/guidelines_fasr_and_fsasr_0.pdf

Fejes, A., Saraiva, A., & Devencos J. (2024). The effect of wearing different types of protection masks on FASR results in a multilingual speech sample database.", *IAFPA 2024 Livre de résumés / Book of Abstracts,* 57-58.

Fraser, H. (2020). Enhancing Forensic Audio: What Works, What Doesn't, and Why. *Griffith Journal of Law & Human Dignity.* 8(1), 85-102.

Pádár, Zs., Nogel, M. & G. Kovács. (2015). Accreditation of forensic laboratories as a part of the "European Forensic Science 2020 concept in countries of the Visegrad Group. *Forensic Science International: Genetics Supplement Series*5 e412-e413.

Rodríguez, D., Apolinário, J., & Biscainho, L. (2010). Audio authenticity: Detecting ENF discontinuity with high precision phase analysis. *IEEE Transactions on Information Forensics and Security*, 5(3), 534–543.

Zampa, F., Bandey, H., Becue, A., Bouzaid, E., Branco, M.J., Buegler, J., Kambosos, M., Kneppers, S., Kriiska-Maivali, K., Mattei, A., & Zatkalikova, L. (2024). ENFSI 2022 multidisciplinary collaborative exercise: organisation and outcomes. *Forensic Science International: Synergy* 8.

Best Practice Manual for Digital Audio Authenticity Analysis, ENFSI-FSA-BPM-002, (2022), [online], Available: https://enfsi.eu/wp-content/uploads/2026/01/FSA-BPM-002_Digital-Audio-Authenticity-Analysis-FIXED.pdf

ENFSI Guideline for Evaluative Reporting in Forensic Science (2015). [online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf

Voice familiarity across speaking styles: Effects on speech intelligibility and voice recognition

Alessandra Metzger[1], *Elisa Pellegrino*[1], *and Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
{alessandra.metzger|elisa.pellegrino|volker.dellwo}@uzh.ch

Within-speaker variability arising from different speaking styles can challenge unfamiliar speech and voice perception (Pellegrino, et al., 2023). In contrast, familiarity with a speaker's voice enhances speech intelligibility (Nygaard, et al., 1994) and supports voice recognition (Lavan, Burston, et al., 2019). Crucially, exposure to within-speaker variability may itself strengthen the development of robust voice familiarity (Lavan, Knight, et al., 2019). This study examines how a highly variable style, infant-directed speech (IDS), modulates the effects of voice familiarity on speech intelligibility and voice recognition. Understanding how speaking style affects listeners' ability to recognise and understand a speaker – and how this interacts with voice familiarity – is crucial for forensic transcription and evaluating earwitness evidence, where style mismatch is common.

40 British English-speaking adults completed an online perceptual learning experiment across four days. Participants were familiarised with a single speaker in either IDS or adult-directed speech (ADS) through a content-based listening task. Following each session, they completed a speech-in-noise transcription task containing utterances from the familiar speaker and four unfamiliar speakers, presented in both IDS and ADS. On the final day, participants additionally completed a yes–no voice recognition task including the familiar voice and five unfamiliar voices in both speaking styles.

Transcription results partially replicated the familiar-voice intelligibility benefit: in ADS, utterances from the familiarised speaker (sp09) were transcribed more accurately than those from the unfamiliar speakers (Figure 1). However, a follow-up experiment without familiarisation revealed that the target speaker of the first experiment was overall clearer in ADS (Figure 2).

Crucially, transcription accuracy for the familiarised voice improved across sessions regardless of familiarisation or test style. Exposure to the target speaker in either IDS or ADS led to comparable gains in intelligibility over time, and familiarity acquired in one speaking style supported intelligibility in both matched and mismatched styles. This indicates that listeners can generalise voice familiarity across speaking style variability in transcription tasks.

Voice recognition results revealed a different pattern (Figure 3). Recognition performance showed an interaction between familiarisation and test style: ADS familiarisation improved recognition accuracy, but only when testing also occurred in ADS. IDS familiarisation produced no recognition benefit. This suggests that voice recognition is more sensitive than transcription to style mismatch and to the high variability characteristics of IDS, at least under the relatively brief familiarisation used in this study.

Overall, the findings demonstrate that both familiarity and speaking style shape listeners' ability to understand and recognise voices, with distinct implications for forensic tasks involving intelligibility versus identity judgments. On the one hand, the robust improvement in transcription accuracy associated with increased exposure to a speaker's voice indicates that it is advisable for forensic transcribers to consult additional material from the target speaker, even when it differs in speaking style from the questioned recording. On the other hand, voice recognition is influenced by both the inherent variability of the speaking style and the match between familiarisation and test style, underscoring the need to account for these factors when evaluating the reliability of earwitness evidence.

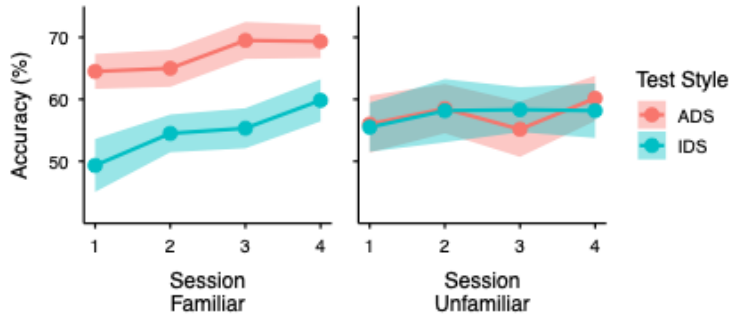


Figure 1. Transcription accuracy across sessions for the familiar and unfamiliar speakers, with test style indicated by colour.

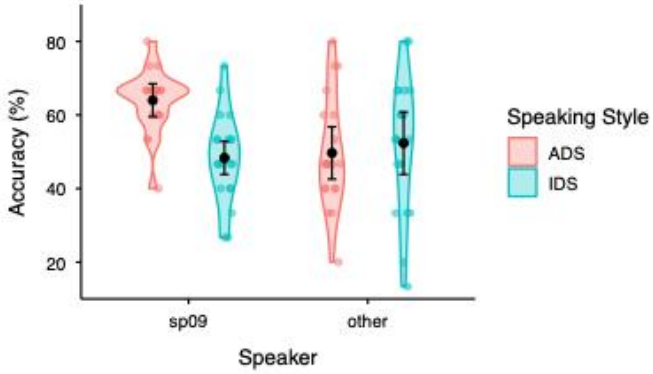


Figure 2. Transcription accuracy in the follow-up experiment for the target speaker (sp09) and four additional speakers, with speaking style indicated by colour.

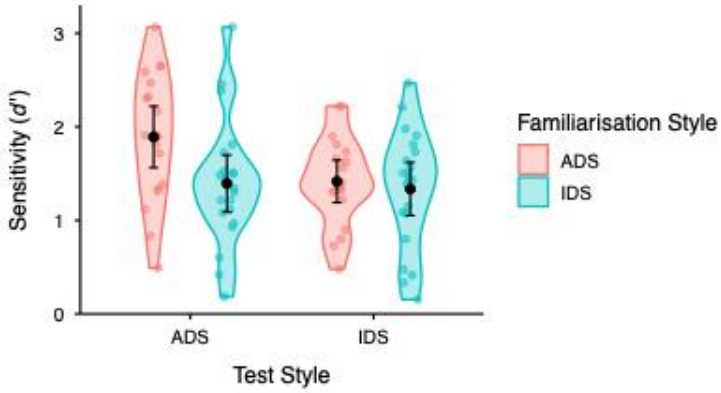


Figure 3. Voice recognition performance as a function of familiarisation style and test style.

References

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British journal of psychology, 110*(3), 576–593. https://doi.org/10.1111/bjop.12348

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition, 193*, 104026. https://doi.org/10.1016/j.cognition.2019.104026

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological science, 5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Pellegrino, E., Asadi, H., & Dellwo, V. (2023). Voice discrimination across speaking styles in Persian. In R. Skarnitzl & J. Volín (Eds.), *Proceedings of the International Congress of Phonetic Sciences*. International Phonetic Association. https://guarant.cz/icphs2023/952.pdf

Forensic Approaches to Voice Perception: Acoustic Variation Across Gender-Diverse Vowel Spaces

Katelyn Taylor and Kirsty McDougall
Department of Theoretical and Applied Linguistics, University of Cambridge, UK
klt56@cam.ac.uk; kem37@cam.ac.uk

Phonetic and sociolinguistic research has long examined how sex influences speech production, yet most studies and reference datasets focus on cisgender speakers, overlooking gender-diverse voices (Zimman, 2018; 2021; Merritt, 2023). This limitation has implications for forensic speaker comparisons involving gender-diverse speakers, alongside the social perception of voices more broadly where assumptions about binary gender are often implicit.

The present research firstly addresses this gap by investigating listener perception of gender identity in Southern Standard British English (SSBE). 61 listeners assessed scripted and spontaneous speech stimuli produced by 18 speakers (6 cisgender, 6 non-binary, 6 transgender). Participants additionally selected the phonetic features they relied on when making their judgements. Results indicated that listeners were generally able to distinguish cisgender from gender-diverse speakers but struggled to attribute specific non-binary or transgender categories correctly, highlighting the systematic categorisation of gender-diverse voices as broadly "gender queer", as can be seen in Figure 1.

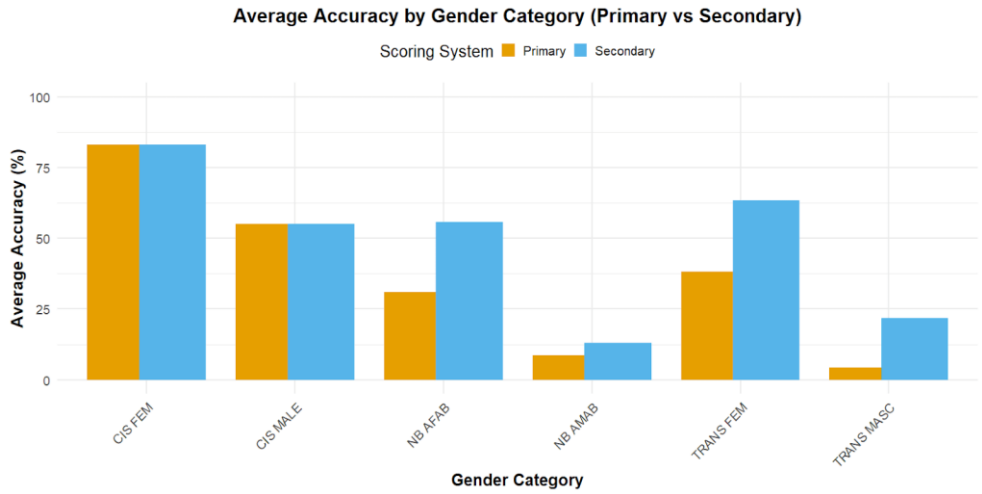


Figure 1. Bar graph showing how often speakers' gender identities were correctly attributed. The Primary scoring system refers to the exact correct gender identity being selected by the listener, and the secondary scoring system refers to speakers being attributed simply as cisgender or gender queer.

Building on these findings, the study also examines vowel space variation across cisgender, non-binary, and transgender SSBE speakers. Twenty-four speakers produced the Rainbow Passage and a reading passage designed to elicit monophthongal vowels in /hVd/ contexts. Recordings were Montreal force-aligned to obtain segmental boundaries, and vowel tokens were measured for dynamic formant values (F1, F2, F3 at 30%, 50%, and 80% of duration). Crucially, the analysis considers both raw (non-normalised) and normalised data. While normalisation is commonly used to reduce anatomical variation, particularly differences in vocal tract length, it remains unclear whether this process may obscure sociophonetic variation relevant to gender identity. This study examines the extent to which normalisation affects observed vowel space patterns, and whether meaningful differences between speaker groups are reduced or preserved through this process.

Preliminary analysis of a subset of six speakers (4 cisgender male, 2 non-binary assigned male at birth (NB AMAB)) is presented in Table 1 and Figure 2. Table 1 illustrates vowel space area (VSA),

calculated using raw data for the corner vowels (/i, æ, ɔ, ɑ/) to approximate each speaker's acoustic vowel space. Figure 2 presents a complementary measure of vowel space peripherality, calculated for each speaker as the mean Euclidean distance of each vowel from that speaker's own schwa centroid, also based on the raw data collected. Together these measures provide insight into both the size and dispersion of individual vowel systems.

Speaker	*Gender Identity*	*VSA (raw data)*
Speaker 5	Cisgender Male	214,168 Hz²
Speaker 3	Cisgender Male	152,691 Hz²
Speaker 1	Cisgender Male	135,632 Hz²
Speaker 2	NB AMAB	129,791 Hz²
Speaker 4	Cisgender Male	113,864 Hz²
Speaker 6	NB AMAB	107,569 Hz²

Table 1. VSA in descending order for six speakers, comprising four cisgender males and two non-binary speakers assigned male at birth. F1 and F2 frequency values were obtained for each of the corner vowels and used to construct a quadrilateral vowel space.

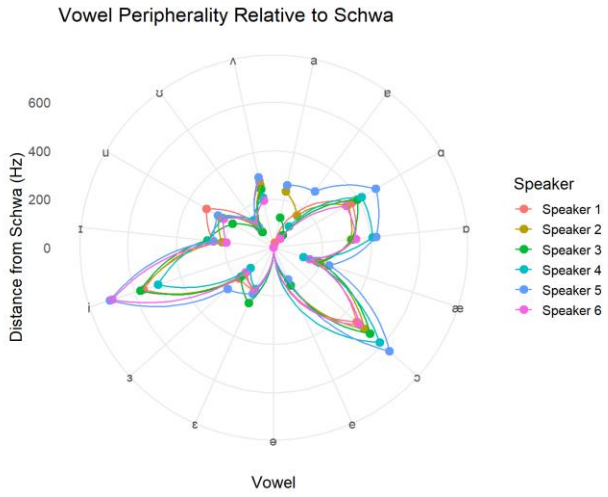


Figure 2. Mean vowel space peripherality, calculated as the average Euclidean distance of each vowel from the speaker-specific schwa centroid in the F1-F2 plane using the raw data.

These results will be used to assess whether vowel space configurations cluster according to sex assigned at birth and/or self-identified gender, and whether gender-diverse speakers display vowel space patterns that diverge from binary norms in terms of size, shape, and vowel positioning. This research contributes to understanding the phonetic representation of gender-diverse voices and informs the construction and interpretation of forensic speaker comparison datasets, ensuring they better reflect social and biological diversity. Ultimately, this has direct implications for forensic practice, where reliance on cisnormative reference data may lead to a mischaracterisation of speakers, affecting evidential weight, speaker comparison outcomes, and the overall fairness and reliability of forensic analyses and formant-based evidence in real-world casework.

References

Merritt, B. (2023). Speech beyond the binary: Some acoustic-phonetic and auditory-perceptual characteristics of non-binary speakers. *JASA Express Letters*, *3*(3).

Zimman, L. (2018). Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, *12*(8), e12284.

Zimman, L. (2021). Gender diversity and the voice. In *The Routledge handbook of language, gender, and sexuality* (pp. 69-90). Routledge.

Perceived voice distinctiveness predicts speaker discrimination but is not described by acoustic voice spaces

Alessandro De Luca[1]*, and Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zurich, Switzerland*
{alessandro.deluca|volker.dellwo}@uzh.ch

The prototype model of voice characteristics predicts that voice distinctiveness is positively correlated with the distance of a voice to the so-called prototype voice (Latinus et al., 2013). The prototype model has also been proposed to explain the process of speaker identification, using a relatively small acoustic feature space (Lavner et al., 2001). It is plausible that a relationship between perceived voice distinctiveness and speaker identity perception exists under the prototype model. Indeed, distinctiveness has already been shown to improve performance in both speaker discrimination and face-voice matching tasks (Stevenage et al., 2017; Stevenage et al., 2018), as well as improving reliability of earwitness' descriptions of voices (Yarmey, 1991). This relationship has so far only been tested with voices represented by controlled experimental stimuli. We were interested in analysing whether such relationship holds with more varied speech stimuli, as well as when considering the distinctiveness of both voice samples in a discrimination trial instead of the distinctiveness of only one voice while keeping the other voice's distinctiveness fixed. Additionally, we explore the relationship between distinctiveness and distance-to-mean in different voice spaces (set of acoustic features and DNN speaker embeddings), which has so far only been analysed using single word representations of voices.

Using a randomly selected sample of naturalistic stimuli from the CSC corpus (Hirschberg et al., 2005) (Standard US English native speakers, N=20, 11:9 M:F speakers ratio) we tested, in a between-listener design (18-35 years old, US English native speakers), one group (N=96) on distinctiveness of 240 spontaneous speech samples in 800 ms frames (Hanning weighing) extracted from continuous utterances and one group (N=78) on their discrimination ability of the same samples in pairs.

Results

We find that sample distinctiveness predicts accuracy in a speaker discrimination task, but that while high distinctiveness aids speaker discrimination in different speaker trials, it does not in different speaker. Furthermore, the relative difference in distinctiveness between the two voice samples presented in a trial is highly influential in determining the probability of receiving a correct response for the discrimination trial. We also report a strong influence of speaker sex on the discrimination performance, with a 40% and 34% reduction in the odds of a correct response in same and different speaker trials respectively, for male speakers compared to female speakers. These results are summarised in Figure 1.
Analysing the relationship between distinctiveness and distance-to-mean in several voice spaces' topologies, we only find a significant weak correlation of median sample distinctiveness with distance-to-mean in a voice space comprised of median F0, HNR, formant dispersion, and voicing ratio, while no correlation is found for all other speaker embedding spaces (ECAPA-TDNN, x-vectors, ResNet50).

Conclusion

Our results provide evidence that perceived voice distinctiveness and discrimination ability of variable speech utterances in two separate groups of listeners correlate well with each other, but that the interaction between distinctiveness of the two voices in the trial is of even more importance.

We can only partially confirm previous findings on the relationship between distinctiveness and the prototype model also in varied stimuli.

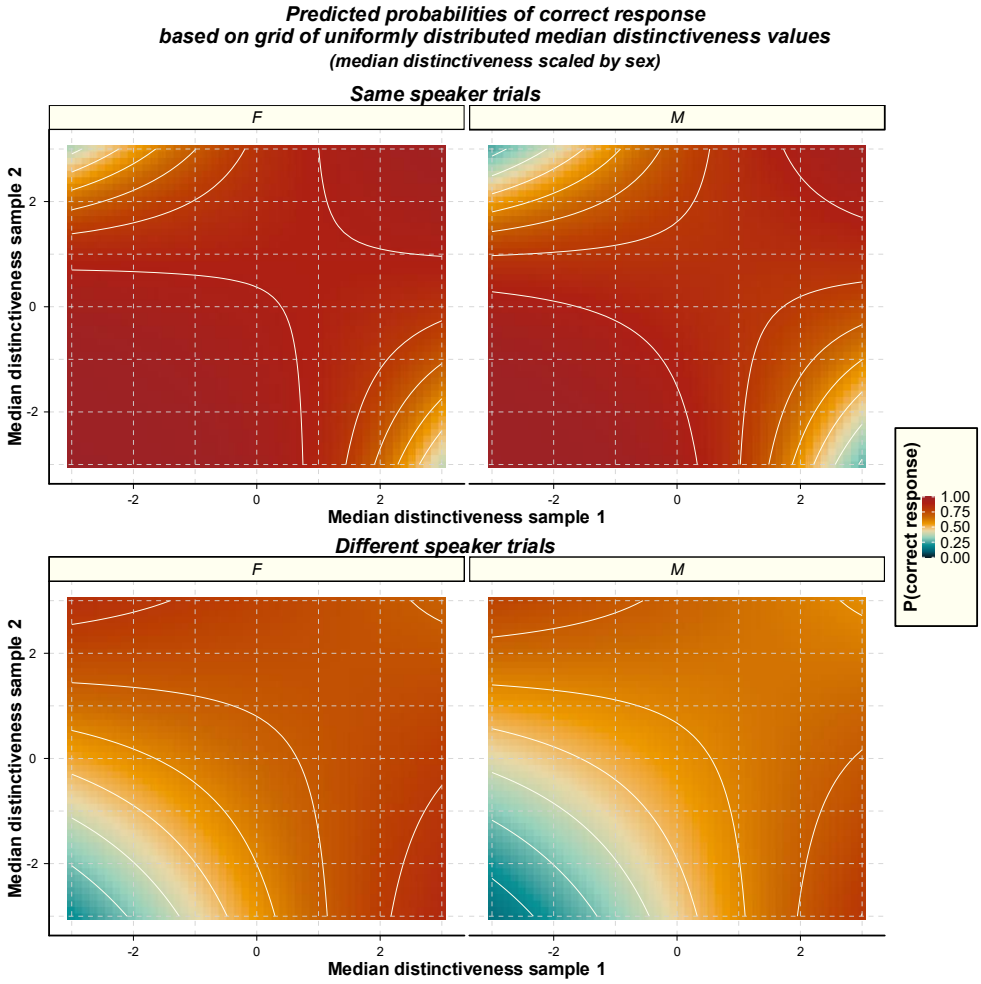


Figure 1. Predicted probability of a correct response (colour gradient) for same (top) and different (bottom) speaker trials split by sex (F: left , M: right) by median distinctiveness of the first sample (x-axis: range [-3sd, +3sd]) and median distinctiveness of the second sample in the speaker discrimination trial (y-axis: range [-3sd, +3sd]. (Median distinctiveness scores are z-scored separately for each sex group).

References

Hirschberg, J. B., Benus, S., Brenier, J. M., Enos, F., Friedman, S., Gilman, S., Girand, C., Graciarena, M., Kathol, A., Michaelis, L., Pellom, B., Stolcke, A., Shriberg, E. (2005). Distinguishing deceptive from non-deceptive speech. doi: 10.7916/D8697C06

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*. 23, 1075-1080. doi: 10.1016/j.cub.2013.04.055

Lavner, Y., Rosenhouse, J., Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology.* 4(1), 63-74. doi: 10.1023/A:1009656816383

Stevenage, S. V., Neil, G. J., Parson, B., Humphreys, A. (2018). A sound effect: exploration of the distinctiveness advantage in voice recognition. *Applied Cognitive Psychology*. 32(5), 526-536. doi: 10.1002/acp.3424

Stevenage, S. V., Hamlin, I., Ford, B. (2017). Distinctiveness helps when matching static faces and voices. *Journal of Cognitive Psychology*. 29(3), 289-304. doi: 10.1080/20445911.2016.1272605

Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Science Society*. 31(4), 421-428. doi: 10.1016/S0015-7368(91)73183-6

The impact of accent imitation on voice identity perception

Emilia Fletcher, Justin J. H. Lo
Linguistics and English Language, Lancaster University, UK
j.h.lo@lancaster.ac.uk

Introduction

Earwitness testimony relies on the ability of listeners to accurately recognise voices, which may be hindered by considerable natural variability of the voice within individual speakers (Lavan, Burston & Garrido, 2019). Speakers may also engage in deliberate disguise to conceal their identity, such as by adopting specific voice qualities or imitating an accent other than their own (Eriksson 2010). Previous studies have found much variability in the ability of listeners to detect imitated accents (Goodman et al., 2024; Markham, 1999), but the extent to which this is related to the perception of the speaker's identity remains under-investigated. In this study, we address the following research questions: (1) How well can listeners detect an accent as natural or imitated? (2) How does accent imitation impact the ability of unfamiliar listeners to perceive voice identity?

Methodology

35 L1 British English participants completed a two-stage online experiment. Participants received a PowerPoint slide with 18 audio stimuli, each containing an English utterance from a mock scam call recorded by one of three female trained actors. Each speaker was represented by two stimuli in their own accent and two stimuli each in imitation of the other two speakers' accent. In the first stage, participants were told that speakers may be disguising their voice, but not the form of disguise or the number of speakers, and asked to freely sort the stimuli into clusters on the PowerPoint slide based on the perceived voice identity. In the second stage, accent was revealed as the form of disguise and participants were asked to classify each stimulus as in a 'natural' or 'imitated' accent.

To investigate how listeners process voice identity from variable stimuli, we analysed participants' ability to 'tell people together' and 'tell people apart' (Lavan, Burston & Garrido, 2019; Lavan et al., 2019), by comparing how often a speaker's own-accent and imitated stimuli were clustered together with their own (*telling together*) or those of other speakers (*telling apart*). Based on the participants' classifications, we calculated measures of sensitivity (*d'*) and bias (*c*) from signal detection theory to analyse their performance in detecting accent imitation.

Results

All but one participant overestimated the number of speakers present in the stimuli (mean: 6.37, range: 3–11). Participants were generally better at telling different speakers apart than telling the same speaker together, and performance was significantly worse in both aspects for imitated stimuli than for own-accent stimuli. *d'* obtained from accent classifications was not significantly different from 0 (0.08, 95% CI [-0.04, 0.20]) while *c* was positive (0.21, 95% CI [0.15, 0.27]), indicating that participants were not able to distinguish between natural and imitated accents and had an overall bias to judge an accent as imitated. Further, there was no evidence that performance in the two tasks was correlated. Together, these findings demonstrate the challenges that accent imitation, as a form of disguise, poses to the formation of reliable speaker percepts for unfamiliar listeners. Variability in performance between voices and possible acoustic-phonetic contributors will be discussed.

References

Eriksson, A. (2010). The disguise voice: Imitating accents or speech styles and impersonating individuals. In C. Llamas & D. Watt (Eds), *Language and identifies* (pp. 86–96). Edinburgh University Press.

Goodman, J. R., Crema, E., Nolan, F., Cohen, E., & Foley, R. A. (2024). Evidence that cultural groups differ in their abilities to detect fake accents. *Evolutionary Human Sciences*, 6, e46:1–18.

Lavan, N., Burston, L. F. K., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, 110, 576–593.

Lavan, N., Burton, A. M., Scott, S. K., McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26, 90–102.

Markham, D. (1999). Listeners and disguised voices: The imitation and perception of dialectal accent. *The International Journal of Speech, Language and the Law*, 6(2), 289–299.

Automatic identification of dialects in the western part of Germany

Vatho Chem[1] *and Reinhold Greisbach*[2]
[1]*Royal Academy of Cambodia, Linguistics Committee, Phnom Penh, Cambodia*
chemvatho@gmail.com
[2]*Institute of Linguistics – Phonetics, University of Cologne, Germany.*
reinhold.greisbach@uni-koeln.de

Until around 150 years ago, most people in Germany grew up speaking only their local dialect until they started school, where they were required to learn Standard German. As Georg Wenker's introduction to his little book impressively demonstrates, a knowledgeable linguist could, at that time, determine the speaker's place of origin to within a few kilometres simply by listening to the pronunciation of a short utterance (Wenker 1877, p.3).

Nowadays, however, hardly anyone in Germany grows up speaking only their local dialect; the influence of the media is too strong, people's mobility is virtually limitless, and through early childhood care and education, children – whose parents may still speak a dialect – come into contact with Standard German from a very young age. Nevertheless, the question arises as to whether it is still possible today, perhaps with technical assistance, to pinpoint a speaker's location based on their pronunciation, provided they still speak their local dialect.

To this end, we draw on the West German language corpus, which was compiled several years ago and is freely available online (Spiekermann et al. 2016ff). As part of the recordings for this language corpus, the 40 sentences from Wenker's second survey (1879–1884) were also recorded in the local dialect, particularly by older speakers. To this end, the sentences were presented in Standard German to the speakers, who were asked to translate them spontaneously into their local dialect.

For our exploratory study, we will limit ourselves to the first sentence of Wenker's corpus, which goes in Standard German: "Im Winter fliegen die trockenen Blätter durch die Luft herum." {In winter, dry leaves flutter through the air.} For the automatic dialect recognition we use a two step hybrid model:

1st step neural net based: Input: audio signal → Output: transcription symbol sequence of the utterance. 2nd step rule based: Based on the recorded dialectal pronunciations depicted in the Wenker maps (Wenker, 1889–1923), which date from the era of monolingual dialect speakers, a set of rules is established that makes it possible to pinpoint the speaker's location (simulating Wenker's auricular-phonetic approach of 1877).

For our experiment we classify the dialects of the region into 10 dialect groups in accordance with standard conventions (Cornelissen & Denkler 2018): Rhine Franconian, Moselle Franconian, Ripuarian, Limburgish German (i.e. Düsseldorf area), Kleverlandish, Bergish, West Münsterlandish, Münsterlandish, South Westphalian and East Westphalian. A total of 1,193 audio signals (one per speaker) drawn from 647 villages in the area are available for the first sentence.

In the hybrid model, a sequence of phonetic symbols is first extracted from the audio signals. The German language model "jonatasgrosman/wav2vec2-large-xlsr-53-german" (Grosman 2021) identifies words in Standard German with near-perfect accuracy and uses "Wav2Vec2ForCTC" (Wolf et al. 2020) to generate a phonetic transcription in IPA. It can also process audio signals that do not conform to the standard, such as those typical of dialects. However, the results are of significantly poorer quality, as the model was trained only on Standard German. Feeding it with dialectal words leads to irregular words and wordforms, which approximates the pronunciation of the dialect speakers in terms of the sound inventory of a Standard German speaker on a phonetic

level. Though no dialectal word will be recognized, this feature makes it possible to subsequently search for dialect markers within these automatically generated symbol sequences.

The outcome of this process was checked manually. No symbol of the automatic process was changed, added or deleted, but spaces were sometimes introduced manually at word boundaries, so that a wordwise comparison between the utterances of the speakers became possible. Word forms that identify a dialect or group of dialects as unambiguously as possible are used as markers. For example, [flaɪçt] {(they) flutter} is mainly used in Westphalian dialects, but not in the Rhineland. Conversely, [dʁyʃ] {dry} is found only in the Rhineland.

The results depend heavily on the markers selected and even more on their number. For the combination of 12 markers chosen here (only from the first sentence of Wenker), there was "no result" in 18.2% of cases, i.e. none of the markers were present in the symbol sequences of the specific speakers. This could be because the speakers spoke close to Standard High German, and thus did not exhibit a single dialect marker. In most cases, however, it was due to poor automatic transcription, which can be attributed to poor recording conditions (background noise and disturbing room acoustics), inaccurate pronunciation by the speakers (due to the advanced age of most dialect speakers, who, judging by the quality of their voices, are almost all over 60 years old ~ and existing language models are not trained primarily to the pronunciation and voice quality of older people), whether the speakers translate word-for-word or freely, whether the speakers are searching for a dialect word while translating and thus hesitating, repeating or stretching their pronunciation and the extent to which the dialect word deviates from Standard German.

For the remaining 976 recordings, the accuracy rate for correct dialect identification was 47.5% (compared to a random probability of 10%), whilst the probability that the correct dialect was among the top three estimated candidates was 86.0%. This means at least a broad location into the two broad groups Low German (Kleverlandish + Westphalian group) vs. Central German (Rhenian fan dialects from Limburgish German to Rhine Franconian) is possible with this procedure, based only on one single sentence.

It has become apparent that our system is not yet capable of automatically localizing speakers reliably. The primary reason for this limitation lies in the poor quality of the automatic transcriptions. If manual transcription were used in the first step, it would be possible to achieve a robust and even more detailed localisation in some regions. At present, the method is therefore not yet suitable for forensic applications. However, substantial improvements can be expected in the future, particularly through advances in speech recognition technology and the development of specialized automatic dialect recognition systems (Chem & Greisbach, to come).

References

Chem, V. & Greisbach, R. (to come): KI-Based Automatic Transcription of the Ripuarian Dialect of German. Talk to be presented at the 11th Congress of the International Society for Dialectology and Geolinguistics, Marburg, September 2026.

Cornelissen, G. & Denkler, M. (2018): Die Dialekte von Nordrhein-Westfalen. Münster

Grosman, J. (2021): *wav2vec2-large-xlsr-53-german*. Hugging Face. [URL: https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-german]

Spiekermann, H. H., Tophinke, D., Vogel, P. M. & Wich-Reif, C. (Eds.) (2016ff.): Dialektatlas Mittleres Westdeutschland (DMW). Siegen: Universität Siegen [URL: https://www.dmw-projekt.de]

Wenker, G. (1877): Das rheinische Platt. Düsseldorf. Self-publishing.

Wenker, G. (1889–1923): Sprachatlas des Deutschen Reichs. Handgezeichnetes Original von Emil Maurmann, Georg Wenker und Ferdinand Wrede. Marburg. [Published as Digital Wenker-Atlas (DiWA); URL <https://www.regionalsprache.de>].

Wolf, T., Debut, L., Sanh, V. et al. (2020): Transformers: State-of-the-Art Natural Language Processing. EMNLP.

The role of temporal information in automatic speech and voice recognition

Erdem Baha Topbas[1], *Srikanth Madikeri, Elisa Pellegrino*[1], *Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zürich, Zürich, Switzerland*
{erdembaha.topbas|srikanth.madikeriraghunathan|elisa.pellegrino|
volker.dellwo}@uzh.ch

Speech signals contain three types of temporal information—amplitude envelope (ENV), temporal fine structure (TFS), and periodicity (PER)—which differ in their dominant fluctuation rates and in how they encode prosodic and segmental contrasts (Rosen, 1992). The periodicity captures the zero-crossing information contained in the waveform, the envelope captures signal intensity across the time domain, and the temporal fine structure focuses on the fluctuations of high rates. The role these cues play in speech intelligibility has been partially researched (Licklider & Pollack, 1948; Shannon, et al, 1995). These works do not explore speaker recognition and automatic speech recognition. This project conducts a systematic analysis to address this gap in the literature.

We utilize systematically distorted signals comprising eight conditions: three versions retaining only one temporal property (ENV, TFS, or PER), three retaining two properties (ENV+TFS, ENV+PER, PER+TFS), and conditions in which all properties were either fully preserved or fully removed (creating a square wave). Signal manipulations included Hilbert transformation, infinite-peak clipping, and a novel method termed isochronous time warping, which alters signal periodicity.

The speaker verification and classification results were explored by training a ResNet18 model and a series of Gaussian Mixture Models (GMMs) respectively. They were trained on clean and undistorted speech from the training subset of the TIMIT corpus (Garofolo et al., 1993) and tested on the test subset with the conditions outlined above. ResNet18 performance was assessed using equal error rate (EER). In the speaker verification task, performance was highest when all features were present and lowest when all were removed. Notably, removing periodicity alone produced the only statistically significant degradation, with EER rising from 17.2% to 48.3% ($p = 0.0006$). Analysis further revealed that while removing ENV or TFS individually had no significant effect, removing both together caused a meaningful drop in performance, with EER increasing from 6.3% in the baseline to 29%.

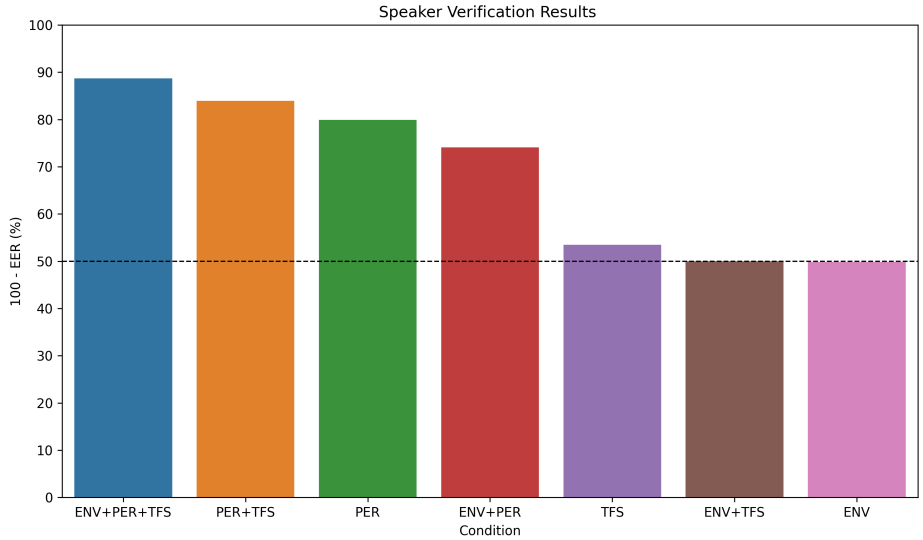


Figure 1: Speaker recognition results across temporal information conditions.

These results can be seen in Figure 1. The GMM classification task yielded a more categorical result: any removal of temporal information—regardless of which property or combination was affected—produced a significant drop in classification accuracy.

Speech intelligibility results were conducted using a pre-trained OpenAI Whisper Automatic Speech Recognition (ASR) model (Radford et al., 2022). This model was given utterances from the Audio MNIST corpus (Becker et al., 2024), distorted in the same condition as above. The corpus contains digits 0-9 being read by a variety of speakers. The transcripts acquired by Whisper were either directly or fuzzily matched to the digit options to get accuracy metrics. The fuzzy matching was done by applying the metaphone algorithm (Philips, 1980) to both the transcripts and the possible outputs to account for homophones and were matched by picking the candidate with the smallest Levenshtein distance (Levenshtein, 1965). The highest performance was obtained with the unmodified utterances, with a transcription accuracy of 96.5%. The ENV, TFS, and ENV+TFS conditions all performed at chance level (~10%). Taking the ENV+PER+TFS case as baseline, the mixed linear model results indicated statistically significant differences from the baseline in all distorted cases.

