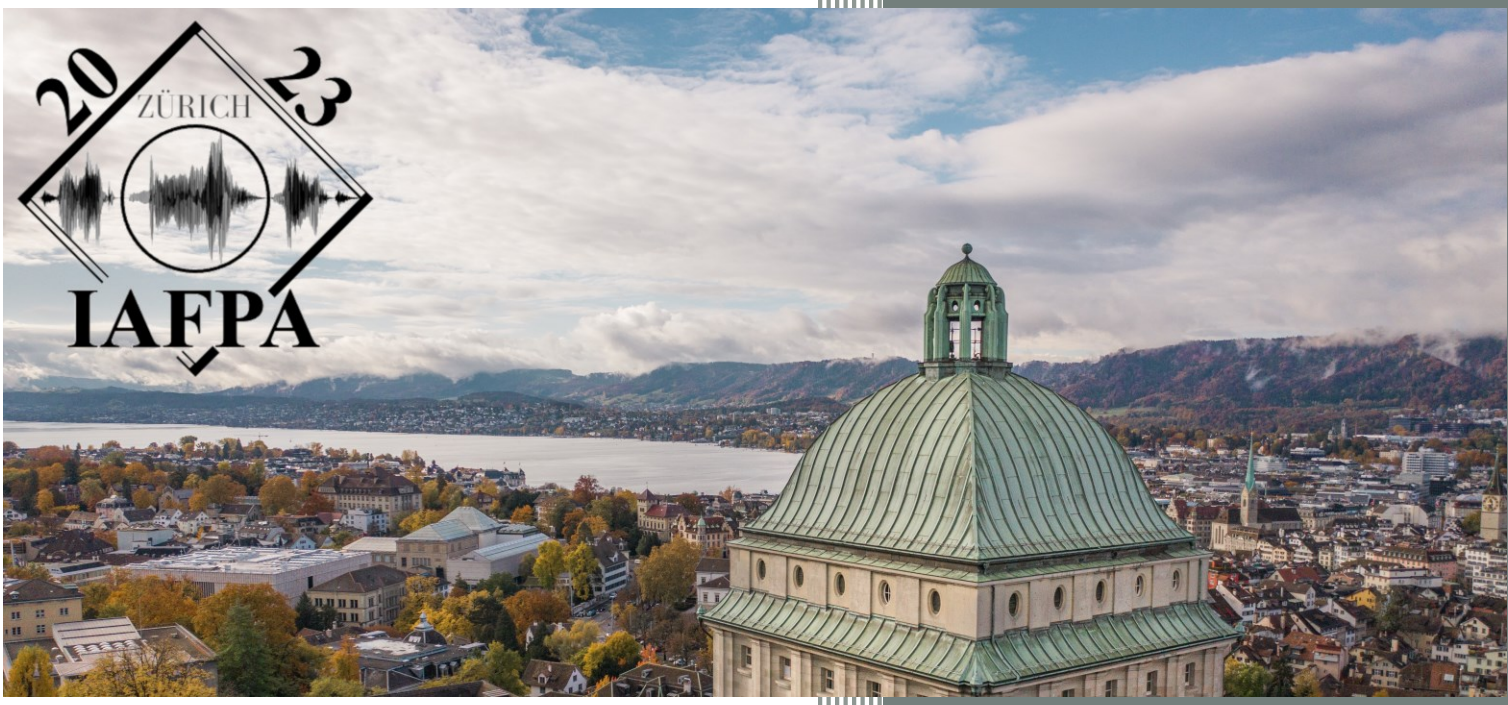
Book of Abstracts

31st IAFPA Conference



CFPA
Centre for Forensic Phonetics and Acoustics (CFPA)

FOR Zürich
Forensisches Institut

July 2023

Invited Speakers

Workshops

Oral presentations

Special Session 1: Case reports

Oral Session 1: Speaker discrimination

Poster sessions

Poster Session 1

Poster Session 2

Invited Speakers

Georgina Brown (Lancaster University, UK)

Synthetic Speech is More Than a Threat

Speech synthesis systems make up a substantial group of "spoofing" techniques that could, in principle, generate speech samples for fraudulent or other malicious purposes. Rapid developments within speech synthesis undoubtedly mean that these technologies will continue to contribute to widespread concerns around "deepfakes". The threat of synthetic speech now has a firm presence in the media, but there is also a body of academic research that has accumulated over the last decade. Research attention has largely been afforded to how automatic speaker recognition systems perform when presented with different forms of spoofed speech. There has also been research on how well spoofed speech can be detected by automatic classifiers and human listeners. As the need for ongoing "spoofing-aware" research is clear, it would not come as a surprise to see further work on spoofed speech in presentations at IAFPA 2023. One aspect that unites much of the existing research literature on spoofed speech is the impressive performance of one particular speech synthesis technique. This talk places this method under the microscope; however, instead of focusing on the threatening aspect of synthetic speech, this talk will zoom in on the opportunities that have become available to forensic speech science. Specifically, this talk demonstrates how speech synthesis methods could help to make automatic speaker recognition systems more explainable.

Catalin Grigoras (UC Denver, USA)

Digital Audio Authentication: Framework, Challenges, and Solutions

The goal of this presentation is to provide a summary review of the latest developments in conducting comprehensive examinations of digital audio authenticity which rely on the results of multiple analyses to inform an ultimate scientific finding or unbiased opinion. Digital audio authentication is a process of establishing the provenance of a questioned recording to determine whether it is consistent with an original one or if there is evidence of editing. This presentation proposes the organization of several techniques in a logical manner for the authentication of digital audio recordings. Special attention has been given to interpreting results from individual analyses and incorporating them into a holistic view of a recording's authenticity where a finding can be corroborated against the results of other analyses. Only in this way can an examiner present a conclusion with confidence and assurance that all possible hypotheses have been exhausted in the execution of this important endeavor. The framework for digital audio authentication that will be discussed involves accurate, repeatable, reliable, unbiased, and scientific analyses derived from peer reviewed publications in order to meet court guidelines or case precedence, best practice recommendations, and the appropriate criteria for international legal systems. The presentation will also include some of the nowadays challenges and solutions.

Petr Motlicek (IDIAP - Speech and Audio Processing Group, Switzerland)

Speaker Verification and Identification for Criminal Investigations and Forensics

The talk will present innovative technologies aiming to support law enforcement agencies become more efficient in processing large-scale diverse cases with a goal to unmask criminal networks and their members as well as to reveal the true identity of perpetrators. Recently closed EC project (ROXANNE), combining capabilities of speech/language technologies and visual analysis with network analysis, will be introduced and project achievements will be demonstrated on various realistic data. Particular focus will be given to speech processing related applications (i.e., speaker identification and verification) deployed by investigators to automatically analyse and cluster lawfully intercepted communication specifically for cross-border organised crime. The talk will also introduce technological solutions to deploy speaker recognition for forensic case.

Workshops

Tina Cambier-Langeveld (Leiden University, NL) and Christin Kirchhübel (Soundscape Voice Evidence, UK)

Let's Talk About Bayes!

Within the forensic field and across many disciplines there has been a shift towards using a Bayesian framework. In this workshop we discuss why that is so. The workshop includes basic training in the principles of Bayes, as well as a discussion on its suitability for non-numerical, experience-based assessments of the evidence as in auditory-acoustic speaker comparisons. We also share experiences with the use of a Bayesian conclusion format in forensic speaker comparison casework, including how it is received and understood by courts (at least in NL and UK jurisdictions).

The workshop is suitable for everyone with an interest in the Bayesian framework. No familiarity with the Bayesian framework is required.

Volker Dellwo (University of Zurich, CH) and Herbert Masthoff (Trier University, DE)

Doing Casework

This workshop is specifically tailored to participants interested in gaining deeper insights into the practical aspects of voice analysis and speaker comparison within the field of forensic phonetics. We will analyze conversational speech samples relevant for forensic cases with auditory and acoustic methods. Participants should bring headphones and equipment for download and playback.

Philip Harrison and Amelia Gully (University of York, UK)

Breaking the Chain: Exploring Limits of Interpreting Audio Evidence

In this workshop, we will explore the elements of the recording chain which are important in forensic casework. We will consider how different sound sources, recording conditions and devices affect the final evidential recording, and how interactions between all of these elements need to be considered when interpreting forensic recordings. We will consider sound source identification and other more common forensic examinations, and exemplify how erroneous conclusions can be drawn if all the relevant factors are not properly taken into account. Using example recordings and interactive demonstrations, we will highlight issues encountered in real cases and discuss the implications for the auditory interpretation of evidential recordings.

You're on the wrong track!

Some disputed utterances in a two-channel phone tap

Vincent J. van Heuven[1, 2, 3]
[1]*Leiden University Centre for Linguistics, Leiden, The Netherlands*
[2]*Multilingualism Doctoral School, University of Pannonia, Veszprém, Hungary*
[3]*Fryske Akademy, Ljouwert/Leeuwarden, The Netherlands*
v.j.j.p.van.heuven@hum.leidenuniv.nl

.

I was asked to evaluate the accuracy of transcripts of phone calls tapped by the Amsterdam police, parts of which were challenged by the accused and his counsel, who argued that the transcripts were (willfully) incorrect and biased against the defense. In my talk I will deal with three disputed passages, which were potentially incriminating for the accused.

1. Diarisation error

The phone calls were tapped such that caller C and receiver R were recorded on separate channels with no crosstalk. The police report claims that the (disputed) R was addressed by C by his first name, Mostafa. It was easy to show, however, that the relevant turn was spoken by R, and that the wrong speaker was assigned to the channel. The transcript was indeed in error here (see also Figure 1).

2. Always listen to channels separately

In the same passage, accused and counsel were certain they heard someone say: *Da's snel* /dɑsnɛl/ 'That's fast!' even though this was not reflected in the transcript. I agreed with them after listening to the recording though a single loudspeaker (as they had done). I then realized that we might be dealing with an artefact. When the two channels are merged, the loudest elements of both channels are dominant and mask the speech on the other channel. When these loud portions are excised from the single channels, and concatenated in their proper time order, there is indeed a quite convincing 'That's fast!' I will demonstrate this in my talk. When the channels are heard separately, the police transcript is correct. The take-home message here is: always listen to channels separately (if you can).

3. Factor prosody in

The police report states that C asks (in Dutch) *Kan ik een broodje bij je kopen?* /kɑn ɪk ən brotjə bɛɪ jə kopə/ > [kɛnəbrocəbɛkopə] 'Can I buy a sandwich from you?', which the prosecution considers incriminating, since *broodje kopen* 'buy a sandwich' would stand for 'buy drugs'. Accused and counsel, however, hear *Kan je naar Broodje Bert komen?* /kɑn jə nar brotjə bɛrt komə/ > [kɛnəbrocəbɛ komə] 'Can you come to Broodje Bert?' (*Broodje Bert* is a respectable sandwich bar in the Amsterdam city centre with no known drug-related activities). I could show that only the reading suggested by the accused is compatible with the signal. For this conclusion, I had to factor in the phonology of fast speech (deletion and coalescence of segments, as documented in Booij, 1995), properties of the Amsterdam city dialect (*kunnen*/*kan* > *kennen*/*ken* 'can'; Berns, 2002), the absence of silence + noise burst in *kopen*/*komen*, and, crucially, the incompatibility of the location of the H*L pitch peak (prepositions do not receive sentence stress in Dutch, e.g., Rietveld & Van Heuven, 2016: 287-291) with the reading suggested by the police transcript (details will be given in the talk, see also Figure 2). The police report was indeed wrong here.

References

Berns, J. (2002). *Taal in stad en land: Amsterdams [Language in town and country: Amsterdam]*. Sdu Uitgevers.

Booij, G. E. (1995). *The phonology of Dutch*. Cambridge University Press.

Rietveld, A. C. M., & Heuven, V. J. van (2016). *Algemene fonetiek. Vierde herziene en uitgebreide druk. [General phonetics. Fourth revised and extended edition]*. Coutinho.

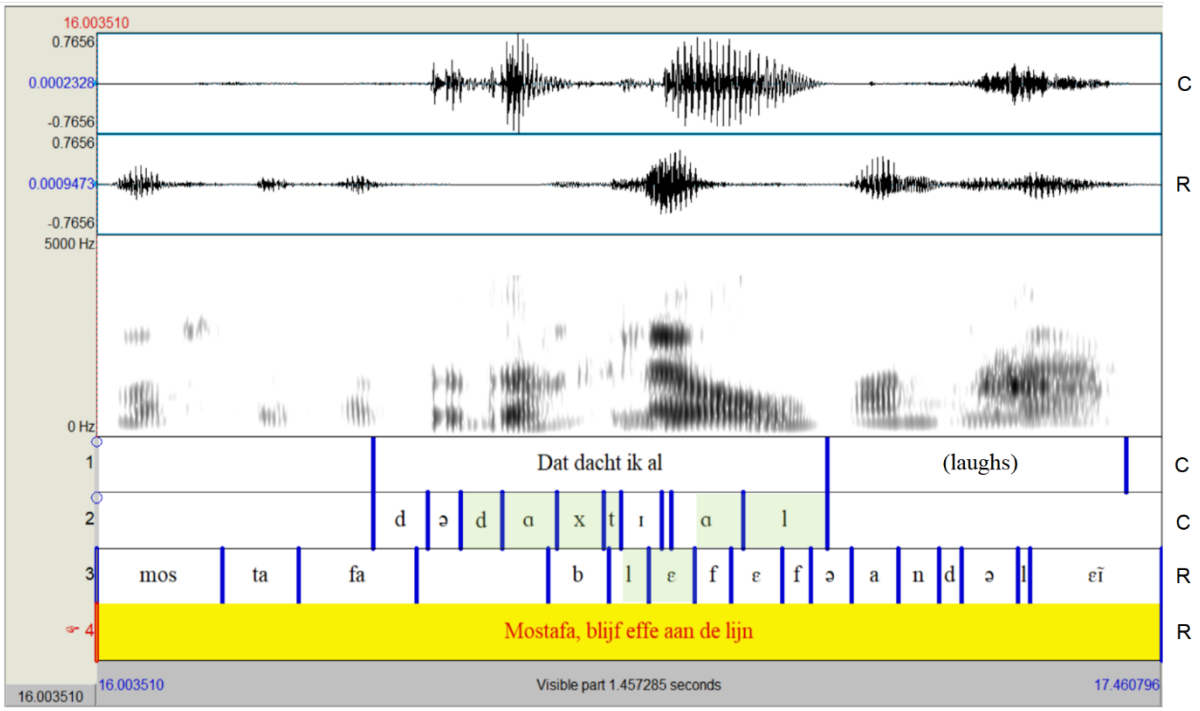


Figure 1. Oscillogram of channel 1 (caller, C) and channel 2 (receiver, R). The third panel from the top contains the spectrogram of the fusion of channels 1 and 2. Annotation tier 2 is a segmentation into phonemes of C's utterance, tier 3 that of R's utterance. When the recording is heard over a single loudspeaker, it is as if one speaker says: *Da's snel* /dɑsnɛl/ 'That's fast!" This is also what is heard when the loud segments (marked in green) are excepted from their separate channels (no crosstalk) and seamlessly concatenated. However, when the channels are heard separately, C clearly says *Dat dacht ik al* /dədɑxɪkɑl/ 'I thought as much' while R says: *Mostafa, blijf effe aan de lijn* /mɔstafa | blɛfɛfandəlɛĩ/ 'Mustafa, just stay on the line'.