Ranking the conditions from best to worst in terms of EER and transcription accuracy shows a similar ordering, with an intraclass correlation coefficient (2,1) score of 85%. This demonstrates their agreement on PER being the most important type of temporal information, with important implications in the design of robust models in both voice and speech recognition, by identifying possible data augmentation methods, and adjustments model architecture. These results show that PER is the most reliable feature, hinting that the models might be over-reliant on them. Therefore, training that involves reducing the emphasis on PER can increase accuracy forensic uses.

References

Becker, S., Vielhaben, J., Ackermann, M., Müller, K.-R., Lapuschkin, S., & Samek, W. (2024). AudioMNIST: Exploring Explainable Artificial Intelligence for audio analysis on a simple benchmark. *Journal of the Franklin Institute*, 361(1), 418–428. https://doi.org/10.1016/j.jfranklin.2023.11.038

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus (LDC93S1)* [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/17gk-bn40

Levenshtein, V. I. (1965). Dvoichnye kody s ispravleniem vypadeniy, vstavok i zameshcheniy simvolov [Binary codes capable of correcting deletions, insertions, and substitutions]. Doklady Akademii Nauk SSSR, 163(4), 845–848. http:// mi.mathnet.ru/dan31411 https://doi.org/10.48550/arXiv.2212.04356

Licklider, J. C., & Pollack, I. (1948). Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *The Journal of the Acoustical Society of America*, 20(1), 42–51. https://doi.org/10.1121/1.1906346

Philips, L. (1990). *Hanging on the metaphone*.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022, December 6). *Robust speech recognition via large-scale weak supervision*. arXiv.org.

Shannon, R.V. et al. (1995). Speech Recognition with Primarily Temporal Cues.*Science270*,303-304. doi:10.1126/science.270.5234.303

Rosen, S. (1993). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philos Trans R Soc Lond B Biol Sci.,* 336(1278):367-73. doi: 10.1098/rstb.1992.0070. PMID: 1354376.

Empirical Validation of Synthetic Speech as a Forensic Proxy for Mismatch and Recording-Condition Research

Peter Milne[1], *Colleen Driscoll*[1], *Emilie Lawrie-Munro*[1]
[1]*Audio & Video Analysis Unit, Technical Investigation Services, Royal Canadian Mounted Police, Ottawa ON, Canada*
{peter.milne|colleen.driscoll|emilie.lawrie-munro}@rcmp-grc.gc.ca

Forensic voice comparison (FVC) routinely requires understanding how mismatches in channel, language, and recording conditions influence the behaviour of speaker comparison systems [6, 7]. However, conducting controlled mismatch studies using authentic speech alone is often impractical: large, balanced datasets spanning multiple channels and languages are rare, and collecting them is costly, time- consuming, and ethically constrained. Modern neural text- to- speech (TTS) systems offer a potential solution by generating high- quality synthetic speech that approximates the acoustic characteristics of a target speaker [3, 4]. If synthetic speech can reproduce the relative patterns of similarity and degradation observed in authentic speech [5], then synthetic datasets could support controlled, reproducible research on mismatch effects.

This study evaluates whether synthetic speech behaves in ways that make it suitable as a proxy for research on channel and language mismatch in FVC. A multilingual synthetic speech corpus was generated for 426 authentic speakers using the XTTS neural TTS model [4]. For each authentic recording, a matched synthetic counterpart was produced using identical text and channel- matched reference conditioning. Authentic and synthetic recordings were processed using an x- vector/PLDA backend [1, 2, 9] adapted and calibrated on disjoint speaker sets. Three comparison systems were evaluated: authentic–authentic (AA), synthetic–synthetic (SS), and authentic–synthetic (AS). The evaluation set consisted of 68 held- out speakers; the remaining 358 speakers were used for backend adaptation and calibration. All systems were examined across SameChannel, CrossChannel, and CrossLanguage conditions in English and French.

Across both languages, AA SameChannel scores exhibited a consistent channel hierarchy (Microphone > Telephone > Bug > Cellphone). SS reproduced this hierarchy in English and showed a closely related pattern in French, with microphone highest and cellphone lowest in both languages. SS scores were shifted downward and showed reduced variance relative to AA, reflecting the more uniform acoustic characteristics of synthetic speech. AS scores, included for completeness, showed further compression but preserved the same relative ordering. These results indicate that synthetic speech preserves the relative channel- dependent structure observed in authentic speech.

CrossChannel comparisons revealed monotonic degradation patterns consistent with expectations based on channel quality. In English AA, microphone- to- bug comparisons yielded the highest scores, followed by microphone- to- telephone and microphone- to- cellphone. SS and AS reproduced this monotonic structure with similar relative spacing. French AA showed a closely related pattern (M–B > M–C > M–T), and SS and AS again preserved the ordering. No inversions or unexpected pairings were observed in either language or system. These findings suggest that synthetic speech responds to channel mismatch in a manner that mirrors the relative behaviour of authentic speech.

CrossLanguage comparisons (English- to- French) showed the expected drop relative to within- language comparisons. This effect appeared in AA, SS, and AS, and the relative spacing between channels was preserved across systems. Although SS and AS exhibited compressed scales, the structure of the

cross- language effect remained stable: channels that degraded more in within- language AA also degraded more in cross- language AA, and the same relationships appeared in SS and AS.

A consistent variance pattern was observed across languages and systems: Var(AA) > Var(SS) > Var(AS). AA exhibited the largest within- speaker variability, reflecting natural variation in authentic recordings. SS showed reduced variance, consistent with synthetic speech's more uniform acoustic properties. AS showed the lowest variance, reflecting both synthetic stability and cross- system mismatch. Despite these scale differences, the relative variability structure was preserved across SameChannel, CrossChannel, and CrossLanguage conditions.

Speaker- level analyses further supported these findings. For each speaker, correlations between AA and SS scores were computed across all comparisons and within each condition subset. English correlations typically ranged from 0.40 to 0.75, with many speakers above 0.60. French correlations were more variable, with a median around 0.48, but showed the same structural trends: higher correlations in SameChannel conditions and lower correlations under channel mismatch. CrossLanguage correlations were often comparable to or higher than within- language correlations, reflecting the fact that both AA and SS degrade in parallel under language mismatch. Variance ratios showed that authentic speech exhibited greater within- speaker variability than synthetic speech for most speakers. Importantly, no evidence suggested that global AA–SS patterns were driven by a small number of outliers; instead, the correlations were broadly distributed across the population.

Overall, the results demonstrate that synthetic speech preserves key relative patterns of system behaviour across channels, languages, and mismatch conditions. Although synthetic speech differs from authentic speech in absolute scale and variability, the structural relationships between conditions are similar. These findings do not establish synthetic speech as a substitute for authentic data in forensic applications. However, they indicate that synthetic speech can reproduce the relative effects of channel and language mismatch and may therefore support controlled, reproducible research on mismatch effects—particularly in scenarios where authentic data are difficult to obtain or systematically manipulate.

References

[1] Snyder, D. et al. (2018). X-vectors: Robust DNN embeddings for speaker recognition, *ICASSP*.

[2] Prince, S. & Elder, J. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity, *ICCV*.

[3] Ren, Y. et al. (2021). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech, *ICLR*.

[4] Casanova, E. et al. (2023). XTTS: Multilingual Zero-Shot Text- to- Speech, *Coqui-AI Technical Report.*

[5] Todisco, M. et al. (2018). ASVspoof 2019: A large-scale evaluation of synthetic and converted speech. *INTERSPEECH*.

[6] Morrison, G. (2009). Forensic voice comparison, *Language and Linguistics Compass*.

[7] Gonzalez- Rodriguez, J. et al. (2007). Robustness in speaker recognition, *IEEE TASLP.*

[9] Kelly, F. et al. (2019). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors," *AES Forensics Conference*.

The inadequacy of statistically-calculated likelihood ratios in forensic voice comparison: the Italian case

Sonia Cenceschi[1], *Chiara Meluzzi*[2], *and Francesco Roberto Dani*[1]
[1]*Digital Forensic Service, Department of Innovative Technologies, SUPSI, Switzerland*
{Sonia.cenceschi|Francesco.dani}@supsi.ch
[2]*Department of Literary Studies, Philology and Linguistics, University of Milan, Italy*
Chiara.meluzzi@unimi.it

Forensic voice comparison has progressively oriented towards adopting the likelihood ratio (LR) as methodological standard, following ENFSI and international scientific community recommendations (Drygajlo et al., 2015; Gold & French, 2011; 2019). The LR framework encompasses distinct approaches: statistically calculated LRs based on reference population models, and verbal LRs expressed on calibrated scales without requiring statistical modelling (Willis et al., 2015; ISO 21043). This paper focuses specifically on the limitations of the former, while acknowledging that verbal LR frameworks represent a legitimate and, in several jurisdictions, widely adopted alternative.

Rigid and uncritical application of statistically-calculated LRs raises epistemological and practical questions meriting serious consideration. While countries such as the UK, Australia, and the Netherlands have integrated the framework more systematically, other contexts present a more complex picture. In Italy, the problem is compounded by a widespread practice of relying exclusively on automatic or semi-automatic speaker recognition software (e.g. Nuance Forensics, IDEM, Impavido) without systematic integration of expert phonetic analysis. This practice effectively reduces the LR to a black-box output applied to data that may not meet the underlying statistical assumptions. Such systems presuppose the availability of representative and statistically robust reference populations (Gonzalez-Rodriguez et al., 2006), a condition rarely met in Italian casework and, crucially, one that cannot be remedied within the constraints of real-case proceedings, where the construction of adequate reference corpora is neither practically nor temporally feasible (Morrison & Enzinger, 2019). The Italian linguistic landscape presents exceptional challenges in this regard: reference corpora must account not only for the profound phonetic divergence between northern and southern varieties, but also for the continuum of regional accents, contact varieties, and sociophonetic features that characterize spontaneous speech in forensic recordings. These dimensions are further compounded by mismatches in channel conditions, and in speech style, from read to spontaneous. When these assumptions are violated, systems produce outputs whose apparent numerical precision conceals significant methodological fragility, generating numerically precise but forensically misleading values (Morrison, 2011; Braun, 2021) that are then presented to courts as objective measures. The risk is particularly acute with ambiguous evidence, where weakly supportive or non-conclusive values may be erroneously interpreted as scientific proof, a failure that automated tools, by design, cannot flag.

This situation amplifies rather than mitigates the epistemological risks of the framework, making the debate particularly relevant for jurisdictions where automated tools have outpaced the development of critical methodological frameworks for their validation and contextualization. Traditional phonetic analysis, by contrast, incorporates deep linguistic knowledge about speaker variability, coarticulation, and idiosyncratic characteristics that statistical models cannot adequately capture (Rose, 2002), and remains indispensable precisely where automated approaches are most likely to fail. Empirical studies further show that judges and juries tend to misinterpret LR values, often treating them as direct probabilities of guilt, a well-documented cognitive bias known as the prosecutor's fallacy (Thompson & Schumann, 2017). However, more recent evidence from jurisdictions with longer experience of LR-based reporting, such as the Netherlands, suggests that professional judges' comprehension may be improving (Rassin et al., 2022), pointing to the role of institutional exposure and training in mitigating these communicative failures. This communicative failure compounds the technical limitations and undermines the reliability of LR-based testimony in court (Morrison et al., 2025). Such analyses are carried out by both court-appointed (CTU) and party-retained (CTP) experts across all stages of criminal proceedings, with highly variable professional backgrounds and no nationally standardised criteria for their qualification or appointment. Such communicative risks are further

compounded in Italy, which still lacks national guidelines for both the selection of forensic voice comparison experts and the standards for expert report writing, leaving significant discretion to individual practitioners and courts. Furthermore, the urgency of reaching decisions within trial deadlines may place excessive pressure on experts in real-case scenarios, particularly in Italophone judicial settings. In our casework experience, judges and law enforcement in Italian proceedings have shown a systematic preference for absolute quantification in expert reports, dispreferring qualitative or more nuanced evaluations, a tendency that further incentivizes the unreflective use of numerical LR outputs regardless of their methodological validity.

Rather than abandoning quantification entirely, we advocate for a hybrid framework combining structured expert evaluation with calibrated verbal scales (Marquis et al., 2016; Drygajlo et al., 2015) and transparent uncertainty reporting, consistent with the verbal LR approach endorsed by Willis et al. (2015) and ISO 21043, which allow LR-based evaluative reporting without requiring statistical models. Our contribution emphasizes the specific challenges faced in jurisdictions such as Italy, where the gap between international best practice and local forensic reality remains particularly pronounced. Obviously, no measure is "bad" or "good" per se. Like many other metrics in all fields of research, when applied to forensic settings it is mandatory to ascertain whether the conditions for using such metrics are maintained not as a general statement, but for that specific case on which an expert report has to be provided.

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.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. The International Journal of Speech, Language and the Law, 18(2), 293–307.

Gold, E., & French, P. (2019). International practices in forensic speaker comparison: Second report. The International Journal of Speech, Language and the Law, 26(1), 1–20.

Gonzalez-Rodriguez, J., Drygajlo, A., Ramos-Castro, D., Garcia-Gomar, M., & Ortega-Garcia, J. (2006). Robust estimation, interpretation and assessment of likelihood ratios in forensic speaker recognition. Computer Speech & Language, 20(2–3), 331–355.

Marquis, R., Biedermann, A., Cadola, L., Champod, C., Gueissaz, L., Massonnet, G., ... & Hicks, T. (2016). Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings. Science & Justice, 56(5), 364-370.

Morrison, G. S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. Science & Justice, 51(3), 91–98.

Morrison, G. S., Bali, A. S., Martire, K. A., Grady, R. H., & Thompson, W. C. (2025). What is the best way to present likelihood ratios? A review of past research and recommendations for future research. Science & Justice, 101342.

Morrison, G. S., & Enzinger, E. (2019). Score based likelihood ratio calculation for forensic voice comparison. In K. Jessen (Ed.), Speaker Classification (pp. 1–52). Springer.

Rose, P. (2002). Forensic speaker identification. CRC Press.

Rassin, E., Arbiyah, N., Boskovic, I., Otgaar, H., & Merckelbach, H. (2022). Likelihood ratios in psychological expert opinion, and their reception by professional judges. The International Journal of Evidence & Proof, 26(4), 325–341. https://doi.org/10.1177/13657127221119545

Thompson, W. C., & Schumann, E. L. (2017). Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. In Expert evidence and scientific proof in criminal trials (pp. 371-391). Routledge.

Willis, S., McKenna, L., McDermott, S., O'Brien, G., & Overill, R. (2015). ENFSI Guideline for Evaluative Reporting in Forensic Science. European Network of Forensic Science Institutes.

Braun, A. (2021). The notion of speaker individuality and the reporting of conclusions in forensic voice comparison. In Bernardasci, C., Dipino, D., Garassino, D., Negrinelli, S., Pellegrino, E., & Schmid, S. (Eds.), L'individualità del parlante nelle scienze fonetiche: applicazioni tecnologiche e forensi. Studi AISV, 8 (pp. 175–191).

Evaluation and Application of Joint Speech Enhancement and Automatic Speaker Recognition in Forensic Science

Wei Li[1], *Wenlin Huang*[1], *Weili Jiang*[1], *and Xin Chen*[1]
[1]Institute of Forensic Science, Ministry of Public Security, Beijing, China
liwei18@cifs.gov.cn

Recently, speech enhancement (SE) and speaker verification (SV) models have seen increasingly widespread application in forensic science. Particular attention has been directed toward generative SE, SV models, and the integrated front-end and back-end relationships within the forensic science (Wu, 2021). However, due to the incomplete understanding of their impact on speaker acoustic features and identity clues, the application of SE and forensic SV models in the judicial domain warrants further research. In this paper, we investigate three representative SE techniques—OM-LSA + MCRA (traditional statistical estimator)(Cohen & Berdugo, 2001), a representative discriminative deep acoustic SE algorithm-DCCRN (Hu, 2020), and a representative generative SE algorithm-FlowMatch-SE (Wang, 2025) on the preservation of identity features across various acoustic-based recognition models (ECAPA-TDNN (Desplanques, 2020), x-vector/PLDA (Snyder, 2018), and WavLM embedding vectors). The composition of the speaker recognition system and the integrated front-end/back-end evaluation framework are illustrated in Figure 1.

We created a large-scale 16kHz labeled speech dataset named LibriSpeech-New by mixing clean LibriSpeech utterances with QUT noise. The noises are collected from cafes, cars, households and streets to improve speech complexity and realism. Consisting of 100 million speech pairs, this dataset facilitates evaluations of speech enhancement and speaker verification. Figure 1 illustrates its construction workflow.

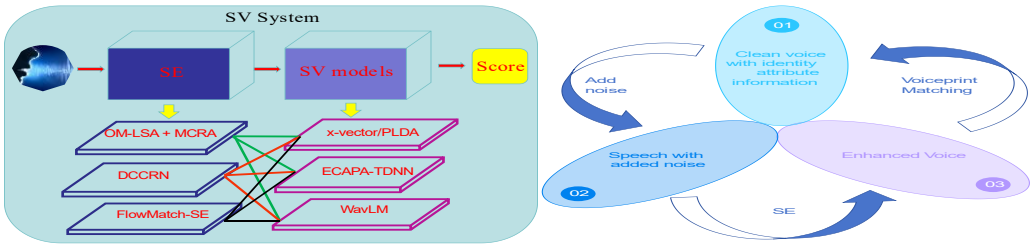


Figure 1. Speaker Recognition System Architecture（On the left is the Joint Front-End/Back-End Evaluation Framework，On the right is Schematic Diagram of Corpus Construction Process）.

This paper conducts research by combining objective evaluation metrics (Equal Error Rate, EER) and score distribution metrics (d'). The separation metric d' indicates how well coupled enhancement techniques and speaker verification models distinguish between the score distributions of the same speaker. A higher d' value signifies better distinguishability.

$$d' = \frac{\mu_{tar} - \mu_{non}}{\sqrt{\frac{1}{2}(\sigma_{tar}^2 + \sigma_{non}^2)}}$$

Here, μ_{tar} denotes the mean of the target score distribution S_{tar}, μ_{non} denotes the mean of the non-target score distribution S_{non}, σ_{tar}^2 denotes the variance of the target scores, and σ_{non}^2 denotes the variance of the non-target scores.

Front-end	*OM-LSA*	*DCCRN*	*FlowMatch*	*Overall*
d'(*x-vector/PLDA*)	3.38	-0.32	2.63	1.32227
d'(*ECAPA-TDNN*)	6.29	4.55	6.61	5.8430
d'(*WavLM*)	-0.0417	0.29	-0.26	0.0660

Table 1. Score Distribution Metrics Based on Different Speaker Verification Models.

As shown in Table 1, different SV systems (combining different SV models and different SE techniques) exhibit varying separation metrics. Compared to x-vector/PLDA and WavLM, ECAPA-TDNN generally demonstrates higher separation metrics. Furthermore, the SV system coupling ECAPA-TDNN with FlowMatch achieves the highest separation metric, signifying optimal capability in distinguishing whether audio samples originate from the same individual.

Figure 3 summarizes the standard performance of SV systems across front-end and back-end components. As shown in Figure 3, the performance of SV systems exhibits a strong coupling relationship with different front-end and back-end configurations. Among these, the Flowmatch-based SE outperforms both OM-LSA and DCCRN when deployed as the SV front-end across three distinct SV back-end scenarios. The ECAPA-TDNN-based SV backend outperforms x-vector/PLDA and WavLM embedding models.

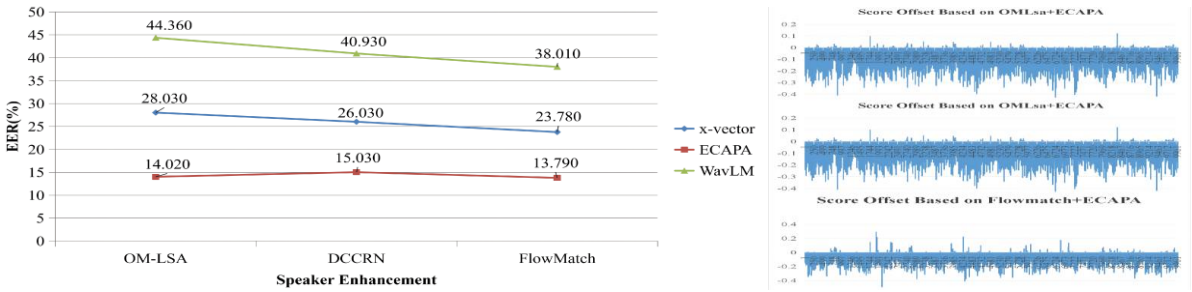


Figure 3. Performance of the Speaker Recognition System (The left figure shows the EER indicator status; The right figure shows Score Distribution Diagram for Different speaker recognition systems)

As shown on the right side of Figure 3, the three speech enhancement methods exert varying degrees of impact on speaker verification scores. OM-LSA+MCRA has the slightest influence on the scores, DCCRN has the greatest impact, and FlowMatch-SE falls in between. This indicates that generative speech enhancement methods generally retain more speaker identity feature cues than discriminative deep enhancement methods, yet their performance is less stable than conventional statistical enhancement methods.

Experiments demonstrate that generative augmentation typically preserves speaker identity clues better than discriminative deep augmentation, though it remains less conservative than traditional statistical augmentation. Simultaneously, the degree of identity transfer heavily depends on the backend model; compared to x-vector/PLDA and WavLM, ECAPA-TDNN generally exhibits superior separation metrics. This conclusion supports evidence-driven deployment decisions. However, it should be noted that generative speech enhancement techniques and automated speaker verification technologies should be used with caution in forensic science applications.

References

Cohen, I., & Berdugo, B. (2001). Speech enhancement for non-stationary noise environments. Signal Processing, 81(11), 2403–2418. https://doi.org/10.1016/S0165-1684(01)00128-1

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech 2020, 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650

Hu, Y., Liu, Y., Lv, S., Xing, M., Zhang, S., Fu, Y., Wu, J., Zhang, B., & Xie, L. (2020). DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement (arXiv:2008.00264). arXiv. https://doi.org/10.48550/arXiv.2008.00264

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-Vectors: Robust DNN Embeddings for Speaker Recognition. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

Wu, Y., Wang, L., Lee, K. A., Liu, M., & Dang, J. (2021). Joint Feature Enhancement and Speaker Recognition with Multi-Objective Task-Oriented Network. Interspeech 2021, 1089–1093. https://doi.org/10.21437/Interspeech.2021-1978

Um, Do These Uhs Matter?
An Investigation into the Influence of Filled Pauses on the Perceptions of the Interviewee

Maya Ely Joshi, James Tompkinson, and Lauren Harrington
Department of Language and Linguistic Science, University of York, United Kingdom
maya.elyj@gmail.com
{james.tompkinson@york.ac.uk | lauren.harrington@york.ac.uk}

In England and Wales, police-suspect interviews must be audio recorded under the Police and Criminal Evidence Act 1984 (PACE) (Haworth, 2018; Tompkinson and Haworth, 2025). However, transcripts can never completely capture the complexity of the spoken interaction and there is a loss of information from altering the evidence (Haworth, 2018) which may affect perceptions of the interviewee. This study investigates whether the manipulation of a linguistic feature, filled pauses *um* and *uh,* in police-suspect interview transcripts alter perceptions of the interviewee.

Two research questions were addressed:

1. Do participants perceive an interviewee differently depending on the modality of the interview record?
2. Will the inclusion of filled pauses in police interview transcripts bring the perception of the speaker closer to the initial, corresponding audio recording?

100 participants were recruited via Prolific, all of whom fulfilled the criteria for UK jury panel eligibility by being 18-75 years old, and having UK residency and nationality. Participants were randomly assigned and presented one of the following conditions:

1. *The audio recording only*
2. *A transcript containing no filled pauses* (No Audio)
3. *A transcript containing all filled pauses* (No Audio)
4. *A Transcript containing filled pauses that were longer than 400 ms in duration* (No Audio)

The audio stimulus was a mock police interview taken from Tompkinson et al's (2025) mock interview corpus. The transcripts were based on the same 2 minute 46 second audio recording and were identical in content except for the representation of filled pauses. Participants were presented with a series of likert-scale questions and asked to provide a response ranging from strongly disagree to strongly agree which captured perceptions of credibility, plausibility and sincerity. Perceptions of *truthfulness* and *guilt,* were also elicited. Unlike previous studies, participants were asked to rate their confidence levels when judging the interviewee as guilty or not guilty. Six emotion attributes: *nervousness, calmness, agitation, relaxation, fear* and *anxiety* were also measured. Open-ended questions were asked about the interviewee's language-use that influenced participants' judgements of the interviewee. A Kruskal-Wallis significance test was conducted and revealed significant effects on all six emotion attributes. A follow-up post-hoc Dunn test with Bonferroni correction identified all six emotion attributes having significant differences between audio and transcript conditions, consistent with previous research (Haworth, 2018; Deamer et al., 2022; Tompkinson and Haworth, 2025). No significant differences were observed between the transcript types, indicating that transcripts do not reliably convey the same perceptual information as the audio recording and thus should not be used as a replacement for audio-based evidence. Including filled pauses does not systematically reduce perceptual differences either. Perhaps surprisingly, out of the audio vs transcript conditions, the transcript condition *containing filled pauses longer than 400 ms,* was actually the least perceptually similar to the audio as the findings showed significant differences in every trait. Chi-square test of independence was conducted on the categorical *truth* and *guilt* judgements. There were no significant differences between the four conditions for these

two traits. However, a notable pattern emerged in ratings of confidence. Participants in the audio condition primarily indicated low to no confidence when assessing the interviewee's guilt. Participants in all three transcript conditions reported greater confidence in their guilt ratings. This suggests that transcripts alone may lead to an overconfident assessment of guilt. These findings suggest that transcript modality may influence evaluative certainty even when the evidential content remains unchanged. The findings of this study contribute to the range of literature (Haworth, 2018; Deamer et al., 2022; Tompkinson and Haworth, 2025) which has sounded caution around the interchangeable treatment of audio recordings and transcripts in policing and legal process. The results highlight the importance of considering how additional linguistic features are represented in transcripts, and support recommendations that audio recordings should be made available alongside transcripts whenever possible. Overall, this research demonstrates that transcripts are not perceptually equivalent to audio recordings and should not be treated as interchangeable forms of evidence.

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. https://doi.org/10.21747/21833745/lanlaw/9_1a2

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof, 22*(4), 428–450. https://doi.org/10.1177/1365712718798656

Tompkinson, J., Eijk, L., Knight, S., & Monteoliva-Garcia, E. (2025, July 20–23). Developing a collection of mock police interviews for use in research and teaching. Paper presented at the 33rd International Association of Forensic Phonetics and Acoustics (IAFPA) Conference, The Hague, Netherlands.

Tompkinson, J., & Haworth, K. (2025). The effect of linguistic detail in police interview transcripts on perceptions of an interviewee.*Language and Law Linguagem E Direito*, *11*(2). Retrieved from https://ojs.letras.up.pt/index.php/LLLD/article/view/14890

Final F_0 Fall in Forensic Speaker Comparison: The Role of Fillers

Leia Weber[1], *Gea de Jong-Lendle*[1] *& Mathias Scharinger*[1]
[1]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*
weberleia@gmail.com; dejong|scharinger@staff.uni-marburg.de

Introduction

Mean F_0 and its variability are well-established parameters in forensic phonetics (Jessen, 2012). This study examines the final F_0 fall (hereafter FFF), a local pitch parameter frequently occurring at the end of declarative utterances (Ladefoged, 2006) and considered relatively robust to changes in intensity (Liberman & Pierrehumbert, 1984). Previous work suggests high speaker-discriminatory potential due to high interspeaker and low intraspeaker variability (De Jong et al., 2005).
This study investigates whether the inclusion of fillers (*ja*, *genau*, *ähm*, *äh*, *oke*) affects the discriminatory power of FFF measurements. Such items are often included due to their high frequency and tendency to exhibit final falls. Their impact on intraspeaker variability, however, remains unclear. Using F-ratios as a measure of discriminatory power, three conditions were compared: (1) a mixed set including fillers and non-fillers reflecting standard practice in forensic analysis, (2) non-fillers only and (3) fillers only. The study aims to determine whether excluding fillers improves the speaker discrimination potential of FFF. Previous work on F_0 baseline, a related pitch parameter that resembles a neutral vocal fold position, demonstrates its strong discriminatory power (Cavalcanti et al., 2024), further indicating the potential of FFF as both parameters operate in the lower pitch range.

Methodology

For 86 young female native speakers of German, FFFs were extracted from approximately 2.5 minutes of spontaneous speech monologues recorded in a sound-treated booth. Topics included everyday conversational subjects (e.g. holidays, favorite books, movies etc.). The measurement point was defined as the lowest F_0 value in the sentence-final syllable of a declarative utterance (using the function “minimum pitch” in Praat). Only speakers with at least 8 measurable FFFs were included, yielding a final sample of 45 speakers (mean age: 24.8 y.).
For the filler-analysis, F-ratios for FFFs were calculated comparing two (different) subgroups (both N=15) extracted from the main group of 45 speakers: (1) non-fillers and (2) fillers only. This analysis was based on 6 FFFs per speaker. Since these items may also occur as non-fillers, classification was based on syntactic/pragmatic context.

Results

Table 1 indicates a reduced discriminatory power when fillers are included in the set of FFF measurements (F=1.01, N=45). In the spontaneous speech segments analysed, 39 % of FFFs were derived from fillers. When only fillers were analysed, the F-ratio decreased further (F=0.79, N=15). In contrast, excluding fillers resulted in the highest F-ratio (F=1.33, N=15), suggesting that speaker identification accuracy based on FFF is maximised when fillers are excluded from the analysis since they lower the interindividual variation.

Conclusions

The findings suggest that fillers should be excluded from FFF measurements, even if they exhibit a falling pitch contour. It should also be noted that the discriminatory power may partly result from monosyllabic versus multisyllabic fillers. The study further shows that FFF measurements are not consistently obtainable in spontaneous speech produced by young German female speakers due to the high prevalence of uptalk (rising pitch contour) and, to a lesser extent, creaky voice. The findings further highlight the need to explore other measurement procedures such as incorporating fall magnitude or including a temporal component. Although the parameter FFF requires further investigation, it already shows potential as a valuable addition to forensic phonetics if future studies replicate these results.

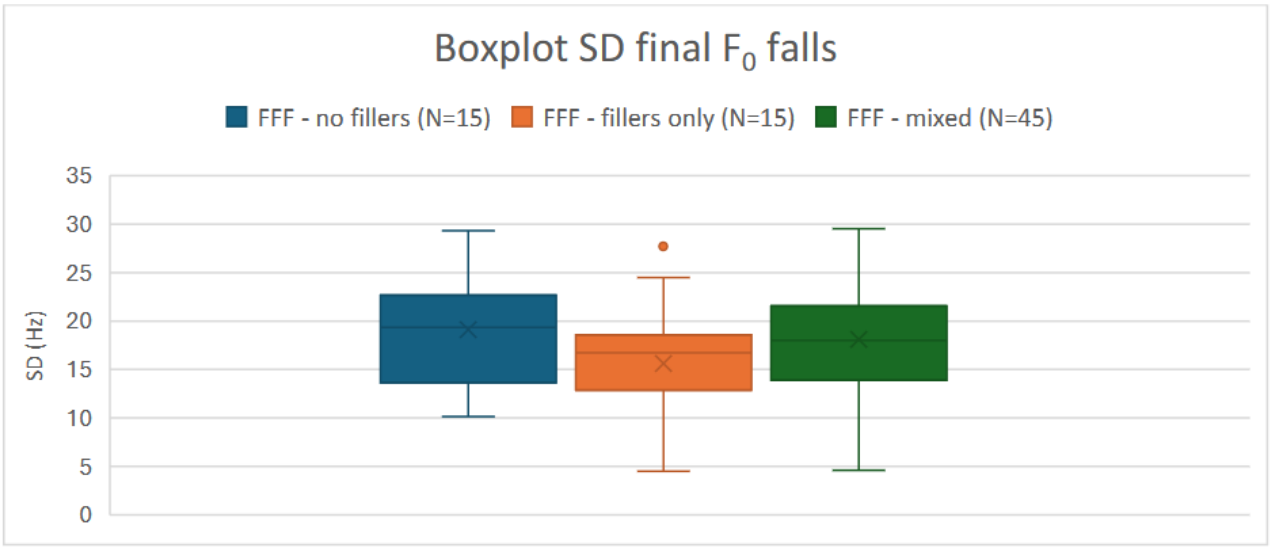


Figure 1. Boxplot of the standard deviation of FFF-mixed (green); (8 measurement points per person), FFF-no fillers (blue) and FFF-fillers only (orange) (both 6 measurement points per person) for young female native speakers of German

Group	*Interindividual Variation. (SD of Means, Hz)*	*Intraindividual Variation (Mean of SDs, Hz)*	*F-ratio*
FFF-mixed (N=45)	18.3	18.1	**1.01**
FFF-no fillers (N=15)	25.4	19.1	**1.33**
FFF-fillers only (N=15)	12.3	15.6	**0.79**

Table 1. Inter- and intraindividual variation and F-ratio for FFF-mixed, FFF-no fillers and FFF-fillers only.

References

Cavalcanti, J. C., Eriksson, A., & Barbosa, P. A. (2024). Multiparametric Analysis of Speaking Fundamental Frequency in Genetically Related Speakers Using Different Speech Materials: Some Forensic Implications. *Journal of Voice*, *38*, 243.e11-243.e29. https://doi.org/10.1016/j.jvoice.2021.08.013

De Jong, G., House, J., Cook, N., & Young, A. (2005). The speaker discriminating power of the final fall: Spontaneous speech. *International Association of Forensic Phonetics and Acoustics.*

Pierrehumbert, J. & Liberman, M. (1984). *Intonational invariance under changes in pitch range and length.* In M. Halle, Mark Aronoff, & R. T. Oehrle (Eds.), *Language sound structure: Studies in phonology* (pp. 157–233). MIT Press.

Jessen, M. (2012). *Phonetische und linguistische Prinzipien des forensischen Stimmenvergleichs*. LINCOM EUROPA.

Ladefoged, P. (2006). *A course in phonetics* (5. ed.). Wadsworth Cengage Learning.

Forensic phonetics before 1980: Some examples

Angelika Braun
Department of Phonetics, University of Trier, Germany
brauna@uni-trier.de

When looking at the list of topics or the session titles at ICPhS conferences, forensic phonetics does not appear until ICPhS 1995 in Stockholm. This suggests that it is a rather young discipline. The present contribution demonstrates, however, that forensic speaker recognition is actually older than phonetics as an academic subject.

The focus is on forensic speaker identification (SI) in Germany, England, and the US before the 1980s. Some early "landmark cases" of using the voice as evidence are covered in detail, indicating their role in the development of the field. The most famous one is contained in the Old Testament[1]; a more recent one arose in conjunction with the beheading of King Charles I in 1649, the trial taking place in 1660 (Howell 1816).

In the early days, forensic phonetics invariably meant speaker identification (SI) by victims and witnesses (Mack vs. State of Fla 1907, McGehee 1937). The reason for this is simply that there were no devices for recording, storing, and analyzing sound. The most important factors which made SI by expert possible were the development of devices for making good quality recordings (which could be stored) and also of the telephone (which could be used for generating recordings). This took until the 1930s. An early case involving a court of law commissioning a forensic report by expert in Germany is documented in 1933. It is part of an exchange between Eberhard Zwirner and Giulio Panconcelli-Calzia. Another major development without which forensic phonetics would not have emerged in the way it did was the development of the sound spectrograph (Braun 2019).

Starting in the 1950s, dialectological expertise was added to the repertoire of forensic analyses (Landesarchiv Baden-Württemberg 2008) in Germany (Knetschke & Sperlbaum 1969)[2], closely followed by a similar development in Britain (Ellis 1990). Ellis (1990) and Baldwin & French (1990) list a large number of cases which they worked on starting in the early 1960s. According to their accounts, they employed auditory analysis exclusively, occasionally (i.e. when that was challenged in court) backed by acoustic analysis. In Germany, voices of offenders were regularly publicized in order to prompt familiar voice recognition. And yet, the 1960s also brought about the infamous voiceprint controversy (Kersta 1962; Braun 2019). The 1960s saw the first publication of a treatise on forensic phonetics (Fährmann 1963) which bore the subtitle *Einführung in die Kriminalphonetik* 'Introduction to Criminal Phonetics'. It was authored not by a phonetician, but by the psychologist Rudolf Fährmann. He outlined a protocol for auditory – and to some extent acoustic – analysis. Fährmann's approach was appreciated behind the Iron Curtain (Koristka 1967; Bijsterfeld 2021) more so than in Western Europe and specifically West Germany, where all research efforts and resources at the *Bundeskriminalamt* (BKA, Federal Criminal Police Office) went into the development of automatic SI (Bunge 1979). Fährmann's work, on the other hand, can be regarded as a precursor of the auditory-acoustic method of voice comparison.

---

[1] Peter French has kindly pointed out in his review that there are quite a few examples of speaker identification by voice in world literature. This is acknowledged. However, the present contribution focuses on non-fictional incidents. I would consider the fictional coverage of speaker ID by voice a different topic. It is, of course, open to debate whether or not one would consider the Bible fictional.

[2] Edeltraud Knetschke and Margret Sperlbaum oversaw the *Spracharchiv* (Speech Archive), a collection of more than 10.000 recordings from all over Germany which had been initiated by Eberhard Zwirner in the 1930. They carried out the auditory phonetic analyses whereas H. Goydtke of the *Physikalisch-Technische Bundesanstalt* (Federal Technical Institute) Braunschweig contributed spectrographic analyses. There is, however, no known case in which the latter has played a major role.

The late 1970s/early1980s saw forensic phonetics as a thriving new field within the repertoire of forensic analyses. In West Germany, a major reason for this was the leftist Red Army Faction committing a series of brutal killings during that time. After several arrests had been made and the trials started in the early 1980s, voice recordings constituted major pieces of evidence, and the judiciary was badly in need of voice experts. A speaker identification unit was formed at the BKA after the automatic approach had failed (Künzel 1987). In England, the demand for forensic phonetic expertise was fostered by a much less dramatic mechanism: The PACE (Police and Criminal Evidence) Act was implemented in 1984. It stated that all police interviews were to be recorded, which provided ample material on which to base SI (French 2017). In the 1980s, the foundations of forensic speaker identification in its present form were laid.

References

Baldwin, John & Peter French (1990). Forensic Phonetics. London etc.: Taylor & Francis,

Bijsterveld, K. (2021). Slicing Sound: Speaker Identification and Sonic Skills at the Stasi, 1966-1989. *Isis 112,* 215-241.

Braun, A. (2019). From *Visible Speech* to *Voiceprints*. The missing link. In M. Pucher, J. Trouvain & C. Lozo (Eds): *HSCR 2019. Proceedings of the Third International Workshop on the History of Speech Communication Research.* Vienna, September 13-14, 2019. Dresden (pp.123-136). DTUpress.

Bunge, E. (1979). Automatic forensic speaker recognition. In *Carnahan Conference on Crime Countermeasures.* University of Kentucky, Lexington, May 16-18 (pp.41-45).

Ellis, St. (1990). ”It’s rather serious…”. Early Speaker Identification. In H. Kniffka (ed.), *Texte zu Theorie und Praxis forensischer Linguistik*. Tübingen (pp.515-521). Niemeyer.

Fährmann, R. (1963). Die Stimm- und Sprechdiagnostik im Dienste der Verbrechensaufklärung (Einführung in die Kriminalphonetik). *Taschenbuch für Kriminalisten* 13, 145-192.

French, P. (2017). A Developmental History of Forensic Speaker Comparison in the UK. *English Phonetics*: 255-270.

Howell, T. B. (1816). *A Complete Collection of State Trials and Proceedings for High Treason and Other Crimes and Misdemeanors from the Earliest Period to the Year 1783, with Notes and Other Illustrations, compiled by T.B. Howell*, Vol. V. London: T.C. Hansard.

Kersta, L. G. (1962). Voiceprint Identification. *Nature* 196: 1253-1257.

Knetschke, E. & Sperlbaum, M. (1979). Das deutsche Spracharchiv im Institut für deutsche Sprache. *Mitteilungen des Instituts für deutsche Sprache* 6. Mannheim.

Koristka, Chr. (1967). Über die Möglichkeit und den augenblicklichen Stand der Verwendung der menschlichen Stimme zur kriminalistischen Identifizierung einer Person (Stimmidentifizierung). In H. Sanke (ed.), *Kriminalistik und forensische Wissenschaften*. Berlin: Institut für Kriminalistik der Humboldt-Universität zu Berlin (pp. 133-149). (Wissenschaftliche Schriftenreihe der Humboldt-Universität zu Berlin.),

Künzel, H. (1987). *Sprechererkennung. Grundzüge forensischer Sprachverarbeitung*. Heidelberg: Kriminalistik Verlag.

MACK vs State of Fla. (1907). 44 So 706.

McGehee, F. (1937). The Reliability of the Identification of the Human Voice. *The Journal of General Psychology* 17: 249-271.

Internet source

Landesarchiv Baden-Württemberg (2008). 'O-Töne eines Kidnappers – Vor fünfzig Jahren: erste Kindesentführung in Deutschland. Akten und Tondokumente im Staatsarchiv Ludwigsburg', *Staatsarchiv Ludwigsburg, Aktuelles der Abteilung*. <http://www.landesarchiv-bw.de/web/42378> (last accessed 22 Nov 2019).

Towards explainable linguistic phonetic methods to authenticate allegedly deepfake audio

Ben Gibb-Reid, Jessica Wormald, and Vincent Hughes
Department of Language and Linguistic Science, University of York
ben.gibb-reid@york.ac.uk

With the increasing development and availability of audio deepfakes (also known as 'voice clones'), a new question is being asked of forensic speech science practitioners: is this questioned recording produced by **an authentic speaker**, or was it **synthetically produced**? This paper addresses the need for science-based complementary human and machine methods which provide forensic experts with robust explainable procedures. Much existing research on deepfake detection relies solely on machine-based detection (Almutairi & Elgibreen, 2022). However, these lack explainability and transparency. This paper builds towards a procedure which incorporates complementary human analytical techniques that rely on segmental and suprasegmental phonetic features (following work by Yang et al., 2026), features relating to the signal (Lee et al., 2023), disfluency patterns, and linguistic features which are contextually inconsistent (e.g. lexico-grammatical, syntactic and interactional features). Various issues are addressed including: i) Defining the question according to the availability of reference material; ii) The relevance of typicality; iii) Using a conclusion framework to assess weight of evidence.

To understand the kinds of features an explainable forensic speech science method might utilise, an exploratory case study is presented which focuses on individual variation across authentic and deepfake samples. Coda /ɹ/ and the monophthongal vowels are analysed for speakers of southern standard British (SB), Yorkshire (YK), Scottish (SC), and Hong Kong Englishes (HK). 16 speakers were recorded doing short monologic tasks, and ElevenLabs was used to create 'voice clones' with two types of synthetic outputs: text-to-speech (TTS; input is a written transcript) and voice conversion (VC; a source input is a recording of one YK male speaker outside of the original 16 speakers). Clones were made using two to six minutes of speech.

VC outputs have segmental realisations close to the source speaker. As shown in Figure 1, though authentic SC and HK speakers produce no intrusive /ɹ/ and some rhotic /ɹ/, the VC samples for these speakers have less rhoticity, and a high rate of intrusive /ɹ/. TTS outputs tend to be closer to authentic speakers for /ɹ/, aside from the introduction of intrusive /ɹ/ for SC and HK speakers. Vowel formant estimates were normalised using the Watt & Fabricius method. In terms of vowel space, VC outputs map much more closely to the source input as displayed in Figure 2. Figure 3 displays TRAP and BATH on a group level. In VC outputs, the SB speakers (who have a phonological split for these vowels) merge them, whereas TTS outputs display slight separation. However, for YK speakers, who maintain a TRAP~BATH merger, TTS outputs introduce a split, aligning them more with SB patterns. The same was true for FOOT~STRUT.

Overall, VC methods inaccurately represented the authentic speakers' segmental realisations. TTS methods were better but tended to move non-SB speakers towards some standard variety; either SB or US English sound patterns. Both methods also introduced artefacts (see Lee et al., 2023) e.g. unnatural pitch shifts, connected speech processes, formant structures. These findings are explainable and transparent in their mapping to authentic patterns of speech. We discuss the future work needed to develop a framework that takes these observations and considers typicality and similarity in a Bayesian framework.

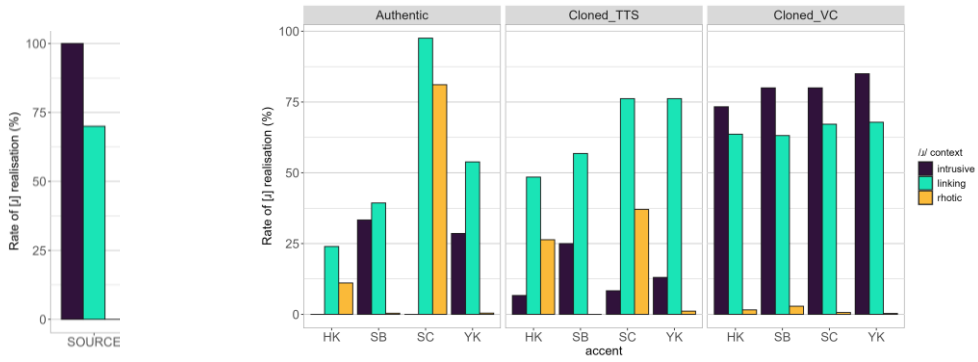


Figure 1. Bar plot of [ɹ] proportions in rhotic/intrusive/linking contexts for authentic & VC samples. proportions are measures of every possible occurrence of [ɹ] in intrusive, linking and rhotic contexts.

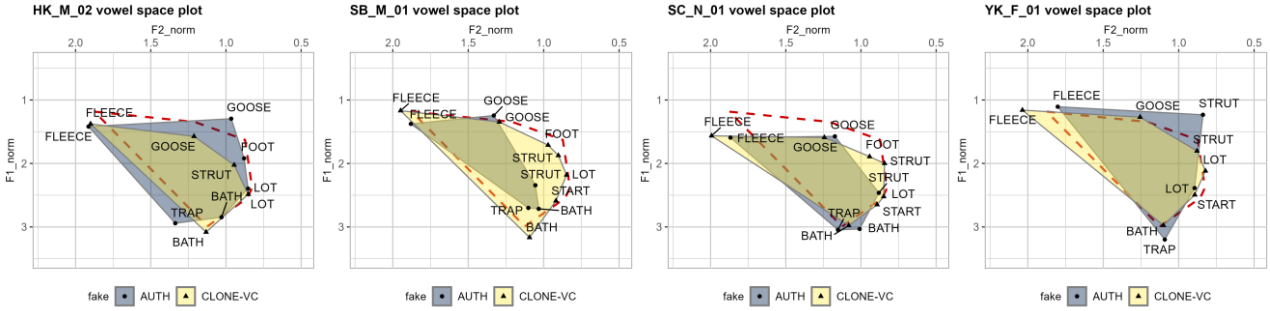


Figure 2. F1~F2 vowel space plots for four speakers showing their authentic sample and their voice conversion output (VC) vowel space. The dotted red line shows the source speaker's vowel space.

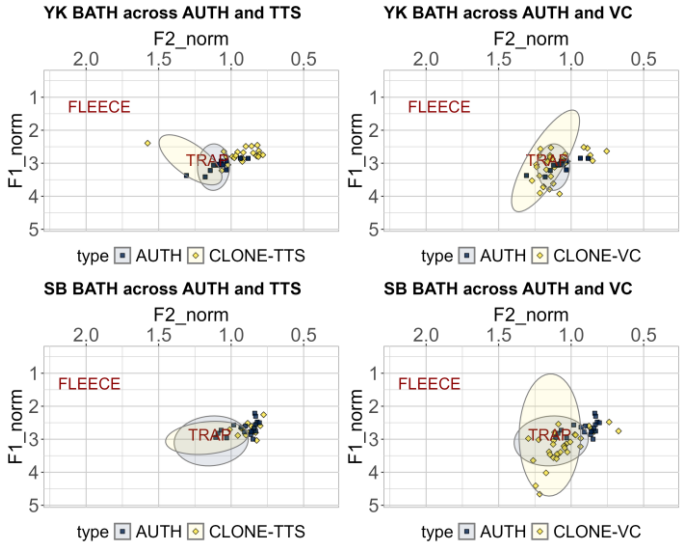


Figure 3. F1~F2 vowel space displaying TRAP (by ellipses) and BATH (via diamond shapes) for TTS (left) and VC (right) methods.

References

Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: Challenges and future directions. *Algorithms, 15(5)*, 155.

Yang, T., Sun, C., Lyu, S., & Rose, P. (2026). Forensic deepfake audio detection using segmental speech features. *Forensic Science International*, *379*, 112768.

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.

Perceiving speaker identity in authentic vs voice-converted speech: An eyetracking study

Justin J. H. Lo
Linguistics and English Language, Lancaster University, UK
j.h.lo@lancaster.ac.uk

Introduction

State-of-the-art AI systems can generate highly human-like speech and bring considerable risks of fraud and disinformation through impersonation. Recent evidence on both text-to-speech and voice conversion systems indicates that human listeners are often unable to reliably distinguish AI-generated speech from authentic human speech (Barrington et al., 2025; Lavan et al., 2025; Warren et al., 2024), especially when the audio quality is degraded (Terblanche et al., 2021). While studies thus far have largely focused on explicit detection or discrimination of AI-generated speech from human speech, much remains unknown about the process of how listeners reach such decisions. This study seeks to explore how listeners process speaker identity by means of eyetracking, in order to investigate voice recognition in real time and address the question of whether – and how – the time-course of voice identity perception differs between authentic human speech and speech generated by voice conversion.

Methodology

L1 English participants completed an eyetracking experiment using a visual-world paradigm (Allopenna et al., 1998), where they listened to authentic and voice-converted samples of short sentences from four female voices that they had been familiarised with alongside associating alien character images. In each case, participants were instructed to look towards the character associated with the voice in the audio stimulus, or a '?' image if the corresponding character was absent. Each participant completed a total of 192 trials, consisting of 96 unique sentences evenly distributed across the four voices, each presented once as an authentic sample and once as a voice-converted sample. The voice-converted samples were generated on ElevenLabs, using sentence recordings from a fifth, unfamiliarised female speaker as source and Instant Voice Clones trained on each target speaker. The time-varying proportions of looks towards the target image were analysed, using growth curve analysis (Forbes et al., 2025; Mirman, 2014), to assess how the time-course of voice identity perception compared between authentic and voice-converted samples.

Results & Discussion

Data from 20 participants are presented here while data collection is continuing. Overall, by the end of the trial, participants correctly recognised the voice from 40% of authentic samples as the target and 30% of voice-converted samples as absent. Across all trials, for both types of samples, fixations on the target character image increase from approximately 400ms after utterance onset (Figure 1, left), indicating early uptake of some indexical information in voices to recognise speaker identity and minimal differentiation between authentic and voice-converted samples. Among only correct trials, target fixations for voice-converted samples diverge from authentic samples from approximately 800 ms, falling from approximately 2000 ms to 0% (Figure 1, right). Together, these findings suggest that, within the first few seconds, listeners generally process voice identity in authentic and voice-converted speech in similar ways. At the same time, perceptual cues to distinguishing authentic and voice-converted samples can be used from early on when they are detected by listeners.

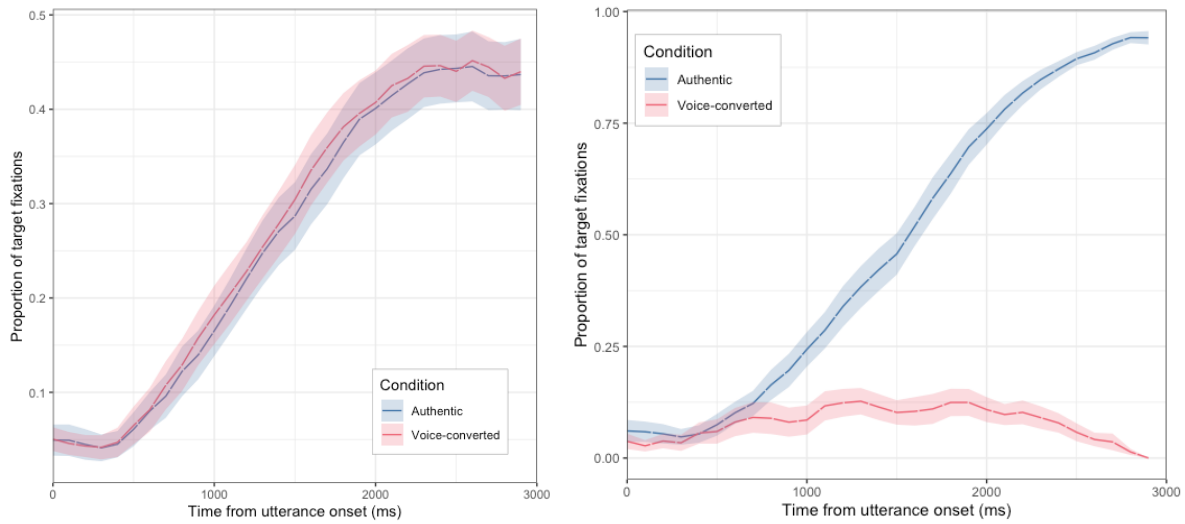


Figure 1. Time-course of fixations on target character image for authentic (blue) and voice-converted (red) samples across all trials (left) and only correct trials (right).

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38(4), 419–439.

Barrington, S., Cooper, E. A., & Farid, H. (2025). People are poorly equipped to detect AI-powered voice clones. *Scientific Reports*, 15, 11004.

Forbes, S., Dink, J., & Ferguson, B. (2025). *eyetrackingR: An R package for eye-tracking data analysis* (v0.2.2). http://www.eyetracking-r.com

Lavan, N., Irvine, M., Rosi, V., & McGettigan, C. (2025). Voice clones sound realistic but not (yet) hyperrealistic. *PLoS One*, 20(9), e0332692.

Mirman, D. (2014). *Growth curve analysis and visualization using R* (1st Ed.). Chapman and Hall/CRC.

Terblanche, C., Harrison, P., & Gully, A. (2021). Human spoofing performance on degraded speech. *Proceedings of Interspeech 2021*, 1738–1742.

Warren, K. Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., Gates, C., & Traynor, P. (2024). “Better be computer or I’m dumb”: A large-scale evaluation of humans as audio deepfake detectors. *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, 2696–2710.

Dialect, voice quality and perceived speaker identity in AI voice conversion

Hope McVean
Department of Linguistics and English Language, Lancaster University, Lancaster, UK
h.mcvean@lancaster.ac.uk

Recent advances in voice conversion (VC) have enabled the generation of synthetic speech that is capable of convincingly impersonating individuals on the basis of only short reference recordings. While such technologies have legitimate applications, they also present emerging opportunities for cybercriminals, particularly in cases involving disinformation, defamation and fraud. Understanding and pre-empting the risks that VC technologies pose takes detailed evaluation of the data they synthesise. Current evaluative practices in the wider VC research (see the Voice Conversion Challenges, 2018-2020) are arguably rather blunt – typically operationalising speaker similarity as a single similarity value and using it to infer how well a voice conversion system produces speech samples based on a target speaker. The current work starts to unpick speaker similarity in voice conversion evaluation by looking at the contribution of dialect similarity on the one hand, and voice quality similarity on the other.

Methodology

The present study investigates how dialect similarity and voice quality pairing might influence listener judgements of speaker identity in converted voices. A 'controlled pairing design' was implemented by voice converting between: a) different speakers within the same dialect group; b) different speakers from different dialect groups; c) different speakers with highly similar voice qualities; d) different speakers with less similar voice qualities. The intention behind this paradigm of voice conversions is to disentangle dialect from voice quality in the evaluation of voice converted output. Speech from two forensic phonetic corpora was used – WYRED (West Yorkshire Regional English Database; Gold et al., 2020) and DyViS (Dynamic Variation in Speech; Nolan et al., 2009). Each corpus represents a distinct regional variety of British English: West Yorkshire and Standard Southern British English (SSBE) respectively. Cross-dialectal speaker pairs were initially screened using an automatic speaker verification (ASV) system to identify acoustically comparable voices. Selected pairs were subsequently validated using auditory vocal profile analysis to establish shared long-term voice quality characteristics across dialect groups. Voice conversion was carried out using ElevenLabs, a commercial GenAI speech synthesis platform, via their 'voice cloning' tool. Listeners then completed an online forced-choice identification task in which, after familiarisation with a reference speaker, they evaluated the speaker similarity of a comparison stimulus. Comparison stimuli systematically varied by stimulus category (bonafide vs converted), dialect alignment (within-dialect vs cross-dialect) and voice quality pairing (paired vs unpaired).

Results

Responses from 212 participants were analysed using mixed-effects logistic regression. Results revealed highly significant effects of dialect alignment and voice quality pairings on the likelihood of "same speaker" judgements. Within-dialect stimuli were more likely to be judged as originating from the reference speaker than cross-dialect stimuli, while congruous voice quality also significantly increased perceived similarity. Notably, voice quality matching increased perceived similarity, even when dialect differed, suggesting that voice quality settings may carry substantial perceptual weight in speaker identity judgements of converted speech.

When considering misuse of voice conversion technologies, the above findings can go on to inform voice analysis in cases involving suspected impersonation attempts. Understanding the consequences of voice quality (mis)matches or dialect (mis)matches in the make-up of voice converted speech samples can better equip an analyst to make assessments about the nature or prospect of AI-generated speech appearing in a given case.

References

Gold, E., Earnshaw, K., & Ross, S. (10 Sep 2020). WYRED - West Yorkshire Regional English Database. UK Data Service. 10.5255/UKDA-SN-854354

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. *arXiv preprint arXiv:1804.04262*.

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.

Toda, T., Chen, L. H., Saito, D., Villavicencio, F., Wester, M., Wu, Z., & Yamagishi, J. (2016, September). The voice conversion challenge 2016. In *Interspeech 2016* (pp. 1632-1636). International Speech Communication Association.

Zhao, Y., Huang, W. C., Tian, X., Yamagishi, J., Das, R. K., Kinnunen, T., ... & Toda, T. (2020). Voice conversion challenge 2020: Intra-lingual semi-parallel and cross-lingual voice conversion. *arXiv preprint arXiv:2008.12527*.

Assessing automatic, acoustic-phonetic, and auditory speaker comparison in the era of open-source deepfakes

Linda Gerlach, Amaia Sagasti, and Anil Alexander
Oxford Wave Research, Oxford, UK
{linda|amaia|anil}@oxfordwaveresearch.com

Commercial systems have made realistic voice cloning widely accessible in recent years, and the providers of these systems have generally introduced some guardrails to reduce the risk of the abuse of this technology. However, more recently, a concerning development has been the availability of open-source toolkits that can operate entirely offline, without any guardrails or ethical safeguards. If such easily accessible deepfake creation tools can realistically mimic characteristics of a target speaker's voice with high precision, the reliability of forensic speaker comparison using auditory, acoustic-phonetic, as well as automatic approaches can be undermined. In this study, we evaluate the similarity between linguistically similar real and cloned speech at an auditory-perceptual level, an acoustic feature level, as well as the level of score distributions within automatic speaker recognition and discuss its implications and potential countermeasures.

The experimental framework uses an in-house dataset of British English speakers, comprising both native and non-native accents to ensure acoustic variety. For the preliminary evaluation, we select a subset of read speech recordings from seven speakers (two females, five males). Source recordings were segmented into 20-second windows and files were checked to contain adequate net speech. To ensure that speaker identity and content are consistent between the 'real' and 'fake' classes, we generated synthetic counterparts for each original recording using the following voice-cloning pipeline: For each speaker, a single real recording served as an acoustic reference to create a voice clone. This clone, together with the transcriptions of each original recording, was used to synthesise the speaker's remaining files using Qwen3-TTS (Hu et al., 2026), a recently released state-of-the-art open-source and offline audio-language model.

To assess the auditory-perceptual impact, we will present a listener experiment where participants are tasked with distinguishing real and fake versions of a speaker's voice. On an acoustic-phonetic level, and as a complementary indicator of perceptual similarity (see Gerlach, 2024), we evaluate the overlap of Long-Term Formant Distributions (LTFD1 to LTFD4) between the real and fake speech sets for each speaker using Praat (Boersma & Weenink, 2024). Figure 1 illustrates substantial overlap across these four LTFDs for Speaker 01, making samples difficult to distinguish. We further examine the score distributions produced by VOCALISE forensic automatic speaker recognition software (Kelly et al., 2019), by comparing a real sample against other real recordings of the same speaker and against the corresponding synthetic clones (same speaker identity and spoken content). As shown in Figure 2, the high degree of overlap between these two distributions for Speaker 01 suggests that the system's scoring remains largely unaffected by the synthetic nature of the samples. Although traditional phonetic speaker comparison including segmental and suprasegmental features, pausing, and hesitation phenomena may provide some additional insights, overall, we find that synthesised speech preserves the biometric identity of the speaker with enough precision to challenge most forms of forensic speaker recognition analysis. These findings suggest that deepfake detection should be an

important consideration in any analysis. To this end, we use Fauxdio (Alexander et al., 2025), a specialised audio deepfake detection tool and explore its effectiveness in flagging potential voice clones. We recommend that deepfake detection analysis becomes a part of the forensic analysis pipeline to address the rapidly evolving landscape of deepfake speech generation.

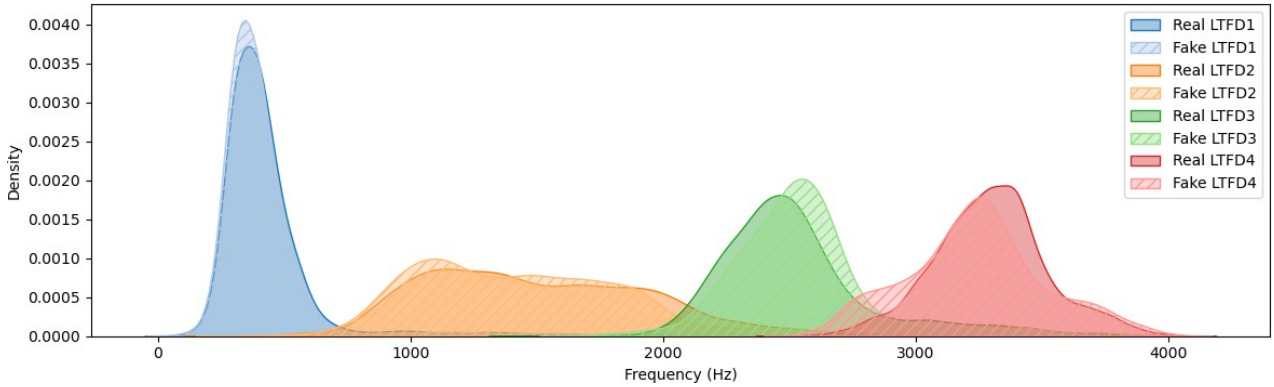


Figure 1. Kernel Density Estimation (KDE) displaying the LTFDs for real and fake samples with the same linguistic content of Speaker 01.

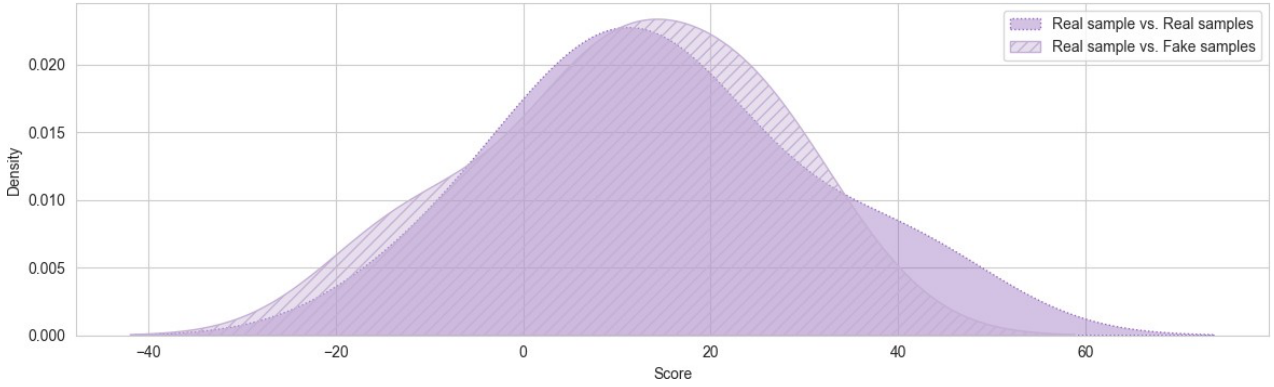


Figure 2. KDE of similarity scores (based on spectral features and PLDA, produced by VOCALISE) comparing a single real sample of Speaker 01 against two distinct populations: real samples (authentic recordings of the same speaker) and fake samples (the corresponding identity- and content-consistent synthetic clones generated with Qwen3-TTS).

References

Alexander, A., Gerlach, L., Coy, T., Forth, O., Lonergan, L., & Kelly, F. (2025). *FAUX*DIO: An audio deepfake detector for law enforcement and forensics. *Proc. of the Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*, Den Haag, The Netherlands, 46-47.

Boersma, P. & Weenink, D. (2024). Praat: doing phonetics by computer [Computer program]. Version 6.4.23, retrieved 27 October 2024 from http://www.praat.org/

Gerlach, L. (2024). Automatic assessment of voice similarity and its implications for forensic applications [Apollo - University of Cambridge Repository]. https://doi.org/10.17863/CAM.115862

Hu, H., Zhu, X., He, T., Guo, D., Zhang, B., Wang, X., Guo, Z., Jiang, Z., Hao, H., Guo, Z., Zhang, X., Zhang, P., Yang, B., Xu, J., Zhou, J., & Lin, J. (2026). Qwen3-TTS Technical Report (arXiv:2601.15621). arXiv.

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. of the AES International Conference*, Paper 27.

Forensic Automatic Speaker Recognition: reflecting on technological progress and casework applications

Finnian Kelly and Anil Alexander
Oxford Wave Research
{finnian|anil}@oxfordwaveresearch.com

In this paper, we will reflect on the continued progress of automatic speaker recognition from a technological perspective and on its growing use in forensic casework. We will draw on our experience in developing a forensic automatic speaker recognition system in use worldwide (Alexander et al. 2024, Kelly et al. 2019a), to provide first-hand insights about the application of the technology in forensic practice and will also make more general observations about the current state of the field.

Over the past few decades, progress in automatic speaker recognition has been driven by major algorithmic developments, from GMM and GMM-MAP, to i-vectors and x-vectors. With each evolution of the technology, performance in forensically relevant conditions has improved (Kelly et al. 2019b). In recent years, automatic speaker recognition research has reflected the trend seen more broadly in AI: a move toward larger DNN architectures to further improve performance, at the cost of greater training data requirements.

To guide a discussion on the current direction of the field from a technology perspective, we consider a performance comparison between established and recently proposed approaches, namely x-vector, ECAPA-TDNN (Desplanques et al. 2020), ResNet34 (Villalba et al. 2020) and ReDimNet b2 (Yakovlev et al. 2024). These approaches are evaluated across three forensically relevant UK databases: GBR-ENG[1], WYRED (Gold, 2020), and SPOKE[1]. From the results in Figure 1, we observe that for comparisons involving telephone, x-vector is competitive with all the newer models. For microphone vs. microphone, and for the mixed conditions in SPOKE, the newer models provide a performance gain, particularly ReDimNet b2.

Improved performance is a product of both model architecture and training data. To satisfy the increased training data requirements of newer architectures, researchers have turned toward web-scraped data to make up a shortfall in existing datasets or alternatively have built on pre-trained foundation models. In both scenarios, the underlying data may not have been obtained from the speakers with appropriate consent. Indeed, all three of the newer models evaluated here were trained using VoxCeleb, which consists entirely of YouTube data. For forensic applications, this poses an ethical and practical dilemma; (potentially small) improvements in accuracy need to be balanced against data provenance.

Reflecting on how automatic speaker recognition is applied in forensic casework, we will share insights gained through our training courses and interactions with VOCALISE users (representing forensic laboratories in over 30 countries). We have observed that although there are differences in the expectations of forensic analysis in different countries, the likelihood ratio, and particularly the equivalent verbal LR scale, works well in practice as an objective and accessible means of expressing results. We note however that difficulties remain in correctly defining the hypotheses (e.g., same speaker vs. different speaker, common source vs. specific source) and in selecting and acquiring appropriate relevant population data that accounts for both speaker demographics and technical conditions. Unfortunately, there are no unified standards for forensic automatic speaker recognition to address these and other challenges. Although the ENFSI guidelines from 2015 (Drygajlo et al. 2015) are very helpful, they require updating, and the more recent NIST OSAC standardisation efforts have been hindered by the dissolution of the speaker recognition group. Development of

[1] Provided by the UK Government. See Harrington et al. (2025) for further VOCALISE results using GBR-ENG.

comprehensive and up-to-date guidelines, ultimately followed by standards, is a pressing need for the field. Finally, we have come to view user training as an ongoing commitment within a network of practitioners who can share expertise, rather than a one-time activity. User training also works both ways, often inspiring and driving the development of new features within the software, for example, upcoming features to aid reporting and audit trails.

We intend our presentation to reflect further on these technical and practical perspectives and prompt more general discussion about the application of automatic speaker recognition to forensic casework.

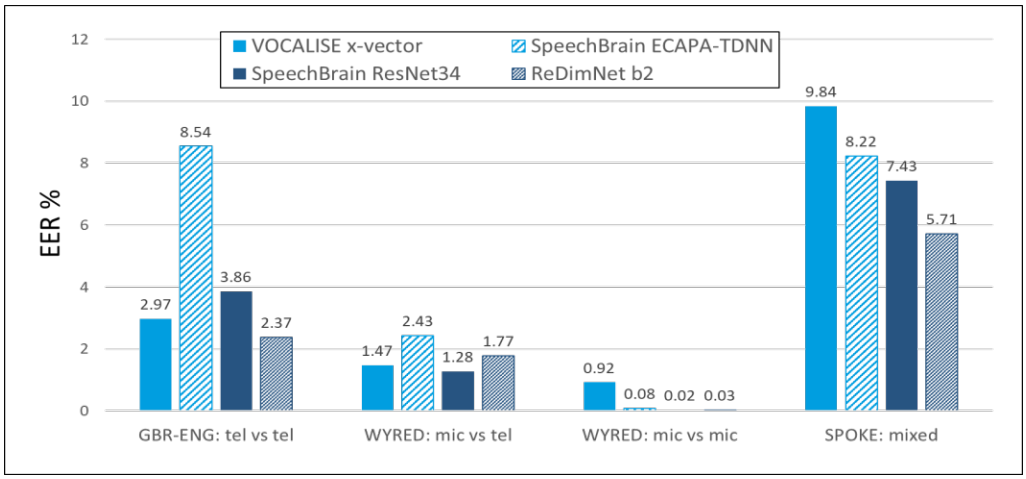


Figure 1: Speaker recognition performance of x-vector, ECAPA-TDNN, ResNet34 and ReDimNet b2 architectures across a range of forensically relevant datasets

References

Alexander, A., Gerlach, L., Coy, T., Forth, O., & Kelly, F. (2024). Handling real-world challenges of variable speech quality and multiple speakers in forensic automatic speaker recognition using VOCALISE, *Audio Engineering Society (AES) Forensics Conference 2024*, Denver, USA.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification, *Interspeech 2020*, 3830-3834.

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, *Frankfurt: Verlag für Polizeiwissenschaft.*

Gold, E. (2020), WYRED - West Yorkshire Regional English Database 2016-2019. [*Data Collection*]. Colchester, Essex: UK Data Service.

Harrington, L., Hughes, V., Harrison. P, Foulkes, P, Wormald, J., Kelly, F., & van der Vloed, D. (2025). Variability in performance across four generations of automatic speaker recognition systems. *Interspeech 2025, 3993-3997.*

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019a). Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors, *Audio Engineering Society (AES) Forensics Conference 2019*, Porto, Portugal.

Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S. & Alexander, A. (2019b). Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval01), *Speech Communication*, Vol 112, September 2019, 30-36.

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Borgstrom, J., García-Perera, L., Richardson, F., Dehak, R., Torres-Carrasquillo, P., & Dehak, N. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and Speakers in the Wild evaluations, *Computer Speech & Language*, Vol 60, March 2020, 101026.

Yakovlev, I., Makarov, R., Balykin, A., Malov, P., Okhotnikov, A., & Torgashov, N. (2024). Reshape Dimensions Network for Speaker Recognition*, Interspeech 2024*, 3235-3239.

Beyond ‘Real’ or ‘Fake’: The challenge of benign AI-based processing in forensic deepfake detection

Amaia Sagasti, Finnian Kelly, and Anil Alexander
Oxford Wave Research, Oxford, UK
{amaia|finnian|anil}@oxfordwaveresearch.com

Accessible tools for synthetic speech generation can now produce deepfakes (i.e., synthetic speech maliciously targeting a specific individual) that are capable of easily fooling the human ear. The research community has responded to this challenge by developing automatic approaches for detecting synthetic speech, which have largely kept pace as the quality of synthetic speech has improved. Effective detection approaches are often based on SSL (Self-Supervised Learning) representations that leverage large acoustic models trained with massive quantities of data.

However, a new challenge has recently emerged, that is as much technical as it is philosophical: the underlying speech generation techniques used for deepfakes are increasingly being applied to ‘benign’ AI-based audio processing tasks, such as speech enhancement, restoration or neural compression (for example, in commercial tools like Adobe *Speech Enhance* and ElevenLabs *Voice Isolator*, or in Meta’s *AudioDec* neural audio codec). Conceptually, ‘fake’ audio samples are defined by a breach of authenticity: either the identity of the speaker is fabricated or stolen, or the content of the speech is invented or diverges from the original. Under this definition, the output of such AI-based processing should be considered ‘real’, since neither the speaker’s identity nor the spoken content has been altered. In practice, however, AI-based processing can introduce subtle artifacts and distributional shifts in the SSL representations (Satish et al., 2026), increasing the risk that these samples may be misclassified as fake in state-of-the-art binary deepfake detectors.

To address this emerging challenge, we first investigate the presence of synthetic artifacts of AI-based processing within the high-dimensional SSL feature spaces often employed for deepfake detection, including standard speech-centric models like Wav2Vec 2.0 and broader audio foundation models such as BEATs (a large-scale audio model that demonstrated effective generalisation and competitive performance in speech-domain tasks; Pianese et al., 2024).

As a preliminary example, we curated a test dataset by applying a commercial ‘benign’ AI-based processing algorithm (Acoustica *DeVerberate*) to an existing dataset of 24 ‘real’ speech samples (3 speakers each with 4 spontaneous and 4 read close-microphone recordings). We generated corresponding synthetic ‘fake’ samples with *Chatterbox* (Resemble AI's open-source TTS algorithm), using the original real speakers as target voices and the transcripts from the original real samples as prompt text. By holding the speech content and speaker identity consistent, we attempt to focus the analysis strictly on the artifacts of both synthetic speech generation and AI-based processing rather than phonetic or speaker-dependent variations.