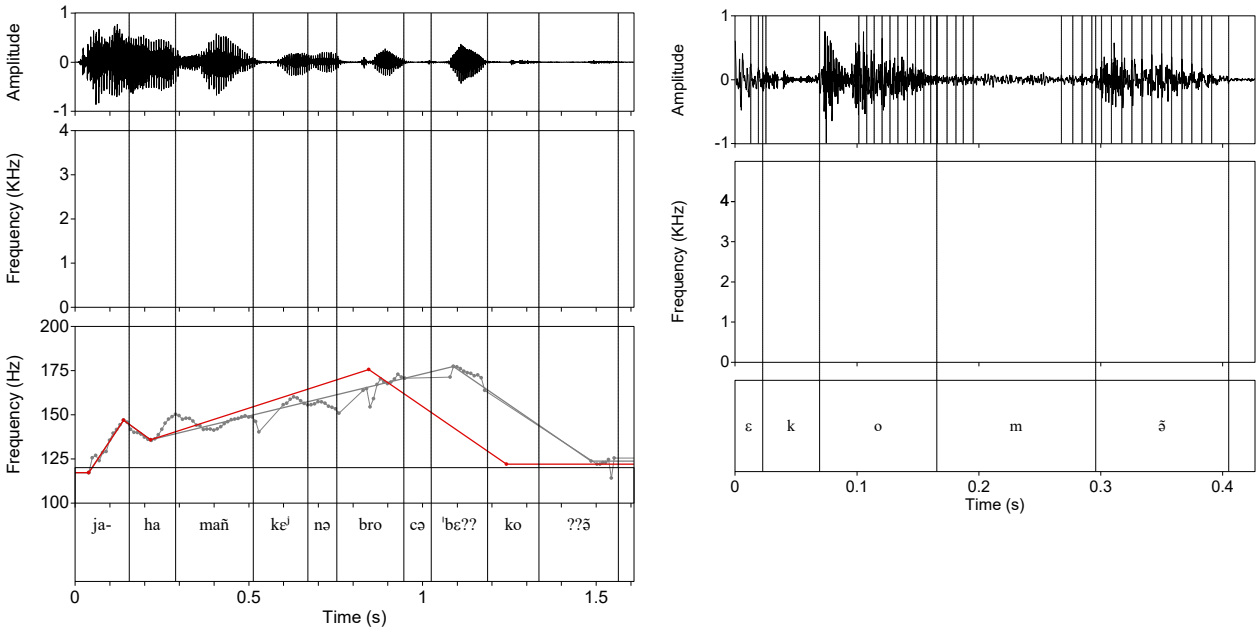


Figure 2. Left: oscillogram, spectrogram and pitch curve (raw = dotted; solid black line = stylized) for disputed utterance (see text). The red stylization is what would be seen if the police transcript had been correct. The right-hand panel is an amplified close up of the utterance-final disputed word *komen*/*kopen*. The intensity below 1.5 kHz in /m/ (but not in /k/) shows that /m/ cannot be a plosive.

Forensic analysis of audio recordings using the Electric Network Frequency method: a case study

Arjan van Dijke
Speech, Language and Audio department, Netherlands Forensic Institute,
The Hague, The Netherlands
a.van.dijke@nfi.nl

In the field of audio forensics, the Electric Network Frequency (ENF) Criterion is known for its use as a tool for authenticating audio recordings (Grigoras, 2005). This method is based on the fact that the electrical grid in a region operates at a consistent, slowly changing frequency, which can be captured in audio recordings as a subtle hum. By extracting this hum and comparing the signal to reference values it is possible to verify the authenticity and integrity of an audio recording. The length of the recording can be a limitation in this method. Under certain conditions, shorter recordings (4 – 10 minutes) can be correctly placed on a specific date and time (Huijbregtse & Geradts, 2009).

Research question

The Netherlands Forensic Institute (NFI) investigated a case about authenticity of police interview recordings. These recordings had multiple points where the recording was suspected to be stopped and restarted. The defense claimed that during these interruptions, the interviewed victim was being influenced by the police interrogators. The court ordered the NFI to investigate the recordings and report about the timings and lengths of these interruptions.

While listening to the audio recordings, it became clear that the recorder had picked up a low frequency hum at 50 Hz. Extraction of the hum showed a slowly changing signal around this frequency as well as at the higher harmonics at 100 Hz and 150 Hz. ENF analysis was used to investigate if the recordings were made at the claimed date and if there were interruptions. The ENF analysis in this case was part of a larger authenticity analysis, which included linguistic conversation analysis and inspection of recording artefacts.

Method and results

The recording dates and place were not disputed, so the ENF reference data for the dates of the recordings were requested from the national electricity transmission system operator of the Netherlands. Five recordings with lengths between 4 minutes 19 seconds and 138 minutes were analyzed with a MATLAB-based ENF toolbox developed at the NFI, based on the method described by Cooper (2008). Three recordings could be placed integrally on the claimed recording dates. Two recordings could not be fitted integrally on the claimed recording dates, but visual inspection of the signals showed that they matched partially. The ENF toolbox was extended with code to search for interruptions in the recordings. This method showed that these two recordings consisted of concatenated parts which were recorded after each other, but not directly following each other in time. This is compatible with the scenario where the recordings were paused and continued multiple times. Figure 1 shows the ENF-signal of one of the recordings compared with the ENF reference data for part of the recording day.

Conclusion

In a case where ENF hum was captured in several audio recordings, it was possible to use ENF analysis to check for interruptions in the recordings. An existing ENF toolbox was extended with code to check for these interruptions and showed useful results. The number of interruptions and their durations were reported to the court.

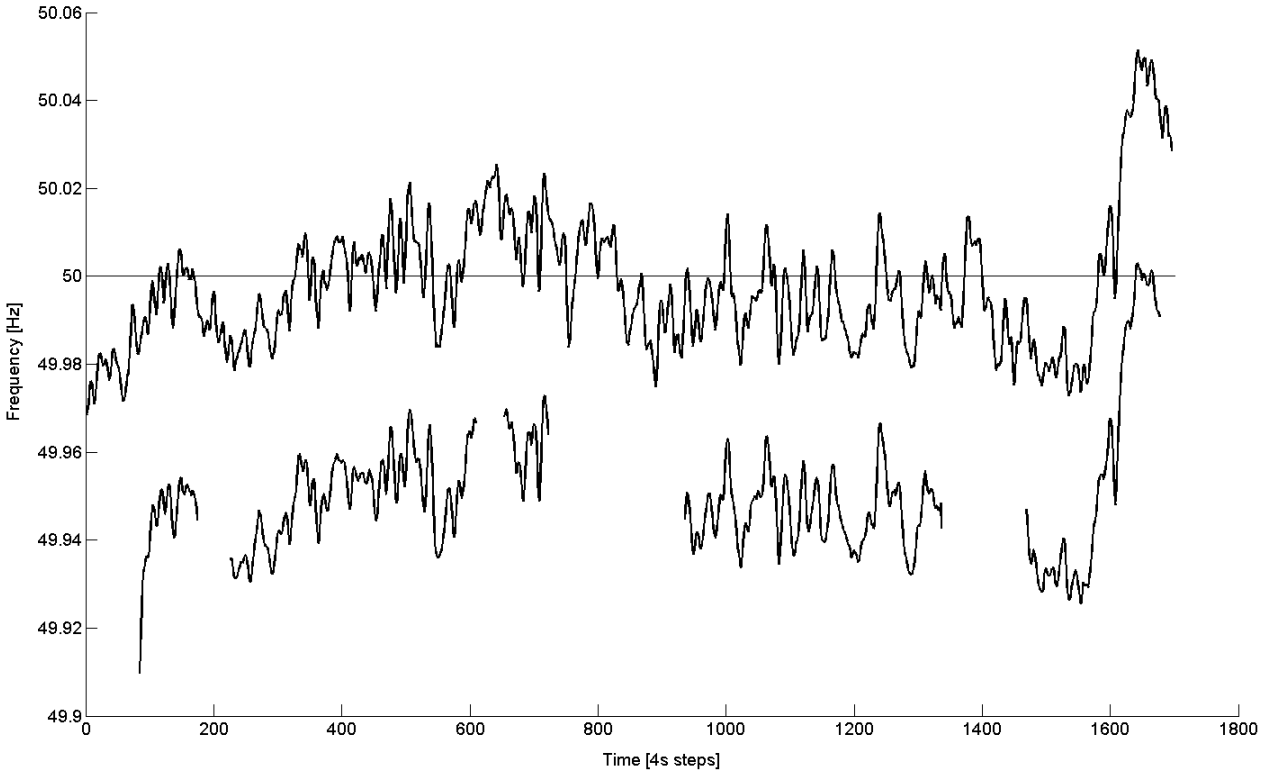


Figure 1. The ENF reference data for a continuous stretch of approximately 2 hours on the recording day (top) and the ENF signal of one of the recordings (bottom). For clarity, the ENF signal from the recording is shifted down by 0.05 Hz. The interruptions in the recording are visible as gaps in the bottom signal.

References

Cooper, A. (2008). The Electric Network Frequency (ENF) as an Aid to Authenticating Forensic Digital Audio Recordings – an Automated Approach. *Audio Engineering Society Conference: 33rd International Conference: Audio Forensics-Theory and Practice.*

Grigoras, C. (2005). Digital audio recording analysis: the Electric Network Frequency (ENF) Criterion. *The International Journal of Speech Language and the Law*, *12*(1), 63-76.

Huijbregtse M., Geradts Z. (2009). Using the ENF Criterion for Determining the Time of Recording of Short Digital Audio Recordings. In: Geradts, Z.J.M.H., Franke, K.Y., Veenman, C.J. (eds) *Computational Forensics. IWCF 2009. Lecture Notes in Computer Science*, vol 5718. Springer, Berlin, Heidelberg.

Case report: how was the disputed audio recording stopped?

Honglin Cao[1,2], Sixue Gao[1,2], Xiaodong Xu[1,2] and Yue Zhong[1,2]
[1]Key Laboratory of Evidence Science (China University of Political Science and Law), Beijing, China.
[2] Fada Institute of Forensic Medicine & Science, Beijing, China.
caohonglin@cupl.edu.cn

Under Chinese law (Supreme People's Court (2022): article 106), in civil cases, if the audio evidence is not formed or acquired by serious infringement upon the lawful rights and interests of others, violation of the law prohibitions or serious breach of public order and good custom, it will be legitimate. According to a recent survey (Cao and Zhang, 2020) based on hundreds of real forensic phonetics-related judgment documents, audio authentication was found to be the second largest number task (31.1%) across the whole forensic-phonetic cases, next only to forensic voice comparison (59.3%).

In this paper we will present a real civil case of forensic audio authentication. The disputed audio recording (DAR) was a face-to-face recording, which was claimed to be made with a pre-installed voice recorder app on a questioned Huawei mobile phone (QMP), of which the model was BLA-AL00 (i.e., Mate 10 Pro). Three speakers were involved in the DAR, of which the duration was about 4 hours (4h 0min 1s 55ms). Speaker A, who was the plaintiff and the owner of the QMP, made the recording. Speaker B and C, who were the defendants, did not contest the speaker identity, but claimed that the DAR had been manipulated by cutting, pasting, and deleting. Several reasons were listed by the defendants for objection, including: (1) the duration of the whole conversation was about 8 hours, other than 4 hours; (2) in the DAR, there were two interruptions, followed by two short self-talking and lower-quality speeches spoken by speaker A, etc. The judge ordered forensic ## Institute to investigate the authenticity of the DAR.

Following the guideline from the *Technical Specification for Forensic Audio Authentication (SF/T 0120-2021)*(Ministry of Justice, 2021), we conducted a systematic examination on the DAR and the QMP, and used another reference Huawei mobile phone (Mate 10 Pro, same model with the QMP; hereinafter referred to as RMP) for experimental and destructive testing.

Several abnormal phenomena were found in the preliminary examination, for example: (1) the "com.android.version" of the DAR is 9, lower than the android version of the QMP (10. Specifically, the build number of the Emotion User Interface (EMUI) of the QMP was 10.0.0.188); (2) the "File last modification date" was "2020-01-15", nine months later than the "Encoded/Tagged date" (2019-04-09); (3) the waveform of the end of DAR is not silence, however, the signals of the end of the other recordings stored in the QMP and some experimental audio recordings (stopped manually in a normal way by the authors) made by QMP (hereinafter EAR_QMP) were all silent segments (the average duration was 12.5 ms); (4) auditory analysis clearly indicated that the conversation was not over when the DAR stopped; (5) the number and type of metadata elements of the DAR and the EAR_QMP were different; (6) the bandwidths (upper frequency of spectrum range) of DAR and EAR_QMP are 16.2kHz and 17.2kHz, respectively, etc.

Normally, we always update the android mobile operating system (EMUI) consciously or unconsciously, and consequently the voice recorder app will also be updated. After updating to some versions, the metadata and the acoustic characteristics of the audio recordings will see some slight changes. As the plaintiff acknowledged that he had updated the system of QMP, so we cannot directly deny that the QMP was not the original recorder of the DAR. In order to investigate the influence of the system/app update, we used RMP to downgrade the android operating system (EMUI) version with the help of the software *Huawei Mobile Assistant*. **Table 1** listed all of the available combinations of EMUI version and voice recorder app version we could use during the period of the examination. Some abnormalities, such as metadata elements and bandwidths, could be interpreted, when the android version was downgraded to 9. The abnormality of date could also be reproduced by re-importing the audio recordings from the computer to the RMP (and also QMP). Those two interruptions and the lower-quality speeches found in DAR, which was emphasized by the

defendants, were also interpreted: during the DAR was being made, the plaintiff played two WeChat audio messages using the QMP; we precisely reproduced this phenomenon using the RMP.

The biggest challenge was to explain why the signal of the end of the DAR was not silent segments. Whether this abnormal acoustic feature was caused by deliberate manipulation (deleting), or other reasons, such as using some abnormal ways to stop the DAR: one phone call (or WeChat audio/video call) was coming, low battery power, battery ran out, accidental power off, the voice recorder app was closed accidentally, or automatically stopped or discontinued because of the limitation of the maximum duration of a single recording? Luckily, after a series of experiments using the RMP, we found that within a special combination (EMUI version 9.0.0.187 + recorder app version 9.0.2.300), the audio recording would stop/discontinue automatically when the duration reached or exceeded 4 hours (see **figure 1**). With this special setting, we repeated the experiment 25 times and found that the signals of the end of the recordings could be either silent segments (17/25) or non-silent noise (8/25). These results provided a plausible explanation to the biggest abnormality found in the DAR.

No.	Android version	EMUI version	Voice recorder app version	Pre-install
1	10	10.0.0.188	10.0.1.563	yes
2		10.0.0.180	10.0.0.516	yes
3		10.0.0.175		yes
4		10.0.0.170		yes
5		10.0.0.156		yes
6	**9**	9.1.0.339	9.1.1.354	yes
7			11.1.1.440	download
8		9.1.0.321	9.1.0.341	yes
9			11.1.1.440	download
10		**9.0.0.187**	**9.0.2.300**	**yes**
11			9.1.1.340	download
12			9.1.1.347	download
13			11.1.1.440	download

Table 1. 13 different experimental settings for the reference Huawei mobile phone (RMP).

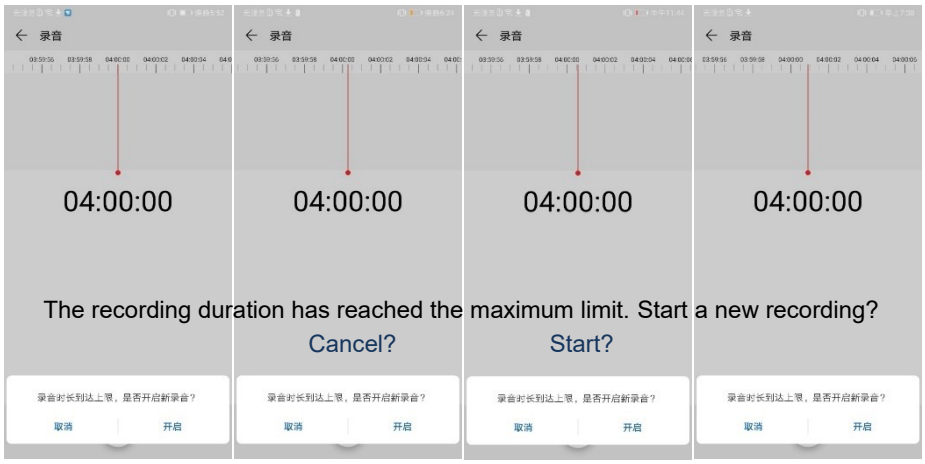


Figure 1. The audio recording would stop automatically, when the duration reached or exceeded 4h.

References

Cao H., and Zhang X. (2020). An Empirical Study on the Present Status of the Application of Evidence of Forensic Phonetics in Courts of China. *Chinese Journal of Phonetics* (01): 90-104.