Initial analysis of the SSL representations from this dataset reveals that AI-based processing induces a drift in the feature space, reducing separability of real and fake samples. This is visualised in Figure 1 via a 3D t-SNE (t-distributed Stochastic Neighbour Embedding) plot, which maps the high-dimensional relationships between three distinct classes: Real, Real-Processed (real samples with AI-based processing), and Fake (synthesised with TTS). The plot clearly illustrates that processed samples occupy a middle ground within the feature space, creating a ‘zone of ambiguity’ that would

compromise binary real/fake classifier performance. We have observed this issue in practice with both commercial and open source deepfake detectors.

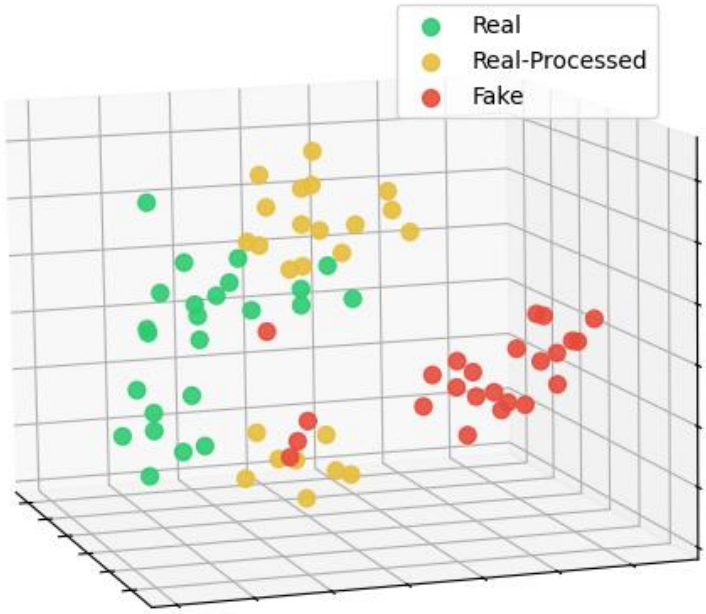


Figure 1. 3D t-SNE visualisation of BEATs embeddings across three domains: Real, Real-Processed (real samples with AI-based processing), and Fake (TTS synthesis). By using identity-consistent samples (same speaker and linguistic content), the plot highlights the distributional shifts caused by artifacts of both benign AI-based processing and of deepfake generation.

Building on this initial investigation, we will further explore the challenges of AI-based processing in forensic deepfake detection and propose mitigation strategies to offset the negative impact on detection performance. Specifically, we will present approaches for the detection of AI-processed samples and will discuss a broader approach to deepfake detection that moves beyond the binary 'real' or 'fake' to a multiclass approach for detecting specific processes that were applied to an audio file, potentially informing a more nuanced understanding of audio authenticity.

References

Pianese, A., Cozzolino, D., Poggi, G., Versoliva, L. (2024). Training-Free Deepfake Voice Recognition by Leveraging Large-Scale Pre-Trained Models. In *Proc. of the 2024 ACM Workshop on Information Hiding and Multimedia Security* (IH&MMSec '24). Baiona, Spain.

Satish, S. H. B., Lameris, H., Gustafson, J., Székely, É. (2026). What Counts as Real? Speech Restoration and Voice Quality Conversion Pose New Challenges to Deepfake Detection. *Under review for Interspeech 2026*. (Preprint available at arXiv:2603.14033).

Special Session
“Expertise in Context”

Special Session: Expertise in Context

This year's Special Session will focus on the implications of the judicial system for practical casework. It specifically addresses the role of the (voice) expert as a core element of many criminal trials and aims at putting it in context by addressing the interplay between experts, judges, and defense lawyers in different judicial systems. In the past, a good part of the discussions within IAFPA e.g. about the phrasing of conclusions or the role of formants in voice comparison has catered to the needs of working within the adversarial system. The ramifications of the inquisitorial system have not received as much attention, for instance in relation to methodology and the format of conclusions.

Since the majority of IAFPA members in general and conference participants in particular come from an adversarial background, they may not fully appreciate what it means to work within the inquisitorial system. The judicial framework, however, has consequences for the role of the expert and, to some extent, even for the manner in which the work is carried out.

Those members of IAFPA who have worked within both the adversarial and the inquisitorial systems of justice will have noted striking differences that will influence not only courtroom testimony but the handling of a case as such. Although there is considerable variation within the jurisdictions that come under the headings of "inquisitorial" and "adversarial", respectively, the session will focus on Germany and the United States as representing these two types of systems.

The session will start with four talks, followed by a panel discussion. The first two talks will outline some principles of the inquisitorial framework and then turn to the role of experts within it. The other two talks will represent the expert's perspective. We have acquired two speakers representing recipients of expert reports. One is Dr. Matthias Quarch, a presiding judge of a Trial Chamber (*Große Strafkammer*) at the Criminal Courts of Law in Aachen, the other is Prof. Jan Bockemühl, a renowned defense lawyer and Honorary Professor at the University of Regensburg. Dr. Quarch will briefly describe the inquisitorial system and then look at the role of expert reports from the court's perspective. What do judges expect of an expert? What makes a good expert and what are the criteria for a good report?

Prof. Bockemühl will bring in the defense point of view. He will cover some structural shortcomings in the present set-up from his perspective and talk about what he considers a good report or a good expert. Prof. Ruth Huntley Bahr (University of South Florida), who has many years of experience as an expert in the United States, will give an account of what it entails to work within the adversarial system. Prof. Angelika Braun (Trier University) will share her personal impressions of the differences between the two frameworks from an expert's perspective.

A panel discussion will then highlight the most salient characteristics of the jurisdictions with respect to expert reports as well as their implications for actual casework.

The Expert in German Criminal Trials

Matthias Quarch
Criminal Courts of Law, Aachen, Germany
Matthias.Quarch@lg-aachen.nrw.de

German criminal trials are governed by the principle of official investigation. The State has the burden of proof in showing that the defendant committed a crime. It is the duty of the prosecution and the court to find the facts which may prove the defendant's guilt. The defendant is considered innocent until proven guilty. S/he does not have to play an active part in a criminal trial, i.e. s/he has the right to remain silent. The court (consisting of one to five judges depending on the sentence to be expected) has to investigate the facts, determines the question of guilt or innocence and does the sentencing.

There are three stages in German criminal trials. The first stage is called "preliminary". This is the prosecutor's sphere. It is important to note that despite the name the prosecutor has to investigate in both directions and is therefore considered to be neutral. The judge's role is only to grant permission for e.g. placing a telephone tap or issue a search warrant, but s/he does not play an active part at this stage. It is the prosecutor's duty to begin an investigation *ex officio* if there is a sufficient suspicion that a crime has been committed.

Upon finishing her or his investigations the prosecutor has to decide whether to uphold or to drop the charges. S/he has to answer the question whether a conviction is more probable than an acquittal. If the answer is **"**yes**"**, that is the beginning of the second stage of the criminal law procedure: The In-between-stage. At this stage the criminal court examines the file in order to find out whether a conviction is more probable than an acquittal (the same question that the prosecutor had to answer at the first stage). If the court also arrives at a „yes", the Main stage of the criminal law procedure will start, and a trial date will be fixed.

In German criminal trials only four kinds of evidence are allowed: witnesses, experts, visual inspections, and documents. The expert is appointed by the prosecutor's office (in stage one) or by the criminal law court (in stage two or three). It is the solitary duty of the prosecutor or of the court to look for a qualified expert for the question at stake and to appoint her or him. An appointment by the court cannot be challenged by the prosecutor or the defense lawyers.

As the German criminal trial is an oral trial, the experts have to deliver their reports orally and in person. Most of the time they will have submitted a preliminary report before the trial starts as a basis for their oral testimony. But the court is only allowed to use the oral testimony for its findings. After having given his or her report, the expert is questioned first by the court, then by the prosecution and finally by the defense. There is nothing like evidence-in-chief or cross-examination. Anyone, including the defendant, may ask all the questions that they have. The expert is not dismissed until all questions have been answered.

The German Federal Court (Bundesgerichtshof) distinguishes witnesses from experts as follows: "Evidence provided by witnesses consists of perceptions of past […] facts and circumstances. This, however, was about providing the specialized knowledge that is necessary for assessing facts. This is typically an expert task." (BGH NJW 1993, 1796; translation A. Braun).

It is the court's duty to decide whether or not to follow the expert's conclusions. There is no obligation for the court to do so. It may appoint a second expert if it is not convinced by the first one's report. The prosecutor and the defense lawyers may also apply for a second expert, but it is up to the court to decide according to the relevant statutory provisions in the code of criminal procedure (*Strafprozessordnung*) whether or not to grant the application. In practice, it is very rarely granted. This decision can be used as an argument in an appeal, but that is seldom successful. There is only one expert in the vast majority of cases.

The Role of the Expert Witness in Criminal Proceedings in the Continental European Legal System

Jan Bockemühl
Attorney and University of Regensburg, Germany
j.bockemuehl@kanzlei-bockemuehl.de

The factual circumstances and forms of evidence in criminal proceedings have become increasingly complex, often touching upon subject areas where the participants in the trial reach the limits of their own knowledge. The range of topics is vast—extending from relatively straightforward blood-alcohol calculations to more difficult questions regarding criminal responsibility in the context of psychiatric diagnoses, accident reconstruction, and issues of document authorship or complex phonetic analysis. The potential subject matter is virtually endless. To answer these questions—some of which are decisive for the outcome of the trial—law enforcement authorities and courts rely on experts to convey the necessary specialized knowledge to the parties involved.

German criminal procedure is governed by the principle of official investigation (*Amtsermittlungs-grundsatz*). Under Section 244, Paragraph 2 of the Code of Criminal Procedure, the court is required to investigate the facts of the case *ex officio* (on its own initiative). If the court requires assistance in areas demanding expert knowledge, it calls upon expert witnesses. In this context, the expert's role is not merely that of an "assistant"; rather, if the expert reports on their own observations, they may also act as a (specialist) witness.

In principle, there are at least three ways an expert can be commissioned to provide an assessment in criminal proceedings: by the court, by the public prosecutor's office, and—less commonly known—by the defense. These various methods will be outlined with reference to the relevant statutory provisions.

This contribution will focus on the defense perspective on expert reports. It will specifically address the – somewhat complex – process by which the defense can introduce an additional expert of its own choosing into the criminal proceedings. The concept of an expert "self-commissioning" (or unilaterally engaging an expert) can be an effective means – alongside the "court-appointed" expert – of compelling the court to hear an additional expert.

Expert Testimony in the U.S.

Ruth Huntley Bahr
University of South Florida, Tampa, U.S,A,
rbahr@usf.edu

Voice experts in the U.S. face a combination of technical, legal, methodological, and ethical challenges that collectively shape the reliability and admissibility of voice evidence. Technologically, experts grapple with non-standardized protocols, variable recording conditions, and algorithmic biases despite the advancements in AI-enhanced voice biometrics.

Legally, two principal standards for the admission of scientific evidence are employed in the U.S. courts: Frye and Daubert. The Frye Standard depends upon establishing the general acceptance of an approach by the scientific community. If the expert's evidence is accepted in the pre-trial hearings, the jurors will be asked to evaluate the quality of the evidence in reaching their decision. In contrast, the Daubert Standard requires judges to function as "gatekeepers" to ensure evidence is scientifically valid and relevant and that the expert is qualified to present the evidence. Federal courts use Daubert, but many states continue to use Frye. These frameworks shape the admissibility of forensic voice evidence, but their inconsistent application across jurisdictions creates uncertainty for analysts and courts alike.

Experts must also communicate sophisticated concepts, such as likelihood ratios and probabilistic reasoning, in ways that are accessible to non-experts while maintaining scientific accuracy and not overwhelming the jury. The inquisitory nature of courtroom proceedings heightens these challenges, as experts must defend methodological choices under cross-examination, manage potential bias, and maintain clarity under pressure. Misunderstandings by judges or jurors can diminish the perceived credibility of expert testimony and increase the likelihood of misinterpretation.

Ethical challenges arise when lawyers are seeking experts to validate their case. The relationship between the expert and the lawyer is critical. The lawyer must understand the analysis process and the findings so that they can prepare their best case. They don't want to be surprised during cross-examination.

Finally, forensic voice examiners lack the large, diverse, and standardized datasets necessary for rigorous validation, and interdisciplinary collaboration remains limited. Without coordinated efforts to strengthen empirical foundations and develop standardized best practices, forensic voice analysis risks not providing the evidentiary rigor demanded by today's courts.

Expert Testimony across Jurisdictions

Angelika Braun
University of Trier, Germany
brauna@uni-trier.de

This contribution represents the perspective of an expert who has worked in eight different jurisdictions and testified in five of them and can thus report from experience within the adversarial as well as the inquisitorial framework. Some impressionistic observations will be discussed.

Both the content of the expert report and the way it is presented in court are likely to vary with the judicial system. Two issues which immediately come to mind are the consequences of having to follow precedents and the number of experts involved in a case. In the adversarial framework, the content of the report is to some extent determined by higher court rulings, e.g. on including formants in every forensic voice comparison. Following these precedents would not be mandatory in the inquisitorial system. In the adversarial framework, the testimony is often carefully orchestrated and may depend on the strategy of the side that commissioned the report more than on a comprehensive rendition of the findings. In the inquisitorial system, the expert usually submits a full report.

In the adversarial system, opposing experts are the rule, which effectively means that there are some built-in checks and balances in assessing the expert's work. In the inquisitorial system, there is generally just one – usually court-appointed – expert who is expressly instructed to be impartial and reminded of that immediately prior to testifying. However, this will not work unless the expert is knowledgeable as well as conscientious. Charlatanry, when it comes under a guise that is rhetorically well-versed and generally well-dressed, may remain undetected.

Roundtable Session on Checking Procedures

Checking Procedures

Bryony Nuttall[1]*, Katherine Earnshaw*[1] *and Richard Rhodes*[1]
[1]*The Forensic Voice Centre, York, UK*
bryony.nuttall@forensicvoicecentre.com

The Forensic Voice Centre is hosting a roundtable discussion between practitioners within the IAFPA community regarding checking and quality assurance procedures. Our aim is to discuss current procedures in the UK and internationally and to provide a platform to develop and share solutions. The session is open to all conference attendees but is tailored to practitioners who are involved in (checking) casework. The session will involve audience participation, as well as a panel discussion between practitioners.

Context

All forensic units performing forensic analysis in the UK must "have a procedure for carrying out appropriate checking and review", according to the UK Forensic Science Regulator's Code of Practice, to ensure that appropriate procedures have been followed and to reduce errors in analyses and reports. This view is supported across international standards and organisations. However, no specific guidance currently exists for forensic speech and audio laboratories on what internal procedures should include. Whilst it could be argued that prescriptivism is neither advantageous nor practical, the lack of standardisation risks inconsistency across forensic laboratories. We believe better communication and information-sharing amongst the IAFPA community regarding this issue can alleviate this risk, so we would like to use this opportunity to learn from other practitioners about their quality assurance procedures and discuss our own methods.

Discussion points

In this session we will discuss the following topics:

- Open, blind, and semi-blind checking approaches [i.e., with or without sight of the first analyst's conclusions/findings, or with sight of some]
- Approaches to reviewing analytical findings and conclusions
- Documentation and recording of corrections / comments
- Differences in checking procedures across case types
- Balancing quality assurance, efficiency, and impartiality
- Opportunities for greater consistency across the forensic speech community

We will also welcome questions and discussion points from the audience during this session. If you would like to submit a question or topic in advance, please contact Bryony Nuttall at bryony.nuttall@forensicvoicecentre.com.

References

[UK] Forensic Science Regulator. (2025). Forensic science activities: Statutory Code of Practice (version 2 – May 2025) - [URL: https://assets.publishing.service.gov.uk/media/68401f4be550203c8209cce9/2025_05_20_Code_of_Practice_final.pdf]

Poster Session I

Individual Variation in the Production of Austrian German Rhotics

Nathalie Elsässer[1], *Hendrik Behrens-Zemek*[1], *and Eva Reinisch*[1]
[1]*Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria*
{nathalie.elsaesser|hendrik.behrens-zemek|eva.reinisch}@oeaw.ac.at

The phonological class of rhotics ("r-sounds") is highly variable, both in terms of place and manner of articulation (POA/MOA). This variation may be due to sociolinguistic or phonological factors, leading to significant individual variation (e.g. Sebregts (2015) for Dutch, Rennicke (2015) for Brazilian Portuguese). For Austrian German (AG) both alveolar and uvular POA have been described (Moosmüller et al., 2015; Ulbrich & Ulbrich, 2007), however, informal observation suggests that speakers produce either alveolar *or* uvular variants. It remains unclear whether there might be differences in MOA or the exact phonetic realization of rhotics (e.g., voicing, number of vibrations in trills etc.) among different speakers of the same POA, and if the distribution of variants across phonetic contexts is individual. The goal of this study is to investigate intra- and inter-speaker variation in AG rhotics to explore whether phonological variation patterns may provide speaker-specific information relevant to forensic speaker identification and verification.

We recorded 62 monolingual AG speakers using audio and Ultrasound Tongue Imaging (UTI). Participants were aged 18–45, resided in Austria and reported AG as their (only) native language. Participants read 319 words containing a rhotic in syllable-onset position either as sole onset or in a cluster (e.g., ***r**ennen/b**r**ennen* (run/burn)), word-initially or medially (e.g., ***R**olle*/*ein**r**ollen* (role/roll up)) and in a stressed or unstressed syllable (e.g., *ver.**r**ei.sen*/*ein.**r**ei.sen* (travel/enter)). Words were embedded in a phonetically neutral pre-context. Audio recordings were annotated automatically by WebMAUS (Kisler et al., 2017) and manually corrected by four phonetically-trained annotators. Rhotics were segmented and classified by POA (uvular/alveolar), MOA (fricative/approximant/tap/trill/fricative-trill), voicing (voiced, unvoiced, voiced-unvoiced, unvoiced-voiced), and, if it was a tap/trill, number of closures. UTI recordings were consulted for verification. So far, 19 participants have been analyzed.

Analysis of the annotated labels confirmed that participants are consistent in their POA, but variation was found in MOA and voicing. Figure 1 shows the proportion of produced variants per speaker (SP). While none of the speakers used only one variant, the distribution of variants varied greatly, even within each POA. For example, SP9 produced almost exclusively uvular fricatives (88.74 %) and approximants (9.37 %), while SP6 used predominantly uvular trills (37.07 %) and taps (18.06 %). For the alveolar POA, SP56 stands out by producing 44.34 % alveolar fricatives, 21.07 % fricative-trills and only 1.89 % taps, whereas most speakers from the alveolar POA tended to produce mainly taps. Pairwise comparisons of a linear-mixed regression model using the occurrence of a tap as response variable, speaker as fixed and word as random effect revealed that the likelihood of SP56 producing a tap differs significantly ($p<0.001$) from all other speakers.

In sum, each speaker produced a range of rhotic variants. While variation was mainly triggered by phonological context, individual patterns of variation significantly differed between speakers. In particular, some individuals stood out by rarely using variants that otherwise were most common among the analyzed speakers. These findings suggest that rhotic realization patterns may carry speaker-specific information and could therefore contribute to forensic speaker comparison. However, further research is required to determine the stability of these patterns across recording sessions and speaking styles.

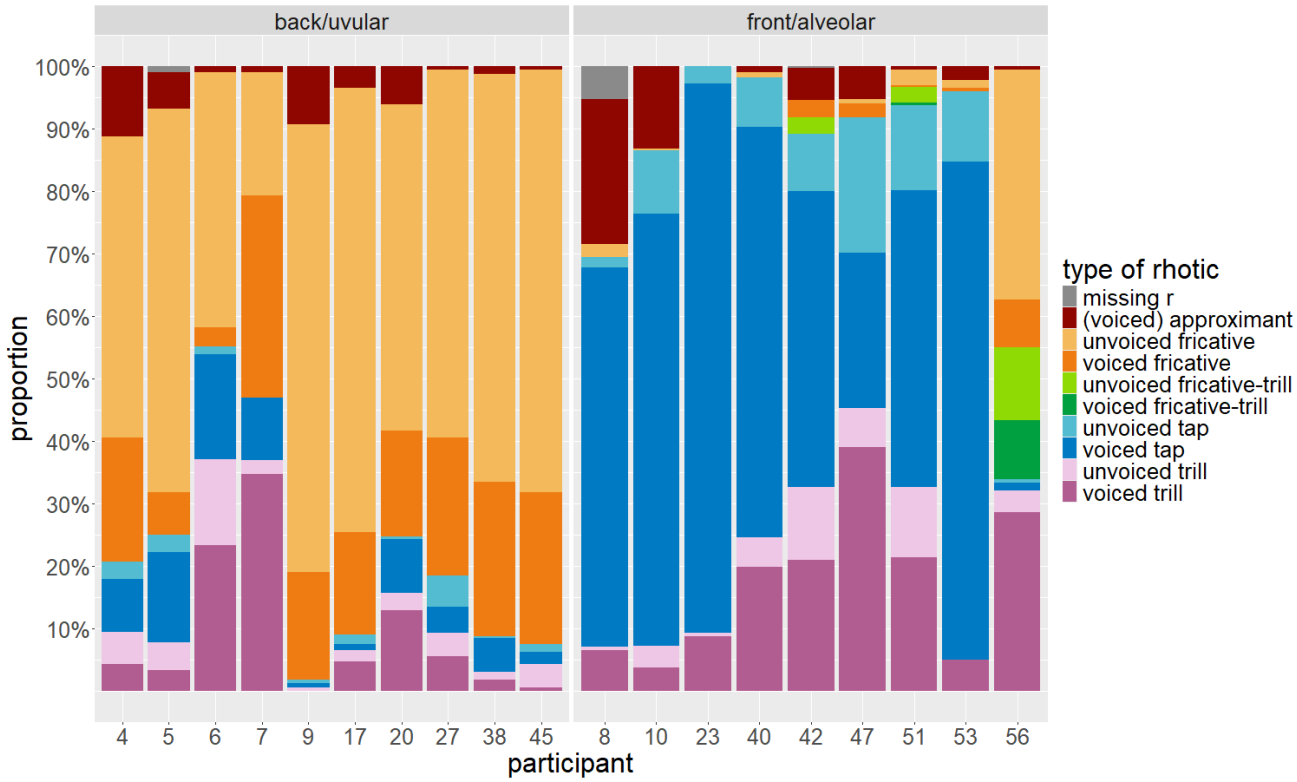


Figure 1. Proportions of produced rhotic variants per speaker. Speakers using back/uvular rhotic variants are pictured on the left, speakers using front/alveolar rhotic variants on the right.

References

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. Computer Speech & Language, 45, 326 – 347. https://doi.org/10.1016/j.csl.2017.01.005

Moosmüller, S., Schmid, C., & Brandstätter, J. (2015). Standard Austrian German. *Journal of the International Phonetic Association*, *45*(3), 339 – 348. https://doi.org/10.1017/ S0025100315000055

Rennicke, I. (2015). *Variation and Change in the Rhotics of Brazilian Portuguese* [Dissertation]. University of Helsinki, Helsinki, Finland.

Sebregts, K. (2015). *The sociophonetics and phonology of Dutch r* [Dissertation]. Netherlands Graduate School of Linguistics, Utrecht, Netherlands.

Ulbrich, C., & Ulbrich, H. (2007). The Realisation of /r/ in Swiss German and Austrian German. In *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007),* Saarbrücken, Germany.

The Effect of Speech Rate on Dynamic Acoustic Features of Mandarin Triphthongs

Kai Gao, Weili Jiang, Xin Chen, Wenlin Huang, Wei Li
huangwenlin@cifs.gov.cn
Institute of Forensic Science, Ministry of Public Security, China

Introduction

Mandarin triphthongs are core acoustic features for forensic voice identification, owing to their stable spectral traits and unique dynamic articulatory properties that provide distinct individual phonetic markers. While existing forensic phonetics research has explored gender differences in mandarin triphthongs' formant trajectories, the modulating effect of speech rate on the dynamic acoustic features of triphthongs remains insufficiently investigated, with unclear gender-related variation rules under different speech rate conditions. To address this limitation, the study systematically examines the impact of slow, normal, and fast speech rates on the dynamic features of four Mandarin triphthongs (/iau/, /iou/, /uai/, /uei/) and analyzes gender-specific differences in these features across speech rate conditions.

Methods

A total of 40 native Mandarin speakers (20 males and 20 females, Beijing natives, mean age 21 years, no vocal disorders) participated in the experiment. Thirty-two target syllables with the four triphthongs as rhymes (covering all four Mandarin tones: level, rising, falling-rising, falling) were recorded in a quiet laboratory with a 44100 Hz sampling rate and 16-bit precision. Praat software was employed for speech data annotation, preprocessing, and vowel segment verification by native phoneticians. The first three formants (F1, F2, F3) of each vowel segment were extracted at 15 equally spaced time points to ensure data comparability.

Results

Speech rate was found to exert a significant and regular influence on the acoustic features of Mandarin triphthongs, with consistent gender differences maintained across all speech rate conditions. For formant features, the fluctuation range and change rate of F1, F2, and F3 all decreased with accelerated speech rate, with F2 (the core indicator of tongue position front-back movement) showing the most pronounced shrinkage in variation range; F3 exhibited the weakest sensitivity to speech rate, maintaining relatively stable fluctuation characteristics. The key time points of formant change were either delayed or advanced with speech rate acceleration, while F1's key variation time point remained highly stable across conditions. Females presented significantly higher overall F1 and F3 values and larger formant fluctuation ranges than males, reflecting more expansive tongue movement in female articulation. For non-formant features, triphthong duration shortened synchronously with faster speech rate, with females having longer original duration and a slightly higher shortening ratio than males, indicating stronger adaptability to speech rate changes. Sound intensity showed a trend of blurred peaks and narrowed fluctuations with accelerated speech rate.

Discussion

This study confirms that speech rate-induced variations in Mandarin triphthongs' acoustic features follow measurable and consistent rules, and gender-specific phonetic characteristics remain stable even in fast speech rate conditions with compressed articulatory movement. The significant shrinkage of F2's fluctuation range with accelerated speech rate is attributed to the compression of tongue position front-back movement, the core articulatory process of triphthongs, while F3's insensitivity to speech rate makes it a robust acoustic marker for forensic voice identification, complementing the findings of gender differences in F3 for triphthong formant trajectories. The longer duration and higher shortening ratio in females reflect more sufficient initial articulatory movements and greater adaptive adjustment to speech rate changes, while the stable 1-2 dB intensity difference between genders is a persistent paralinguistic feature independent of speech rate.

References

Cao Honglin, Wang Yujing, Li Jingyang.(2017). Influence of speech rate on dynamic formant characteristics of triphthongs [J]. *Journal of Tsinghua University (Science and Technology)*, 57(9): 958-962.

Cao Honglin, Kong Jiangping (2013). Forensic speaker comparison by using longterm formant distribution. *Chinese Journal of Forensic Sciences*, 66(1):62-67 (in Chinese).

McDougall K. (2002). Speaker characterising properties of formant dynamics: a case study [C]//Proc of 9th Australasian International Conference on SST. Melbourne, Australia, 403-408.

The Social Salience of tapped /r/ [ɾ] in Dublin Accents and Linguistic Attitudes Towards it in a Forensic Context

Elin Cat Stüer[1], *Grace Wenling Cao*[2],
[1, 2] *School of Languages, Cultures and Linguistics, University College Dublin*
[1]elin.stuer@ucdconnect.ie [2]grace.cao@ucd.ie

Dublin English has been researched with different approaches identifying linguistic attitudes and phonetic features (Hickey, 2005; Lonergan, 2013). Studies have found that North Dublin varieties are more stigmatised and associated with a lower social class, whereas South Dublin accents are perceived as higher status (Hickey, 2005). Lonergan (2013) therefore claims that the terms "North" and "South" do not accurately represent geographical locations but are proxies for social class. Research into North Dublin English has identified vowel realisations in detail, but research into consonants falls short, despite claims that realisations of /r/ are quite variable (Lonergan, 2013). The present study is interested in the /r/-sound in North Dublin English and asks the following research question: Is the tapped [ɾ] a linguistic marker of North Dublin English and what social meaning does it carry in forensic contexts?

To answer this, two experiments were conducted through SurveyMonkey with 34 participants (20 females, 13 males, 1 non-binary). To mimic a forensic-realistic scenario, participants were told to imagine that they were hearing people on the witness stand and were primed by a court photo at the beginning of the experiments. Experiment 1 included a verbal-guise study with ten different guises, representing four Dublin accents: North Dublin, South Dublin, New D4 and Old D4 and two filler guises of American and British English. The participants rated these guises on nine traits on a six-point Likert scale. The second experiment was inspired by Labov et al. (2011): a male and a female speaker of North Dublin English read a pre-designed dialogue containing six r-words twice, once pronouncing all the r-words with a tap [ɾ] and once with an alveolar approximant [ɹ]. Then three audio clips were prepared, each of which included 0%, 50% and 100% of the tapped-r. The same group of listeners rated these audio clips on four traits on a semantic differential scale from 0-10: *North Dublin*, *posh*, *trustworthy* and *natural*.

By performing a Principal Component Analysis (PCA) on the data from the first experiment, two dimensions *Status* and *Solidarity* were gained (see Table 1). The results mostly confirmed findings of previous studies (Hickey, 2005), that North Dublin accents are perceived as significantly lower in status ($p < .001$) and slightly higher in solidarity ($p = 0.14$; see Fig. 1). Recordings with a higher percentage of [ɾ] were rated as more *North Dublin* and less *posh* (see Fig. 2). While *trustworthy* was not significant, the 0% tapped version of the female speaker was rated slightly higher than the others ($p = 0.51$).

This study has valuable forensic implications. Firstly, the identification of [ɾ] as a linguistic variable of North Dublin accents is valuable for speaker identification and speaker profiling. Secondly, the forensic-related traits *trustworthy* and *reliable* should be examined separately to more accurately simulate a forensic context. The slightly higher rating of trustworthiness for the 0% version of the female North Dublin speaker might indicate that women are more likely to be affected by their North Dublin accents in a forensic context compared to men. (Words: 500)

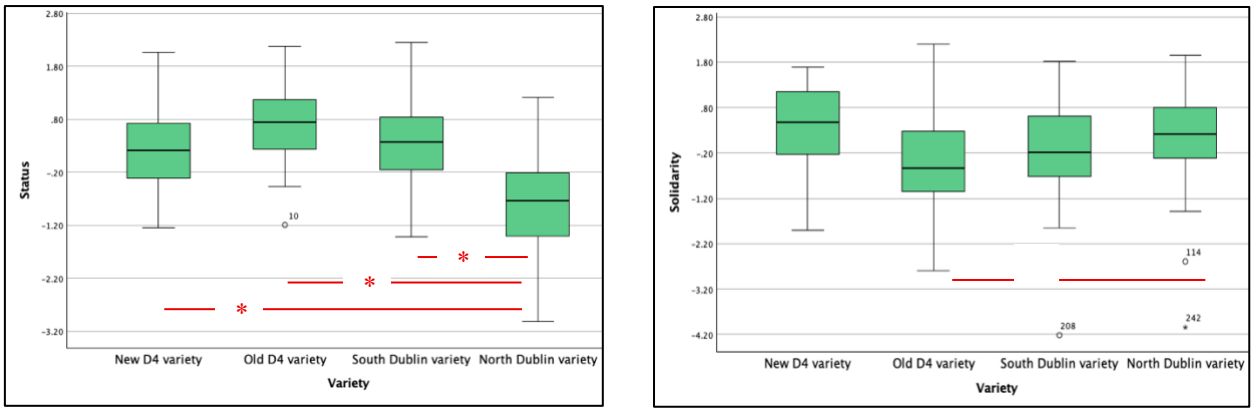


Figure 1: Results from the Verbal-Guise experiment- A higher value indicates a more positive

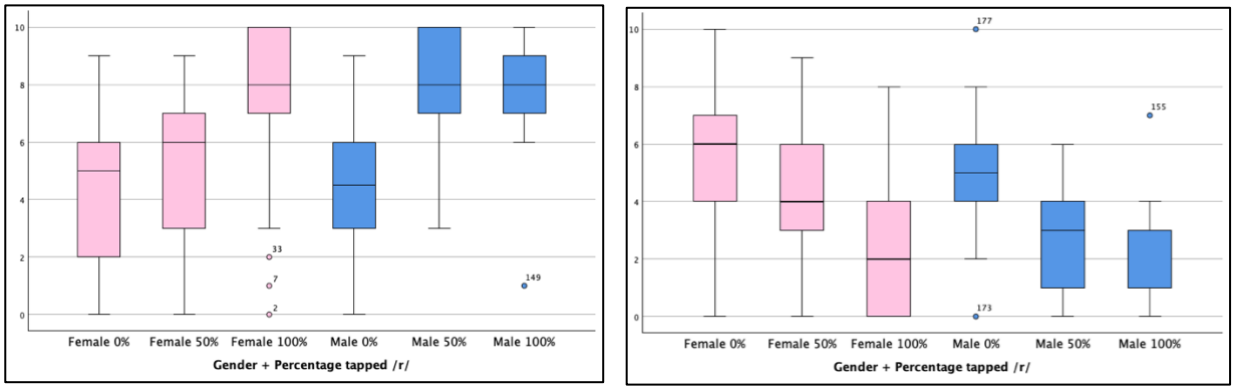


Figure 2. Results from Experiment 2 for traits 'North Dublin'(left panel) and 'posh'(right panel). A higher value indicates a stronger attitude toward the trait.

	1 (Solidarity)	2 (Status)
approachable	.845	.002
trustworthy	.816	.190
reliable	.733	.283
friendly	.733	-.109
funny	.644	-.269
posh	-.201	.845
well educated	.166	.842
higher social class	-.163	.840
ambitious	.208	.582

Note: Factor loadings over 0.40 are highlighted

Table 1. Results from the PCA

References

Hickey R. (2005). *Dublin English: Evolution and Change* (1st ed.). J. Benjamins.

Labov, W., Ash, S., Ravindranath, M., Weldon, T., Baranowski, M. and Nagy, N. (2011). Properties of the sociolinguistic monitor. *Journal of Sociolinguistics*, 15(4), 431-463. https://doi.org/10.1111/j.1467-9841.2011.00504.x.

Lonergan, J. (2013). *An Acoustic and Perceptual Study of Dublin English Phonology*. [PhD thesis]. University College Dublin.

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 Communications, 9(1462013), 1-16.

/s/ and /ʃ/ Variation Across Gender-Diverse Speakers: Implications for Forensic Casework Involving Minority Speaker Populations

Katelyn Taylor and Kirsty McDougall
Department of Theoretical and Applied Linguistics, University of Cambridge
klt56@cam.ac.uk; kem37@cam.ac.uk

Phonetic and sociolinguistic research has extensively explored how sex-based factors shape speech production, yet most existing studies and corpora remain centred on cisgender speakers, leaving gender-diverse voices underrepresented (Zimman, 2018, 2021; Merritt, 2023). This gap poses challenges for forensic speaker comparison or identification where transgender or non-binary individuals could become involved, and it affects broader work on voice perception where assumptions about binary gender categories are often taken for granted.

Sibilant consonants such as /s/ and /ʃ/ have long been recognised as salient markers in the negotiation of gender and sexuality. Research has shown that variation in these sounds can index social identity across dimensions including gender, class, region, and ethnicity. However, the origins of these associations remain debated. Phonetically oriented studies often attribute gendered differences in /s/ production to sex-based anatomical variation (Fuchs & Toda, 2010), while sociolinguistic research emphasises learned articulatory habits and the role of social practice (Avery & Liss, 1996; Zimman, 2017). These competing perspectives highlight the complex interplay between biological and social factors in speech production.

Building on prior perceptual findings that listeners can somewhat successfully attribute speakers' gender identities from speech cues, the present study investigates whether acoustic variation in /s/ and /ʃ/ aligns more closely with gender identity or biological sex, contributing to broader discussions of how phonetic detail encodes social meaning.

Twenty-four speakers participated in this study, representing six gender identity groups, outlined in Table 1. Each speaker was recorded reading the Rainbow Passage twice (Fairbanks, 1960). The recordings were processed using Montreal Forced Alignment and manually checked to ensure accurate boundary alignment and reliable IPA transcription. Acoustic measurements of Centre of Gravity (CoG), intensity, skewness, and kurtosis will be extracted using the Vargo (2025) Praat Mass Analyzer script. These variables will be analysed to determine whether variation in /s/ and /ʃ/ productions aligns more closely with previous phonetic or sociolinguistic findings.

Gender Identity	*Number of speakers*	*Sex assigned at birth*
Cisgender Male	4	Male
Cisgender Female	4	Female
Non-binary	8 (4+4)	Male and Female
Transgender Male	4	Female
Transgender Female	4	Male
Sum:	24	

Table 1. An outline of the sex assigned at birth and gender identity of the 24 speakers recorded reading the Rainbow Passage for the present study.

As this experiment is ongoing, the acoustic analysis and statistical modelling are still being completed. Findings and emerging patterns and their implications for forensic casework involving speakers from minority populations will be presented at the conference.

References

Avery, J. D., & Liss, J. M. (1996). Acoustic characteristics of less-masculine-sounding male speech. *The Journal of the Acoustical Society of America*, *99*(6), 3738-3748.

Fairbanks, Grant (1960). Voice and articulation drillbook. New York: Harper & Row.

Fuchs, Susanne & Martine Toda. 2010. Do differences in male versus female /s/ reflect biological or sociophonetic factors? In Susanne Fuchs, Martine Toda & Marzena Zygis (eds.), An interdisciplinary guide to turbulent sounds,281–302. Berlin & New York: Mouton de Gruyter.

Merritt, B. (2023). Speech beyond the binary: Some acoustic-phonetic and auditory-perceptual characteristics of non-binary speakers. *JASA Express Letters*, *3*(3).

Vargo, Julian (2025). Praat Mass Analyzer [Praat Script]

Zimman, L. (2017). Variability in /s/ among transgender speakers: Evidence for a socially grounded account of gender and sibilants. *Linguistics*, *55*(5), 993-1019.

Zimman, L. (2018). Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, *12*(8), e12284.

Zimman, L. (2021). Gender diversity and the voice. In *The Routledge handbook of language, gender, and sexuality* (pp. 69-90). Routledge.

Phoneme Inventory Effects on Speaker Specificity in /n/: A Cross-Linguistic Study

Annie Baker[1], *Eleanor Chodroff*[1]
[1]*Department of Computational Linguistics, University of Zurich, Switzerland*
annie.baker@uzh.ch eleanor.chodroff@uzh.ch

Introduction

Nasal segments have been shown to exhibit relatively high levels of speaker specificity across numerous languages including Standard Dutch (Smorenburg & Heeren, 2021), Japanese (Amino & Arai, 2009), Standard Southern British English and Leeds English (Kavanagh, 2012), as well as Spanish and Catalan (Gavalda & Cicres, 2025). The present study investigates whether the phoneme inventory influences the extent of speaker-specific variation by examining the degree of speaker specificity in the alveolar nasal /n/ across languages with considerably different nasal inventories. We compare Abkhaz, whose phoneme inventory includes two nasals /n/ and /m/, with Tamil, which has six (/n, m, n̪, ɲ, ɳ, ŋ/). We hypothesise that speakers of languages with few nasal contrasts may have greater freedom for individual variation, being that they are less constrained by the need to preserve a large number of linguistic contrasts, in this case along the place of articulation dimension. As nasal place contrasts are cued acoustically by spectral properties arising from vocal-tract filtering and antiresonance structure (Stevens, 2000), the analysis focuses on acoustic measures which may encode both phonological and speaker-specific information.

Methods

Speech data for this study come from the Mozilla Common Voice corpus (Ardila et al., 2020) version 17, with corresponding time-aligned TextGrids from the VoxCommunis Corpus (Ahn & Chodroff, 2022; Zhang et al., 2025a; Zhang et al., 2025b). Grapheme-to-phoneme conversion was implemented using the XPF Corpus for Abkhaz (Cohen Priva et al., 2021) and Epitran for Tamil (Mortensen et al., 2018). For each language, 200 speakers with at least 10 tokens of /n/ were chosen at random from the broader corpus. The following acoustic properties were then measured in hertz from the middle 20 ms of each /n/ segment following multitaper spectral analysis (Shadle et al., 2012; Chodroff & Wilson, 2014): centre of gravity (COG), spectral standard deviation (SD) (square root of spectral variance) and the spectral minimum between the estimated F1 and F2 values, representing the first anti-formant (Stevens, 2000; Kavanagh, 2012; Gavalda & Cicres, 2025). F1 and F2 of the nasal were estimated using Praat with the utterance-specific parameters for formant extraction based on the mean f0 of the inputted utterance. The Jensen-Shannon divergence (JSD) was then calculated (Lin, 1991) between each pair of speaker distributions for each measure. JSD provides a symmetric distance metric between two distributions, with higher JSD values representing greater distance between the distributions, and therefore higher speaker specificity.

Results and Discussion

For all three acoustic properties (COG, SD, spectral minimum), a linear mixed effects model was fitted to predict JSD, with the fixed effect of language and a random by-speaker intercept. The preliminary results show that, contrary to expectations, the JSD values were significantly higher for Tamil than for Abkhaz across all three acoustic properties, with the most marked difference being JSD for COG (see Figure 1). The Tamil data also showed higher variance for all JSD measures, as indicated by the standard deviations. These findings suggest that languages differ in speaker specificity for nasal segments, but phoneme inventory size may not reliably predict this variation, at least along these dimensions. Next steps for this project include further investigation into establishing relevant acoustic dimensions for place of articulation contrasts, and analysis of further spectral measures such as amplitude measures.

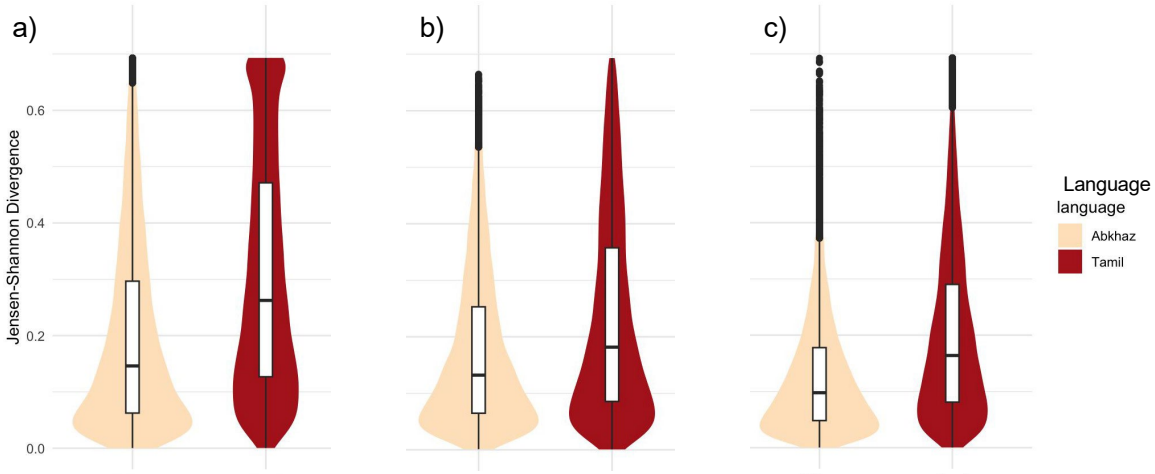


Figure 1. Violin and boxplots showing the Jensen-Shannon divergence values for a) centre of gravity,b) spectral standard deviation, and c) spectral minimum for Abkhaz and Tamil.

References

Ahn, E., & Chodroff, E. (2022). *VoxCommunis: A corpus for cross-linguistic phonetic analysis*. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 5286–5294).

Amino, K., & Arai, T. (2009). Speaker-dependent characteristics of the nasals. *Forensic Science International, 185*(1–3), 21–28.

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. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 4218-4222).

Chodroff, E., & Wilson, C. (2014). Burst spectrum as a cue for the stop voicing contrast in American English. *The Journal of the Acoustical Society of America, 136*(5), 2762–2772.https://doi.org/10.1121/1.4896470

Cohen Priva, Uriel and Strand, Emily and Yang, Shiying and Mizgerd, William and Creighton, Abigail and Bai, Justin and Mathew, Rebecca and Shao, Allison and Schuster, Jordan and Wiepert, Daniela. (2021). The Cross-linguistic Phonological Frequencies (XPF) Corpus

Gavaldà, N., & Cicres, J. (2025). An exploration of the language independence and speaker-specificity of /m/ and /n/ in speakers of Catalan, Spanish, and English. *International Journal of Bilingualism*. Advance online publication. https://doi.org/10.1177/13670069251380015

Kavanagh, C. (2012). *New consonantal acoustic parameters for forensic speaker comparison* [Doctoral dissertation, University of York].

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory, 37*(1), 145–151. https://doi.org/10.1109/18.61115

Mortensen, D. R., Dalmia, S., & Littell, P. (2018, May). Epitran: Precision G2P for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).*

Shadle, C. H., Cohn, A., Fougeron, C., & Huffman, M. (2012). Acoustics and aerodynamics of fricatives. In A.C. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 511–526). Oxford University Press.

Smorenburg, B., & Heeren, W. (2021). Forensic value of acoustic-phonetic features from Standard Dutch nasals and fricatives. In *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications – Book of Abstracts* (pp. 46–47).

Zhang, M., Ahn, E., Baker, A., & Chodroff, E. (2025, February). *VoxCommunis Corpus*. OSF Repository. https://doi.org/10.17605/OSF.IO/T957V

Zhang, M., Farhadipour, A., Baker, A., Ma, J., Pricop, B., & Chodroff, E. (2025). Quantifying and reducing speaker heterogeneity within the Common Voice corpus for phonetic analysis. In *Proceedings of Interspeech 2025*, 3933-3937, doi: 10.21437/Interspeech.2025-2027

Fundamentally challenging: The role of f_0 in cross-language speaker discrimination

Leah Bradshaw[1,2]*, Julien Plante-Hebert*[3] *and Elizabeth Allyn-Smith*[3]
[1]*Department of Computational Linguistics, University of Zurich, Switzerland*
{leah.bradshaw}@uzh.ch
[2]*JP French International, UK.*
{leah.bradshaw}@jpfrench.com
[3]*Département de Linguistique, Université du Québec à Montréal, Montréal, QC, Canada*
{plante-hebert.julien| smith.elizabeth_allyn}@uqam.ca

This project investigates how bilingualism affects speaker discrimination, focusing specifically on how language-dependent variation in fundamental frequency (f_0) influences cross-language indexical processing. Indexical information conveys speaker-specific traits such as identity, accent, and emotional state, enabling listeners to recognise and distinguish individuals (Abercrombie, 1967). However, indexical processing is highly variable and context dependent, particularly when speakers shift language or speaking style (e.g., (Bartle & Dellwo, 2015; Perrachione, 2017; Stevenage et al., 2021). Despite bilingual speakers representing approximately 43% of the global population (Jayanath, 2021), research on cross-language voice processing has largely focused on monolingual listeners encountering unfamiliar languages, leaving bilingual speech processing comparatively understudied. This gap is especially relevant in forensic phonetics where multilingual speakers are increasingly common yet difficult to reliably analyse.

Previous research demonstrates a robust Language Familiarity Effect (LFE), whereby listeners recognise and discrimination voices more accurately in familiar than unfamiliar languages (overview papers: Levi, 2019; Perrachione, 2017). Studies across several languages show that even when listeners become familiar with bilingual speakers, recognition accuracy decreases when those speakers switch languages (Goggin et al., 1991; Orena et al., 2019; Winters et al., 2008). One possible explanation lies in bilingual speech production itself: bilingual speakers often exhibit systematic language-dependent shifts in acoustic properties, particularly f_0. Such differences have been reported for Japanese–English (Graham, 2015), Korean–English (Cheng, 2020), Russian–English (Altenberg & Ferrand, 2006), and French–English bilinguals (Schwab & Goldman, 2016). It is known that f_0 is a well-established cue for speaker identity (e.g., Dellwo et al., 2018; Nolan et al., 2011; Skarnitzl & Vaňková, 2017), therefore, these shifts may directly impair listeners' ability to identity speakers across languages.

The present study will examine whether variation in f_0 affects cross-language speaker discrimination. Specifically, we ask: (1) how naturally occurring f_0 variability contributes to indexical processing, and (2) whether consistent versus inconsistent f_0 patterns across languages facilitate more accurate speaker discrimination when bilingual speakers switch languages. Read speech recordings were collected from 6 early French–English bilinguals in Montréal who acquired both languages before school age. Participants completed two recording blocks, one in French and one in English, with randomised order across speakers. Using controlled acoustic manipulations in Praat, these recordings were used to generate stimuli for a perceptual experiment whereby individual speaker stimuli was created with both consistent and inconsistent f_0 patterns across languages. The perceptual experiment will be conducted with three listeners groups: Canadian English monolinguals, Canadian French monolinguals and Quebec French–English bilinguals. Participants will complete a cross-language voice discrimination task in which they will judge whether the speech was produced by the same or two different speakers. In all trials, listeners will be presented with a pair of voices with one snippet presented in English and one in French. For same-speaker trials, stimuli will either have the speakers' natural f_0 variability across the two languages, or their median f_0 will be manipulated to be the same in the two samples. It is predicted that listeners will discriminate speakers more accurately when f_0 is held consistent across the two languages.

References

Abercrombie, D. (1967). *Elements of general phonetics.* Edinburgh University Press.

Altenberg, E. P., & Ferrand, C. T. (2006). Fundamental Frequency in Monolingual English, Bilingual English/Russian, and Bilingual English/Cantonese Young Adult Women. *Journal of Voice*, *20*(1), 89–96. https://doi.org/10.1016/j.jvoice.2005.01.005

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech Language and the Law*, *22*(2), 229–248. https://doi.org/10.1558/ijsll.v22i2.23101

Cheng, A. (2020). Cross-linguistic f0 differences in bilingual speakers of English and Korean. *The Journal of the Acoustical Society of America*, *147*(2), EL67–EL73. https://doi.org/10.1121/10.0000498

Dellwo, V., Kathiresan, T., Pellegrino, E., He, L., Schwab, S., & Maurer, D. (2018). Influences of Fundamental Oscillation on Speaker Identification in Vocalic Utterances by Humans and Computers. *Interspeech 2018*, 3795–3799. https://doi.org/10.21437/Interspeech.2018-2331

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

Graham, C. (2015). Fundamental Frequency Range in Japanese and English: The Case of Simultaneous Bilinguals. *Phonetica*, *71*(4), 271–295. https://doi.org/10.1159/000381627

Jayanath, T. (2021). A comparative study on bilingualism and multilingualism. *KnowEx Social Sciences*, *1*(02), 79–91.

Levi, S. V. (2019). Methodological considerations for interpreting the Language Familiarity Effect in talker processing. *WIREs Cognitive Science*, *10*(2), e1483. https://doi.org/10.1002/wcs.1483

Nolan, F., McDougall, K., & Hudson, T. (2011, August). Some acoustic correlates of perceived (dis) similarity between same-accent voices. In *ICPhS* (Vol. 17, pp. 1506-1509).

Orena, A. J., Polka, L., & Theodore, R. M. (2019). Identifying bilingual talkers after a language switch: Language experience mattersa). *The Journal of the Acoustical Society of America*, *145*(4), EL303–EL309. https://doi.org/10.1121/1.5097735

Perrachione, T. K. (2017). *Speaker recognition across languages*. Oxford University Press.

Schwab, S., & Goldman, J.-P. (2016). Do speakers show different F0 when they speak in different languages? The case of English, French and German. *Speech Prosody 2016*, 6–10. https://doi.org/10.21437/SpeechProsody.2016-2

Skarnitzl, R., & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *AUC PHILOLOGICA*, *2017*(3), 7–17.

Stevenage, S. V., Tomlin, R., Neil, G. J., & Symons, A. E. (2021). May I Speak Freely? The Difficulty in Vocal Identity Processing Across Free and Scripted Speech. *Journal of Nonverbal Behavior*, *45*(1), 149–163. https://doi.org/10.1007/s10919-020-00348-w

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. https://doi.org/10.1121/1.2913046

One Speaker, Multiple Dialects: Insights from the Language Familiarity Effect for Forensic Speaker Recognition

Sophie Möller[1], *Radek Skarnitzl*[1], *Volker Dellwo*[2], *Elisa Pellegrino*[2], *Tyler Perrachione*[3]

[1]*Institute of Phonetics, Charles University, Czech Republic*
{sophie.moeller|radek.skarnitzl}@ff.cuni.cz

[2]*Department of Computational Linguistics, University of Zurich, Switzerland;*
{elisa.pellegrino|volker.dellwo}@uzh.ch

[3]*Department of Speech, Language and Hearing Sciences, Boston University, United States*
tkp@bu.edu

Introduction

In forensic speaker comparison, practitioners may encounter language mismatch conditions involving standard and dialect speech across recordings. During auditory assessments by native-language experts, similarities in vocal characteristics may suggest that recordings originate from the same speaker, whereas language- or dialect-related differences may instead indicate different speakers.

Background

The present study investigates how linguistic and voice quality characteristics influence native (naïve) listeners' judgments of speaker identity under dialect mismatch conditions. This question is particularly relevant in sociolinguistic contexts such as German-speaking Switzerland, where dialects play a central role in social identity and perception, and where unexpected shifts in salient dialectal features may interfere with voice recognition by native listeners.

In such contexts, it is therefore plausible that familiarity with a language or one of its dialects may bias native listeners toward perceiving different speakers under mismatch conditions, whereas not knowing about the dialect differences (as a non-native speaker) may reduce such biasing effects by encouraging greater reliance on more invariant acoustic cues to talker identity. This possibility is particularly interesting in relation to the Language Familiarity Effect (LFE) – the well-documented advantage listeners show when recognizing voices speaking their native language compared to a non-native one (McLaughlin et al. 2019; Perrachione 2017). Importantly, this does not necessarily imply an overall disadvantage for native listeners but rather predicts a stronger mismatch-related decrement between matched and mismatched conditions. While previous studies have investigated the impact of dialectal variation on talker identification, less is known about whether native and non-native listeners are affected differently by dialect mismatch conditions within a single language. We refer to this predicted pattern as a "native familiarity disadvantage".

Methodology

Using an online training-identification paradigm, the study investigates how mismatch between Swiss Standard German and regional Swiss German dialects influences perceptual speaker recognition by native (Swiss) and non-native (Czech) listeners. During training, listeners learn six target speakers, half presented exclusively in Swiss Standard German and half in their regional dialect, through repeated exposure to different full-sentence recordings. They repeatedly associate each voice with a speaker identity and complete immediate speaker-identification trials by selecting the learned target speaker from two foil speakers. During the test phase, listeners complete the same identification task using new sentence material, with target speakers presented either in matched (e.g., dialect–dialect) or mismatched speaking-mode (Standard–dialect) conditions relative to training. Stimuli are drawn from the DACH corpus (Schmid et al. 2019).

Hypothesis

In line with our expectations about a "native familiarity disadvantage", we hypothesize that dialect mismatch between training and test recordings will reduce speaker recognition accuracy by native Swiss German listeners. In contrast, we expect non-native Czech listeners to show comparable performance across conditions, indicating reduced susceptibility to dialect mismatch.

The findings may provide insights into whether dialect mismatch leads native listeners to underestimate speaker similarity across speaking varieties, with potential implications for the interpretation of auditory speaker comparisons in forensic contexts. More broadly, the study may contribute to our understanding of how top-down linguistic expectations and bottom-up acoustic cues interact during speaker identity judgements.

References

McLaughlin, D. E., Carter, Y. D., Cheng, C. C., & Perrachione, T. K. (2019). Hierarchical contributions of linguistic knowledge to talker identification: Phonological versus lexical familiarity. *Attention, Perception, & Psychophysics*, 81(4), 1088–1107. https://doi.org/10.3758/s13414-019-01778-5

Perrachione, T. K. (2019). Recognizing speakers across languages. In S. Frühholz & P. Belin (Eds), *The Oxford Handbook of Voice Perception* (pp. 515–538). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780198743187.013.23

Schmid, S., Kleber, F., & Moosmüller, S. (2019). *Typology of vowel and consonant quantity in Southern German varieties: Acoustic, perception, and articulatory analyses of adult and child speakers* (DACH corpus). Swiss National Science Foundation project 164377. Accessible at: https://data.snf.ch/grants/grant/164377

Grammatical Gender in Bidialectal Speakers of Swiss and Standard German

Niels O. Schiller[1]
[1]*Department of Linguistics and Translation, City University of Hong Kong, Kowloon, Hong Kong SAR*
Niels.Schiller@cityu.edu.hk

Forensic case work often involves multilingual or multidialectal speakers because multilingualism is a wide-spread phenomenon. It is estimated that more than half of the world's population is proficient in more than one language (Grosjean, 2010). Languages as well as dialects influence each other and may result in speaker-specific idiosyncrasies. This study will focus on grammatical features, i.e., grammatical gender, in two closely related linguistic varieties. In many language varieties, a noun's grammatical gender is expressed overtly via its definite determiner (Corbett, 1991). The varieties studied here are Swiss German (participants' L1), an understudied language variety, and Standard German (participants' L2). While both Standard German and Swiss German have a high lexical similarity and share the same grammatical gender system, i.e., both dialects distinguish *masculine*, *feminine* and *neuter* gender, some nouns do not share the same grammatical gender (see below; Gallmann, 2015). Grammatical gender as a lexico-syntactic feature is an intrinsic property of the noun as it is often considered to be arbitrary. This becomes apparent by comparing the grammatical gender of the word *snail* in Standard German (*die*$_{fem}$ *Schnecke*; feminine) and Swiss German (*de*$_{masc}$ *Schnägg*; masculine). Bidialectal speakers must therefore have different lexical representations for these words. This provides an opportunity to investigate the representation of lexico-syntactic features (e.g., grammatical gender) in the mental lexicon of bidialectal speakers (see Schiller & Sá Leite, 2026). Forty Swiss German (L1) – Standard German (L2) speakers from the Zürich area named 30 target pictures in both dialects (blocked presentation). Results of this picture-naming experiment showed faster naming latencies in Standard German when the grammatical gender of picture and distractor word matched (the so-called congruent condition) than when they did not match (the so-called incongruent condition; see Figure 1 and Table 1). In Swiss German, this effect was not observed, possibly because gender ambiguity diluted the effect (see Figure 1 and Table 2). Possible implications of the results for forensic work will be discussed.

Gender congruency		**Target type**	
Congruent	986	Different	971
Incongruent	1,004	Same	925

Table 1. Mean naming latencies for gender congruency and target type in Standard German (in ms).

Gender congruency		**Target type**	
Congruent	1,035	Different	998
Incongruent	1,041	Same	978

Table 2. Mean naming latencies for gender congruency and target type in Swiss German (in ms).

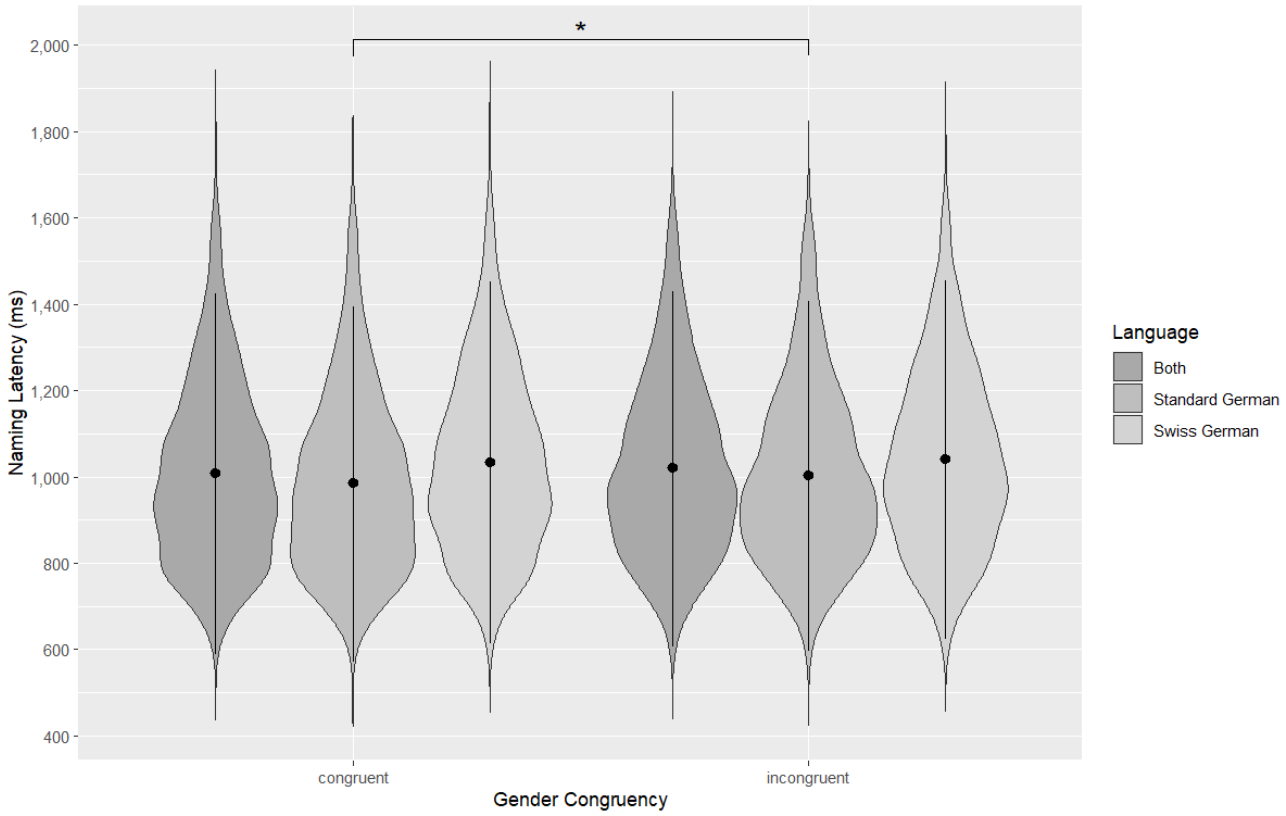


Figure 1. Naming Latencies (in ms) in Relation to Gender Congruency in Standard German, Swiss German, and Across Both Dialects.

References

Corbett, G. G. (1991). *Gender*. Cambridge University Press.

Gallmann, H. (2015). *Zürichdeutsches Wörterbuch* (3., überarbeitete und erweiterte Auflage). Verlag Neue Zürcher Zeitung.

Grosjean, J. (2010). *Bilingual. Life and Reality*. Harvard University Press.

Schiller, N. O., & Sá Leite, A. R. (2026). The psycholinguistics of grammatical gender. In N. O. Schiller & T. Kupisch (Eds.), *The Oxford Handbook of Gender and Classifiers* (pp. 271-292). Oxford University Press.

On the speaker-discriminatory potential of breath pauses in speech

Sascha Schäfer and Jürgen Trouvain
Language Science and Technology, Saarland University, Saarbrücken, Germany
{sschaefer|trouvain}@lst.uni-saarland.de

In contrast to most breathing at rest, speakers usually produce noises when inhaling while speaking. Yet, the role of breath noises in forensic phonetic casework and research remains opaque. While an international survey of speaker comparison practices (Gold & French, 2011) indicates that some of the practitioners who include paralinguistic parameters also consider audible breathing, the overall proportion of experts using such parameters was not disclosed. Furthermore, there is a general lack of research on breath noises from a forensic perspective, with a few notable exceptions (Kienast & Glitza, 2003; Link, 2012; Werner et al., 2022).

Audible breathing nonetheless holds considerable promise for speaker discrimination as early research suggests breathing patterns to be individual and relatively stable throughout a speaker's lifetime (Benchetrit, 1989). Perhaps more importantly, breathing co-occurs with speech rather than being part of it, thus being under less linguistic control. Non-verbal vocalisations (NVVs) like breath noises might therefore be relatively robust to changes in language or speaking style, or to disguises (for an overview of NVVs in a forensic context cf. Braun, 2020).

While Werner et al. (2022) found little evidence for lay listeners' ability to discriminate between speakers based on breath noises alone, an empirical study by Kienast and Glitza (2003) offers a more promising outlook for forensic speaker comparisons by taking an acoustic angle. Ten male speakers who produced short samples of spontaneous speech differed in the spectral characteristics of their audible oral inbreaths. The speakers further varied in the proportion of their breath noise patterns (oral, nasal, and consecutive nasal-oral), with a few speakers showing clear preferences. Metrics like the absolute duration of inhalation noises were, however, characterised by high within-speaker and low between-speaker variability, hampering forensic application. By contrast, a similar study with eight female speakers and read speech stimuli (Link, 2012) found considerably higher between-speaker differences in the duration of audible inbreaths. Physical and psychological stress were further shown to have an impact on multiple parameters, painting an overall complex picture.

The present study aims to unpack this complexity by synthesising findings from both the forensic and the wider phonetic literature. We evaluate the speaker-discriminatory potential of breath pause-related parameters and discuss challenges for annotation and interpretation. Crucially, we draw attention to the often-overlooked structural complexity of breath pauses, which frequently contain further NVVs, such as clicks, smacks, sighs, and glottal hesitations (cf. Gold et al., 2013; Trouvain et al., 2014; Trouvain & Werner, 2022). We additionally argue that a comprehensive account of breathing behaviour requires better integration with the growing body of work on disfluencies (Braun & Elsässer, 2023; McDougall & Duckworth, 2018). For instance, Figure 1 shows an example of a complex breath pause from the DyViS corpus (Nolan et al., 2009) that also comprises a filler particle as well as periods of complete silence and a click sound.

Based on our observations, we propose novel parameters for future forensic research that consider the entire composition of breath pauses. These include, e.g., (a) the pause-internal duration ratio of inhalation noises and the silent "edges" that commonly surround them, (b) the pause-internal duration ratio of the silent edge preceding an inhalation noise and the silent edge following it, and (c) speaker-specific preferences for ordering the various NVVs within complex breath pauses. We underpin our line of argumentation with qualitative examples of spontaneous and read speech from the DyViS (Nolan et al., 2009) and IFCASL (Trouvain et al., 2016) corpora.

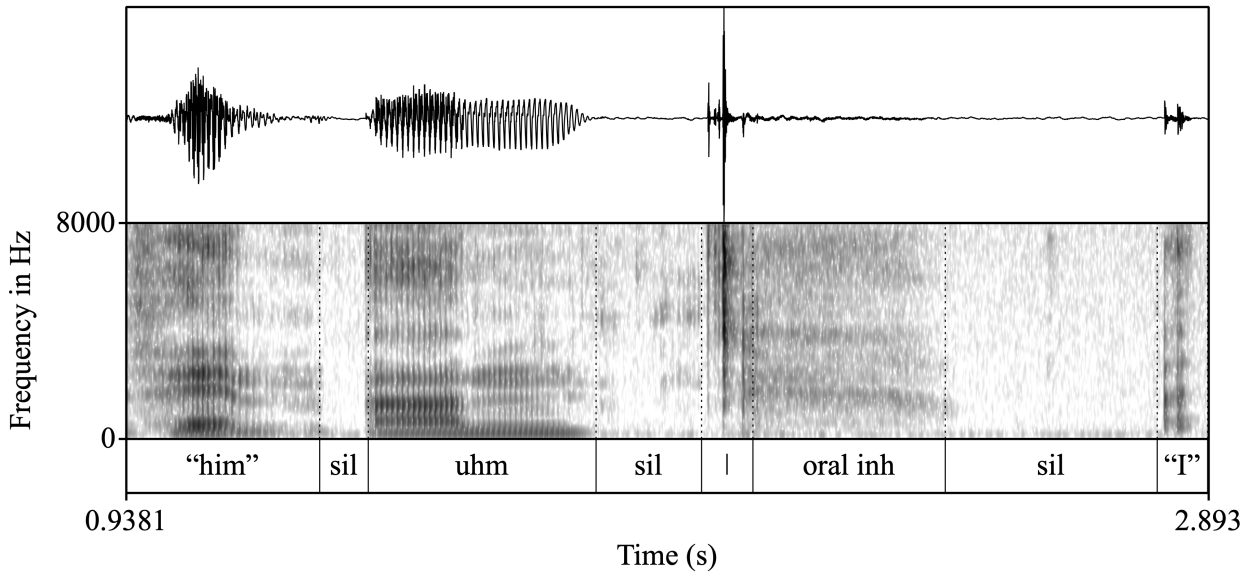


Figure 1. An annotated example of a complex inhalation pause between two intonation phrases consisting of a period of silence followed by a filler particle, another period of silence, a dental click, audible oral inhalation noise and a final period of silence.

References

Braun, A. (2020). Nonverbal vocalisations: A forensic phonetic perspective. *Laughter and Other Non-Verbal Vocalisations Workshop*, *Bielefeld*, *Germany*. 19–23.

Braun, A., & Elsässer, N. (2023). Are there individual disfluency patterns? *International Congress of Phonetic Sciences*, *Prague, Czech Republic,* 3854–3858.

Benchetrit, G., Shea, S.A., Pham Dinh, T., Bodocco, S., Baconnier, P., Guz, A. (1989). Individuality of breathing patterns in adults assessed over time. *Respiratory Physiology,* 75, 199-210.

Gold, E., & French, P. (2011). International practices in forensic speaker comparison. *The International Journal of Speech, Language and the Law*, *18*(2), 293–307.

Gold, E., French, P. & Harrison, P (2013). Clicking behavior as a possible speaker discriminant in English. *Journal of the International Phonetic Association*, *43*(3), 339-349

Kienast, M., & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. *International Congress of Phonetic Sciences, Barcelona*, *Spain*, 1607–1610.

Link, L. (2012). *Individualtypische Aspekte des Atemgeräusches: Eine experimentalphonetische Untersuchung [Unpublished MA dissertation, Philipps-Universität Marburg, Germany].*

McDougall, K., & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles. *The International Journal of Speech, Language and the Law*, *25*(2), 205–230.

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.

Trouvain, J. (2014). Laughing, breathing, clicking: The prosody of nonverbal vocalisations. *Speech Prosody*, *Dublin*, *Republic of Ireland*. 598–602.

Trouvain, J., & Werner, R. (2022). A phonetic view on annotating speech pauses and pause-internal phonetic particles. In Schwarze, C. & Grawunder, S. (Eds.), *Transkription und Annotation gesprochener Sprache und multimodaler Interaktion: Konzepte, Probleme, Lösungen* (pp. 55-70). Narr.

Trouvain, J. et al. (2016). The IFCASL corpus of French and German non-native and native read speech. *International Conference on Language Resources and Evaluation, Portorož, Slovenia,* 1333–1338.

Werner, R., Trouvain, J., & Möbius, B. (2022). *Speaker discrimination and classification in breath noises by human listeners [Conference Presentation].* Annual Conference of the International Association for Forensic Phonetics and Acoustics, Prague, Czech Republic.

The role of voice quality in voice memorability: an experiment

Alice Paver[1] *and Kirsty McDougall*[1]
[1]*Phonetics Laboratory, University of Cambridge, UK*
{aegp2|kem37}@cam.ac.uk

Research has shown that perceptually distinctive voices are easier for lay-listeners to recall in voice memory experiments (Babel et al., 2020; Foulkes & Barron, 2000; Mullennix et al., 2011; Skuk & Schweinberger, 2013; Stevenage et al., 2018). However, the relationship between distinctiveness and recognisability is not straightforward, and the phonetic characteristics underlying these effects remain poorly understood. Much of the existing work on voice memorability has focused on overall impressions of distinctiveness rather than the contribution of specific phonetic variables, yet other research shows that phonetic features such as f0, long-term formant measurements and voice quality (VQ) contribute to listener perceptions of voice similarity and distinctiveness (Paver & McDougall, in press).

Some studies have attempted to control for phonetic typicality; Sørensen (2012) observed that speakers with extreme f0 were more accurately identified in voice parades than speakers with middling f0. Conversely, Paver et al. (submitted) demonstrated that this was not the case, and illustrated that VQ may play a part in determining which voices are more memorable. VQ is of particular interest as it is highly speaker-specific and thus plays an important role in distinguishing between speakers. In Paver and McDougall (2024), it was shown that some VQ settings, e.g. lowered larynx, fronted tongue body, yield judgements of greater distinctiveness among pairs of speakers matched for that particular VQ. The hypothesis being tested in the present study is that speakers with such a VQ setting will also be more memorable within a line-up of matched-VQ speakers. The present study will therefore use stimuli controlled for VQ to explore whether VQ indeed affects listeners' ability to encode and recognise unfamiliar voices.

An online listening experiment is being conducted to explore listener identification accuracy for speech stimuli with different VQ. Stimuli have been constructed from the PASR dataset (Hughes et al., 2022). Four male British English speakers each produced six different VQ settings, creating 24 distinct voice recordings of the Rainbow Passage. For each recording, one target-exposure sample (~20s) and two short parade samples (each ~3s) has been created; these samples have been normalised for mean f0 (107Hz) and intensity (70dB). 240 listeners will be recruited online via Prolific and randomly assigned to one of 24 voice parades (one parade per target voice stimulus). After hearing a target-exposure stimulus, listeners will complete a distractor task before hearing a voice line-up consisting of eight stimuli matched to the exposure voice for VQ (two samples per speaker). For each sample in the line-up, listeners will indicate whether the voice belonged to the speaker they heard previously, creating both target-present and target-absent identification decisions. This design enables exploration of voice memory accuracy for each of the speaker × VQ combinations, allowing for an understanding of the impact of VQ on voice memorability across different speakers and ensuring that memorability effects are not confounded with individual speaker effects, as they may have been in previous studies. This presentation will discuss the implications of the results for the understanding of voice memory, particularly as it relates to earwitness behaviour.

References

Babel, M., McGuire, G., & Willis, C. (2020). The role of voice evaluation in voice recall. In B. Weiss, J. Trouvain, M. Barkat-Defradas, & J. J. Ohala (Eds), *Voice Attractiveness: Studies of Sexy, Likable and Charismatic Speakers* (pp. 101–124). Springer Singapore Pte. Limited. https://doi.org/10.1007/978981-15-6627-1_6

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of a close social network. *International Journal of Speech, Language and the Law*, *7*(2), 180–198. https://doi.org/10.1558/sll.2000.7.2.180

Hughes, V., Foulkes, P., Harrison, P., Wormald, J., Xu, C., van der Vloed, D., & Kelly, F. (2022). *Personspecific ASR* [Data set]. http://pasr.york.ac.uk/

Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. M. (2011). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology*, *25*(1), 29–34. https://doi.org/10.1002/acp.1635

Paver, A., & McDougall, K. (2024). *The role of voice quality in perceived voice distinctiveness: An experiment*. International Association for Forensic Phonetics and Acoustics, Montréal, Canada.