Ministry of Justice of the People's Republic of China (2021), Technical specification for forensic audio authentication (SF/T 0120-2021).

Supreme People's Court (2022), Interpretation of the Supreme People's Court on the Application of the Civil Procedure Law of the People's Republic of China, https://www.court.gov.cn/fabu-xiangqing-353651.html.

The influence of mismatch conditions on LRs

Angelika Braun and Jacek Kudera
Department of Phonetics, Trier University, Trier, Germany
{brauna|kudera}@uni-trier.de

A frequently cited advantage of ASR over auditory-acoustic voice comparison is that it will provide the results in LR format (Morrison 2009, 2018; Morrison & Enzinger 2016, 2018, 2019; Champod & Meuwly 2000). This is said to be preferred over the "traditional" verbal conclusions, because it includes the typicality aspect beyond the similarity criterion by assessing similarity against a suitable reference population (Ajili 2017; Becker 2012). Still, some forensic practitioners have responded to the less-than-enthusiastic "customer" reactions to LRs (Sjerps & Biesheuvel 1999; de Keijser & Elffers 2012; Braun 2021) by reconverting numeric LRs into verbal probability statements (Rose 2002; Champod & Evett 2000). And yet LRs are not carved in stone. The variability may be the result of the type of material submitted and/or decisions made by the human operating the ASR system (Braun 2021; Hansen & Hasan 2015).

Three principal causes for the variability of LRs can be distinguished:

- The nature of the materials, specifically mismatch conditions of different kinds, e.g. channel mismatches and situational mismatches. This is a group of factors upon which the expert does not usually have any influence.
- The decisions taken by the expert on which materials to use and which to discard. This is a genuinely subjective element which largely depends on the skills and discretion of the expert. This includes preprocessing of materials by the expert which may serve to remove background noise or nonlinguistic elements like laughter or coughing.
- The choice of background materials which are used to assess the typicality of the voices in question.

The present contribution explores some of these issues and presents a range of verbal ratios yielded in examining the three variables. The mismatch conditions were tested using a leading ASR system on the market. The ASR environment has been proved suitable to use in the tested case conditions on the basis of the comparison of multiple files for known speakers. The following variables were tested, all other things being equal:

- Experiment 1: A re-examination of case materials involving situational mismatch which had previously been analyzed using two different ASR systems;
- Experiment 2: Same speaker with different quantities of material (N = 2);
- Experiment 3: Same speaker sober and inebriated (N = 10).

Materials consisted of spontaneous speech in experiments 1 and 2, and readings of "The North Wind and the Sun" as well as picture descriptions in experiment 3. First results indicate that LRs in the re-examination differ from those of the initial analysis mostly in degree, but sometimes (for some recordings) also in kind. The magnitude of this variability may amount to several steps on the reconverted verbal scale. Increasing the amount of material will not necessarily lead to higher LRs. Intoxication causes a decrease in likelihood ratios.

These findings may become an issue especially in jurisdictions that usually work with one expert only (Braun et al. 2005; Broeders 1999; Margot 1998). If there is no counter expert to challenge the results and point out to the courts that the LR would be different under other circumstances, there is a danger that the triers of fact will erroneously treat the numbers (and the related verbal conclusion) as "God's Truth", which they are not.

References

Ajili, M. (2017). Reliability of voice comparison for forensic applications. PhD Dissertation, University of Avignon.

Becker, T. (2012). *Automatischer Forensischer Stimmenvergleich*. Verlag Book on Demand.

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, D., Pellegrino, E. e St. Schmid (eds.): 'Lindividualità del parlante nelle scienze fonetiche: applicazioni tecnologiche e forensic. Speaker individuality in phonetics and speech sciences: speech technology and forensic applications. Milano. 174-191.

Braun, A., Köster, J.-P., Künzel H.J. & Odenthal, H.-J. (2005). Speaker Recognition in Germany. In Wolf, D. (Ed.), *Beiträge zur Geschichte und neueren Entwicklung der Sprachakustik und Informationsverarbeitung, Werner Endres zum 90. Geburtstag*. Dresden: w.e.b. Universitätsverlag, 78-86.

Broeders, A.P.A. (1999). Some observations on the use of probability scales in forensic identification. In *The International Journal of Speech, Language, and the Law* 6(2), 228-241.

Champod, C., Evett, I. W. (2000). Commentary on APA Broeders (1999) 'Some observations on the use of probability scales in forensic identification', Forensic Linguistics 6 (2): 228–41. In *The International Journal of Speech, Language and the Law*, *7*(2), 239-243.

Champod, C., Meuwly, D. (2000). The inference of identity in forensic speaker recognition. In *Speech communication,* 31(2-3), 193-203.

De Keijser, J., Elffers, H. (2012). Understanding of forensic expert reports by judges, defense lawyers and forensic professionals. In *Psychology, Crime & Law,* 18(2), 191-207.

Hansen, J. H., Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. In *IEEE Signal processing magazine,* 32(6), 74-99.

Margot, P. (1998). The role of the forensic scientist in an inquisitorial system of justice. In *Science & Justice* 38(2), 71-73,

Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. In *Science & Justice*, 49(4), 298-308.

Morrison, G. S. (2018). The impact in forensic voice comparison of lack of calibration and of mismatched conditions between the known-speaker recording and the relevant-population sample recordings. In *Forensic Science International*, 283, e1-e7.

Morrison, G. S., Enzinger, E. (2016). What should a forensic practitioner's likelihood ratio be? In *Science & Justice*, 56(5), 374-379.

Morrison, G. S., Enzinger, E. (2018). Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. In *Science & Justice*, 58(1), 47-58.

Morrison, G. S., Enzinger, E. (2019). Introduction to forensic voice comparison. In Katz, W. F., Assman, P. F. (Eds.), *The Routledge Handbook of Phonetics*. London: Routledge, 599-634.

Rose, P. (2002). *Forensic speaker identification*. London, New York: Taylor & Francis.

Sjerps, M. & Biesheuvel, D.B. (1999). The interpretation of conventional and 'Bayesian' verbal scales for expressing expert opinion: a small experiment among jurists. In *Forensic Linguistics* 6(2), 214-227.

Blind Grouping: Practical Implementation

Mirjam J.I. de Jonge and Tina Cambier-Langeveld
Speech, Language and Audio Group, Netherlands Forensic Institute, The Hague, Netherlands
m.de.jonge@nfi.nl

Blind grouping has been part of the forensic speaker comparison toolbox at the Netherlands Forensic Institute (NFI) since 2002 as a complement to traditional auditory-acoustic analyses (Cambier-Langeveld et al., 2014). It has been presented at IAFPA several times, starting with Cambier-Langeveld & van der Torre (2004), and with validation results presented in Cambier-Langeveld (2016). Despite its added value according to those who use it (see also Schreuder 2011), it has not been widely adopted by practitioners outside the Netherlands. We hope the idea of blind grouping may fall on more fertile grounds given the current need for validation and assessing performance. In this Special Session on casework practice, we will focus on the practical aspects when implementing this additional analysis method.

A blind grouping task consists of fragments (of 10-20 seconds) selected from a case's disputed material, suspect material, and – where possible, but not necessarily – distractor material (foils), which the blind analyst sorts into clusters. A blind grouping analysis at the NFI results in three dimensions of information: the clusters of audio fragments themselves, and judgements of the level of coherence *within* and distinctiveness *between* each cluster; if the analyst feels unable to make a judgement, 'no judgement' is recorded. Reasoning within a Bayesian framework, we will discuss how to combine outcomes from different analyses, especially when they can't be considered fully independent.

Including a blind component in our analyses serves two partially independent purposes: dealing with context bias, and assessing our performance as expert practitioners. In non-blinded normal conditions, the simple fact that the analyst knows which speech fragments have been assigned to a certain speaker before invites confirmation bias. Having a parallel type of analysis where this knowledge is absent serves as a control to whether the material is sufficiently distinctive to draw conclusions, while allowing the analyst to draw on perceptual Gestalt information that typically doesn't have a place in traditional analyses. Regarding performance assessment, jurisdictions differ in the extent to which they require forensic speech analysts to complete formal training or examinations, but even without formal requirements, there is value in self-assessment. Beyond information on the performance of an individual analyst in a specific case, consistently collecting and analysing blind grouping outcomes also serves to validate the method and contribute to gaps in the scientific understanding of voice perception (see Lavan et al., 2019).

It is an inevitable necessity to have someone available who can prepare the blind grouping setup, as the blind analyst cannot have any prior knowledge of the case or the materials. We will illustrate our practices in information management, selection of materials, and interpretation of outcomes. We propose the blind grouping task is a possible and feasible option for inter-laboratory collaboration between independent practitioners.

References

Cambier-Langeveld, T. (2016). Validation data from the Blind Grouping Task. *Presentation at IAFPA 25th Annual Conference*. York, United Kingdom.

Cambier-Langeveld, T., M. van Rossum and J. Vermeulen (2014). Whose voice is that? Challenges in forensic phonetics. In: J. Caspers, Y. Chen, W. Heeren, J. Pacilly, N.O. Schiller and E. van Zanten (eds), *Above and Beyond the Segments. Experimental linguistics and phonetics*. Amsterdam: John Benjamins Publishing Company, 14-27.

Cambier-Langeveld, T. and E.J. van der Torre (2004). Fighting the Confirmation Bias: Blind Grouping. *Presentation at IAFPA 13th Annual Conference*. Helsinki, Finland.

Lavan, N., Burston, L. F. K., and 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.

Schreuder, M. (2011). Expectancy bias and forensic speaker identification. *Presentation at IAFPA 20th Annual Conference*. Vienna, Austria.

Is pitch equally powerful for the auditory discrimination of low-, mid- and high-pitched voices?

Alice Paver, Kirsty McDougall and Francis Nolan
Phonetics Laboratory, University of Cambridge, United Kingdom
{aep58|kem37|fjn1}@cam.ac.uk

Some voices are more distinctive sounding than others, but the phonetic underpinnings of distinctiveness are not yet well understood. Previous studies have identified that voice distinctiveness is a consideration in earwitness identification (Stevenage et al. 2018, Orchard and Yarmey 1995, Yarmey 1991). McDougall (2013) found a correlation between speaker f0 and perceived voice similarity, and Sørensen (2012) observed that target speakers with more extreme f0 were more easily recognised than those with middling f0. Although f0 is clearly salient to listeners when distinguishing between voices, the extent to which the perceived distinctiveness of a voice is linked to a speaker's pitch remains unclear.

This study explored this question through two online listening experiments. In Experiment 1, three groups of four speakers with low (L), medium (M) and high (H) pitch were chosen from a corpus of 100 male SSBE speakers. For each speaker, two three-second samples containing short sections of semi-spontaneous speech were extracted. 35 participants heard all 78 speaker pairings in a randomized order and rated the similarity of each pair of voices on a 9-point Likert scale. Raw Likert-scale judgements were subjected to multi-dimensional scaling which illustrated that speakers largely fell into three clusters along dimension 1 depending on their pitch group (Figure 1). When observing Likert scores, speakers in the high pitch group were rated as more dissimilar from one another (median=7) than those in the medium (median=3) and low (median=3) pitch groups. A Kruskal-Wallis test confirmed that pitch group had a statistically significant effect on average Likert scale ratings (χ^2=127.71, df=2, $p<0.001$).

In Experiment 2, all stimuli were resynthesised into low-, medium- and high-pitched versions, and rated by 60 different listeners as in Experiment 1. When stimuli were grouped by speakers' original pitch, the H group were again rated most dissimilar from one another, regardless of the pitch manipulation applied to the stimuli. When grouped by manipulated pitch, the same twelve speakers heard at medium-pitch were consistently rated as most dissimilar (Figure 2). A mixed-effects linear regression model confirmed that original pitch ($p<0.001$) and manipulated pitch ($p<0.001$) both had a statistically significant effect on z-scored Likert ratings.

The results of both experiments illustrate that pitch is relevant to listeners in discriminating among all speakers. Experiment 1 showed that the H group were heard as more different from one another than the M or L groups. Concrete conclusions cannot be drawn due to the small sample size, however, results suggest listeners may put more emphasis on other vocal properties to distinguish between similarly-pitched speakers. Furthermore, when all voices were resynthesised to be medium-pitched (Experiment 2), listeners perceived the whole group of 12 speakers as more distinct from one another than when the same voices were heard as high- or low-pitched. This suggests listeners may be better at using properties other than f0 in distinguishing between voices that have a frequently encountered f0, and may habitually rely more exclusively on f0 for voices with rarer f0 values. The implications of these results on earwitness evidence will be discussed.

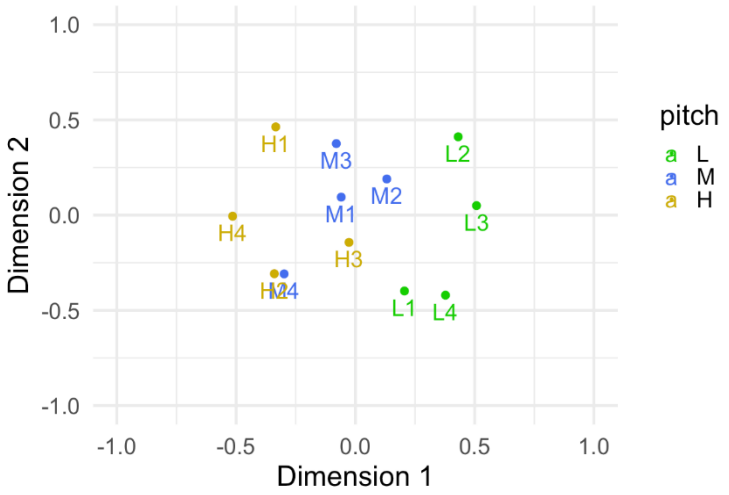


Figure 1. Scatterplot of coordinates for the first two dimensions from the MDS in Experiment 1, with speakers labelled and coloured according to their pitch group.

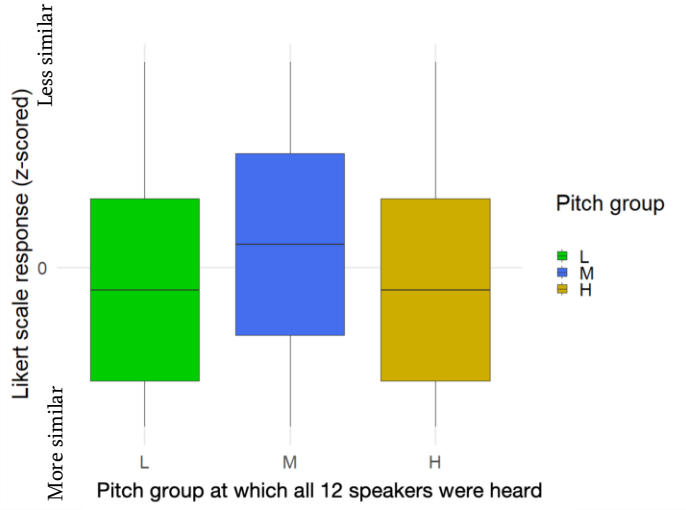


Figure 2. Boxplot of z-scored Likert scale judgements of the similarity of speaker pairings from Experiment 2, grouped by the pitch at which they were heard.

References

McDougall, K. (2013). Earwitness evidence and the question of voice similarity. *British Academy Review, 21*, 18-21. https://www.thebritishacademy.ac.uk/documents/805/BAR21-06-McDougall.pdf

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9*(3), 249–260. https://doi.org/10.1002/acp.235090306

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

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

Own-age bias in voice recognition by younger and older adults

Valeriia Perepelytsia[1] *and Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
valeriia.perepelytsia@uzh.ch, volker.dellwo@uzh.zh

Introduction

In face recognition, different own-group biases have been reported. For example, the so called own-race bias (Meissner & Brigham, 2001) shows that faces from familiar face populations are better recognizable than faces of non-familiar populations. The own-age bias (OAB) was also studied in face recognition, but the results are mixed: some studies report no OAB, while others report it only in one of the age cohorts (see Rhodes & Anastasi, 2012, for a review). We investigated the OAB in voice recognition in two age cohorts: younger and older adults. We hypothesized that if an OAB exists in voice recognition, then older adults should be significantly poorer in recognizing voices of younger adults and vice versa.