Paver, A., & McDougall, K. (in press). Perceived voice similarity. In F. Nolan, K. McDougall, & T. Hudson (Eds), *The Oxford Handbook of Forensic Phonetics*. Oxford University Press.

Paver, A., McDougall, K., & Nolan, F. (submitted). *Voice memory, acoustic-phonetic typicality and perceptual distinctiveness: The role of f0 distribution in a population*.

Skuk, V. G., & Schweinberger, S. R. (2013). Gender differences in familiar voice identification. *Hearing Research*, *296*, 131–140. https://doi.org/10.1016/j.heares.2012.11.004

Sørensen, M. H. (2012). Voice line-ups: Speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech Language and the Law*, *19*(2), 145–158. https://doi.org/10.1558/ijsll.v19i2.145

Stevenage, S. V., Neil, G. J., Parsons, B., & Humphreys, A. (2018). A sound effect: Exploration of the distinctiveness advantage in voice recognition. *Applied Cognitive Psychology*, *32*(5), 526–536. https://doi.org/10.1002/acp.3424

Transcription in noise: investigating how errors produced by lay listeners vary across increasing noise levels

Lauren Harrington
Department of Language and Linguistic Science, University of York, York, UK
lauren.harrington@york.ac.uk

Transcripts of speech evidence are used for investigative and evidential purposes in the UK criminal justice system. A range of people may be responsible for producing transcripts of forensic audio materials, i.e. recordings which typically suffer from poor quality audio conditions. These include forensic phonetics experts, private transcription agencies, police detectives or other police employees such as interview transcribers (Tompkinson et al., 2022). Outside of linguistics, transcription is widely viewed as an 'administrative burden' and its complexities are often underappreciated. Given this, and the practicalities and costs associated with hiring an expert, it is likely that the majority of forensic audio materials are transcribed by lay people (i.e. transcribers without linguistic or phonetic training), without systematic methods for dealing with challenging audio. It is well documented that transcription accuracy declines for poorer quality audio (e.g. Lange et al., 2011). This study investigates *how* transcription performance declines across conditions, by exploring the types of errors produced by transcribers, the impact of those errors, and how their distribution varies across different noise conditions. It extends previous work, which tends to focus on accuracy rates rather than taking into account the actual content of the transcripts.

135 participants from across the UK were recruited via Prolific to take part in an online transcription experiment. Each participant was presented with 24 stimuli, comprising short extracts from the mock police interview task from the DyViS corpus (Nolan et al., 2009). Each extract contained between 3-6 seconds of speech from an individual male speaker of Standard Southern British English. The stimuli were mixed with speech-shaped noise in Praat (Boersma & Weenink, 2022) at signal-to-noise ratios (SNR) of 6 dB, 0 dB and -3 dB, which represent 'fair', 'moderate' and 'poor' intelligibility respectively. Participants were presented with 8 stimuli at each SNR and were instructed to transcribe all of the speech in the audio file. Participant transcripts were compared at the word level with a reference transcript, which was produced by the author using the studio quality version of the files. Word pairings were marked as a match or assigned an error category - substitution, insertion, or deletion.

As expected, performance declines with increased noise levels (i.e. as the SNR decreases). An average word accuracy rate of 84.4% is achieved by participants in the 'fair' condition (6 dB SNR), which drops to 69.2% in the 'moderate' condition (0 dB SNR) and 49.4% in the 'poor' condition (-3 dB SNR). An ANOVA comparing a null model to a model including SNR as a fixed effect revealed a statistically significant effect of noise on transcription accuracy ($\chi^2 = 177.72$, $df = 4$, $p < .001$).

To better understand the nature of the decline in performance, the specific types of transcription errors were analysed. Deletions account for the majority of transcription errors in all conditions, comprising an increasing proportion of the total errors as noise increases (57.2% in 6 dB SNR, 64.1% in 0 dB SNR, 73.1% in -3 dB SNR). Substitution errors, whereby the listener has transcribed an incorrect word, are also frequently produced in all noise conditions, accounting for around a quarter to a third of the total errors (34.4% in 6 dB SNR, 29.8% in 0 dB SNR, 22.8% in -3 dB SNR). Insertion errors are relatively rare and account for only 4-9% of errors within each condition. Overall, these findings suggest that listeners adopt a more conservative approach to transcription when noise levels are high, opting to omit words entirely rather than offering an interpretation. However, it should be noted that although deletions make up an increasingly bigger proportion of errors as the noise level increases, substitution errors are still produced more frequently in the noisier conditions.

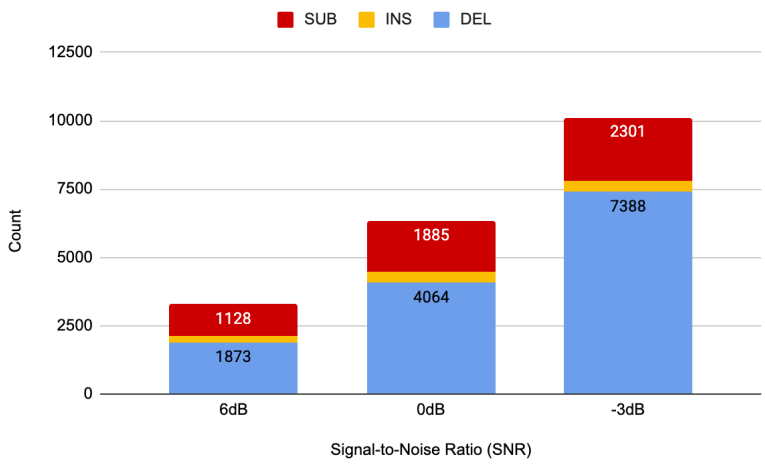


Figure 1. Raw counts of each error type (substitution, insertion deletion) within each noise condition.

Substitutions are of particular interest given that they may prime a reader of the transcript to hear the mistranscribed content (e.g. Fraser & Kinoshita, 2021). This poster will further explore the types of substitution errors produced across noise conditions, and the impact of those errors on the resulting transcription. As demonstrated by the examples in Table 1, many of the participant transcriptions in the least noisy condition tend to be either entirely correct or contain relatively minor grammatical substitutions, but as noise increases the transcriptions can become more nonsensical or result in a change in meaning. These findings will offer critical insights into the specific types of transcription errors lay listeners make when attempting to transcribe poor quality audio materials.

SNR	*Participant transcriptions*
6 dB	I've been in there a couple of times but I've never got to know any of the people that work there
	I've been in there a couple of times but I've never got to know anybody who works there
0 dB	I've been in there a couple of times but I didn't particularly know any of the people who work there
	I've been [...] there a couple of times but I think it's [...] only the people who work there
-3 dB	I really don't have time but I do [...] know people that work there
	[...] It's hard when you spend all day working

Table 1. Example participant transcriptions within each noise condition for the utterance "*I've been in there a couple of times but I've never got to know any of the people who work there*". Substitution errors are highlighted in grey and deletion errors are marked with [...].

References

Boersma, P., & Weenink, D. (2020). Doing phonetics by computer [Computer program] (Version 6.1.30).

Fraser, H., & Kinoshita, Y. (2021). Injustice arising from the unnoticed power of priming: How lawyers and even judges can be misled by unreliable transcripts of indistinct forensic audio. *Criminal Law Journal, 45*(3), 142–152.

Lange, N. D., Thomas, R. P., Dana, J., & Dawes, R. M. (2011). Contextual biases in the interpretation of auditory evidence. *Law and Human Behavior, 35*(3), 178–187.

Nolan, F., McDougall, K., De Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *IJSLL, 16*(1), 31-57.

Tompkinson, J., Haworth, K., & Richardson, E. (2022). *For the record: assessing force-level variation in the transcription of police-suspect interviews in England and Wales*. iIIRG Conference, Winchester.

Vowel Distribution and the Stability of Long-Term Formants in German

Philipp Hutmacher, Angelika Braun, and Katharina Zahner-Ritter
Department of Phonetics, Trier University, Germany
{s2phhutm|brauna|k.zahner-ritter}@uni-trier.de

Introduction

The forensic utility of LTFDs (long-term formant distributions) is based on the fact that formants depend on a speaker's characteristic features, such as the dimensions of the vocal tract or voice quality parameters, such as lip rounding or palatalisation (Nolan & Grigoras, 2005). Previous research shows that LTFDs can indeed reliably distinguish between speakers (Moos, 2010; Gold et al., 2013). In particular, it is the third and fourth formants (LTF3 & LTF4) that tend to show the greatest speaker specificity. F1 and F2 are more closely linked to linguistically driven modifications of the vocal tract, such as those involved in vowel distinction (Clermont & Mokhtari, 1998). Previous studies have produced ambiguous findings regarding the influence of linguistic factors on LTFDs (Tomić & French, 2019; Cho & Munro, 2017; Heeren et al., 2014; Jessen & Becker, 2010). Since phoneme inventories vary across languages, and assuming that the respective languages exhibit at least some divergence in vowel distribution, LTF 1 and 2 are expected to show language-specific differences. This raises the question of the extent to which vowel quality influences LTFDs, compared with the influence of individual characteristics relating to articulation and anatomy.

In this contribution, we test whether manipulating the proportions among the vowels in a given sample affects the LTFs 1-4. We hypothesize that changes in the proportions among the vowels affect the shape of LTFs 1 and 2 in accordance with the vertical and horizontal position of the tongue as well as the degree of lip rounding for the respective vowels. We further expect that the presence of rounded front vowels will lower the values of all LTFs, particularly that of LTF3.

Methodology

In order to study this question, we are conducting an experimental study which manipulates the distribution of the vowel inventory within the sample sets. This is done in such a way that, across five vowel conditions, a disproportionately large number of instances of a particular vowel is produced in each case. We constructed 42 German declarative sentences, each of which was recorded by ten male native speakers of German (mean age = 34.7 years). The study examines the vowels /iː/, /aː/, /uː/, /yː/ and /øː/, and a neutral control condition approximating the vowel distribution typically found in German (cf. Pätzold & Simpson, 1997). The sentences fulfil the criterion that the target vowel in each condition should account for more than 50% of the stressed vowels in that condition, whilst the phonetic context and sentence structure are kept as constant as possible across conditions. The recordings were made in a professional recording studio and lasted an average of 323 seconds per speaker in their entirety, i.e. before extracting the vocalic stream. For the analysis, the first four formants of all vocalic segments are extracted in Praat using the LPC method at 10ms intervals and corrected manually (cf. Barreda, 2021). The statistical analysis includes linear mixed models to investigate the effects of the conditions, as well as correlation analyses to determine the intra-individual stability of the LTF values.

Expected results

Based on the hypotheses put forward, we expect the results to show an intra-individual effect of vowel condition on the LTFs 1 and 2. In other words, the influence of the rate of occurrence of the different vowels is large enough to make a speaker's LTFs vary with the height and fronting of the respective vowel condition. A disproportionately high degree of, e.g., /iː/ will lead to a lower LTF1 than if there were predominantly /aː/-vowels present. We expect an effect of speaker individuality on LTF 3 and

4 but no effect of vowel condition on these two LTFs except that they may be lowered if there are predominantly rounded front vowels present. This study aims to contribute to our understanding of the validity of long-term formants in forensic speaker comparison and the factors influencing them.

References

Barreda, S. (2021). Fast Track: fast (nearly) automatic formant-tracking using Praat. *Linguistics Vanguard*, 7(1), 20200051.

Cho, S., & Munro, M. J. (2017). F0, long-term formants and LTAS in Korean-English bilinguals. *31st General Meeting of the Phonetic Society of Japan*, Tokyo, 188-193.

Clermont, F., & Mokhtari, P. (1998). Acoustic-articulatory evaluation of the upper vowel-formant region and its presumed speaker-specific potency. *5th International Conference on Spoken Language Processing (ICSLP 98)*, Sydney, 359.

Gold, E., French P., & Harrison, P. (2013). Examining long-term formant distributions as a discriminant in forensic speaker comparisons under a likelihood ratio framework. *Proceedings of Meetings on Acoustics*, 19, 060041.

Heeren, W., van der Vloed, D., & Vermeulen, J. (2014). Exploring long-term formants in bilingual speakers. *International Association for Forensic Phonetics and Acoustics (IAFPA 2014)*, Zürich, 39-40.

Jessen, M., & Becker, T. (2010). Long-term formant distribution as a forensic-phonetic feature. *ASA 2nd Pan-American/Iberian Meeting on Acoustics*, Cancún.

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. *The Phonetician*, 101/102, 7-24.

Nolan, F., & Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. *The International Journal of Speech, Language and the Law*, 12(2), 143-173.

Pätzold, M., & Simpson, A.P. (1997). Acoustic analysis of German vowels in the Kiel Corpus of Read Speech. In K. J. Kohler (ed.), *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung*, Kiel University, Vol: 32, 215-247.

Tomić, K. D., & French, P. (2019). Long-term formant frequencies in cross-language forensic voice comparison under likelihood ratio framework. *International Association for Forensic Phonetics and Acoustics (IAFPA 2019)*, Istanbul.

Signs of Partner Proximity in Formant Measurements: A First Look at the Relevance of Entrainment on Voice Identification Analysis

Ellen Grand and Georgina Brown
Linguistics and English Language, Lancaster University, UK
{e.grand|g.brown5}@lancaster.ac.uk

Members of Organised Criminal Groups (OCGs) are regularly persons of interest to security and law enforcement agencies, and so may be subjected to identification and investigation efforts. Voice analysis is one such identification method and is increasingly recognised by security agencies as a vital piece of intelligence.

This work examines one form of intra-speaker variation that is anticipated to be particularly relevant to recordings obtained in investigations of OCGs. Specifically, this work inspects how the relationships between speakers affect their communicative strategies, and how, in turn, these effects could go on to influence voice identification analyses. In order to address this, this study connects existing scientific literature on accommodation and entrainment to navigate their consequences on widely used voice identification methods. *Partner proximity* (Levitan & Hirschberg, 2011:3081) has been identified as a suitable framework through which to measure these effects. Partner proximity tracks the similarity between two speakers while they converse with one another, while comparing the degree of that similarity to when those speakers converse with others.

Building on the work of Earnshaw (2020), phase one of the current thesis examines signs of partner proximity in the formant frequency measurements taken from the FACE, TRAP and DRESS vowels in 15 paired tasks (30 speakers total) from the West Yorkshire Regional English Database (WYRED; Gold, Ross & Earnshaw, 2018). The vowels were examined collectively to ensure there were a suitable number of vowels to run statistical analysis on, and additionally, to look at a representation of the vowel space. Each task consisted of approximately 20 minutes of speech from which the first 10 clear realisations of each relevant vowel token were extracted; from these, formant frequency measurements were taken at the midpoint of the vowel. Euclidean distances and t-tests have been combined to infer the proximity between speakers and assess whether entrainment is taking effect in the speech of those speakers. The way in which this is approached within this methodology is to determine the similarity between speakers across different speaking sessions, and then assess the similarity between speakers when they speak to one another.

In general, the entrainment literature makes use of acoustic features that are frequent, and some experiments have incorporated the time sequencing of the acoustic features (Levitan & Hirschberg, 2011; Borrie, Lubold & Pon-Barry 2015; Wynn & Borrie, 2022). As the aims of this thesis are influenced by forensic applications, the features used in this study (formant frequencies of a small selection of vowels) align with those used to carry out forensic voice comparison analyses. Therefore, the current thesis adapts some of the methodology seen in entrainment studies to bring these features into focus whilst dealing with real life conundrums such as some loss of temporal detail.

The key findings from this first phase showed indications of some phonetic movement between speakers when conversing with one another. When considering the results seen in Earnshaw (2020), the current work yielded results that align with those findings. In addition, F0 measurements have been processed to produce Euclidean distances and t-test scores to contribute towards validation of the methodology as well as replicate previous works in entrainment and in forensic scenarios (Levitan & Hirschberg, 2011; Kruyt, 2023).

This initial study lays the groundwork to extend measures of partner proximity to other voice identification features and methods - specifically to speaker embeddings used within automatic speaker recognition.

References

Earnshaw, K. (2020). *A forensic phonetic investigation of regional variation and accommodation in West Yorkshire* (Doctoral thesis, University of Huddersfield).

Kruyt, J., de Jong, D., D'Ausilio, A., & Beňuš, Š. (2023). Measuring prosodic entrainment in conversation: A review and comparison of different methods. *Journal of Speech, Language, and Hearing Research, 66*(11), 4280–4314.

Borrie, S. A., Lubold, N., & Pon-Barry, H. (2015). Disordered speech disrupts conversational entrainment: A study of acoustic-prosodic entrainment and communicative success in populations with communication challenges. *Frontiers in Psychology, 6*, 1187

Levitan, R., & Hirschberg, J. (2011, August). Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions. In *Proceedings of Interspeech 2011* (pp. 3081–3084).

Gold, E., Ross, S., & Earnshaw, K. (2018, September). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. In *Proceedings of Interspeech 2018: Speech Research for Emerging Markets in Multilingual Societies* (pp. 2748–2752).

Can listeners learn to estimate speaker age?

Johanna Lentz, Katharina Zahner-Ritter and Angelika Braun
Department of Phonetics, Trier University, Germany
{s2jolent|zahnerritter|brauna}@uni-trier.de

It has been shown in multiple studies that listeners are able to estimate speaker age within certain limits (Braun, 1996; Braun & Cerrato, 1999; Shipp & Hollien, 1969; Skoog Waller et al., 2015; Worstbrock & Braun, 2024). This is due to the fact that the bodily structures used for speech production (brain, lungs, larynx, vocal tract) change with time, and these changes are audible (Braun & Friebis, 2009). However, they are not uniform, resulting in varying discrepancies between perceived age (PA) and calendar age (CA). Some specific findings were that

- the discrepancy between PA and CA increases with advancing CA (Braun & Rietveld, 1995);
- the age of smokers is generally overestimated, but the difference between PA and CA is smaller than that for nonsmokers (Braun, 1994);
- slower speaking tempo is the best predictor for advancing age (Braun & Rietveld 1995);
- experienced listeners do not necessarily outperform lay listeners in age estimation tasks, although targeted expertise and practice may lead to improved accuracy (Braun, 1996).

Since speaker age estimation can provide important information about an unknown speaker in forensic voice analysis, it would be advantageous if phonetic experts could be trained to make more systematic use of acoustic age cues. As a first step, the present study examined whether such training effects can already be observed in phonetically naive listeners. A total of 40 male voices with a CA between 25 and 59 years were used, 20 of which were used for testing and 20 for training. The speech material was taken from an existing corpus previously used in age perception research (Braun, 1994). The recordings consisted of "The North Wind and the Sun" read out loud. The age distribution across decades and the smoking status were balanced between the two groups. 41 phonetically naive subjects, 20 of them younger (i.e. between 25 and 34 years) and 21 older (i.e. between 55 and 64 years) served as listeners. In a first listening session (T1), listeners had to estimate speaker age to the exact year based on one sentence of read speech. Different sentences of similar duration were used in T1 and T2 in order to minimize memory effects between sessions. After T1, listeners took part in a 90-minute training session consisting of a short theoretical introduction and a practical exercise phase. The training focused on two key cues for age estimation, namely speaking rate and smoking-related voice quality, and provided participants with explicit strategies for adjusting their initial estimates (e.g., increasing age estimates for slower speech and decreasing them for faster speech). During the practice phase, participants applied these strategies to additional speech samples and received feedback on the speakers' CA. In a second listening session (T2), they had to estimate speaker age again based on a different sentence. The time difference between the two listening sessions was 25 to 31 days.

Results indicate a significant positive relationship between CA and PA both times (Figure 1; T1: ß = 0.68, 95% CI [0.30, 1.06]; T2: ß = 0.49, 95% CI [0.11, 0.88]), i.e. increasing CA was associated with a higher PA. In the first session, the discrepancy between perceived and chronological age tended to increase with speaker age, in line with previous findings ($\beta = 0.10$, 95% CI [-0.07, 0.27]). Following training, this tendency was reversed ($\beta = -0.14$, 95% CI [-0.31, 0.03]), resulting in a significant interaction between speaker age and test session ($\beta = -0.24$, $p < .001$).

However, overall listener accuracy (defined as the mean absolute difference between PA and CA in years) did not improve with the training. It amounted to 10.2 years before and 10.8 years after. We explain this finding as follows: The training was based on the assumption that listeners tend to underestimate the age of slow speakers and non-smokers and overestimate the age of fast speakers and smokers. Accordingly, participants were instructed to adjust their estimates by increasing PA for slow speech and non-smokers and decreasing it for fast speech and smokers. However, analyses of the pre-training data did not exhibit the patterns described above. Therefore, the training could not

have the desired effect. We then tested whether listeners did follow the instructions given to them at all. This was the case for both tempo and smoking. A significant interaction between speaking rate and test session indicated an increased reliance on speaking rate as a cue for age estimation following training (Figure 2; β = -3.80, p = .013), suggesting that slower speech was increasingly associated with higher perceived age estimates after training. A similar pattern emerged for smoking-related voice cues. As expected, smokers were overestimated in the pre-training session (ß = 5.30 years, *p* = .065). This overestimation remained largely unchanged after training (β = 4.92 years, p = .084). In contrast, a significant increase in age estimates was observed for non-smokers (Δβ = 5.31 years, p < .001), who showed no systematic overestimation prior to training (β = 0.88 years, p = .749) but were significantly overestimated following training (β = 6.19 years, p = .034). In summary, although overall accuracy did not improve, training changed the listeners' estimation strategies by increasing reliance on speaking rate and smoking-related voice cues. If this works for lay listeners, there is reason to assume that it will work for trained phoneticians all the more. It therefore seems worthwhile to train listeners to make use of cues like tempo and smoking. However, the present findings suggest that the training needs to be adapted to the individual listener's behavior, for example by first identifying whether individual listeners systematically over- or underestimate certain speaker groups (e.g., smokers or slower speakers) and adjusting training instructions accordingly. Future studies should investigate whether more individualized forms of training or alternative cue combinations lead to greater improvements in accuracy. At the same time, the role of smoking as an acoustic cue may be decreasing due to declining smoking prevalence, raising the question of whether other lifestyle-related voice characteristics (e.g., those associated with newer forms of nicotine consumption) may become more relevant in future.

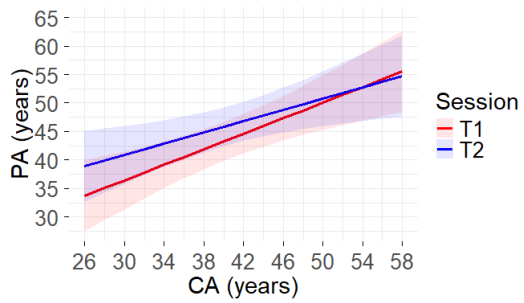


Figure 1. PA as a function of CA

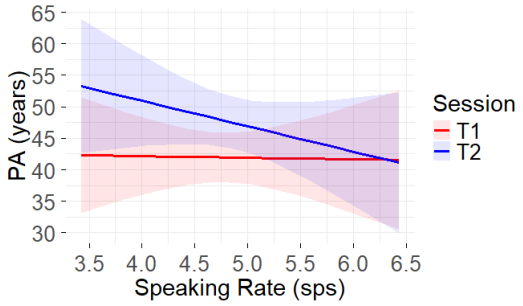


Figure 2. PA as a function of SPS

References

Braun, A. (1994). The effect of cigarette smoking on vocal parameters. *Proceedings ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 161–164.

Braun, A. (1996). Age estimation by different listener groups. *International Journal of Speech, Language and the Law*, 3(1), 65–73. https://doi.org/10.1558/ijsll.v3i1.65

Braun, A., & Cerrato, L. (1999). Estimating speaker age across languages. *Proceedings of ICPhS*, 1369–1372.

Braun, A., & Friebis, S. (2009). Phonetic Cues to Speaker Age: A Longitudinal Study. *Formal Linguistics and Law* (S. 141–162). Mouton deGruyter.

Braun, A., & Rietveld, T. (1995). The Influence of Smoking Habits on Perceived Age. *Proceedings of the XIIIth International Congres of Phonetic Sciences* (Bd. 2, S. 299–297). Stockholm University.

Shipp, T., & Hollien, H. (1969). Perception of the Aging Male Voice. *Journal of Speech and Hearing Research*, 12(4), 703–710. https://doi.org/10.1044/jshr.1204.703

Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology*, 6. https://doi.org/10.3389/fpsyg.2015.00978

Worstbrock, C., & Braun, A. (2024). Estimating Speaker Age from Voices. *Conference on Corpora for Language and Aging Research*, 6, 75–76.

The Construction of a Voice Parade
Stimulus playback and identification accuracy

Pauline Winkler[1], Gea de Jong-Lendle[1], Ulrike Stühler[2] and Mathias Scharinger[1]
[1]Institut für Germanische Sprachwissenschaft, Philipps-Universität Marburg, Germany
{winklerw|gea.dejong|mathias.scharinger}@students|staff.uni-marburg.de

[2]Fachbereich Psychologie, Philipps-Universität Marburg, Germany
ulrike.stuehler@staff.uni-marburg.de

Introduction

In some cases, where victims have only heard the voice of the perpetrator but not seen the person, a voice parade may be commissioned. The construction of a voice parade requires specific expertise. Research and past cases have shown that identification based on eye- and earwitnesses' statements may be prone to error. Voice parades should therefore be constructed in a way that seeks to improve earwitness accuracy and minimises witness error (de Jong-Lendle, in prep.). In this study the effect of the number of permitted listening opportunities on earwitness accuracy is studied: should the witness be allowed to hear each voice more than once, in analogy to a visual parade? Broeders and Amelsvoort (2001) warn that repeated playback may give the witness the possibility to do cross-referencing; playing samples more often may encourage relative judgments to be made. Voice parade guidelines vary considerably, also in the number of listening opportunities afforded to earwitnesses: whereas some jurisdictions, e.g. The Netherlands (Van Amelsvoort, 2018), permit only a single exposure, countries like the UK and Germany allow unlimited replay (Home Office Circular 057/2003; Gfroerer and Jessen, 2022).

	Germany	**Netherlands**	**England & Wales**
Number of Foils (+ suspect)	5+1	5+1	8+1
Length of audio material	60 sec.	20 sec.	60 sec.
Procedure	sequential	sequential	serial
Number of playbacks	unlimited	single	unlimited
Pre-test	yes (4-6 Pers.)	yes (min. 2 Pers.)	yes

Table 1. Comparison of Voice Parades guidelines in Germany, Netherlands, and England & Wales (de Jong-Lendle, in prep.).

Methods

This study investigates the effect of audio playback on voice identification accuracy. Participants completed an online voice parade experiment using PsychoPy. The methodology for constructing and conducting the parade was based on an adapted version of German guidelines.

Participants were native speakers of German aged 18 to 65y. and required to use headphones and complete the experiment in a quiet environment. To simulate real-life conditions and incidental memory processes, participants were not informed that the study concerns forensic voice identification; instead, it was presented as a study on voice perception.

Prior to the experiment, participants provided metadata including age, sex, dialect background, education, and profession. Their headphone volume was then calibrated. They were subsequently exposed to a 60-second recording of the "perpetrator's" voice: a young woman engaged in an

emotional telephone call with an accomplice discussing the critical condition of their kidnapped victim. Following a filler task introducing a retention interval of approximately 7 minutes, participants completed a six-voice lineup. The experiment employed a between-subjects 2 × 2 factorial design, manipulating stimulus playback (single vs. unlimited) and target presence (present vs. absent).

All target and foil samples were produced by female students at the University of Marburg who spoke a variety close to Standard German. The lineup stimuli were approximately 30 seconds in length and consisted of excerpts of spontaneous speech from interviews on neutral topics. Participants were explicitly informed that the perpetrator might or might not be present in the lineup.

Results

Playback condition	Target presence	N	Correct	Accuracy	Mean confidence (0-10)
All voices once	Target-absent	33	13	39.4%	7.26
	Target-present	35	5	14.3%	7.33
At least one replay	Target-absent	4	1	25.0%	6.88
	Target-present	4	0	0.0%	8.17
The play-once group in the target-absent condition showed the best performance: 39.4% Only 8 of 41 unlimited-playback listeners used replay; only foils were replayed. The target voice was only heard once. The highest mean confidence (8.17) was observed for the listening group which performed the worst (0.0%)					

Table 2. Comparison of accuracy (%) and mean confidence at parade level for four Voice Parade Playback conditions (Playback once/unlimited x Target present/absent)

Conclusion

Overall parade-level accuracy was low (max. 39.4%). Generally, replay conditions were not associated with a better performance – it was especially poor in target-present parades (0.0%). Accuracy was higher in target absent conditions at both parade- and voice-level. At voice level, target-present parades showed high percentages of false rejections (>85%) and longer decision time was associated with lower confidence ($p < .001$). Additional replays were associated with lower voice-level confidence ($p = .006$) and confidence was significantly higher for the target-present condition ($p = .048$) with higher confidence predicting greater accuracy ($p < .001$).

References

Amelsvoort, A.G. van (2018). Handleiding Confrontatie. Handleidingen voor de politie. Sdu Overheid, Den Haag.

Broeders, A., van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. Problems of Forensic Sciences 47, 238–241.

Gfroerer, S., Jessen, M. (2022). Sprechererkennung und Tonträgerauswertung, in: Müller, E., Schlothauer, R., Knauer, C. (Eds.), Münchener Anwaltshandbuch Strafverteidigung (3. Überarbeitete Auflage). Beck, München, pp. 2862–2890.

Home Office (2003). Advice on the use of voice identification parades (UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group). Police Leadership and Powers Unit.

de Jong-Lendle, G. (in prep.). Voice parade construction and conduct, Oxford Handbook of Forensic Phonetics, Ch. 24, Oxford.

Nolan, F. (2003). A recent voice parade. *The International Journal of Speech, Language and the Law*, *10*(2), 277–291. https://doi.org/10.1558/sll.2003.10.2.277

Poster Session II

From Absence to Construction: The Rise of *Fonoaudiologia Forense* in Forensic Speaker Comparison in Brazil and Paraguay

Pablo Vourliotis Alvarenga[1], *Ana Paula Sanches*[2], *and Samantha Moraes*[3]

[1]*VPS Laboratorio de Biometría Vocal / Universidad Autónoma del Paraguay, Paraguay*

pablovourliotis@gmail.com

[2]*InFForense, Brasil*

anasanches@fonoforense.com

[3]*Fonoaudiologia Forense RS – FFRS, Brasil*

samantha@ffrs.com.br

Forensic Speaker Comparison (FSC) has become an increasingly relevant evidentiary practice within judicial systems in the Southern Cone, particularly in cases involving organized crime, extortion, and drug trafficking. Despite its operational importance, the field remains unevenly developed when comparing Brazil and Paraguay, and scientific literature on FSC in South America is still limited. Institutional practices reveal considerable methodological variability, reflecting the coexistence of analytical traditions arising from Engineering, Phonetics, Linguistics, and the speech sciences.

The term ***Fonoaudiologia Forense*** (Portuguese) / ***Fonoaudiología Forense*** (Spanish) — provisionally rendered as *Forensic Speech Science* for international audiences — designates the professional field of speech-language pathologists specialised in acoustics and phonetics applied to forensic and judicial purposes. One of its subareas is FSC, which integrates four examination types: voice, speech (phonetic focus), language (sociolinguistics and phonology), and acoustics — all under a forensic perspective. The term functions as a professional-legal category substantially broader than its English translation implies, and is not reducible to the clinical connotations of *"speech-language pathology"* in the anglophone tradition. This distinction is essential for interpreting the findings presented here.

A further dimension of the observed variability concerns the professional backgrounds of FSC practitioners. In Brazil, within official forensic institutes, FSC is conducted by specialists with diverse training — Information Technology, Engineering, and Biology, among others. In Paraguay, the task falls almost exclusively to criminalistics experts, validated at undergraduate level. A foundational study (Lima & Marinho, 2016) documented that 87% of judicial professionals in Asunción attributed forensic voice analysis exclusively to criminalistics experts — a stark contrast with Brazil, where the role has been progressively taken up by phoneticians and speech-language pathologists with specialised training.

Method

The study adopts an exploratory comparative design comprising: (1) a bibliographic review of indexed publications (2000–2026) in Scopus, Web of Science, and SciELO; (2) documentary analysis of professional regulations, criminal procedural codes, and institutional protocols in both countries; and (3) a structured survey of FSC practitioners, focusing on training profiles, methodologies in use, available tools, and perceived barriers to standardisation.

Results and Discussion

The findings reveal two differentiated but interconnected contexts, linked through academic exchanges evidenced by the 2016 study itself, conducted by Brazilian researchers in Paraguay. Brazil presents a consolidated institutional ecosystem with professional regulation and methodological protocols adopted by federal and state forensic laboratories. Paraguay represents a rapidly evolving context: in 2025, the first Procedures Manual for Forensic Voice Identification from a speech-

language pathology perspective was formally presented (Vourliotis, 2025). This nine-phase protocol covers chain of custody, quality control, perceptual-auditory evaluation (EVAPEFOR), acoustic-instrumental analysis, and likelihood ratio estimation, aligned with ENFSI-FSAAWG (2015), SWG-Speech (2010), and ISO/IEC 17025:2017, and adapted to the bilingual Spanish–Guaraní phonetic context, including *jopara* code-switching. The manual frames FSC within a probabilistic paradigm based on hypothesis testing and likelihood ratio evaluation (Morrison, 2011), in contrast with categorical identification models.

Conclusion

FSC in Brazil and Paraguay is undergoing institutional and methodological maturation, marked by progressive adoption of probabilistic, standardised, and linguistically informed approaches. The Paraguayan case demonstrates that significant advances are achievable within a decade when grounded in international standards and adapted to regional sociolinguistic realities. Together, both cases illustrate how *Fonoaudiologia Forense* has broadened methodological diversity in South American FSC and aligned forensic practice with evidence-based forensic science — offering a relevant model for other Latin American countries.

References

ENFSI Forensic Speech & Audio Analysis Working Group. (2015). *Best practice guidelines for forensic speaker comparison.* European Network of Forensic Science Institutes.

Lima, I. L., & Marinho, R. R. (2016). Avaliação dos conhecimentos das pessoas que trabalham no âmbito judicial com relação à análise forense da voz na cidade de Assunção, Paraguai. *Unpublished manuscript.*

Morrison, G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice,* 51(3), 91–98.

Scientific Working Group for Speaker Recognition (SWG-Speech). (2010). *Guidelines for best practice in forensic speaker recognition.* FBI / Department of Justice.

Vourliotis, P. M. (2025). *Manual de procedimentos para a identificação forense por voz sob perspectiva fonoaudiológica.* VPS Laboratorio de Biometría Vocal / Universidad Autónoma del Paraguay.

Willis, S.M. et al. (2015). *ENFSI guideline for evaluative reporting in forensic science.* Dublin: European Network of Forensic Science Institutes.

It's the Evidence That Matters. The Forensics of LOID

Andreea Dragu, Mihaela Stan
Verified AB, Sweden
{andreea.dragu|mihaela.stan}@verified.se

Verified AB's LOID (Linguistic Origin IDentification) method has been developed based on the needs and requirements of end users of language analysis reports. Thus, it has been built to stand up to scrutiny during the asylum process, which essentially means integrating a forensic approach in order to answer a natural question anyone might direct at our analyses: 'why should we believe you?'.

Thus, in our presentation, we will argue for the forensic legitimacy of LOID, while also explaining our approach in managing the authentic, practical challenges of fitting dialectology into the box of asylum forensics.

Since LADO is yet to be fully regulated, we will show what means LOID has chosen in order to integrate principles such as duty of care to good science, confining experts to their area of expertise, transparency, impartiality, balance, clarity, suitability for intended purpose, declared limitations.

The LOID method rests on three pillars: the cooperation between a qualified linguist and a native-speaker analyst (one of the two methods accepted by IAFPA), a transparently presented reasoning that makes the analysis replicable and verifiable, and reliance on published research as the main source of dialectological information.

The challenges that LADO is faced with come from multiple directions and Verified has experienced most of them: the inherently imprecise nature of dialectology, the lack of (up-to-date) research, understudied dialects, the strict requirements of forensics, the limited accessibility of scientific language for non-specialists, the specific needs of end users. Thus, we will show some of the solutions incorporated into LOID as a result of Verified AB's over two decades of experience, based on practical examples.

We will focus on two categories of challenges: lacking dialectological research and adapting the LOID output to forensic requirements and users' needs. For the former, we will show several approaches to acquiring up-to-date dialectal information by means of examples such as Uyghur, Kinyamulenge, Lower Juba Somali or Lingala, based on using native-speakers in two ways: instructing them to extract information from published sources in their native language and/or extracting their first-hand knowledge through customised questionnaires. For the latter, we will present how specific components of a typical LOID report were designed to meet forensic requirements and users' needs.

Our presentation constitutes Verified AB's contribution to integrating LADO into the broader world of forensic linguistics and sparking constructive discussions around good practices, improvements and the future of language analysis for asylum purposes.

References

Cambier-Langeveld, T. (2010). The Validity of Language Analysis in the Netherlands. In K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and Origin: The Role of Language in European Asylum Procedures: A Linguistic and Legal Survey* (pp. 12–33). Wolf Legal Publishers.

Clopper, C. G., & Pisoni, D. B. (2006). Effects of Region of Origin and Geographic Mobility on Perceptual Dialect Categorization. *Language Variation and Change, 18*(2), 193–221.