Method

Database and speakers. The materials were drawn from the TEVOID corpus (Dellwo et a., 2012), containing read and spontaneous sentence recordings from younger adults (YAs, aged 20–30 years) and older adults (OAs, aged 66–81 years) in Zurich German. 10 younger speakers (5 female) and 10 elderly speakers (5 female) were included in this study.
Stimuli. To create same- and different speaker pairs for a speaker discrimination task, we used read sentences from the TEVOID corpus. The sentences were resampled to 10 kHz, and 800 ms snippets were extracted from each sentence midpoint (Hanning window, 80–5,000 Hz, 40 Hz slope) before stimuli pairs were created. Each pair consisted of two 800 ms snippets extracted from sentence midpoint separated by a 500 ms silent interval. Each listener received a unique subset of 80 stimuli pairs with equal number of younger and older voice pairs, as well as female and male voice pairs.
Listeners. In total, 74 listeners completed the experiment, including 42 YAs and 32 OAs. OA group included normal-hearing participants (pure-tone audiometric (PTA) thresholds ≤ 25 dB, frequencies 0.5–4 kHz) and participants with slight hearing impairment (PTA thresholds ≤ 40 dB, frequencies 0.5–4 kHz) (WHO, 2008).
Procedure. All listeners completed a speaker discrimination (same/different judgement) task, whereby in each trial they heard a pair of stimuli and had to decide whether they stemmed from one speaker or from two different speakers. The experiment consisted of 80 trials and was administered via the Gorilla experiment builder (Anwyl-Irvine et al., 2020). All listeners performed the test at the Linguistic Research Infrastructure laboratory at the University of Zurich to assure equal experimental conditions for all participants.

Results

We analyzed listeners' performance with signal detection theory (Stanislaw & Todorov, 1999). Listeners' sensitivity was quantified with d', and listeners' bias was measured with criterion location *c*. The two-way interaction between listener age and speaker age on d' was not significant, suggesting that, contrary to our hypothesis, young and elderly listeners did not reveal different recognition behavior for younger and older speakers. However, the main effect of listener's age on d' was significant ($F(1, 72) = 13.18$, $p = 0.0005$), whereby YA listeners performed significantly better compared to the OA listeners in the voice discrimination test (Figure 1a). This was expected given that YAs also outperform OAs in face recognition (Rhodes & Anastasi, 2012). Speaker age was not significant, suggesting that both young and elderly voices were discriminated equally well. However, in terms of listeners' bias, it was found that young voices sounded significantly more similar for all listeners ($F(1, 72) = 39.4$, $p < 0.0001$) compared to elderly voices (Figure 1b).

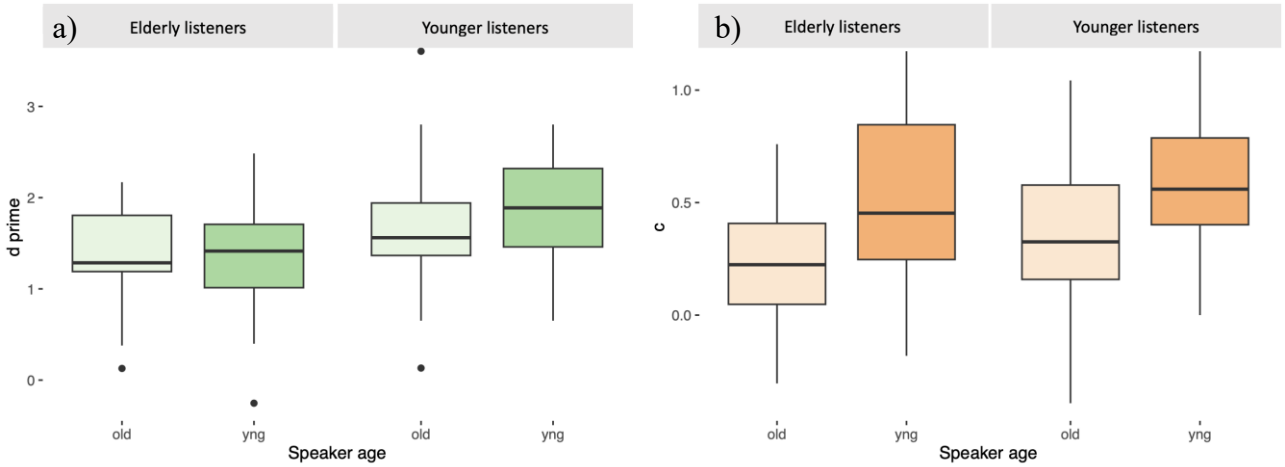


Figure 1a. Boxplots showing median, range, and interquartile range for d' values per listener group ("elderly" = older adult listeners, younger = young adult listeners). **1b.** Boxplots showing median, range, and interquartile range for *c* values per listener group ("old" = elderly speakers, "yng" = young speakers).

Results suggest that age-related changes to the voice make elderly voices more discriminable, but further research is needed to understand this effect fully. For example, which specific changes lead to more discriminability of the older voices and how hearing loss influences voice recognition performance in elderly listeners. The results are relevant for the general understanding of voice perception in different age groups, as well as factors contributing to the nonexpert listeners' performance in forensic cases

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. https://doi.org/10.3758/s13428-019-01237-x

Dellwo, V., Leemann, A., & Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. Interspeech Conference Proceedings, 1–4. https://doi.org/10.5167/UZH-68554

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law*, 7(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Rhodes, M. G., & Anastasi, J. S. (2012). The own-age bias in face recognition: A meta-analytic and theoretical review. *Psychological Bulletin*, 138(1), 146–174. https://doi.org/10.1037/a0025750

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, 31(1), 137–149. https://doi.org/10.3758/BF03207704

World Health Organization. (2008). World Health Organisation Grades of Hearing Impairment. https://ec.europa.eu/health/scientific_committees/opinions_layman/en/hearing-loss-personal-music-player-mp3/figtableboxes/table-4.htm

The effect of other forensic evidence and expert opinions on lay listener perceptions in voice comparison tasks

Vince Hughes[1], Carmen Llamas[1] and Thomas Kettig[2]
[1]University of York, York, UK
[2]York University, Toronto, Canada
{vince.hughes|carmen.llamas|thomas.kettig}@york.ac.uk

A crucial step in forensic voice comparison is the presentation of evidence to an end-user. In many jurisdictions, a jury of lay people are responsible for evaluating voice evidence presented by an expert. Work on cognitive bias in forensics has largely focused on experts rather than jurors (e.g. Dror 2011, Rhodes 2016). However, the nature of other evidence in the case or the criminal context itself could bias jurors towards perceiving the similarity of voices in different ways. While previous work has investigated how lay people understand expert forensic conclusions (Martire et al. 2014), little is known about how lay listeners' interpretations of an expert opinion can influence their evaluation of voices. In this paper, we assess the effects of other forensic evidence and expert opinions on lay listener sameness judgements in a forensic voice comparison task.

We created a bespoke, game-like tool which immerses participants in a jury context (Hughes & Llamas 2021-23). 1505 UK participants were recruited via Prolific. Participants were each presented with 24 same- and different-speaker voice pairs (of 120 pairs in total) across three levels. For each comparison, they provided similarity and sameness judgements on a 0 to 100 scale. The first level used a Qualtrics-like interface, while the second level introduced graphics which placed participants on a jury. In the third level, participants were presented with either additional forensic evidence (DNA, fingerprint, footprint) or the conclusion of a forensic expert in the form of a verbal or numerical likelihood ratio of low (LR = 10), medium (LR = 1000), or high (LR = 100000) magnitude. In all cases, the opinion of the expert was consistent with the ground truth but the magnitude was randomised. For the purposes of this study, we focus on participants' sameness ratings within and between levels.

We found no significant differences in sameness ratings provided by participants in the initial 'jury' level compared with the 'other evidence' level – perhaps unsurprising, given that participants were only shown a visual representation of the other evidence, but not told whether it provided support for the prosecution or defence. Effects in the 'expert evidence' level depended on whether participants encountered verbally or numerically expressed conclusions. Responses to verbal conclusions followed expected patterns: sameness ratings significantly differed from the 'jury' level and got stronger (in the correct direction) as the magnitude of evidence increased from low to high – except for low-magnitude evidence in same-speaker pairs, which exhibited no difference from the jury level (cf. the 'weak evidence effect' found by Martire et al. 2014). Sameness responses to numerical conclusions, however, only differed significantly from the jury level in different-speaker pairs at medium magnitude, with no other significant differences; in contrast with the findings of Martire et al. (2014), participants appeared to struggle with both the smallest and largest numerical LRs, with the high-magnitude numerical conclusions producing sameness ratings that were comparable with low magnitude conclusions. We discuss the implications of these findings for forensic voice comparison and the presentation of evidence to courts.

References

Dror, I.E. (2011) The paradox of human expertise: Why experts can get it wrong. In N. Kapur (Ed.) *The Paradoxical Brain*. Cambridge: Cambridge University Press. pp. 177-188

Hughes, V. And Llamas, C. (2021-23) *Novel Methods for Assessing Speaker Recognition Performance*. AHRC-funded project, AH/T012978/1.

Martire, K. A., Kemp, R. I., Sayle, M. and Newell, B. R. (2014) On the interpretation of likelihood ratios in forensic science evidence: presentation formats and the weak evidence effect. *Forensic Science International* 240: 61-68.

Rhodes, R. (2016) Cognitive bias in forensic speech science: risks and proposed safeguards. IAFPA conference, York, UK.

Exploring the relationship between acoustic-phonetic and perceived voice similarity

Leah Bradshaw[1], Eleanor Chodrof[1] and Volker Dellwo[1]
[1]*Department of Computational Linguistics, University of Zurich, Switzerland*
leah.bradshaw@uzh.ch

Speaker and linguistic information show a bidirectional relationship on their influence of spoken language processing. For instance, findings show speech perception accuracy improves when speakers are familiar to the listeners (e.g., Levi et al. 2011; Souza, et al., 2013). Further, processing speaker information improves when speakers are native speakers of the corresponding language (the Language Familiarity Effect e.g., Goldstein, et al., 1981; Hollein et al., 1982; Thompson, 1987). Recent studies have demonstrated that linguistic variability can strongly influence speaker discrimination performance. Narayan et al. (2017), later replicated by Quinto et al. (2020), explored the effects of varying levels of phonologically and semantic relatedness on speaker discrimination performance. Both studies observed improved accuracy in same-speaker trials if the words produced by the speaker were phonologically or lexically related (e.g., producing a phonological rhyme, day-bay, or a lexical compound, day-dream). Comparatively, no significant effects were found for different-speaker trials. The effect observed in same-speaker trials was argued to occur due to speakers sounding more similar when producing phonologically related words. However, this relationship between voice similarity and phonological similarity is not explicitly explored.

The influence of linguistic variability on speaker identity processing is a concern for forensic phonetics. Namely, samples used in forensic speaker comparisons, or voice line-up tasks are rarely perfectly matched for linguistic content, motivating further exploration into the effect of acoustic-phonetic similarity on voice judgements. Here, we ran an experiment in which we explored the effect of varying degrees of acoustic-phonetic similarity on voice similarity judgements. Based on previous findings, we hypothesise that increased acoustic-phonetic word similarity in same-speaker trials will lead to higher perceived voice similarity. Comparatively, for different-speaker trials, it is not expected that acoustic-phonetic similarity will play a large role.

Methods

Twenty-five native Swiss-German listeners (10 M) conducted a voice similarity judgement task. Participants completed pairwise comparisons of 8 female voices producing either the same word (*Gastwirt-Gastwirt*; same-word condition), two different words with some acoustic-phonetic overlap (*Köchin-Kinder*; similar-word condition) and two different words with no acoustic-phonetic overlap (*Vögel-Schneider*; different-word condition). Voice similarity was rated on a scale from 1 (Very Dissimilar) to 6 (Very Similar). Responses were z-scored to account for bias in scale usage and analysed using a generalised linear mixed-effects model with a beta distribution. *Response* was predicted from the independent variables *Condition* (same/similar/different words), *Speaker Match* (same/different speaker), *Sex* (male/female), and the interaction between *Condition~Speaker Match*, as well as a by-*participant* random intercept.

Results

Findings showed a nearly mirrored effect across the same, similar, and different word conditions between same and different speaker trials (Figure 1). Among same speaker trials, the similarity rating increased for same words and decreased for similar and different words; whereas, among different speaker trials, the similarity rating decreased for same words and increased for similar and different words. Model outputs reflected this trend within the data, showing the interactions between *Condition* and *Speaker Match* to be significant ($p < 0.001$). Post-hoc analyses confirmed the directionality and significance of the relationship between the same-word condition, compared to the other two conditions within each *Speaker Match* setting. In comparison with previous findings, within each *Speaker Match* setting, acoustic-phonetic similarity in the different word conditions did not have a

significant effect. Likely this finding is a result of substantially less acoustic-phonetic overlap in our similar word condition, compared with the phonological rhyme condition in the previous studies. It is plausible that the degree of overlap and/or the types of overlapping linguistic information, e.g., vowel vs. plosive, are contributing to a more complex relationship than initially understood. Further research is necessary to explore the exact effects of this on naïve listener similarity judgements.

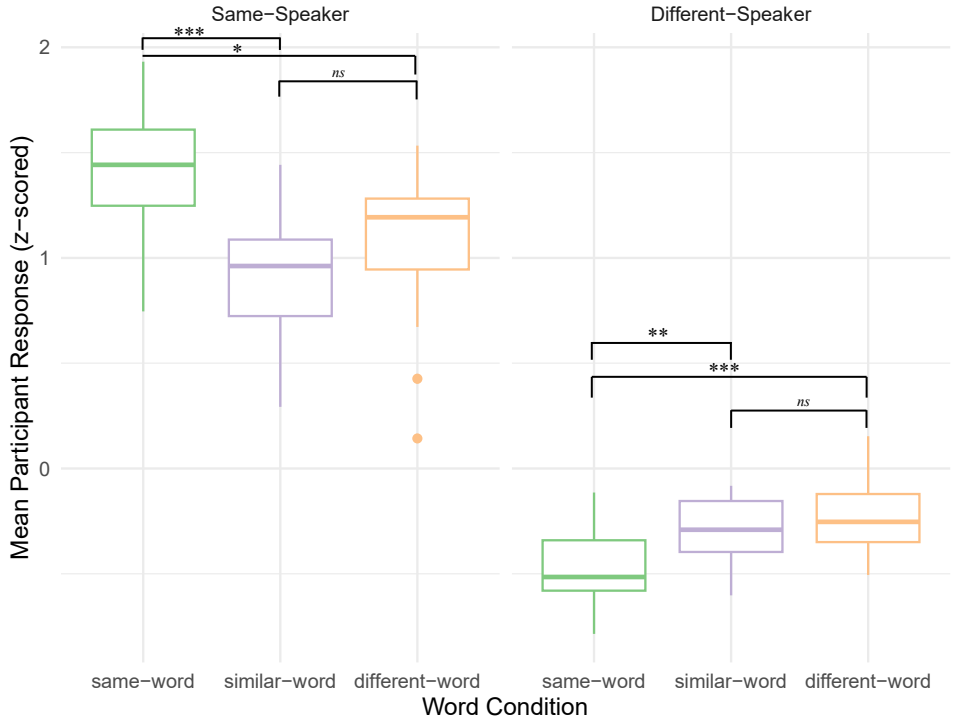


Figure 1. Boxplot of scaled mean participant responses by condition for same and different speaker comparisons. Asterisks represent significant differences between conditions at the following alpha thresholds: * = 0.01, ** = 0.01, *** = 0.001. Higher scores indicate greater perceived voice similarity.

References

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. Bulletin of the Psychonomic Society, 17(5), 217–220.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress, and disguise speaking conditions. Journal of Phonetics, 10, 139–148.

Levi, S. V., Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: Whose familiar voices are more intelligible? Journal of the Acoustical Society of America, 130 (6), 4053–4062.

Quinto, A., Abu El Adas, S., & Levi, S. V. (2020). Re-Examining the Effect of Top-Down Linguistic Information on Speaker-Voice Discrimination. *Cognitive Science*, *44*(10), e12902.