Code of Practice (2020). *IAFPA*. https://www.iafpa.net/about/code-of-practice/

de Graaf, A., ten Thije, J., & Verrips, M. (2011, July 11–14). *Eliciting Spontaneous and Natural Speech for the Purpose of LADO.* IAFL10 Colloquium on LADO, Birmingham, UK.

Dragu, A., Lundberg, L. J.& Björsten, S. (2022). One Bias Particular to Asylum Courts: The Native-Speaker Bias. *International Journal of Language and Linguistics, 10*(2), 163–165. https://doi.org/10.11648/j.ijll.20221002.24

Eriksson, A. (2008, July 23–24). *Guidelines? What guidelines?* IAFPA 17th Annual Conference, Lausanne, Switzerland.

Foulkes, P., & Wilson, K. (2011). Language Analysis for the Determination of Origin: An Empirical Study. In *Proceedings of the International Congress of Phonetic Sciences*, 691–694.

Hoskin, J. (2018). Native Speaker Non-Linguists in LADO: An Insider Perspective. In I. Nick (Ed.), *Forensic Linguistics: Asylum-seekers, Refugees and Immigrants* (pp. 23–40). Vernon Press.

Language and Determination of National Identity Cases Resolution (2009). *IAFPA*. Retrieved 3 April 2026 from https://www.iafpa.net/about/resolutions/

Language and National Origin Group. (2004). Guidelines for the Use of Language Analysis in Relation to Questions of National Origin in Refugee Cases. *International Journal of Speech Language and the Law, 11*(2), 261–266.

Markham, D. (1999). Listeners and Disguised Voices: The Imitation and Perception of Dialectal Accent. *Forensic Linguistics, 6*(2), 289–299.

Masthoff, K., Boubaker, Y., & Köster, O. (2010). The Tell-Tale Dialect: Analysis of Dialectal Variation of German Native Speakers in Telephone Conversations. IAFPA 2010, Trier, Germany.

Pfeifer, M. (2023). "The Native Ear". Accented Testimonial Desire and Asylum. In P. Rangan, A. Saxena, R. T. Srinivasan, & P. Sundar (Eds.), *Thinking with an Accent* (pp. 192-207). University of California Press.

Scharenborg, O., & van Os, M. (2019). Why Listening in Background Noise Is Harder in a Non-Native Language than in a Native Language: A Review. *Speech Communication, 108*, 53–64. https://doi.org/10.1016/j.specom.2019.03.001

Utlåtandeskala forensisk resultatvärdering NFC (n.d.). *Polisen*. Retrieved 3 April 2026 from https://polisen.se/tjanster-tillstand/bestall-forensisk-undersokning/utlatandeskala/

Verrips, M. (2011). LADO and the Pressure to Draw Strong Conclusions. A Response to Tina Cambier-Langeveld. *International Journal of Speech, Language and the Law, 18*(1), 131-143.

A survey of forensic practitioners on speaker attribution methods

Tallulah Buckley[1]*, Kirsty McDougall*[1]*, and Richard Rhodes*[2]
[1]*University of Cambridge, Cambridge, UK*
{tbb32|kem37}@cam.ac.uk
[2]*The Forensic Voice Centre, York, UK.*
richard.rhodes@forensicvoicecentre.com

In transcribing an evidential recording of a spoken interaction involving two or more speakers, speaker attribution is the process of assigning each utterance to a speaker. Traditionally forensic speech science (FSS) has focused on voice parades, speaker comparison and more recently transcription. However, there is currently little research into speaker attribution within FSS. Speaker attribution is a highly important aspect of casework; it can be requested as a stand-alone task, but it is often embedded within casework tasks such as transcription and speaker comparison. The more unique challenges demonstrate the importance of considering speaker attribution independently. These challenges include identifying the number of speakers in a recording which as shown in Loakes, Fraser and McDougall (2023) trained transcribers often do not identify reliability. This present work is a survey of forensic practitioners which aims to gain insight into current methods and inform future research into speaker attribution.

Previous surveys (Harrington and Rhodes 2024; Gold and French 2019; Rhodes 2016) that were conducted to explore respectively transcription, speaker comparison and cognitive bias methodology provided information about the variety of approaches used by practitioners at that moment in time. The present survey aims to evaluate the current range of uses and methods of speaker attribution to contextualise future research. In order to explore current practices, the survey addressed a variety of themes, such as, countries, affiliations and experience, understanding and context, speaker attribution methods (including contributing factors, contextual information and presentation), use of automatic approaches, and finally accreditation and validation.

Preliminary results

In the initial stage of data collection the survey received responses from a variety of affiliations across Europe, North America and Australia. The survey was distributed on several mailing lists including IAFPA, IAFLL and ENFSI. The preliminary findings of the study provide several interesting observations. It is notable that the frequency of speaker attribution being undertaken within casework is high compared with its frequency as a stand-alone task.

The survey also inquired about which linguistic or phonetic features a practitioner most frequently draws on when attributing speech to an individual. Whilst this is highly case-specific, the most commonly identified aspect of speech was voice quality; these findings can inform subsequent research.

An ever-evolving aspect of casework is automatic approaches. Initial results show, in contrast to the rising use of automatic speaker recognition systems in forensic speaker comparisons, the majority of respondents do not currently use automatic speech recognition when completing forensic transcriptions or speaker diarisation to attribute speech in casework.

The accreditation and validation section yielded reassuring results with the majority of respondents stating they had systems in place to mitigate bias. Further, all initial respondents stated compliance with both ENFSI guidance (2021) and the IAFPA code of practice (2020). In this paper we will expand on these preliminary results to gain a wider range of feedback internationally.

References

European Network of Forensic Science Institutes. Best Practice Manual for the Methodology of Forensic Speaker Comparison. (ENFSI-FSA-BPM-003 Issue No. 001 (2022).

Gold, E., & French, P. (2019). 'International practices in forensic speaker comparisons'. *The International Journal of Speech, Language and the Law* 26(1):1–20.

Harrington, L. & Rhodes, R. (2024). 'Survey on forensic transcription practices'. *The International Journal of Speech, Language and the Law* 31(2):236–66.

International Association for Forensic Phonetics and Acoustics. Code of Practice. (2020).

Loakes, D., Fraser, H. & McDougall, K. (2023). 'A preliminary investigation of the acoustic factors impacting decision making in speaker attribution.' Acoustics 2023 conference of the Acoustical Society of America, 4-8 December 2023, Sydney, Australia. https://pubs.aip.org/asa/jasa/article/154/4_supplement/A159/2924018/A-preliminary-investigation-of-the-acoustic

Rhodes, R. (2016). 'Cognitive Bias in Forensic Speech Science: A Survey on Risks and Proposed Safeguards.' IAFPA conference, York.

Audio Enhancement and Transcription Interlaboratory Comparison Exercises

Katherine Earnshaw[1], *Bryony Nuttall*[1], *and Richard Rhodes*[1]
[1]*The Forensic Voice Centre, York, UK*
Katherine.earnshaw@forensicvoicecentre.com

We would like to use the poster session to inform practitioners within the IAFPA community about two interlaboratory comparison exercises in audio enhancement and transcription taking place later this year. Our aim is to share our plans for how these exercises will run and to invite practitioners who currently undertake audio enhancement and/or transcription casework to express their interest in participating. We also welcome advice from other organisations that have run similar exercises on best practice procedures.

Context

The UK Forensic Science Regulator's Code of Practice requires all forensic units performing forensic science activities (FSAs) in the UK to assure the quality of their results. Accordingly, it is recommended that forensic units must "determine the availability and evaluate the appropriateness of schemes for proficiency tests (PT), interlaboratory comparisons (ILC) and collaborative exercises (CE) that are relevant to its FSAs" (UK FSR Code of Practice 30.1.1). While there are institutions that provide speech and audio related PT schemes for its members such as The European Network of Forensic Science Institutes (ENFSI); there are still relatively few opportunities to participate in exercises such as these across the UK and internationally. For this reason, we are proposing to run two relatively small-scale ILC exercises: one in audio enhancement and one in forensic transcription. The purpose of these exercises is to share processes and methods and to discuss best practices for these FSAs. The results may also assist in demonstrating practitioner competence, method development and method validation.

Summary of process

The two ILCs will be facilitated by The Forensic Voice Centre using recordings made by our partners in the Department of Language and Linguistic Science at The University of York. The recordings are designed to be forensically relevant in terms of their content and quality, but importantly they are also accompanied by high-quality 'ground truth' recordings which can be consulted after the exercises are completed. This is particularly useful for the transcription ILC where transcripts of the 'ground truth' recordings can be consulted to assess accuracy. Interested participants will be asked to complete a data sharing agreement with the University of York and then they will be provided with formal instructions and the relevant recordings for analysis. We anticipate that the audio enhancement ILC will take place first, followed by the forensic transcription ILC during Autumn of 2026; however, exact timescales will be agreed once the list of participants has been determined.

Audio Enhancement ILC

Participating practitioners will be asked to enhance the audio in some forensically relevant recordings, to document their process followed, and to submit a report reflecting their standard casework practices. This will be followed by a roundtable discussion to examine inter-laboratory variation in approaches and outcomes. Although the speech in the recordings is spoken in English, we anticipate that this exercise will be suitable for any practitioners within the IAFPA community who routinely carry out enhancement casework.

We invite feedback on how to make this exercise most useful, e.g., how do we assess the extent to which the intelligibility of the speech in the recording has been improved? Should we ask lay listeners or experts to rate the improvement? What criteria should we use?

Forensic Transcription ILC

Participating practitioners will be asked to transcribe the audio in a forensically relevant recording, to document their process followed, and to submit a transcript and report reflecting their standard casework practices. This will be followed by a roundtable discussion to examine inter-laboratory variation in methods and transcript content. There will also be an opportunity to review the transcripts of the forensically relevant material against the 'ground truth' recording and its corresponding transcript. As the speech in the recording is spoken in English, we anticipate that this exercise will only be suitable for practitioners within the IAFPA community who routinely transcribe English speech in their casework.
We invite feedback on how to make this exercise most useful, e.g., what is the best method to 'grade' transcripts and/or the types of errors that occur?

All interested parties are invited to express their interest in participating in either of the two ILC exercises via a link which will be shared during the conference.

References

[UK] Forensic Science Regulator. (2025). Forensic science activities: Statutory Code of Practice (version 2 – May 2025) - [URL: https://assets.publishing.service.gov.uk/media/68401f4be550203c8209cce9/2025_05_20_Code_of_Practice_final.pdf]

DIVERSE: a corpus of cisgender female speakers for phonetic and forensic research

Alice Paver[1], *Chloe Patman*[1], *Kirsty McDougall*[1], *Paul Foulkes*[2] *and Jay Paulson*[1]
[1]*Phonetics Laboratory, University of Cambridge, UK*
{aegp2|cep72|kem37|ep716}@cam.ac.uk
[2]*Department of Language and Linguistic Science, University of York, UK*
paul.foulkes@york.ac.uk

Approximately 5-10% of casework involving forensic audio analysis includes female speech (Forensic Voice Centre, personal communication, 2023). Although female speakers therefore constitute a smaller proportion of casework data, women still represent half of the population, and their voices do occur in forensic investigations. Yet most research in forensic speech science has focused on male voices. Both biological and social differences between male and female speakers may influence the realisation of phonetic features and mean that different analytical approaches are appropriate for each group.

DIVERSE (Database of Individual Variation in English by Recording style and SEx) is the first forensically motivated database including non-contemporaneous, style-controlled speech from female speakers of British English. This database, which will be made publicly available, is modelled on DyViS (Nolan et al., 2009), a larger-scale resource including 100 male speakers of a comparable demographic who completed similar tasks. Speech recordings were collected from 32 cisgender adult female speakers of Standard Southern British English (SSBE), aged between 18 and 30 years. Each participant completed two recording sessions, held at least one week apart. Across both sessions, tasks were designed to elicit a range of speaking styles (Table 1). These included a reading passage, a simulated police interview and a voicenote. During the second session, participants additionally completed a telephone call with their 'accomplice'. All recordings were made in a sound-treated recording booth at studio quality, while the other tasks were also simultaneously recorded in voicenote/telephone quality respectively.

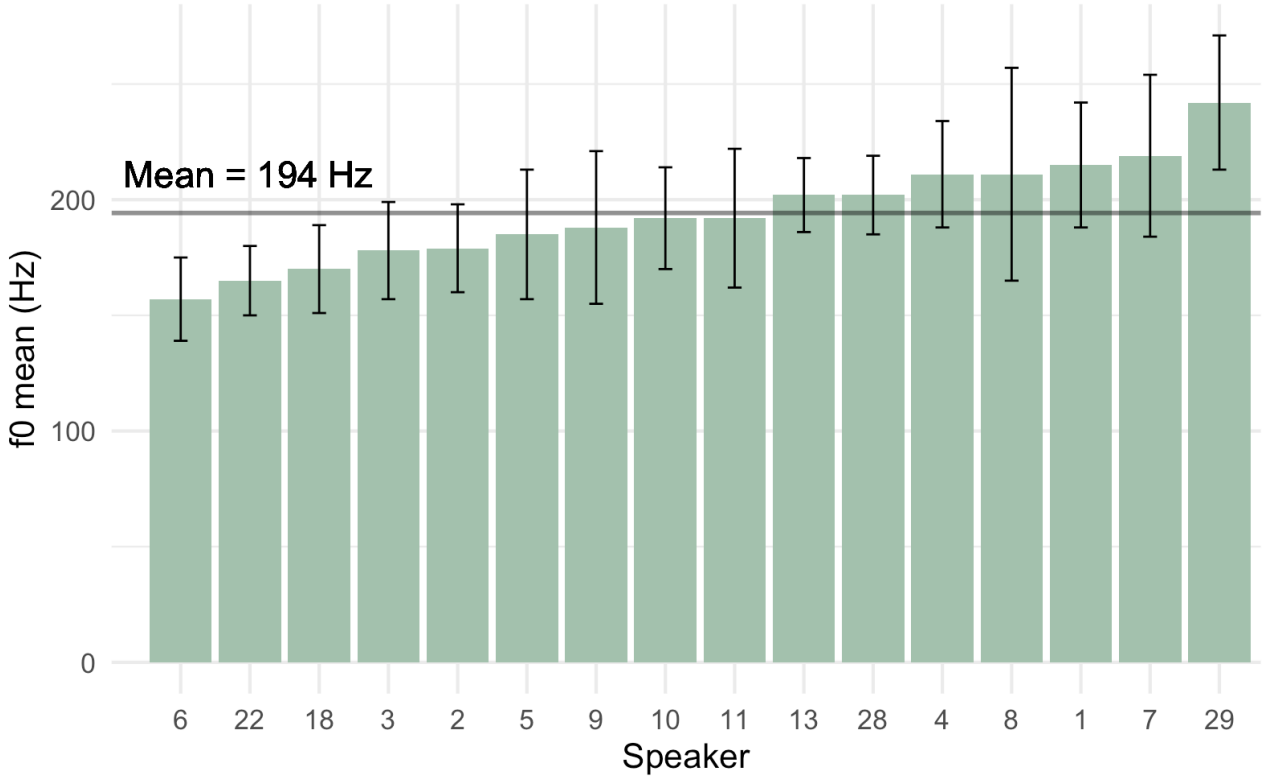
Task	Session 1	Session 2	Spontaneous speech	Read speech	Interlocutor
Read passage	✓	✓		✓ (SQ)	
Simulated police interview	✓	✓	✓ (SQ)		Police officer (Author 1)
Voicenote	✓	✓	✓ (SQ, VNQ)		
Telephone conversation with an accomplice		✓	✓ (SQ, TQ)		Accomplice (Author 2)

Table 1. An overview of the tasks undertaken in the DIVERSE dataset. SQ = Studio quality; VNQ = Voice note quality and TQ = Telephone quality.

This presentation outlines the motivation and structure of the database, alongside fundamental frequency (f0) statistics for all 32 speakers, with measurements taken from the Session 2 Interview task. Analysis was conducted using Praat (Boersma & Weenink, 2025). Measurements were taken after the first 60s to allow the speaker to settle into their natural speech, and analysis excluded sections of overlapping speech, non-speech vocalisations, or poor/inaccurate f0 tracking induced by e.g. extreme creaky or whispered voice. We aimed for ~120s net speech per speaker. Pitch range settings were specified per speaker depending on the accuracy of f0 tracking, but the default setting of 100-300Hz was used for the majority of speakers. Preliminary results for the group distribution (based on analysis of 16 speakers of 32) found that speakers' mean f0 values were normally distributed with a group mean f0 value of 194Hz. The lowest-pitched speaker (speaker 6) had a mean of 157Hz, and the highest-pitched speaker had a mean of 242Hz (speaker 29; Figure 1). Standard deviations were

also calculated, and ranged from 15Hz (speaker 22) to 48Hz (speaker 8). Further results for all 32 speakers will be presented.

Figure 1. f0 values of 16 DIVERSE speakers, with columns indicating means and vertical bars indicating ±1 standard deviation. Group mean is indicated by the horizontal line. Speakers are arranged in ascending order of mean f0.



These preliminary findings provide a first reference for f0 distributions in spontaneous speech of young adult female speakers of SSBE. The overall mean of speakers' f0 means is slightly lower than that in existing reports for read speech in SSBE (Mennen et al., 2012) and German (Künzel, 1989), and spontaneous speech for Swiss German (Hedegard et al., 2023). This database and associated findings contribute to the limited body of forensically motivated research on female speech.

References

Boersma, P., & Weenink, D. (2025). *Praat: Doing phonetics by computer* (Version 6.4.42) [Computer software]. http://www.praat.org

Hedegard, H., Fröhlich, A., Tomaschek, F., Steiner, C., & Leemann, A. (2023). Filling the population statistics gap: Swiss German reference data on F0 and speech tempo for forensic contexts. *Proceedings of Interspeech 2023*, 2558–2562. https://doi.org/10.21437/Interspeech.2023-1042

Künzel, H. J. (1989). How well does average fundamental frequency correlate with speaker height and weight? *Phonetica*, *46*(1–3), 117–125. https://doi.org/10.1159/000261832

Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: A comparison of English and German. *The Journal of the Acoustical Society of America*, *131*(3), 2249–2260. https://doi.org/10.1121/1.3681950

Nolan, F., McDougall, K., Jong, G. de, & 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/ijsll.v16i1.31

F0 corpus for case-adapted reference populations

Almut Braun, Michael Jessen, Anja Moos and Isolde Wagner
Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{Almut.Braun|Michael.Jessen|Anja.Moos|Isolde.Wagner}@bka.bund.de

Motivation and procedural method

Fundamental frequency is one of the parameters most frequently analysed in cases of forensic speaker comparison (Gold & French, 2011). In day-to-day casework, a need was felt for a realistic reference corpus of mean f0 that is case-specifically adaptable (e.g., a mean f0 reference population of male speakers of German in neutrally loud but agitatedly spoken voice messages). Therefore, f0 data and corresponding meta data – such as, for example, perceived sex of the speaker, spoken language, spontaneity, loudness, intonational markedness, formality – have been extracted from already processed cases and saved in a pivot table with customisable filtering options. Whenever possible, several mean f0 measurements per speaker were collected to obtain a corpus not only suitable for assessing interspeaker variability but also suitable to gain insight into intraspeaker variability in different speaking contexts. For the latter, data came either from cases with several recordings per (previously identified) suspect, or from comparisons of questioned and suspect speaker recordings of particular cases where the comparison result was in the higher positive ranges of the conclusion scale.

Advantages of a case-adaptable mean f0 reference population corpus

The main advantage of the presented corpus is that it contains real-case mean f0 measurements in a variety of different speaking contexts, with a focus on inter- as well as on intraspeaker variability of mean f0. Casework-approximated corpora with regard to mean f0 exist, but those often contain mean f0 measurements only for particular speaking contexts (e. g., Jessen et al., 2003 for German neutrally loud and Lombard speech, male speakers; Hudson et al., 2007 for Standard Southern British English – speech elicited in a simulated police interview, young male speakers). In many other, especially non-forensically-oriented studies, mean f0 measurements were not based on spontaneous speech samples (e.g., Natour & Wingate, 2009; Aghajanzadeh et al., 2012). For some languages – as far as we know – no f0 reference data exist, e.g., for Albanian. A further advantage of the corpus presented here is the enhancement of intersubjective consistency with regard to mean f0 measurements in a variety of different speaking contexts. Joint training sessions of different forensic phonetic experts ensure sufficiently similar classifications in terms of particular speaking conditions in which mean f0 was measured (e.g., in speech classified as neutrally loud with marked intonation). Since we experience a broad range of different speaking contexts in our casework and some conditions occur much more frequently than others, the corpus will contain more mean f0 data for common speaking contexts compared to speaking contexts that are less common (e.g., particular foreign languages, female voices). The corpus has been designed in an extendable way allowing it to grow over time, such that more mean f0 data can be collected systematically – also for less common speaking contexts.

Number	*Speaker*	*perceived sex*	*...*	*loudness*	*intonational markedness*	*mean f0*
1	1_a	male	...	neutral	slightly marked	129
2	1_b	male	...	slightly louder	strongly marked	189
3	2_a	male	...	neutral	not marked	120
4	3_a	male	...	neutral	not marked	89
...	...	...	...	...	...	...

Table 1. Example of the case-adaptable mean f0 reference population corpus. Loudness = vocal effort in the passage(s) used for mean f0 measurement; intonational markedness = frequency and intensity of pitch accents in the passage(s) used for mean f0 measurement.

References

Aghajanzadeh, M., Ghorbani, A., Torabinezhad, F., & Keyhani, M. R. (2012). Comparing of two Farsi passages in laboratory evaluation of mean and fundamental frequency variations range. *Auditory and Vestibular Research,* 21(1), 62-68.

Gold, E., & French, J. M. (2011). International Practices in Forensic Speaker Comparison. *International Journal of Speech, Language and the Law,* 18(2), 293-307.

Hudson, T., de Jong, G., McDougall, K., Harrison, P., & Nolan, F. (2007). F0 statistics for 100 young male speakers of Standard Southern British English. *Proceedings of the International Congress of Phonetic Sciences,* 16 (Saarbrücken), 1809-1812.

Jessen, M., Köster, O., & Gfroerer, S. (2003). Effect of increased vocal effort on average and range of fundamental frequency in a sample of 100 German-speaking male subjects. *Proceedings of the International Congress of Phonetic Sciences,* 15 (Barcelona), 1623-1626.

Natour, Y. S., & Wingate, J. M. (2009). Fundamental Frequency Characteristics of Jordanian Arabic Speakers. *Journal of Voice,* 23(5), 560-566.

Effect of combining supralaryngeal with laryngeal information in a forensic telephone-based dataset

Michael Jessen and Canan Konrat
Department of Text, Speech and Audio, Bundeskriminalamt, Germany
{Michael.Jessen|Canan.Konrat}@bka.bund.de

Both supralaryngeal and laryngeal voice acoustics carry speaker-discriminatory information, but the supralaryngeal contribution is stronger, especially when MFCC (Mel Frequency Cepstral Coefficients) are included, but also with formant frequencies alone (Jessen, 1997; Hughes et al., 2023; Chan and Wang, 2024). There is mixed evidence as to whether combining both supralaryngeal and laryngeal information leads to better speaker discrimination than supralaryngeal information alone. Improvement is possible but is limited to factors such as matching (vs. mismatched) conditions (Chan and Wang, 2024 for formants when fused with laryngeal voice quality and optionally f0; MFCC is not improved upon) or dependent on re-assigning the roles of training, test and reference speakers (Hughes et al., 2023: Fig. 3).

The current study expands upon the one of Jessen et al. (2023). In that study, laryngeal voice quality using the VoiceSauce software was measured in a dataset of 23 male adult speakers of German from telephone conversation in forensic cases. Speaker discrimination performance was investigated using single voice quality parameters as well as various combinations thereof. Performance was limited, with most results above EER of 30% and Cllr above 0.9. The best-performing voice quality parameter was CPP (Cepstral Peak Prominence) with results between about 25-30% EER and 0.8-0.9 Cllr. F0 and F1, F2, F3 were measured for comparison purposes and formants, especially F3, outperformed the voice quality parameters and their combinations. In the present study, all three formants are combined (with multivariate modeling as well as with fusion of the univariate parameters; see Jessen et al., 2003 for further methodological details relevant here as well) in order to establish a baseline of supralaryngeal performance. This is compared against combining the formants with two kinds of laryngeal parameter sets. The first set is as close as possible to the one proposed in Lee et al. (2019) and Afshan et al. (2022), referred to here as UCLA set. The set used here contains (along with F1, F2, F3) the following parameters: H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz, CPP, f0. The second set just contains CPP (again, along with F1, F2, F3), which was the most successful parameter in the Jessen et al. study. As in that study, separate tests are performed here by varying number of Gaussians (3 and 17 Gaussians) as well as enabling or disabling MAP (Maximum A Posteriori) adaptation. There were also six different conditions in that study varying as to whether all vowels were used (corresponding to Long-Term Formant Analysis) or just /a/ and whether the formant and f0 information was manually corrected or not, in two combinations. Here, three of those conditions are used (Conditions 2, 3 and 5), which are all the ones containing manual correction of the formants. Results for Condition 2 are presented in Table 1.

Table 1 shows that in seven of the eight methodological variations there is improvement in EER when formants are combined with at least one of the two laryngeal parameter sets. Improvement is stronger and more consistent with the CPP set than the UCLA set. Sometimes improvement in EER with the CPP set is quite remarkable (EER below 10% with real case data). The contribution of the UCLA set is essentially random, sometimes leading to improvement, but sometimes quite the opposite, especially with the multivariate models. One explanation for the greater success of the CPP set is that it is simpler, whereas the UCLA set contains several parameters that had low performance in Jessen et al. (2023) and this might have led to instability of the models. It makes sense that the multivariate models are more susceptible to instability than the univariate models followed by fusion. The use of MAP was probably able to compensate for some of the instability, when considering from the upper part of the table that deterioration from formants alone to formants with the UCLA set is smaller with MAP than without it. Again, it makes sense that MAP has that effect. MAP is also of benefit for the fused models, in which case using the UCLA set consistently leads to improvement.

When instead of EER we look at Cllr, the results are different. Combining with laryngeal parameters most of the time leads to no improvement over formants alone and sometimes the effect is quite

deteriorating. Examination of the corresponding Tippett plots shows that limited Cllr performance or even Cllr values larger than 1 probably result from outliers of same-speaker comparisons with values much below LLR zero, which were not found or not to the same extent among formants alone. Those outliers generally would have no effect on EER but do affect Cllr.

When examining the results of the other two conditions investigated (not shown), the patterns are similar for Condition 3, but differ to some extent for Condition 5, the one based on /a/ rather than long-term formants. In Condition 5, there is no asymmetry in performance between the parameters EER and Cllr. Like in Condition 2, seven of the eight methodological variations lead to improvement in the combination process, but now improvement is shown about as often by EER as Cllr. There is also less of an asymmetry between the combination abilities of CPP and UCLA. Whereas in Condition 2 the performance of CPP combined with formants was always better than the one of UCLA combined with formants (EER in 8 of 8 methodological variations), in Condition 5 this drops to 5/8 (both for EER and Cllr). When applying fusion along with MAP, UCLA consistently outperforms CPP. Generally, overall performance in Condition 5 is lower than in 2 or 3, with the best EER of all at 14.4 and the best Cllr at 0.66.

In conclusion, the results show that combining formant frequencies with laryngeal features can lead to improvement, but only for some feature combinations (simpler model preferred) and only for some of the methodological variations, though the clear majority. Hughes et al. (2023) found similar patterns as in the present study (improvement on average, but only in terms of EER, not Cllr; see Fig. 3). Chan and Wang (2024) also showed a combination benefit, but only in the most favorable conditions. In their study, only Cllr was successful, not EER (see their Table 2, mean Cllr vs. mean EER in LTFDs vs. LTFDs + LTLVQ or in LTFDs vs. LTFDs + LTF0 + LTLVQ).

	3 Gaussians no MAP		17 Gaussians no MAP		3 Gaussians MAP		17 Gaussians MAP	
Multivariate	EER	Cllr	EER	Cllr	EER	Cllr	EER	Cllr
F123	16.1	0.59	16.2	0.60	16.3	0.61	16.2	0.61
F123 + UCLA	27.2	0.93	23.2	0.78	18.2	0.74	21.7	0.75
F123 + CPP	18.8	0.75	12.7	0.64	11.5	0.60	13.0	0.62
Fused	EER	Cllr	EER	Cllr	EER	Cllr	EER	Cllr
F123	16.4	0.61	14.3	0.63	14.6	0.64	15.3	0.65
F123 + UCLA	15.0	0.68	16.7	1.16	13.1	0.78	12.7	0.60
F123 + CPP	11.7	0.61	9.0	0.66	9.5	0.65	12.6	0.61

Table 1. Result for Condition 2 (long-term formants, manually corrected). Upper part: All investigated parameters combined in a multivariate GMM-UBM model. Lower part: The individual parameters in univariate GMM-UBM models were subsequently fused using logistic regression fusion. For further details see Jessen et al. (2023). EER: Equal Error Rate; Cllr: Log-likelihood-ratio cost. Lower values of these performance parameters indicate better speaker discrimination.

References

Afshan, A., Kreiman J., & Alwan, A. (2022). Speaker discrimination performance for “easy“ versus “hard” voices in style-matched and -mismatched speech. *J. Acoust. Soc. Am*., 151, 1393–1403.

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 Sci. Int.*, 363, 112199.

Hughes, V., Cardoso, A., Foulkes, P., French, P., Gully, A., & Harrison, P. (2023). Speaker- specificity in speech production: The contribution of source and filter. *J. Phonetics*, 97, 101224.

Jessen, M. (1997). Speaker-specific information in voice quality parameters. *The Intern. J. of Speech, Lang. and the Law (Forensic Linguistics)*, 4, 84–103.

Jessen M., Konrat, C., & Horn, J. (2023). Voice comparison analysis of forensic recordings using the VoiceSauce program. In *Proceedings ICPhS*, Prague, 3845–3849.

Lee, Y., Keating, P., & Kreiman, J. (2019). Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am*., 146, 1568–1579.

Choosing an appropriate reference population can mitigate limitations in automatic cross-language forensic speaker comparison

Kristina Tomić[1], *Dominic Watt*[1,2], *Valeriia Vyshnevetska*[1,4], *Volker Dellwo*[1,3] *and Peter French*[1,2,3]

[1]*JP French International, Forensic speech and audio laboratory, York, UK*
ktomic|dominic.watt|peter.french@jpfrench.com

[2] *Department of Language and Linguistic Science, University of York, UK*

[3] *Department of Computational Linguistics, University of Zurich, Switzerland*

[4] *Linguistic Research Infrastructure, University of Zurich, Switzerland*
valeriia.vyshnevetska|volker.dellwo@uzh.ch

In the past two years, JP French International has seen a rise in the number of requests for cross-language forensic speaker comparison (FSC). Presently, such requests amount to 8% of all those received. Given the continuing increase in multilingualism and polyglossia within societies, together with the steep growth of international communication, one would expect to see this proportion rise further. Automatic speaker recognition (ASR) systems, which rely on deep neural networks to analyse acoustic features across a range of frequencies at a frame level, offer a promising complementary tool for cross-language speaker comparison (see Cao et al., 2024; Kahil et al., 2018; Lo, 2021; Van der Vloed et al., 2017). Still, one of the challenges of utilising such systems in casework on a regular basis, apart from their limited performance on samples with poor and mismatched quality, is the availability of relevant reference population datasets. The present study utilises Phonexia's Voice Inspector 5.2.0 (Phonexia s.r.o., 2026), an ASR system trained on large amounts of data from various languages, to assess system performance under cross-language speaker comparison with varying reference populations.

Research structure

The corpus used for the research was recorded in 2022 (see Tomić & French, 2023). It consists of 50 female native speakers of Serbian, speaking Serbian and Serbian-accented English in an interview recorded over a cellular network directly from the sound card on the receiving end. For comparison and compilation of reference populations (RPs), we used 60-second net speech samples without pauses. A matrix of 50 Same-Speaker (SS) and 2,450 Different-Speaker (DS) comparisons was executed for each RP. In addition to the RPs compiled directly from the corpus recordings (only Serbian, only English, and a combination of Serbian and English), we ran tests using RPs with a mismatched language (Czech) and accent (British English) available as ready-made voice profiles within the software. Comparisons where both samples were in Serbian served as baseline performance data.

Results

Same-language comparisons: While the system performed at ceiling level for the samples with matched language and RP, a slight increase in false alarms was noticed when the language of the RP did not match the samples, with the language of the same family (Slavic) performing slightly better. The results are comparable to Watt et al.'s (2020) study where introduction of accent mismatch in RP led to higher same-speaker LR scores because compared samples appeared atypical against a population with a different accent. Evidence (E) is a system-calibrated logarithmic score that presents distance between two samples and is calculated without reference population (see Jessen et al., 2019). The score is on average higher for SS pairs in same-language comparisons, but the difference is not as prominent in DS pairs.

Cross-language comparisons introduced a substantial number of misses (SS errors) and yielded significantly weaker positive average LLRs when using mismatched RPs. The most severe degradation occurred with Serbian-accented English RP, which generated a critical 36% miss rate

and a low average SS LLR of 0.61. The system seems to perform better when the reference population matches the sample from the known speaker (Serbian), as already observed in previous literature, albeit for different systems (González-Rodríguez et al., 2006; Morrison et al., 2012). Finally, system performance was highly resilient when using the mixed reference data from the corpus (Serbian + English) or the in-built British English reference data (BrE), both achieving a 0% SS error rate and low DS error rates (≤ 0.12).

QS-KS	RP	Av distance (E)		Min/Max E		Av LLR		ER (%)		Cllr
		SS	DS	SS	DS	SS	DS	SS	DS	
Sr-Sr	Serbian	5.04	-1.47	3.87/ 5.42	-4.79/ 2.61	4.15	-11.09	0	0	0.016
Sr-Sr	BrE					4.17	-4.02	0	0.24	0.043
Sr-Sr	Czech					4.12	-6.05	0	0.08	0.021
En-Sr	Serbian	3.96	-1.55	2.34/ 4.69	-5.12/ 2.32	1.62	-11.25	10	0	0.233
En-Sr	English					0.61	-16.03	36	0	0.512
En-Sr	Sr + En					3.08	-3.77	0	0.08	0.067
En-Sr	BrE					3.74	-4.12	0	0.12	0.051

Table 1. Speaker comparison using Phonexia's Voice Inspector 5.2.0; QS-KS – sample from the questioned and known speaker, RP – reference population, E – Evidence or distance score between tested samples, LLR – log-likelihood ratio, ER – error rates.

The analysis indicates that while a poorly chosen reference database can severely undermine system reliability (leading to high miss rates), an appropriate reference population can effectively mitigate language-mismatch penalties. These initial results suggest that cross-language speaker comparison can be reliably supported by an ASR system provided the reference population is carefully optimised. Further validation across a wider typological range of language groups is required to confirm this.

References

Alexander, A., & Drygajlo, A. (2004). Scoring and direct methods for the interpretation of evidence in forensic speaker recognition. *Proceedings of Interspeech* 2004 (ICSLP), (pp. 2397-2400).

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 ISCSLP*, Beijing, China, 2024, (pp. 279-283).doi:10.1109/ISCSLP63861.2024.10800518

Jessen, M., Bortlík, J., Schwarz, P & Solewicz, Y. A. (2019). Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication* 111: 22-28. doi: 10.1016/j.specom.2019.05.002

Kahil, Y., Zergat, K. Y., & Amrouche, A. (2018). Forensic Voice Comparison with ALIZE/LIA_RAL toolkit. *Proceedings of the 4th IEEE International Conference on SIVA, Guelma, Algeria*.

Lo, J. H. (2021). Issues of Bilingualism in Likelihood Ratio-based Forensic Voice Comparison. *Doctoral dissertation*, University of York. https://etheses.whiterose.ac.uk/id/eprint/30007/

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. *Proceedings of the Odyssey 2012, Singapore*, (pp. 62-77).

Phonexia s.r.o. (2026). *Phonexia Voice Inspector* (Version 5.2.0) [Computer software]. https://docs.phonexia.com/products/voice-inspector

Tomić, K., & French, P. (2023). Vocal Profile Analysis as a Tool in Cross-Language Forensic Speaker Comparison. *A paper presented at the 31st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Zurich, July 9–12, 2023*.

van der Vloed, D., Jessen, M., & Gfroerer, S. (2017). Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language. *A paper presented at the Conference on Audio Forensics 2017, June 15–17 Arlington, VA, USA.*

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. doi:10.1558/ijsll.41466

AI-ccent judgements: Accent stereotypes persist with AI-generated voices

Jun Liu[1], *Leah Bradshaw*[1,2], *and Eleanor Chodroff*[1]
[1]*Department of Computational Linguistics, University of Zurich, Switzerland*
{jun.liu|leah.bradshaw|eleanor.chodroff}@uzh.ch
[2]*JP French International, UK.*
{leah.bradshaw}@jpfrench.com

Recent advances in neural Text-to-Speech (TTS) systems have enabled the creation of highly realistic synthetic voices capable of reproducing human-like intonation, rhythm, and conversational fluency (Do et al., 2022). While existing research has largely focused on the technological development and detection of synthetic speech (Almutairi & Elgibreen, 2022; Khurana et al., 2023), AI-generated voices also provide a valuable framework for investigating how listeners evaluate social identity, credibility, and authenticity through speech. Because synthetic voices allow accent, gender, and vocal style to be manipulated while holding semantic content constant, they create new opportunities for examining whether traditional sociolinguistic stereotypes persist in human-AI interaction. This study explores how regional accent and perceived gender shape listener evaluations of synthetic speech. Previous studies demonstrate that listeners rapidly attribute personality traits and social identities based solely on accent, often overriding objective acoustic evaluation (Shamina, 2016). Indeed, empirical evidence demonstrates that listeners form several biases around a speaker based on their accent alone (e.g., Garrett et al., 2003; Coupland & Bishop, 2007; Paver et al., 2025). AI-generated speech therefore provides a useful model for investigating whether entrenched social evaluations persist even when listeners know that a voice is synthetic. The study additionally considers how listeners judge "humanness" and "naturalness" in AI-generated speech. Understanding which cues listeners prioritise has important implications for forensic speech science, where synthetic speech is becoming an increasingly important consideration in fraud investigations, disputed recordings, and the evaluation of voice evidence.

To investigate these questions, 40 listeners ages 21-40 participated in a Likert-scale rating task, comprising 20 British English and 20 American English participants (21 F, 18 M, 1 preferred not to report gender) in which they were informed that all voices were AI-generated. Participants evaluated 16 AI-generated voices created using ElevenLabs; two male and two female voices for four accent backgrounds (Yorkshire English, SSBE, General American, Southern American). Participants rated each voice from 1 to 6 on perceived trustworthiness, professionalism, human-likeness of the voice and naturalness of the overall audio, alongside providing qualitative comments regarding their judgements. Findings demonstrate that traditional accent stereotypes persist within evaluations of synthetic speech. Across listener groups, Yorkshire English voices received the highest overall trustworthiness ratings, while SSBE voices were consistently ranked the most professional (Figures 1 and 2). Furthermore, ratings of "human-like" voices and "natural" audio quality were found to be highly correlated ($r = 0.88$, $p < 0.001$). Qualitative analysis indicated that listeners frequently relied on the same vocal cues when making both judgements, focusing primarily on characteristics of the voice itself rather than on broader audio artefacts. This suggests that lay listeners may evaluate AI speech authenticity through socially conditioned perceptions of vocal performance rather than cues that reliably distinguish synthetic from human speech. These findings have important implications for forensic speech science, highlighting that intuitive judgements about whether a voice is "real" may be influenced by accent stereotypes and perceptions of humanness rather than diagnostic acoustic evidence. As AI-generated speech becomes increasingly prevalent in criminal investigations, fraud, and disputed recordings, understanding these perceptual biases will be essential for interpreting listener judgements and informing best practice in the evaluation of voice evidence. Although forensic speech science increasingly relies on acoustic and computational analyses, human perceptual judgements remain central to many stages of casework, including the assessment of recordings, forensic transcription, earwitness testimony, and the interpretation of disputed voice evidence.

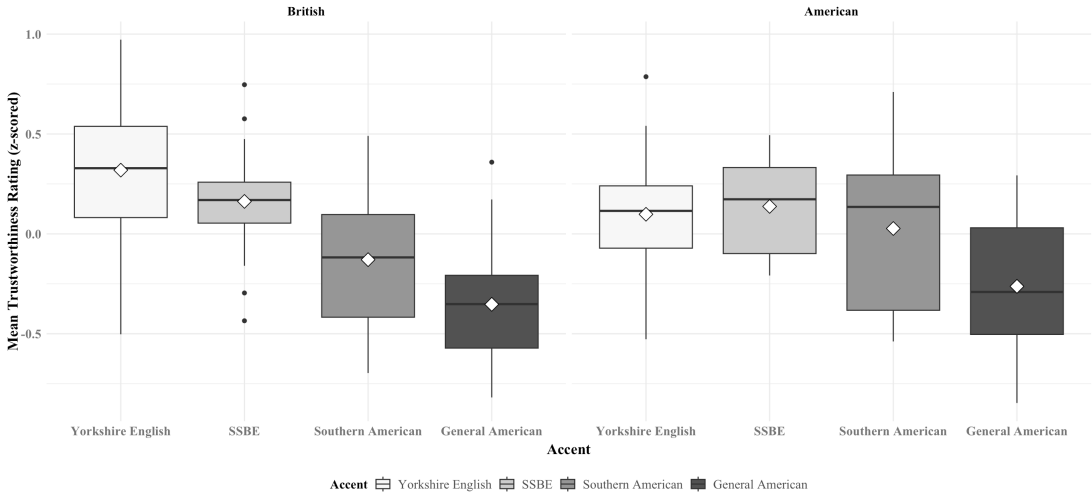


Figure 1. Mean trustworthiness ratings for British and American English listeners faceted by dialect type. Higher ratings indicate higher perceptions of trustworthiness from the voices.

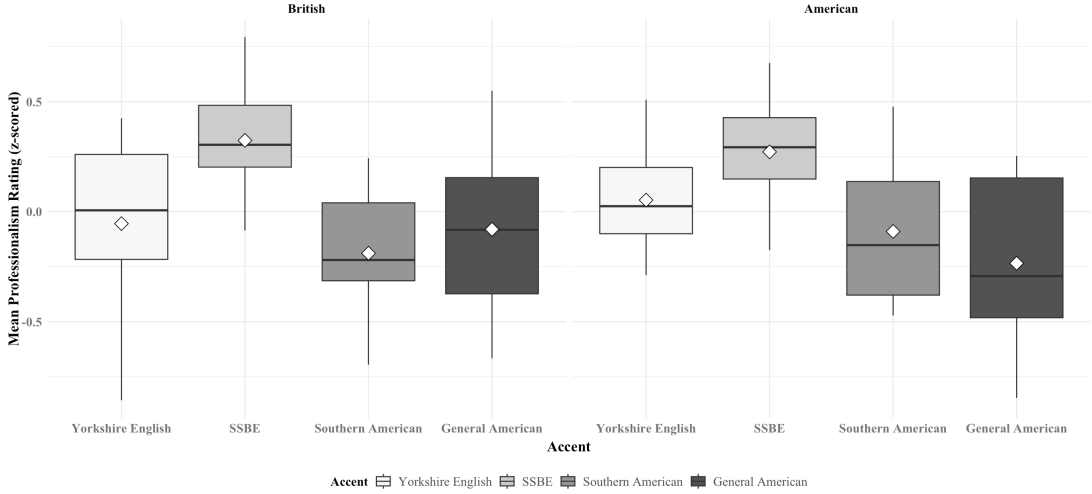


Figure 2. Mean professionalism ratings for British and American English listeners faceted by dialect type. Higher ratings indicate higher perceptions of professionalism from the voice.

References

Almutairi, Z., & Elgibreen, H. (2022). A review of modern audio deepfake detection methods: challenges and future directions. *Algorithms*, *15*(5), 155.

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. Guilford Press.

Coupland, N., & Bishop, H. (2007). Ideologised values for British accents 1. *Journal of sociolinguistics*, *11*(1), 74-93.

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2022). Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (pp. 16–22).

Garrett, P., Coupland, N., & Williams, A. (2003). *Investigating language attitudes: Social meanings of dialect, ethnicity and performance*. University of Wales Press.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications*, *82*(3), 3713-3744.

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology*, *33*(4), 378-392.

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.

Shamina, E. (2016). An experimental study of English accent perception. *ExLing 2016*, 159.

Zahn, C. J., & Hopper, R. (1985). Measuring language attitudes: The speech evaluation instrument. *Journal of language and social psychology*, *4*(2), 113-123.

Developing a forensic methodology for deepfake audio detection

Anna Bartle[1], *Matthew Boyle*[1], *Ben Gibb-Reid*[2], *Ronan Given*[1], *Amelia Gully*[2], *Philip Harrison*[2], *Vincent Hughes*[2], *Lucy Morville*[2], *Luca Resti*[3], *Andrew Robinson*[2] *and Jessica Wormald*[2]

[1]*Metropolitan Police Service, Audio Lab, UK.*

anna.bartle@met.police.uk

[2]*Department of Language and Linguistic Science, University of York, UK.*

[3]*Department of Computer Science, University of York, UK.*

In recent years, there has been huge growth in the performance and availability of AI-based systems capable of producing realistic-sounding synthetic voices. These systems can be used to generate synthetic speech from a non-specific individual or, perhaps more worryingly, be used to mimic the voice of a real person (via voice cloning or voice conversion) to produce speech that person never said. Deepfake audio therefore presents an emerging threat within law enforcement. In the UK, the Office of the Policing Chief Scientific Adviser and the National Police Chiefs' Council have identified deepfakes as a primary concern in ensuring the integrity of evidence passing through the criminal justice system. However, at present, there is no established and validated framework for forensic speech and audio experts to address questions about deepfakes.

In this paper, we describe a collaborative project between the Metropolitan Police Service (MPS) and the University of York to develop human-based audio deepfake detection methods to complement AI-based automatic detection tools. The aim is to effectively integrate human expertise into deepfake detection to improve reliability and provide explainability. The project utilises a large set of genuine and deepfake recordings made by MPS. The deepfakes were generated using a range of state-of-the-art commercial text-to-speech (i.e. voice cloning) and speech-to-speech (i.e. voice conversion) systems. The project consists of two strands of method development:

(1) Auditory and acoustic phonetic analysis: This involves adapting established forensic voice comparison methods (see ENFSI BPM 2021) to the task of deepfake detection in cases where there is a reference recording of a known speaker. Initially, we conducted auditory analysis across a cross-section of real and deepfake voices to identify potentially useful features, such as consonant and vowel realisations, long-term pitch and voice quality, as well as non-linguistic features, such as hesitations, pauses and breathing patterns. More systematic, data-driven, analysis will then be conducted on a larger set of recordings in order to quantify the performance of different features.

(2) Signal processing analysis: This involves taking methods from the signal processing domain to identify properties of the signal that contribute the most to deepfake detection. In this strand, we will focus on bi-spectrum analysis (detecting internal signal correlations), sub-band entropy (measuring signal smoothness), spectral feature trajectories (assessing the variability within a recording) and the envelope modulation spectrum (detecting periodicity in noise).

Once relevant features have been determined through the two strands of work, testing will be conducted with a set of previously unseen real and deepfake samples to assess the overall efficacy of the methods. These results will then be combined with the output of AI-based deepfake detection

systems being tested in-house by MPS to produce a three-pronged analysis framework integrating phonetic, automatic, and signal processing methods.

References

ENFSI (2021) Best Practice Manual for the Methodology of Forensic Speaker Comparison. ENFSI-BPM-FSC-01.

Can Spoofed Voices Compromise an ASR system?

Julia Roegner[1], *Simon Werner*[1], *Angelika Braun*[1]
[1]*Trier University, Germany*
s2juroeg@uni-trier.de

Given the availability and the quality of voice clones, they can be expected to show up in forensic casework on a regular basis (Kudera et al., 2024; San Segundo and Delgado, 2025).

This contribution addresses the question of whether a SOTA ASR system can deal with voice clones. To this end, one male native speaker of German provided the original speech samples. He was recorded on three consecutive days, producing six genuine samples. Depending on the recommendation of the manufacturer, one or two samples were used to clone his voice with two different models: a commercial one (ElevenLabs Inc., 2026) and an open-source one (Zyphra Technologies Inc., 2026). 30 clones per model were generated.

For testing, we compared the four remaining genuine samples to one cloned sample each. The first two testing modes (TMs) offered by Phonexia™ Voice Inspector (Phonexia, 2024) were used. In TM1, we applied one-to-one comparisons (4x60=240 genuine-cloned comparisons & 4x2=8 genuine-genuine comparisons), while in TM2, we combined the four genuine samples as material for the suspected speaker (1x60=60 genuine-cloned comparisons & 1x2 genuine-genuine comparisons).

The 300 comparisons (240+60) involving a cloned sample were judged to originate from different speakers 81% of the time (Fig. 1)[1]. The clones produced by the commercial model were more often classified correctly as coming from different speakers than those from the open-source model. The commercial model yielded correct rejection rates of 100% (TM1) and 50% (TM2), whereas the open-source model showed correct rejection rates of 90% (TM1) and 0% (TM2). The all-genuine comparisons had a correct rejection rate of 100% each.

We then looked at the probability ratings in terms of absolute log-likelihood ratios (LLRs, Fig. 2).

Between the clones, the absolute values were higher on average for the commercial model in TM1, and higher for the open-source model in TM2 despite that model's correct rejection rates of 0.0%. In TM1, the absolute LLR values of genuine sample lie in between those of the clones, and they show the highest values of the three models. Mann-Whitney U tests revealed significant differences of absolute values between correct and incorrect answers for samples of the same model (TM1 – open-source: $p<0.001$; TM2 – commercial: $p<0.05$). In both cases, the absolute values for incorrectly recognized samples are lower than the absolute values for correct answers. Further Mann-Whitney U tests showed a significant difference between clones depending on the model ($p<0.001$ in each testing mode).

In addition to the between-model variability, we observed some intra-model variability (see Fig. 2). In other words, two clones based on identical material may well produce categorically different results in terms of both correctness and range of LLR values. This as well as the influence of the testing method makes it difficult to draw general conclusions with respect to the effectiveness of cloned samples in deceiving ASR systems. The experiments show a relatively high probability for a correct result in one-to-one comparisons, but produced a large number of false positives in TM2. We would therefore emphasize the need to verify the authenticity of samples before applying ASR. To improve accuracy, the inclusion of synthesized speech in the training data of ASR might be useful. As a next step, we are planning to test the performance of human listeners as compared to ASR.

[1] Please note that a no-match answer was considered correct in the genuine-clone pairs, whereas a match was considered correct in the genuine-genuine pairs.

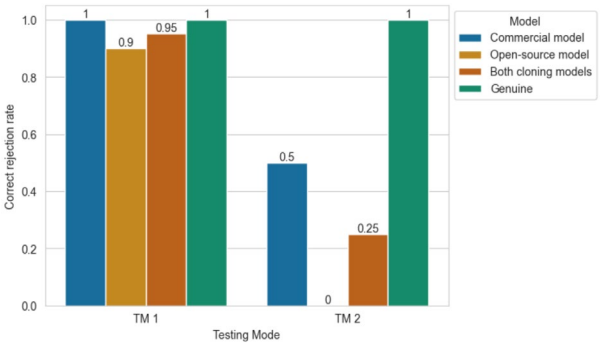
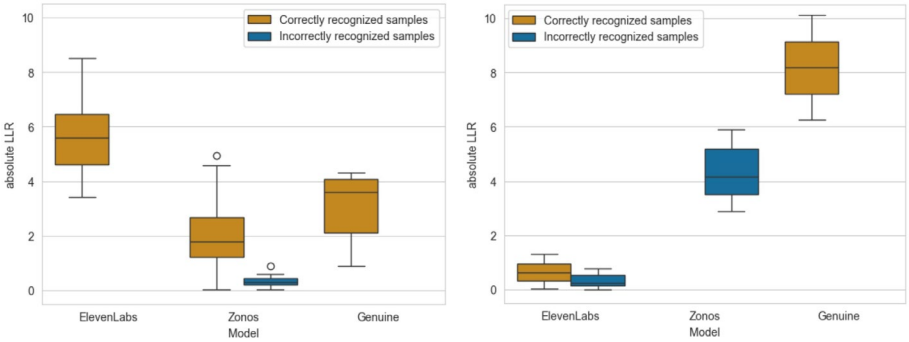


Figure 1. The correct rejection rates of samples by model, sorted by testing mode.



(a) TM1 (b) TM2

Figure 2. Boxplots for the absolute log-likelihood ratio (LLR) values throughout the two testing modes for each model, divided by correctly and incorrectly recognized samples.

References

Kudera, J., Coccia, M., Fadaeijouybari, S., Preidt, T., Ranjan, A., & Braun, A. (2024). Voice cloning and mismatch conditions in forensic automatic speaker recognition. In *International Conference on Speech and Computer* (pp. 171-184). Cham: Springer Nature Switzerland.

Phonexia (2024). *Voice inspector* (Version 5.2.0) [Computer software]. https://docs.phonexia.com/products/voice-inspector/download-voice-inspector-5-2

San Segundo, E., & Delgado, J. (2025). A qualitative systematic review of intra-speaker variation in the human voice. *Journal of Speech Sciences*, *14*(00), e025019.

Voice Cloning Models

ElevenLabs Inc. (2026): *ElevenCreative* (26 Jan Version) [AI-Platform to generate and alter audio and video]. https://elevenlabs.io/app/home

Zyphra Technologies Inc (2026): Zonos (26 Jan Version) [AI-model for voice cloning]. https://playground.zyphra.com/audio

Comparing humans and neural networks in speech deepfake detection

Manuel Otero-Gonzalez[1,2], *Aurora López-Jareño*[1], *Eugenia San Segundo*[1] *and Joaquin Gonzalez-Rodriguez*[2]

[1]*Phonetics Laboratory*, *Spanish National Research Council (CSIC), Madrid, Spain*

{manuel.otero|aurora.lopez|eugenia.sansegundo}@cchs.csic.es

[2]*AUDIAS, Universidad Autónoma de Madrid, Spain*

joaquin.gonzalez@uam.es

Audio deepfakes generated by modern text-to-speech systems have reached a level of realism that challenges both human listeners, whose accuracy in recent perceptual studies approaches chance levels (Lavan et al., 2025), and automatic detection systems, which often generalize poorly out of domain. This poses significant challenges for forensic phonetics, where speech is used as evidence in forensic speaker comparison (FSC) and authentication tasks.

The few studies comparing automatic detectors and human consensus judgments report comparable performance under realistic conditions (Mai et al., 2023; Warren et al., 2024), with collective human performance matching out-of-domain models, although humans and machines rely on different cues and misclassify different audio samples (Warren et al., 2024).

In this work, we evaluate ResNet18, a state-of-the-art residual deep neural network, using Spanish and Japanese audio samples to compare its performance with human listeners. We hypothesized that a hybrid human-machine approach could enhance classification accuracy. To test this, human and model scores were converted to log-likelihood ratios (LLRs) and both were combined into a hybrid human-machine system using logistic regression. The three systems were further evaluated by measuring their calibration through the log-likelihood ratio cost function (C_{llr}). Finally, to refine the comparative analysis, we qualitatively analysed the participants' reasons for their classification decisions to identify salient characteristics perceived in the audio samples based on concordant and discordant classifications between humans and ResNet18.

Methodology

To evaluate human and model performance, a multispeaker corpus was developed (San Segundo et al., 2025), consisting of 40 natural samples from audiobooks and media interviews in Spanish and Japanese, and 40 synthetic counterparts cloned using the ElevenLabs TTS Multilingual v2 model.

In total, 54 individuals (28 native Spanish and 26 Japanese speakers) were recruited for the detection task. For the automatic detection of deepfakes, we used a ResNet18 (residual network) architecture developed by one of the participating systems (Rohdin et al., 2024) in the ASVspoof5 challenge (Wang et al., 2024). This system was trained and validated on the ASVspoof 5 dataset and subsequently evaluated on the dataset employed in this study.

For metrics comparison, the human performance was measured collectively. The receiver operating characteristic (ROC) curves and their corresponding area under the curve (AUC) were calculated for humans and ResNet18 using the same 80 samples. Additionally, systems were calibrated using logistic regression employing a cross-validation approach (Ramos et al., 2011), and the fusion approach combined both LLRs using the same method

For the qualitative analysis, the methodology of San Segundo et al. (2025) was applied, following the speech classification proposed by Leeman et al. (2024), which is commonly used in FSC.

Results

The hybrid system human-machine showed better performance and the best calibration compared to the individual systems (Figure 1A), with an AUC of 0.8413, C_{llr} of 0.7078 and min-C_{llr} of 0.6867.

This suggests that despite the superior performance of the hybrid system, deepfake detection remains challenging for humans, machines, and fused systems alike.

Following the qualitative study (Figure 1B), we identified perceptual features that stand out particularly in some of the following situations: in audio clips correctly classified by both human listeners and ResNet18 (pattern A), by humans only (pattern B), and by ResNet18 only (pattern C). The preliminary results suggest that laughter, emotion, segmental features, pitch anomalies, and a monotonous intonation/rhythm/intensity/speed/tone may serve as key discriminative clues.

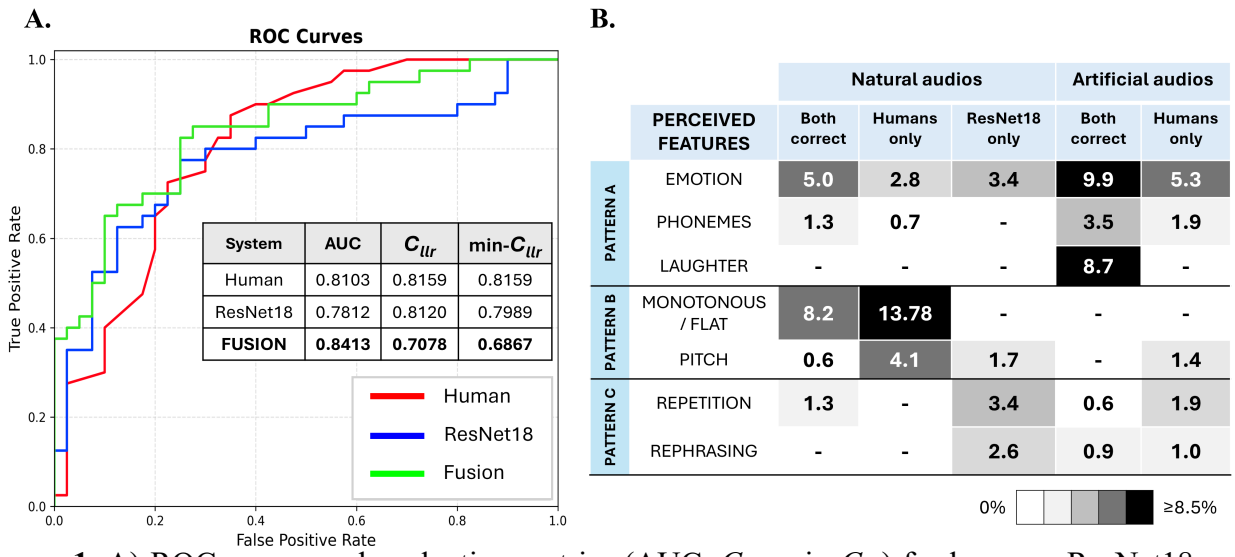


Figure 1. A) ROC curves and evaluation metrics (AUC, C_{llr}, min-C_{llr}) for humans, ResNet18, and the hybrid system (N = 80 audio samples); B) Comparative analysis of human-perceived features across classification outcome groups. Note: Subsets are categorized by whether the human and the model reached the same correct conclusion or diverged. Shading intensity corresponds to normalized frequency levels as indicated in the legend.

References

Lavan, N., Irvine, M., Rosi, V., & Mc Gettigan, C. (2025). Voice clones sound realistic but not (yet) hyperrealistic. *PLoSOne*, *20*(9), e0332692. https://doi.org/10.1371/journal.pone.0332692

Leemann, A., Perkins, R., Buker, G. S., & Foulkes, P. (2024). *An Introduction to Forensic Phonetics and Forensic Linguistics.* Routledge.

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. PLOSONE, 18(8), e0289567. https://doi.org/10.1371/journal.pone.0285333

Ramos, D., Franco-Pedroso, J., & Gonzalez-Rodriguez, J. (2011). Calibration and weight of the evidence by human listeners: The ATVS-UAM submission to NIST HUMAN-aided speaker recognition 2010. Proceedings of the IEEE ICASSP, 5908–5911. https://doi.org/10.1109/ICASSP.2011.5947706

Rohdin, J., Zhang, L., Oldřich, P., Staněk, V., Mihola, D., Peng, J., Stafylakis, T., Beveraki, D., Silnova, A., Brukner, J., & Burget, L. (2024). BUT systems and analyses for the ASVspoof 5 Challenge. Proceedings of the Automatic Speaker Verification Spoofing Countermeasures Workshop (ASVspoof 2024), 24–31. https://doi.org/10.21437/ASVspoof.2024-4

San Segundo, E., López-Jareño, A., Wang, X., & Yamagishi, J. (2025). Human perception of audio deepfakes: The role of language and speaking style [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2512.09221

Wang, X., Delgado, H., Tak, H., Jung, J., Shim, H., Todisco, M., Kukanov, I., Liu, X., Sahidullah, M., Kinnunen, T., Evans, N., Lee, K. A., & Yamagishi, J. (2024). ASVspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *Proceedings of the ASVspoof Workshop 2024*.

Warren, K., Tucker, T., Crowder, A., Olszewski, D., Lu, A., Fedele, C., Pasternak, M., Layton, S., Butler, K., Gates, C., & Traynor, P. (2024). "Better Be Computer or I'm Dumb": A Large-Scale Evaluation of Humans as Audio Deepfake Detectors. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (pp. 2696-2710). https://doi.org/10.1145/3658644.3670325

Calibration of ASV Systems and Robustness Thereof

Torbjörn Onshage[1, 2], *Aron Paulsson*[1], *Greta Öhlund Wistbacka*[1, 3], *Susanna Whitling*[1, 4], *and Andreas Jakobsson*[1, 5]

[1]*Voice Print Sweden AB, Sweden.*
{aron.paulsson|torbjorn.onshage}@voiceprint.se
[2]*Centre for Languages and Literature, Lund University, Sweden.*
[3]*Department of Public Health and Caring Sciences, Uppsala University, Sweden.*
greta.ohlund.wistbacka@uu.se
[4]*Department of Logopedics, Phoniatrics and Audiology, Lunds University, Sweden.*
susanna.whitling@med.lu.se
[5]*Div. Of Mathematical Statistics, Lund University, Sweden.*
andreas.jakobsson@matstat.lu.se

The use of automatic speaker verification (ASV) systems in forensic speech analysis is increasing rapidly. However, to convert the "similarity scores" provided by such systems to likelihood ratios (LRs) used in a forensic speaker comparison setting, access to a dataset similar to that of the investigated material is needed on a per case basis (Leegwater et al., 2024). This study investigates how the choice of calibration data affects this conversion. Using the feature model ECAPA-TDNN (Desplanques et al., 2020) with cosine-similarity scoring, we evaluate calibration across multiple dataset configurations on the VoxCeleb1 dataset (Nagrani et al., 2017) and examine the influence of two factors: recording quality degradation, and demographic partitioning. This is then compared to the results of van der Vloed (2024).

Quality Degradation

Applying a GSM telephone filter to the VoxCeleb1 dataset introduces a significant difference on the resulting calibration curves, as compared to the unfiltered data, as seen in Figure 1. This difference is at worst approximately 4 log-likelihood units with KDE calibration at the same score. These shifts could correspond to substantially different conclusions under standard verbal LR scales, cf. Association of Forensic Science Providers (2009). Down-sampling to 8 kHz produced a more moderate but still practically significant degradation.

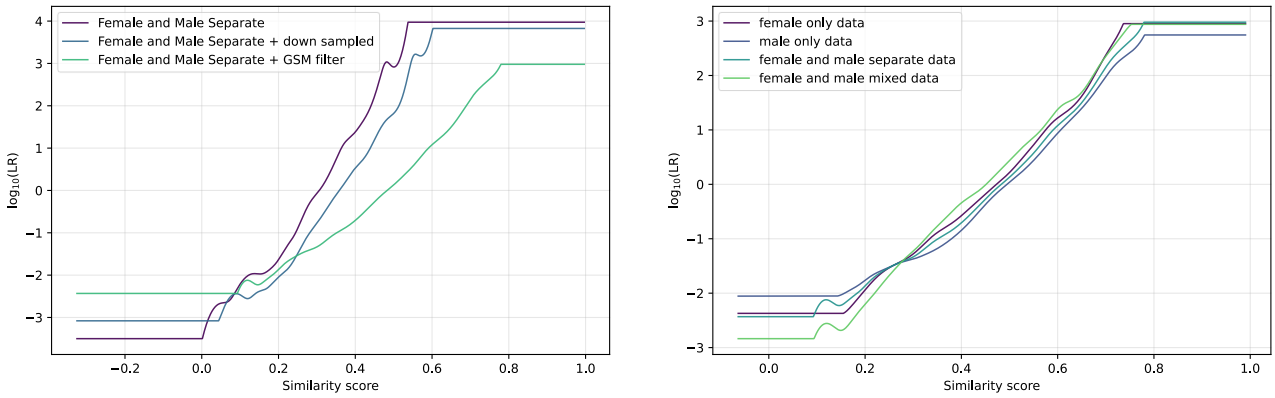


Figure 1. KDE based calibration curves with ELUB-bounds (Vergeer et al., 2016) based on ECAPA-TDNN scoring of the VoxCeleb1 dataset, quality degradation (left), and demographic partitioning (right).

Demographic Partitioning

The effect of gender-based sub-grouping was tested using four configurations: male-only, female-only, separate same-gender pairs, and mixed pairs. The results of such division produced only marginal differences in calibration output, as illustrated in Figure 1. Even though gender dimorphism in speech is generally perceptually noticeable, Figure 1 suggests that fine-grained demographic partitioning yields limited practical gains. This is likely due to the score-based calibration approach, i.e., calibration depends on the difference between the two speakers and not their extracted feature values. This is in line with the results of van der Vloed (2024), who, on their data, found that language affected the calibration less than recording conditions.

Discussion

The results indicate that the quality and acoustic environment have a larger effect on the calibration process, vis-à-vis narrowing the demographic partitioning. While calibration against female speakers in male cases might be extreme, it raises the question of relative importance between matching the demographic of the speaker, and the setting of the recording, to the calibration. Furthermore, how can practitioners identify the existence of non-interchangeable factors, and what they are? Especially, from a dataset with a wide variety of mixed conditions, when is a dataset sufficiently representative, and the calibration sufficiently reliable to support evidence reporting.

The property of interchangeability as suggested by van der Vloed (2024) provides a way of measuring this relative importance. However, with this approach, to assess the need for more granular data matching certain conditions, one needs to gather said data. Despite these shortcomings, can interchangeability be used to assess and improve the robustness of scoring systems, and to validate augmentation techniques for simulating acoustic conditions such as telephone recordings?

References

Association of Forensic Science Providers. 2009. “Standards for the Formulation of Evaluative Forensic Science Expert Opinion.” Science & Justice 49(3):161–64. doi:10.1016/j.scijus.2009.07.004.

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. https://doi.org/10.1016/j.forsciint.2024.111994

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. Interspeech 2020, 3830–3834. https://doi.org/10.21437/Interspeech.2020-2650

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

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? Science & Justice, 56(6), 482–491. https://doi.org/10.1016/j.scijus.2016.06.003

van der Vloed, D. (2024). Interchangeability of Calibration Audio Datasets for Forensic Automatic Speaker Recognition. 2024 12th International Workshop on Biometrics and Forensics (IWBF), 1–6. https://doi.org/10.1109/IWBF62628.2024.10593938

The Risk of Data Contamination in Forensic Voice Comparison: a Perception Experiment using Voice-mixed Speech Samples

Scarlett Vanroose-Peer
Linguistics and English Language, Lancaster University, UK
s.vanroose-peer@lancaster.ac.uk

Data contamination arising from sample-mixing is a widely acknowledged phenomenon across many forensic disciplines, but the risk it poses to Forensic Voice Comparison has not yet been explored. In speech recordings containing multiple speakers, the typical first data preparation step is to separate these recordings into single-speaker samples, i.e., a form of speaker attribution. Inaccuracies in speaker attribution would result in 'voice-mixing', and the consequences on downstream analyses are yet to be uncovered. Moreover, there is evidence in the neighbouring literature that speakers who have highly similar voices have great potential to be mistaken for one another (Gerlach et al, 2023). The precise nature of how those findings extend to the prospects and potential for 'voice-mixing' is not currently known, but it suggests that the risk of 'voice-mixing', in at least some contexts, could be high. Taken together, this paper addresses two main questions: (1) To what extent can listeners detect voice-mixed from authentic voice samples? (2) To what extent does the degree of voice-similarity in the voice-mixed samples impact detection?

Methodology

118 naïve listeners completed an online perception experiment, where they listened to voice-mixed and authentic speech samples and rated the extent to which they believed voice-mixing to be present. Stimuli were created using Task 4: 'voicemail scenario', of the WYRED (Gold, Ross & Earnshaw, 2018). *Wespeaker* (Wang et al, 2024) automatic similarity scores were retrieved to create a subset of 6 speakers consisting of high-, mid- and low-similarity pairs. Each speaker in the subset was mixed with every other speaker in the subset to create 6 'authentic' and 15 'voice-mixed' stimuli, each ~15s long. The voice-mixed samples contained 1-3s of contaminated speech. Participants listened to all 21 samples in a randomised order and rated each sample on a 5-point Likert scale ranging from definitely/likely voice-mixed, to unsure, to likely/definitely authentic. Two Cumulative Link Mixed Models were performed to determine whether exposure to a given sample group significantly predicted the listeners' answers. Krippendorf's alpha coefficient was used as a measure of inter-rater reliability, to establish the extent of agreement amongst listeners.

Results

Results revealed that there was not a significant relationship between whether the sample was voice-mixed or authentic, and listener responses ($p = 0.86$). However, there was a significant relationship between the degree of similarity within the voice-mixed samples and the responses obtained. Listeners were more likely to rate voice-mixed samples with highly similar-sounding speakers as authentic than for true authentic samples, whereas listeners were less likely to rate mid- and low-similarity voice-mixed samples as authentic. Results also showed a relatively large variation between listeners' responses ($\alpha = 0.42$). Taken together, the results revealed that listeners were generally poor at distinguishing voice-mixed from authentic recordings. As expected, the risk of voice-mixing is larger when speakers are highly similar-sounding, but the risk does not entirely diminish when the speakers are less similar-sounding.

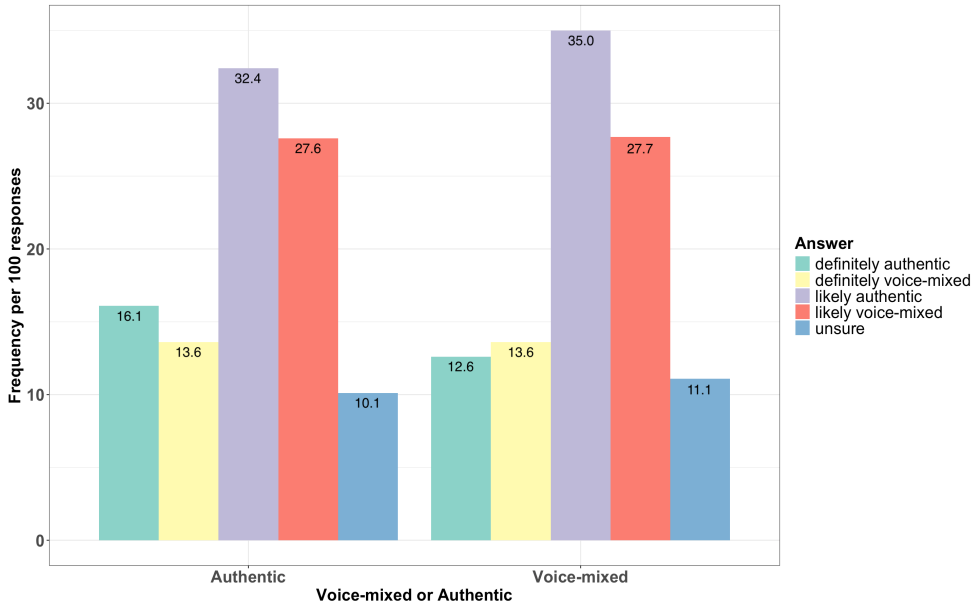


Figure 1: Listener responses for all authentic and voice-mixed samples

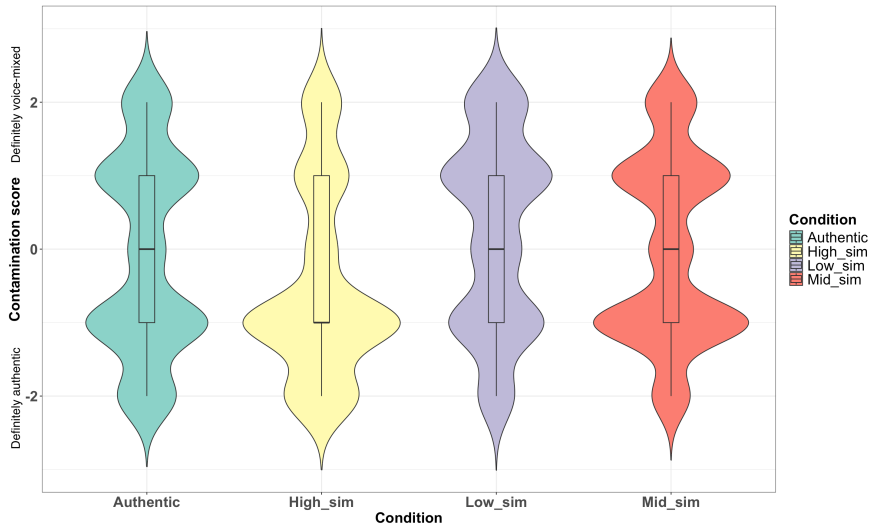


Figure 2: Distribution of ratings for authentic and voice-mixed samples by speaker-similarity

References

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2023). Voice twins: discovering extremely similar-sounding, unrelated speakers. *Proceedings of Interspeech 2023*, Dublin, Ireland.

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. *Proceedings of Interspeech 2018, Hyderabad, India*, 2748-2752.

Wang, S., Chen, Z., Han, B., Wang, H., Liang, C., Zhang, B., ... & Li, H. (2024). Advancing speaker embedding learning: Wespeaker toolkit for research and production. *Speech Communication*, *162*, 103104.