Narayan, C. R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. Cognitive Science, 41(5), 1361–1376.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. Journal of the American Academy of Audiology, 24(8), 689–700.

Thompson, C. P. (1987). A language effect in voice identification. Applied Cognitive Psychology, 1(2), 121131.

Forensic transcription: a survey of expert transcription practices in Europe and North America

Lauren Harrington[1] *and Richard Rhodes*[1,2]
[1]*Department of Language and Linguistic Science, University of York, York, UK*
[2]*The Forensic Voice Centre, York, UK*
{lauren.harrington/richard.rhodes}@york.ac.uk

This talk presents the results of a survey on the transcription practices employed by 28 forensic practitioners from multiple jurisdictions across Europe and North America. We will present how our respondents produce forensic transcripts, and what they consider are relevant bias factors and mitigation strategies.

Forensic transcription is an under-explored area within forensic speech science. Most research on the topic has been carried out by Fraser (e.g. 2022) concerning priming and the legal context for forensic transcripts to appear in court. In recent years, surveys on cognitive bias within forensic speech science (Rhodes, 2016) and on forensic translation (Lai, 2023) have been published, but little work has addressed expert practitioners' methods for forensic transcription. This leads to our three main motivations for this survey and presentation. Firstly, this work provides a forum for practitioners to discuss their methods and see how their own approach fits within the array of practices used worldwide. Secondly, it provides information on the methods commonly used by practitioners which can be extremely helpful for those researching forensic transcription. Thirdly, it highlights key differences in practices or methods that helps to identify areas for further research.

Respondents

There were 28 respondents to the survey; it was targeted to IAFPA members, ENFSI speech and audio group members, relevant mailing lists, and to forensic practitioners directly. The countries or jurisdictions in which respondents primarily work, and the number of respondents from each, are as follows: United Kingdom (8), the Netherlands (6), Switzerland (3), United States of America (3), Germany (2), Canada (1), Croatia (1), Italy (1), Romania (1) and Ukraine (1). 13 respondents were affiliated with government laboratories, 3 were affiliated with an independent facility with multiple staff members, 5 were individual private practitioners, 2 were affiliated with a private provider or a research institute, and the rest reported affiliation with a combination of the above.

Preliminary results

50% of respondents reported that they carry out forensic transcription frequently or very frequently (rather than rarely or occasionally). 12 of the 28 respondents reported that 2 transcribers typically work on a transcript, while 8 reported 3 or more transcribers. The remaining 8 respondents reported that a single expert works on the transcript (these tended to be sole private or government-lab practitioners). Of the 20 respondents who have multiple transcribers working on a transcript, 10 typically work in parallel (i.e. analysts work independently on separate transcripts), 7 typically work sequentially (i.e. analysts building on a previous version of the transcript), and 3 typically use a combination of these methods. Those who work alone reported that they either produce only one draft or develop previous versions of their own transcript.

The number of drafts typically made varied, but over 85% respondents reported more than one draft: 12 respondents reported 2-3 drafts, 10 respondents reported 4-6 drafts and 2 respondents reported 8 or more drafts. It is probably safe to assume that the number of drafts varies according to the factors and recordings in the case, and the type of drafting process used. Almost all respondents represent different levels of confidence in the transcript (normally 2 or 3 levels). No experts used speech-to-text or automatic speech recognition systems in the transcription drafting process (one used a system to detect speech vs silence).

Cognitive bias

The survey also focused on awareness of cognitive bias. 24 of 28 respondents considered that cognitive bias plays a significant role in transcription. We asked respondents how influential they believed the following factors can be on the perception and transcription of speech in forensic recordings. The factors are ranked from most to least influential according to average rating, with mean and mode scores at the end:

Factor	Mean	Mode
Poor audio quality (e.g. background noise)	5.29	6
Experience with the speaker's accent/dialect	4.82	5
Information from instructing party (e.g. incriminating evidence about speakers)	4.36	5
Information about a suspect (e.g. criminal record)	3.46	5
Content of recordings (e.g. highly emotive speech)	3.36	3
Information about the type of offence	3.36	3
Expectations of instructing party	3.32	1

Table 1. Factors that may influence transcription, ordered from most to least influential according to respondents' ratings. Factors were rated on a scale of 1 to 6, where 1 represents no effect at all on transcription and 6 represents a great effect on transcription.

20 out of 28 respondents reported having some form of protocol or procedures in place to protect against the effects of cognitive bias, which shows a shift from 2016 when only 50% of survey respondents had bias mitigation policies in place (Rhodes, 2016). These included appointing a case/information manager who liaises with the client and provides contextual information to the transcribers when appropriate, or withholding of all contextual information. 22 respondents typically receive transcripts from their instructing party; half reported that they do not refer to these at any point during the process, and the other half reported that these are referred to only after blind draft(s) have been produced.

The results show a broad consensus in some areas, but they also raise interesting questions we will discuss in the presentation, including: should transcribers refer to relevant case information and existing transcripts, or should evidential transcripts be produced maximally independently? Should transcript drafts by different analysts be produced in series or in parallel, or when is each approach best used? Are transcripts better when produced by multiple analysts (as suggested in Tschäpe and Wagner (2012)); and how many analysts is optimal? Could automatic speech transcription play a role in forensic processes?

These questions warrant further research exploration to determine what impact these factors have on the quality of forensic transcripts.

References

Fraser, H. (2022) A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. Front. Commun. 7:898410.

Lai, M. (2023) Transcribing and translating forensic speech evidence containing foreign languages—An Australian perspective. Front. Commun. 8:1096639.

Rhodes, R. (2016, July 24-27) *Cognitive bias in forensic speech science* [Conference presentation]. 25th Annual IAFPA Conference, York, UK.

Tschäpe, N. & Wagner, I. (2012, August 5-9) *Analysis of disputed utterances: A proficiency test* [Conference presentation]. 21st Annual IAFPA Conference, Santander, Spain.

Towards accountable evidence-based methods for producing reliable transcripts of indistinct forensic audio

Helen Fraser[1]*, Debbie Loakes*[1]*, Ute Knoch*[2] *and Lauren Harrington*[3]
[1]*Research Hub for Language in Forensic Evidence, University of Melbourne*
{helen.fraser|dloakes}@unimelb.edu.au
[2]*Language Testing Research Centre, University of Melbourne.*
uknoch@unimelb.edu.au
[3]*Department of Language and Linguistic Science, University of York.*
lauren.harrington@york.ac.uk

Covert recordings provide powerful evidence in criminal trials, but are often of extremely poor quality. Many jurisdictions allow the court to be assisted by a police transcript, but these can be unreliable (French and Fraser, 2018). The law has developed safeguards intended to ensure that triers of fact are not misled by inaccurate transcripts. However, these safeguards are ineffective, as they rely on lawyers and judges checking the transcript against the audio (Fraser and Kinoshita, 2021). Even if experts are consulted, responsibility for evaluating their findings typically rests with lawyers and judges. Multiple cases of actual and potential injustice have been documented. In 2017, Australian linguists raised a Call to Action, asking the judiciary to review and reform the legal handling of indistinct forensic audio.

The present paper starts with a brief update on the progress of the Call to Action, which gives reason to hope that police transcripts will eventually be disallowed. This makes it important for linguists to be able to provide reliable transcripts of indistinct forensic audio via accountable, evidence-based methods – that do not start from a police version (though investigators' opinions may well be sought at an appropriate point in the process).

The paper then shares results of an experiment which forms part of an investigation into how transcription methods can best be evaluated, drawing on insights from language testing research (Knoch and Macqueen 2020).

Forty participants transcribed a three-minute sample of forensic-like audio, without contextual information. Each transcript was divided into intonation phrases (IPs) and each IP was scored against the reference transcript, with a global rating, and three analysis ratings, showing how much was *misinterpreted*, *missing* or *added*.

Overall scores were relatively low, and highly variable – e.g., exact matches to the reference transcript varied from 47% to 12%. A wide range of demographic data were collected, but the only one to show a significant correlation with score was language background, with L2 speakers of English, on average, scoring lower than L1 speakers – even though all L2 speakers were highly proficient in English (cf. de Boer, 2016). Other factors that might have been expected to correlate with scores (e.g., a background in phonetics or forensic speech science) did not. Participants' confidence was not a reliable indication of performance.

These results, following those of Tschäpe & Wagner (2012) and Love & Wright (2020), confirm that it is unrealistic to expect individual transcribers with no contextual information to produce demonstrably reliable transcripts (Fraser 2022a). Rather it is necessary for accredited transcribers to follow an evidence-based method, designed and managed by experts – as is done for other responsible forensic sciences. Further, it is important to ensure that transcripts are used in trials in a manner that minimises opportunities for the court to be misled about the content of indistinct forensic audio (see Haworth, 2018; Fraser, 2022b).

The paper concludes by outlining a proposed method for producing and evaluating transcripts, seeking discussion and input from IAFPA members (see Harrington and Rhodes, in prep).

References

de Boer, Meike. *Expectancy bias in forensic speech transcriptions: Can a forensic context change what is heard in an audio fragment?* Masters Thesis, Maastricht

Fraser, H. (2022a). Forensic transcription: Legal and scientific perspectives. In C. Bernardasci, et al (Eds.), *Speaker Individuality in Phonetics and Speech Sciences: Speech Technology and Forensic Applications* (pp. 19–32). Milano: Officinaventuno.

Fraser, H. (2022b). A framework for deciding how to create and evaluate transcripts for forensic and other purposes. *Frontiers in Communication*.

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.

Fraser, H., & Loakes, D. (2020). Acoustic injustice: The experience of listening to indistinct covert recordings presented as evidence in court. *Law Text Culture*, *24*, 405–429.

French, P., & Fraser, H. (2018). Why “ad hoc experts” should not provide transcripts of indistinct forensic audio, and a proposal for a better approach. *Criminal Law Journal*, *42*, 298–302.

Harrington, L. & Rhodes, R. (in prep) Forensic transcription: A survey of expert transcription practices in Europe and North America.

Haworth, K. (2018). Tapes, transcripts and trials. *International Journal of Evidence and Proof*, *22*(4), 428–450.

Knoch, U., and Macqueen, S. (2020). *Assessing English for Professional Purposes*. Routledge.

Love, R., & Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. *Frontiers in Communication*: 6:797448. doi: 10.3389/fcomm.2021.797448

Marzi, T. et al. (2021). Mapping the featural and holistic face processing of bad and good face recognizers. *Behavioural Sciences*, *11*(5), 75.

Tschäpe, N., & Wagner, I. (2012). Analysis of disputed utterances: A proficiency test. *IAFPA*.

Clustering a large number of unknown voices

Hanna Ruch[1, 3]*, Andrea Fröhlich*[1, 2, 4] *and Sarah Lim*[1]
[1]*Zurich Forensic Science Institute, Zurich, Switzerland*
phonetik@for-zh.ch
[2]*Department of Computational Linguistics, University of Zurich, Switzerland*
[3]*University Research Priority Program Language and Space, University of Zurich, Switzerland*
[4]*Applied Face Cognition Lab, University of Lausanne, Switzerland*

Introduction

In a traditional forensic speaker comparison case (FSC), two recordings –a questioned recording and a reference recording– are compared to find out if they are spoken by the same person. However, in recent years, such one-to-one comparisons have become scarcer as an increasing number of cases involve several questioned and comparison recordings. These cases are challenging traditional auditory-phonetic and acoustic FSC methods, which can be very time-consuming. One of these scenarios could be a case, for instance, which contains a large collection of recordings with an unknown number of potential speakers, in which investigators are interested in finding out how many speakers might be present within this collection. One possible approach to dealing with this challenge is the implementation of automatic speaker comparison (ASR) systems to create (dis)similarity matrices, which are then further statistically post-processed to approximate the right amount of potential speakers.
Currently, most research on voice similarity or clustering of unknown speakers in ASR contexts has focused on existing corpora or databases (e. g. Lukic et al. 2017, Gerlach et al. 2020) and has rarely investigated casework data. We have therefore started exploring the implementation of ASR systems alongside statistical clustering methods to obtain clusters which have to be auditorily post-processed by phoneticians because ASR systems' performance is highly dependent on the quality and duration of the audio recordings (Kelly et al., 2019). Initial tests (see Ruch et al. 2021) suggest a potential new approach to tackle these big data challenges.

Case data

In the present paper, we report on a case with 57 questioned recordings by an unknown number of offenders. The recordings are all spoken in Standard German but differ in length (from 7 to 68 seconds net speech), communication setting, channel, and acoustic quality. The question to be addressed is: *How many individual speakers are present in this group of audio files?*

Method

These case-specific audio files are highly variable and very challenging to process automatically. We have therefore decided to apply a mixed approach by combining auditory, automatic, and statistical clustering methods. For validation purposes, known voices from police officers were also included to assess the effect of net speech, acoustic quality, and session variability on the ASR system. VOCALISE is used (version 2021A-xVector with MFCCs as features, Kelly et al. 2019), and (dis)similarity matrices are then used as the dependent variable in several statistical methods (e.g., multi-dimensional scaling, hierarchical clustering etc.) using *R* (R Development Core Team 2022). The analysis is ongoing, but a clustered heat map (Figure 1), combined with auditory post-processing on a feature level, appears to be the most appropriate approach to cluster the questioned voices.

Our presentation will further report the pre-processing procedures necessary for such cases, explain and discuss methods and results, and describe methodological challenges. The presentation will be completed by desiderata for future research (and validation work) in voice clustering.

First Results

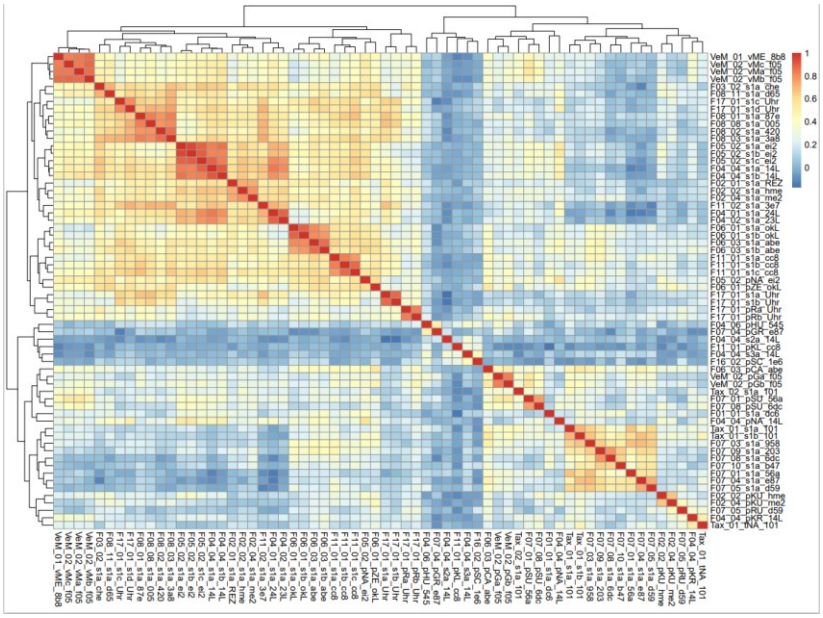


Figure 1. The clustered heatmap visualising the clustering results shows different groups of potential speakers. The colours indicate how similar the compared recordings are. Starting from blue (dissimilar) to red (similar).

References

Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. Speech Communication, 124, 85-95.

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

Kelly, F., Fröhlich, A., Dellwo, V., Forth, O., Kent, S., & Alexander, A. (2019). Evaluation of VOCALISE under conditions reflecting those of an actual forensic voice comparison case (forensic_eval_01). *Speech Communication*, *112*, 30-36.

Lukic, Y., Vogt, C., Dürr, O., and Stadelmann, T. (2017). Learning embeddings for speaker clustering based on voice equality. IEEE International Workshop on Machine Learning for Signal Processing, September 25-28 2017, Tokyo, Japan. https://digitalcollection.zhaw.ch/bitstream/11475/7088/1/MLSP_2017.pdf

R Development Core Team (2021). *R*: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ruch, H., Fröhlich, A., and Lory, M. (2021). Clustering of unknown voices. Paper presented at the XVII AISV Conference, February 2-4 2021, Zurich, Switzerland.

An investigation of the effect of warning strength on voice parade performance

Kirsty McDougall[1], *Nikolas Pautz*[2], *Peter Goodwin*[2], *Francis Nolan*[1], *Katrin Müller-Johnson*[3], *Alice Paver*[1] *and Harriet M.J. Smith*[2]

[1]*University of Cambridge, UK*

[1]{kem37|fjn1|aep58}@cam.ac.uk

[2]*Nottingham Trent University, UK*

[2]{nikolas.pautz|peter.goodwin|harriet.smith02}@ntu.ac.uk

[3]*University of Oxford, UK*

[3]katrin.mueller-johnson@crim.ox.ac.uk

Earwitness identification evidence collected through a voice parade can play a crucial role in a legal case. In studies of earwitness behaviour, listeners are typically exposed to the voice of a target speaker then later asked whether they are able to recognise that speaker's voice in a line-up of voice samples. Alongside voice parades including the target speaker ('target-present'), researchers include among their experimental conditions parades in which the target speaker is not present ('target-absent') to simulate the situation in which an innocent suspect has been apprehended. Listeners' accuracy rates in such experiments are lower in target-absent than target-present parades (e.g. Pautz et al. 2023, Smith et al. 2022), with listeners often choosing an incorrect foil rather than rejecting the parade when the target is not heard. Calderwood et al. (2019) hypothesises whether (for children in particular, who show greater false alarm rates than adults) listeners may feel a social pressure to choose a speaker rather than reject the line-up. The present study explores this further for adult listeners by manipulating the levels of warning participants receive pre-parade about the consequences of an incorrect identification.

Nine-person target-present and target-absent voice parades were prepared for six target speakers of English: SSBE (3), York, Bradford and Wakefield. Eight foil demographically-matched speakers were chosen per target using multidimensional scaling of listener judgements (see McDougall *et al.* 2015). Using mock police interview material, parades were constructed following UK 2003 Home Office guidelines (see de Jong-Lendle *et al.* 2015), except voice samples were 15s long instead of the previously prescribed 60s (see Pautz *et al.* 2023).

272 participants recruited via Prolific were randomly allocated to a target-present or target-absent parade with one of three warnings (Table 1):

Standard	Remember that the voice you heard at the beginning of the experiment may or may not be present.
Strong	Remember that the perpetrator may or may not be present. Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction.
Very strong	Remember that the perpetrator may or may not be present. Please consider your response carefully. In a real case, selecting someone from the lineup when the perpetrator is not present could lead to a wrongful conviction. Voice recognition can be very difficult. Only make a positive identification if you are very sure.

Table 1. Warnings given to participants prior to undertaking the voice parade.

Participants heard a target voice for 60s then undertook a cognitively demanding five-minute filler task, before receiving their warning. Participants heard all nine voices in the parade, then received their warning again with additional instructions to choose 'none' if they thought the target speaker was not present.

Figure 1 shows mean results for each warning type in target-absent and target-present parades. Target-absent accuracy did not meaningfully differ between the standard and strong warning

conditions, but there was a meaningful increase in accuracy between the standard and very strong warning conditions. However, signal detection theory analysis suggests that this happened because increasingly strong warnings led earwitnesses to adopt a more conservative decision-making criterion. That is, they were less likely to make a positive identification, regardless of whether the target was present or absent. Signal sensitivity, (i.e., the ability to detect a target voice from foils) was meaningfully (BF > 100) above chance level for participants in the standard warning conditions, but not in other warning conditions (BF < 3). Response criterion showed that listeners in the standard warning condition were meaningfully (BF = 99) more likely to respond 'present' compared to a more neutral response criterion in the stronger warnings (BF < 3). Overall it appears that stronger warnings reduce false alarms, but also reduce the overall ability of participants to identify the target by apparently making them too cautious. Therefore, although the results suggest that strengthening the warning improves target-absent accuracy, in order not to compromise target-present accuracy, it is recommended that the standard warning is maintained.

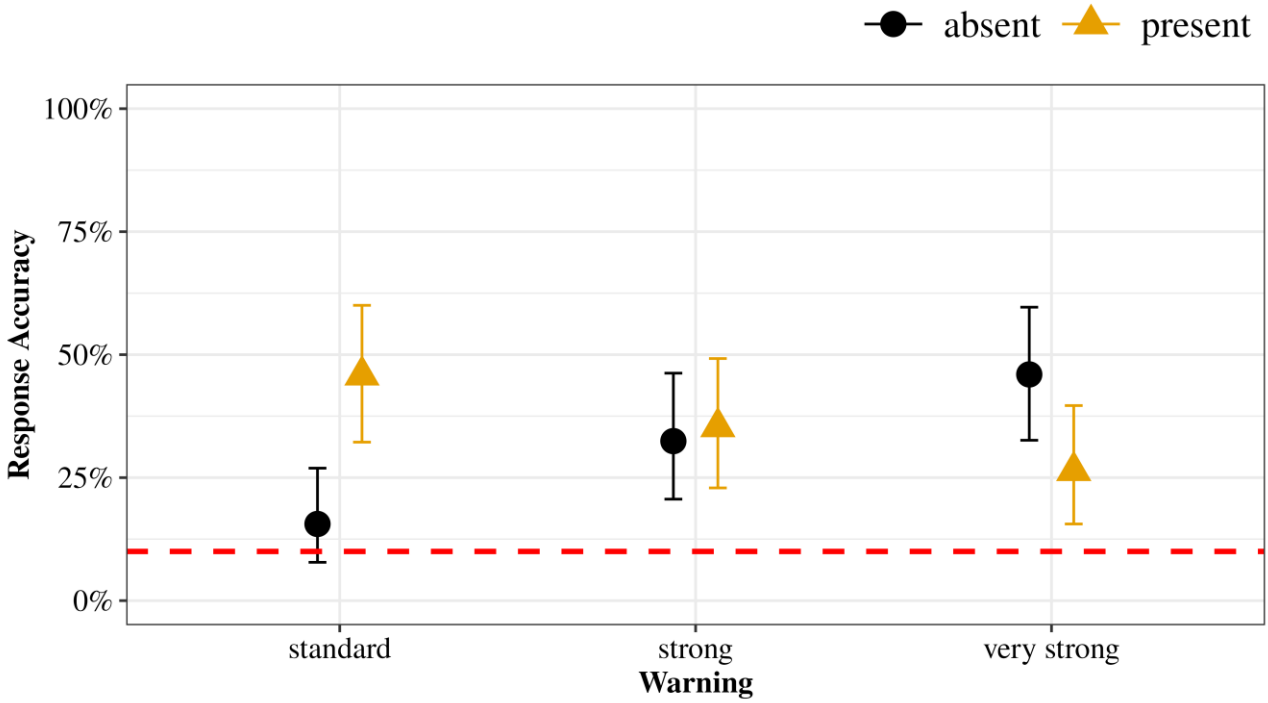


Figure 1. Voice identification accuracy for listeners given the standard, strong or very strong warnings in target-absent and target-present conditions. Error bars indicate 95% confidence intervals for the condition means. Chance level is 10%, shown by the red dotted line.

References

Calderwood, L., McKay, D. R., & Stevenage, S. V. (2019). Children's identification of unfamiliar voices on both target-present and target-absent lineups. *Psychology, Crime & Law*, 25(9), 896-910. doi:10.1080/1068316X.2019.1597090

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). 'Voice lineups: a practical guide.' In The Scottish Consortium for ICPhS 2015 (ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, 10-14 August 2015, Glasgow. Paper number 0598. 1-5.

Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A. & Smith, H.M.J. (2023). Identifying unfamiliar voices: the influence of sample duration and parade size. *Quarterly Journal of Experimental Psychology*. 17470218231155738. doi:10.1177/17470218231155738

Smith, H. M. J., Roeser, J., Pautz, N., Davis, J. P., Robson, J., Wright, D., Braber, N., & Stacey, P. C. (2023). Evaluating earwitness identification procedures: adapting pre-parade instructions and parade procedure. *Memory*, 31(1), 147-161. doi:10.1080/09658211.2022.2129065

McDougall, K., Nolan, F. & Hudson, T. (2015). Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica* 72: 257-272.

Acoustic, perceptual and ASR analysis of targeted voice manipulations

Radek Skarnitzl[1], *Tomáš Nechanský*[1] *and Alžběta Houzar*[1]
[1]*Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czech Republic*
{radek.skarnitzl|tomas.nechansky|alzbeta.houzar}@ff.cuni.cz

The human voice is a remarkably flexible tool which we make use of in everyday communication. The flexibility of our voice is enabled by the plasticity of our speech production mechanism and its many degrees of freedom in combining various gradient modifications (Nolan, 2012). The plasticity of the human voice becomes perhaps most obvious in disguised voices. While voice disguise in forensic casework seems to involve relatively simple changes (Figueiredo & Britto, 1996; Masthoff, 1996), studies outside of the forensic casework context have revealed quite sophisticated voice disguise strategies (Růžičková & Skarnitzl, 2017; Smith et al., 2019). It is such modifications of the voice that we are focusing on in this study.

We analyzed 15 speakers of Czech (10 male, 5 female), all phonetically trained, who were asked to read the Czech version of the Rainbow Passage first in their habitual voice, and then in fifteen different modified voices. These included articulatory changes (e.g., spread and rounded lips, open and closed jaw, palatalization and pharyngealization, nasalization and denasalization), phonatory changes (pressed, breathy, whispery, and creaky voice), as well as combined modifications (spread lips with breathy voice, rounded lips with whispery voice, open jaw with creaky voice).

Acoustic analyses focused on shifts, with respect to the speakers' habitual voice, in fundamental frequency (f_0), formants ($F1$–$F3$), harmonicity and smoothed cepstral peak prominence (CPPS), as well as several measures of spectral slope. Most significant shifts were revealed in the phonatory and combined modifications (see examples in Figures 1 and 2). On the other hand, $F3$ and f_0 characteristics turned out to be most stable across the individual voice disguise strategies; however, the stability of *F3* values appears to be speaker-specific (*cf.* Disner & Benítez, 2018).

In the perceptual analysis, 120 Czech respondents were asked to assess the similarity of articulatory and phonatory modifications from the same speaker's habitual voice using a visual analogue scale. We used an online experiment on the Gorilla platform. Palatalization and pharyngealization as articulatory modifications, and creaky and pressed voice as phonatory ones were perceived as most different from the respective habitual voices (mean score of 70 or more on a 0–100 scale), while lip spreading was perceived as most similar (mean score of 25); nevertheless, the results also point to considerable variability in the ratings of voice similarity.

Finally, we examined the performance of Phonexia Speaker Identification (SID4 XL5) system, a DNN-based automatic speaker recognition (ASR) system *vis-à-vis* the manipulations. The overall comparison of every recording to all others, which contained nearly 28,000 comparisons, resulted in a rather low equal error rate (EER) of 6.26%, with a log-likelihood-ratio cost (C_{llr}) of 0.25. In the same-speaker comparison of each speakers' habitual voice with the 15 changed versions, the performance of the ASR system on the same-speaker recordings was also better than we expected, with log-likelihood ratios (LLRs) ranging from 3.47 for the combination of creaky voice with open jaw articulation and 9.56 for lip-spreading.

Overall, pressed phonation turned out as the most effective voice setting for voice disguise, as it yielded the lowest LLR values in the ASR analysis, the highest perceived difference from the habitual voice, and the most significant shifts of acoustic parameters.

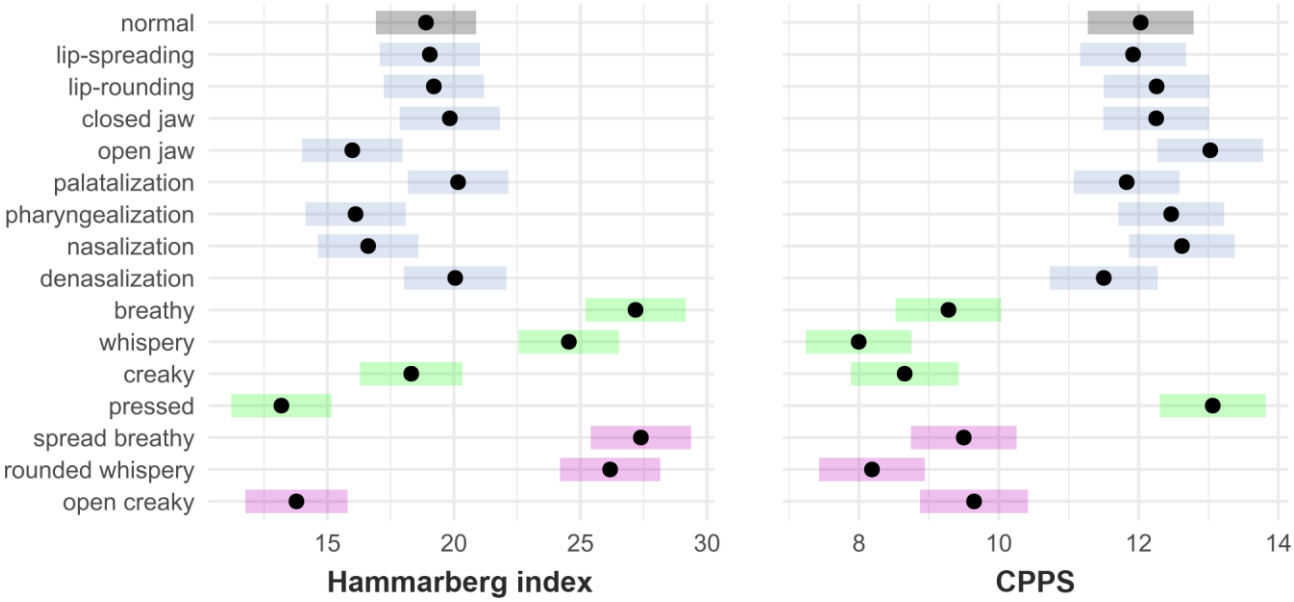


Figure 1. Shifts of the Hammarberg index with a speaker-dependent pivot (left) and smoothed cepstral peak prominence (CPPS, right) in individual targeted voice manipulations.

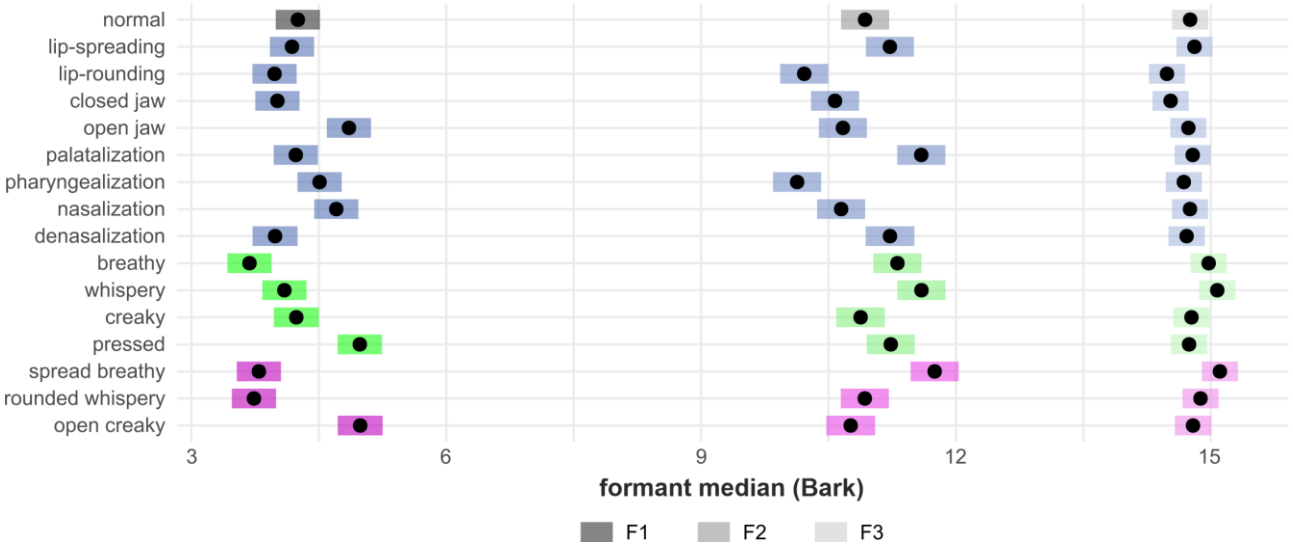


Figure 2. Shifts of *F*1–*F*3 in individual targeted voice manipulations.

References

Disner, S. & Benítez, A. (2018). F2 and F3 covariance as evidence of speaker identity. In *Proceedings of IAFPA 2018*: 86.

Figueiredo, R. M. & Britto, H. S. (1996). A report on the acoustic effects of one type of disguise. *Forensic Linguistics, 3*, 168-175.

Masthoff, H. (1996). A report on a voice disguise experiment. *Forensic Linguistics, 3*, 160-167.

Nolan, F. (2012). Degrees of freedom in speech production: An argument for native speakers in LADO. *International Journal of Speech, Language and the Law, 19*, 263-289.

Růžičková, A. & Skarnitzl, R. (2017). Voice disguise strategies in Czech male speakers. *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, 19–34.

Smith, A. B., Mason, N., Browne, M. E. & Sullivan, B. (2019). Acoustic characteristics of disguised speech: Speaker strategies and listener error patterns. *International Journal of Speech, Language and the Law, 26*, 85-95.

A convincing voice clone? Automatic voice similarity assessment for synthetic speech samples

Linda Gerlach[1, 2]*, Finnian Kelly*[2]*, Kirsty McDougall*[1] *and Anil Alexander*[2]
[1]*Phonetics Laboratory, University of Cambridge, Cambridge, UK.*
{lg589|kem37}@cam.ac.uk
[2]*Oxford Wave Research, Oxford, UK.*
{finnian|anil}@oxfordwaveresearch.com

The rise of deep neural networks has allowed speech synthesis to be taken to a new level and voices built to mimic a target speaker (aka voice clones, spoofed speech, deepfakes) have shaken confidence in cybersecurity. Challenges such as ASVspoof have been created to gauge the risk of spoofed speech for automatic speaker verification systems and to develop spoofing countermeasures. Researchers have begun evaluating the performance of expert and naïve listeners in the detection of spoofed speech (Kirchhübel & Brown, 2022; Terblanche et al., 2021). Nevertheless, spoofed speech may also present an opportunity, for example: to protect the identities of witnesses, or officers in undercover investigations. Natural-sounding spoofs that perceptually sound like a fictitious or real intended target speaker and are different from the source speaker (in this case the witness, or the officer) would provide a useful capability for protecting identities.

Current research methods in speech synthesis and voice conversion systems are regularly evaluated in the Blizzard Challenge and the Voice Conversion Challenge. Their focus is on comparing intelligibility of spoofed speech samples across different systems, as well as assessing sample naturalness and the spoof's similarity with respect to a source or target speaker. These criteria are judged based on human listeners' mean opinion scores (MOS). It is widely known, however, that the recruitment of listeners for such studies is expensive and time-consuming, and particularly in forensic settings as outlined above, listener experiments may not be feasible.

Recent research has highlighted the possibility of assessing voice similarity automatically by training models with MOS. The first VoiceMOS Challenge 2022 (Huang et al., 2022) showed promising MOS predictions of naturalness across a number of models but it also indicated limited adaptability of the models to new speakers and listeners due to the small number of listener ratings available for model training. With respect to unseen listeners, this means that the predictions may be accurate for some listener groups, e.g. synthetic speech experts, but not for others, e.g. the general public. Das et al. (2020) explored the use of an automatic speaker recognition (ASR) system based on x-vectors for similarity assessments with encouraging results. Previous research by Gerlach et al. (2020, 2023) also suggested the use of a pre-trained ASR system to assess the similarity of natural voices based on phonetically relevant acoustic features (long-term formant distributions).

The present study applies this latter system to estimate the similarity of synthetic and natural speech samples from the Blizzard Challenge 2020 (Zhou et al., 2020) and then correlates automatically obtained scores with the MOS provided. The selected data stem from the Mandarin 'hub' task and consist of samples of about 1 min duration from a 'news' task from 16 different speech synthesis systems (n = 68 samples per system), in addition to natural speech samples (n = 17). A pilot experiment was conducted comparing all natural speech samples to all spoofed samples per synthesis system using x-vectors, resulting in sixteen 17×68 score matrices. Phonetic features and spectral features were considered separately, and scores were calculated using PLDA. Scores were averaged across the natural speech samples to create 68 values for each synthesis system. Next, the average of those 68 values was calculated to obtain one mean score for each system against the natural speech. Preliminary results indicate that correlation of the system scores with the corresponding MOS yield higher correlation coefficients for spectral features than for phonetically relevant features, with Spearman $rho_{phonetic}$ = .406 (p = 0.119) and $rho_{spectral}$ = .797 (p < .001). This suggests that spectral features may be better suited for assessing the performance of synthetic systems in terms of similarity.

The potential impact of factors affecting ASR-based similarity assessment, such as artefacts of the different synthesis systems will be considered, and other datasets will be explored. Implications of the findings for the use of spoofed speech to protect the identities of witnesses and officers will be discussed.

References

Das, R. K., Kinnunen, T., Huang, W.-C., Ling, Z., Yamagishi, J., Zhao, Y., Tian, X., & Toda, T. (2020). Predictions of Subjective Ratings and Spoofing Assessments of Voice Conversion Challenge 2020 Submissions. *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 99–120. https://doi.org/10.21437/VCC_BC.2020-15

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (forthcoming 2023). Automatic assessment of voice similarity within and across speaker groups with different accents. *Proceedings of the International Congress of Phonetic Sciences (ICPhS), August 2023*.

Gerlach, L., McDougall, K., Kelly, F., Alexander, A., & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication*, *124*, 85–95. https://doi.org/10.1016/j.specom.2020.08.003

Huang, W.-C., Cooper, E., Tsao, Y., Wang, H.-M., Toda, T., & Yamagishi, J. (2022). The VoiceMOS Challenge 2022. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *September 2022*, 4536–4540. http://arxiv.org/abs/2203.11389

Kirchhübel, C., & Brown, G. (2022). Spoofed speech from the perspective of a forensic phonetician. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *September 2022*, 1308–1312. https://doi.org/10.21437/Interspeech.2022-661

Terblanche, C., Harrison, P., & Gully, A. J. (2021). Human Spoofing Detection Performance on Degraded Speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, August/September 2021, 1738–1742. https://doi.org/10.21437/Interspeech.2021-1225

Zhou, X., Ling, Z.-H., & King, S. (2020). The Blizzard Challenge 2020. *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 1–18. https://doi.org/10.21437/VCC_BC.2020-1

PASS (Phonetic Assessment of Spoofed Speech): Towards a human-expert-based framework for spoofed speech detection

Daniel Denian Lee[1], *Kirsty McDougall*[1], *Finnian Kelly*[2] *and Anil Alexander*[2]
[1]*Phonetics Laboratory, University of Cambridge, UK.*
{ddl26|kem37}@cam.ac.uk;
[2]*Oxford Wave Research, UK.*
{finnian|anil}@oxfordwaveresearch.com

As technology for generating spoofed (aka deepfake) speech has become more powerful and more accessible, the threat of spoofing 'attacks', e.g. unauthorised bank account access, and spread of disinformation (Halpern & Kelly, 2022; Mai et al., 2023), has increased. To complement existing automatic tools for spoofing detection, here we introduce a first version of the PASS (Phonetic Assessment of Spoofed Speech) framework (Table 1), which is proposed as a holistic tool for human-expert-based spoofed speech detection. The development of PASS was initiated by an auditory-phonetic and acoustic (AuPhA) assessment of the Blizzard Challenge 2021 (BC2021; Ling et al., 2021) dataset, which consists of genuine (native European Spanish) speech from a female speaker, along with 12 spoofed (synthesised) versions of her speech, each created by a different team participating in the challenge. Impressionistic assessments of BC2021 synthesised speech (illustrated in Figures 1 and 2) corroborated Kirchhübel and Brown's (2022) evaluation of 'fake speech' features and also revealed further insights. Therefore, a pilot study of the dataset was conducted. Three utterance contexts from 12 teams' submissions were examined using an AuPhA method, comparing them against their genuine counterparts, yielding a set of candidate features. The resulting PASS framework proposes 'auditory', 'visual', and 'acoustic-phonetic' categories of potential spoofed audio features. The auditory labels are described as perceptual 'qualities'—inspired by the Vocal Profile Analysis Scheme and Laverian (1980) voice quality theory—while visual labels refer to visibly atypical features in audio spectrograms and waveforms. Acoustic-phonetic labels refer to features that can be detected auditorily and acoustically, and invoke linguistic-theoretic knowledge.

A phonetically trained listener applied the PASS framework to a blind test involving 10 samples of genuine and spoofed audio from BC2021. Preliminary findings show that the majority of judgments were correct, and that FORMANT ATTENUATION and FOGGING were particularly effective across different speech synthesis methods. A post-hoc review of PASS categories was conducted under additional utterance, speaker, and language conditions. This paper demonstrates the potential of PASS as a practical aid for human experts to discriminate machine from human speech. Future work will target additional languages to improve the cross-linguistic generalisability of PASS categories, as well as test real-world applicability in forensic casework and voice anti-spoofing contexts, and will compare human perception with machine evaluation in the detection of fake speech.

	Description (Auditory)
TINNY QUALITY	An auditory label for 'hollow' or 'thin'-sounding audio.
CRACKLY QUALITY	An auditory label for 'bubbling' or 'crackling' sounds that occur constantly or frequently in the audio background.
MUFFLED QUALITY	An auditory label for the overall attenuation of segmental sounds, with dampening effects particularly pronounced for obstruent consonants.
RHYTHMIC QUALITY	An auditory label for the impression of an artificial rhythm, tempo, and metrical feet.
	Description (Visual)
FOGGING	A visual label for the 'smearing' or 'blurring' of otherwise distinctive structural features in the spectrogram for vowels and consonants.
FORMANT ATTENUATION	Refers to the loss of formant structure definition, particularly for vowels in the higher frequency regions.
PSEUDO-FORMANTS	Formant-like structures in the spectrogram that occur during the articulation of approximant consonants, which behave differently in spoofed audio depending on the specific segment.

CONCATENATEDNESS	Visible overly 'neat' segmental chunking and relative lack of dynamic between-segment features in the acoustic signal.
HYPERNEATNESS	Overly 'neat' linear predictive coding (LPC) points and tracks for formants, with unusually minimal errors in the spectrogram.
	Description (Acoustic-Phonetic)
HYPERFLAT PROSODY	An auditorily perceptible and acoustically analysable property that may be described as an overly level or flat prosodic pattern that is characteristic of 'robotic' speech.
COARTICULATORY DEFICIT	The deficit of between-segment coarticulatory features, which can result in the speech sounding overly 'neat' due to the concatenation of cleanly spliced segment content.

Table 1. The proposed Phonetic Assessment of Spoofed Speech framework.

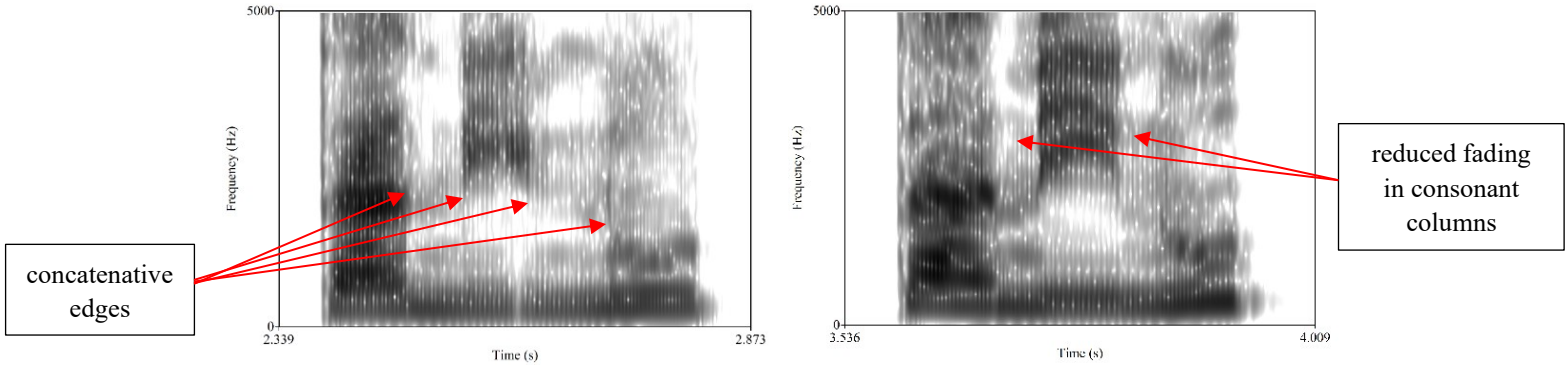


Figure 1. CONCATENATEDNESS artefact of spoofing (left) versus natural speech (right). The natural sample shows less visual contrast between alternating vowel and consonant regions.

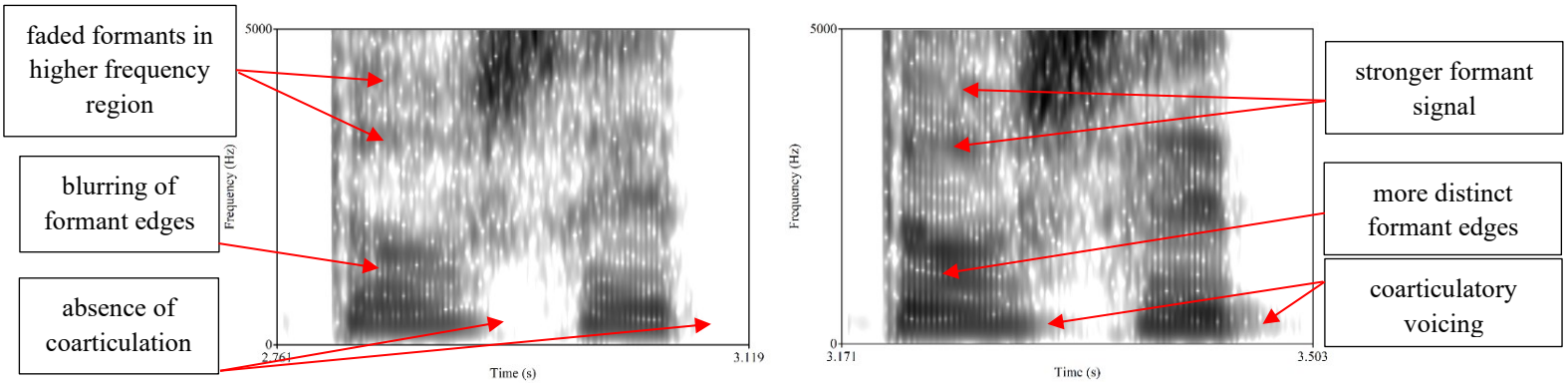


Figure 2. FOGGING/COART. DEFICIT artefacts of spoofing (left) versus natural speech (right).

References

Halpern, B. M., & Kelly, F. (2022). Can DeepFake voices steal high-profile identities? 30th *IAFPA*, 80–81.

Kirchhuebel, C., & Brown, G. (2022). Spoofed speech from the perspective of a forensic phonetician. *Interspeech 2022*, 1308–1312. https://doi.org/10.21437/Interspeech.2022-661

Laver, J. (1980). *The phonetic description of voice quality* (Vol. 31). Cambridge University Press.

Ling, Z.-H., Zhou, X., & King, S. (2021). The Blizzard Challenge 2021. *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 1–18. https://doi.org/10.21437/VCC_BC.2020-1

Mai, K. T., Bray, S. D., Davies, T., & Griffin, L. D. (2023). *Warning: Humans cannot reliably detect speech deepfakes*. arXiv. https://doi.org/arXiv:2301.07829v1 [cs.HC]

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

Chenzi Xu[1], *Paul Foulkes*[1], *Philip Harrison*[1], *Vincent Hughes*[1], *Poppy Welch*[1], *Jessica Wormald*[1], *Finnian Kelly*[2] *and David van der Vloed*[3]

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

{chenzi.xu|paul.foulkes|philip.harrison|vincent.hughes|poppy.welch|jessica.wormald}@york.ac.uk

[2]*Oxford Wave Research, Oxford, UK*

finnian@oxfordwaveresearch.com

[2]*Speech and Audio Research, Netherlands Forensic Institute, The Hague, The Netherlands*

d.vandervloed@nfi.nl

Automatic speaker recognition (ASR) systems usually take a pair of speech recordings as input, extract their speaker embeddings using deep learning (e.g. x-vectors; Snyder et al. 2018), and output through a classifier a speaker similarity score, which is in turn calibrated to a likelihood ratio. Despite the increasing accuracy of the ASR prediction, relatively little is known about the relationship between voice properties and ASR outputs. It has thus been a challenge to explain the output to an end-user in forensic context. This study aims to improve the interpretability of the scores by an ASR system by assessing how acoustic mismatches related to speech production impact different-speaker scores on a given evaluation corpus. Hautamäki and Kinnunen (2020) identified the most prominent factor in explaining low same-speaker scores as the difference in long-term f0 mean. This study focuses on the different-speaker scores in forensically realistic data and explores how differences in a range of acoustic features contribute to the discrimination of speakers. In particular, which acoustic similarities between speakers contribute to more difficult discrimination?

In this experiment, we model the impact of acoustic distance on the ASR score in discriminating speakers with similar demographic profiles. The study utilised a subset of the Home Office Contest corpus[1] containing 155 mobile phone recordings, all from different male speakers of London English. Each recording is a single channel of a mobile phone conversation, about 15 minutes long, with 8kHz sampling rate. Different-speaker (DS) comparisons were conducted using the pre-trained VOCALISE 2021 ASR system (version 3.0.0.1746; Kelly et al. 2019) with x-vectors and PLDA to generate scores. The scores were calibrated using a dataset of mobile phone recordings (8kHz, 16 bit, and single channel) from 20 speakers with a similar demographic profile – male London speakers – from the GBR-ENG corpus. We randomly selected two recordings per speaker for calibration. Bayesian calibration with Jeffreys non-informative priors was used due to the relatively small calibration set (Brümmer & Swart, 2014). The C_{llr} based on the DS likelihood-ratio values was 0.0152, 0.15% of the pairs (18/11925) had a positive calibrated score (i.e. lend contrary-to-fact support to a same-speaker decision). A range of acoustic features including f0, formants, formant bandwidths, jitter, shimmer, spectral tilts and so on were extracted automatically using Praat (Boersma & Weenink, 2022) and the OpenSMILE toolkit (Eyben et al. 2013). In our regression models, the dependent variable is the calibrated scores and the predictor is the acoustic distance between speakers in each comparison, represented by the absolute differences of the statistics of the selected long-term acoustic features or ensemble differences of feature groups. In general, the larger the acoustic distance the lower the calibrated score. Specific pairs that were difficult to discriminate in the ASR system are further examined and discussed. The findings will help us to flag or predict difficult voices for the ASR system to discriminate, and facilitate further exploration on how the discrimination may be improved with score calibration based on a dataset with acoustically similar speakers.

[1] Both GBR-ENG corpus and Home office Contest corpus belong to a telephonic speech database collected for the UK Government for evaluating speech technologies. Further details on application.

References

Brümmer, N., & Swart, A. (2014). Bayesian calibration for forensic evidence reporting. arXiv preprint arXiv:1403.5997.

Eyben, Florian, Felix Weninger, Florian Gross, and Björn Schuller. 2013. “Recent Developments in openSMILE, the Munich Open-Source Multimedia Feature Extractor.” In *Proceedings of the 21st ACM International Conference on Multimedia*. New York, NY, USA: ACM. https://doi.org/10.1145/2502081.2502224.

Hautamäki, Rosa González, and Tomi Kinnunen. 2020. “Why Did the X-Vector System Miss a Target Speaker? Impact of Acoustic Mismatch upon Target Score on VoxCeleb Data.” In *Interspeech 2020*. ISCA: ISCA. https://doi.org/10.21437/interspeech.2020-2715.

Kelly, F., Forth, O., Kent, S., Gerlach, L. and Alexander, A. (2019) Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proceedings of the Audio Engineering Conference: 2019 AES International Conference on Audio Forensics*.

Boersma, P. & Weenink, D. (2022) Praat: doing phonetics by computer [Computer program]. Version 6.2.06, retrieved 23 January 2022 from https://www.praat.org.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018) X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, 5329–5333.

Effects of vocal variation on the output of an automatic speaker recognition system

Vincent Hughes[1], *Jessica Wormald*[1], *Paul Foulkes*[1], *Philip Harrison*[1], *Poppy Welch*[1], *Chenzi Xu*[1], *Finnian Kelly*[2] *and David van der Vloed*[3]

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

{vincent.hughes|jessica.wormald|paul.foulkes|philip.harrison|poppy.welch|chenzi.xu}@york.ac.uk

[2]*Oxford Wave Research, Oxford, UK*

finnian@oxfordwaveresearch.com

[2]*Speech and Audio Research, Netherlands Forensic Institute, The Hague, The Netherlands*

d.vandervloed@nfi.nl

Automatic speaker recognition (ASR) is increasingly used in forensic voice comparison cases. State-of-the-art systems utilise deep learning to convert acoustic features into compact speaker embeddings (e.g. x-vectors; Snyder et al. 2018). Embeddings from known and unknown voice samples are compared to generate a score, which in turn is calibrated to compute a numerical likelihood ratio (LR). Impressive performance of state-of-the-art systems has been reported with forensically realistic data (Morrison and Enzinger 2019), with marked improvements over previous generations of systems (e.g. i-vectors and GMM-UBM). Despite this progress, still relatively little is known about why certain voices perform better or worse within an ASR system, in part due to the abstract relationship between input and output, especially in state-of-the-art DNN-based systems. This issue is particularly important in the context of forensic voice comparison, where it is important for the practitioner to understand whether the output of an ASR system is reasonable given the input, and to explain the output to an end-user (e.g. a court).

In this study, we examine the effects of vocal variation on ASR output. We collected controlled recordings of six phoneticians reading the same text whilst systematically varying aspects of their speech production. Variations included modal voice, a range of laryngeal voice qualities and supralaryngeal vocal settings, high and low pitch, accent guises, and miscellaneous disguise techniques. Each speaker produced three repetitions of each vocal condition in each of three recording sessions, separated by at least one week. Analysis was conducted using the VOCALISE 2021 ASR system (version 3.0.0.1746; Kelly et al. 2019). X-vectors were generated for each sample from each speaker. Cross-session same-speaker (SS) and different-speaker (DS) comparisons were then conducted using PLDA to generate scores. Scores were converted to log LRs using calibration coefficients generated from condition-matched, cross-session SS and DS scores for 20 DyViS speakers (Nolan et al. 2009). Bayesian calibration with Jeffreys non-informative priors was used to account for the relatively small calibration set. Overall performance was evaluated using the log LR cost function (C_{llr}) and its two constituents: calibration loss (C_{llr}^{cal}) and discrimination loss (C_{llr}^{min}).

System performance was generally excellent across all matched-condition comparisons, with almost all vocal conditions producing C_{llr}s equivalent to the modal-modal condition. The exception was the whisper condition, which produced a markedly higher C_{llr}^{cal} and a marginally higher C_{llr}^{min}. Unsurprisingly, condition mismatch had a much greater effect both in terms of calibration and discrimination loss. Whisper again had the largest effect on system output. In addition, vocal settings that substantially alter the supralaryngeal vocal tract (e.g. backed tongue body and lowered larynx) were found to have marked effects on system performance. Comparisons involving high pitch also generated relatively high C_{llr} values (whereas low pitch did not), although interestingly this was most evident for speakers who achieved high pitch through modification of the vocal tract (e.g. through raising the larynx) rather than solely increasing the rate of vocal fold vibration. We discuss the implications of these findings for the use of ASR in forensic voice comparison casework.

References

Kelly, F., Forth, O., Kent, S., Gerlach, L. and Alexander, A. (2019) Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proceedings of the Audio Engineering Conference: 2019 AES International Conference on Audio Forensics*.

Morrison, G. S. and Enzinger, E. (2019) Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion. *Speech Communication*, 112, 37–39.

Nolan, F., McDougall, K., de Jong, G. and 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, 31–57.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018) X-vectors: robust DNN embeddings for speaker recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),* Calgary, 5329–5333.

CON(gruence)-plots for assessing agreement between voice comparison systems

Michael Jessen[1], *Anil Alexander*[2], *Thomas Coy*[2], *Oscar Forth*[2] *and Finnian Kelly*[2]
[1]*Department of Text, Speech and Audio, Bundeskriminalamt, Germany*
michael.jessen@bka.bund.de
[2]*Oxford Wave Research Ltd., Oxford, U.K.*
{anil|tom.coy|oscar|finnian}@oxfordwaveresearch.com

In forensic voice comparison, a practitioner may use multiple systems to compare recordings. Traditional performance measures such as EER and Cllr, or representations such as Tippett plots, can be used to select the 'best' system in the conditions of the case. However, these measures do not inform of the agreement, or congruence, between the different systems for each single comparison; for example, they do not inform about how often System-1 and System-2 both output LRs that support speaker identity. In order to capture such information in an accessible manner, and exploit benefits from multiple systems, a representation called the CON(gruence)-plot has been developed, and implemented as a component of the software BIO-METRICS (https://oxfordwaveresearch.com/products/bio-metrics/).

A CON-plot (Fig. 1) consists of the LR scores output by two voice comparison systems for the same set of comparisons, with same-speaker (H0) and different-speaker (H1) scores clearly differentiated. The plot is divided into four quadrants based on either a log LR value of zero or the Equal Error Rate (EER) score threshold. The relative occupancy of each of the quadrants is indicative of the potential errors introduced by each of the systems; in an ideal scenario, all same-speaker scores would exist in the upper right quadrant (high scores on both methods), and all different-speaker scores in the lower left (low scores on both methods). Entries in the remaining quadrants indicate disagreement, or incongruence, between the two systems, i.e. one of them supporting speaker identity the other non-identity. The level of congruence between systems is also expressed numerically in terms of a (Spearman) rank correlation metric.

CON-plots were applied here to a scenario in which System-1 is high in speaker discrimination but limited in explainability, whereas System-2 is lower in discrimination but higher in explainability (based on phonetic theory). Specifically, an automatic speaker recognition system using x-vector technology was used as System-1 and a semiautomatic system based on long-term formant analysis (LTF) was used as System-2. These systems were applied to a test set called GFS (German Forensic Speech; Solewicz et al. 2017). Further details about the x-vector system applied to this set are presented in Klug et al. (2021) and the LTF system in Jessen (2021).

The resulting CON-plot is presented in Fig. 1. As shown, the level of correlation between the two systems is relatively high. This is expected because both the MFCC features (Mel Frequency Cepstral Coefficients) of the x-vector system and the formant frequencies of the LTF system are strongly or entirely influenced by vocal tract shape. It can be seen that the x-vector system performs much better in terms of speaker discrimination than the LTF system. For example, if LogLR=0 is used as a decision threshold, there are many more false acceptances for the LTF system (red dots above line y=0) than the x-vector system (red dots right of line x=0). The exact performance indices are: x-vector EER 3.0%, Cllr 0.13; LTF EER 17.3%, Cllr 0.66; fused EER 3.3%, Cllr 0.14.

Interestingly, despite the performance difference between the systems in terms of EER and Cllr, there is a high level of congruence among the same-speaker comparisons. In only 2 of 23 comparisons is there disagreement, and in 20 comparisons both systems correctly support speaker identity (in one they both incorrectly support nonidentity, but barely). This prompts the idea of an “explainability-enhanced” mode, where LRs are taken from the more discriminant system, and any incongruent results are declared inconclusive. This and further casework implications will be discussed at the presentation.

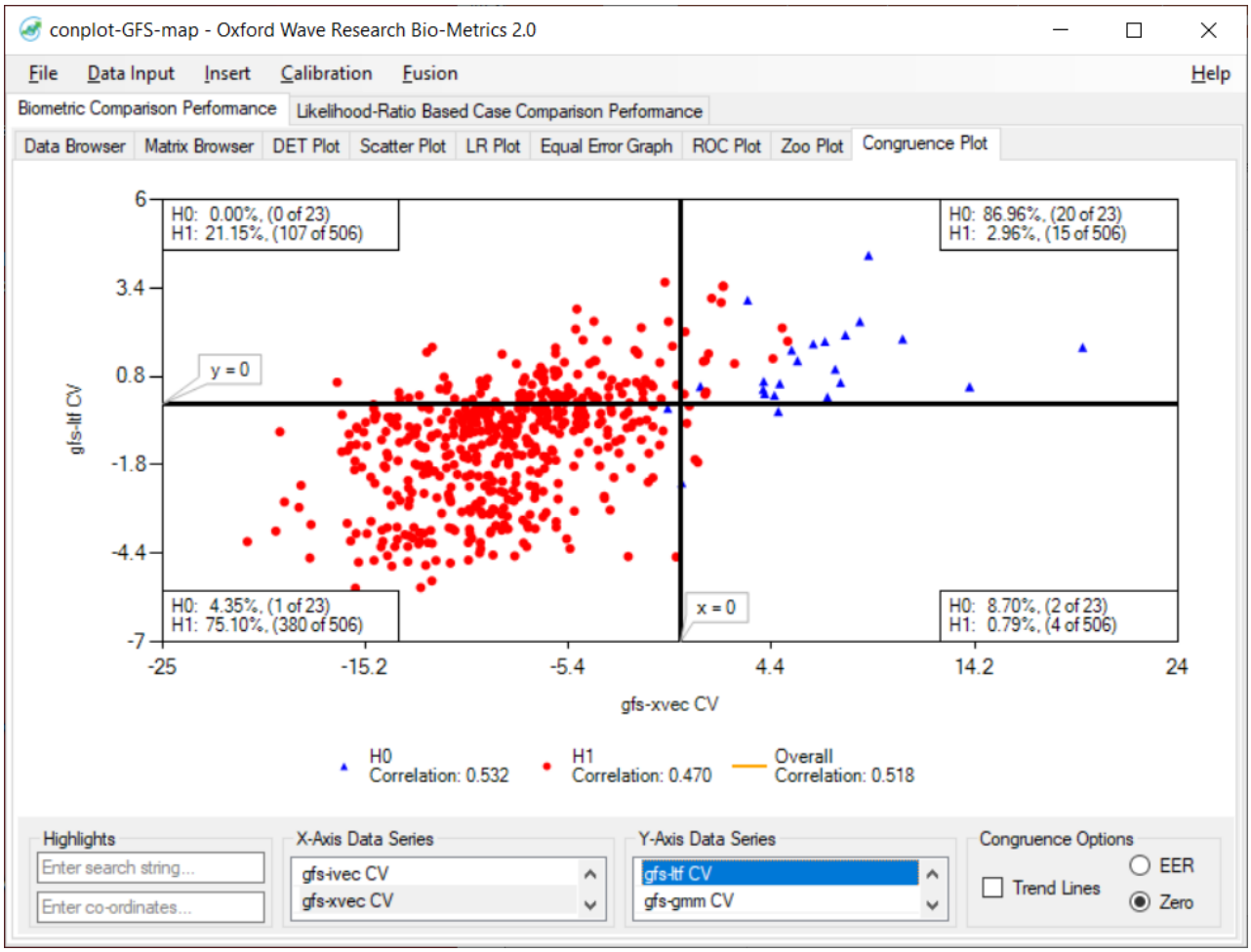


Figure 1. CON-plot for comparison of an x-vector system (X-axis) with an LTF system (Y-axis). Each point on the plot represents the output scores (in terms of Log_eLR after logistic regression cross-validation calibration) of both systems for a single comparison. The blue triangles represent the 23 same-speaker comparisons (H0) and the red dots the 506 different-speaker comparisons (H1). The horizontal and vertical lines represent LogLR=0. Each quadrant shows the number of H0 and H1 comparisons it contains, both in absolute terms and as a percentage of the total number of H0 and H1 comparisons. Also shown in the plot are the Spearman's rank correlation coefficients for H0 and H1 comparisons, and for all comparisons. Further features of the CON-plot will be explained in the presentation.

References

Jessen, M. (2021). MAP adaptation characteristics in forensic long-term formant analysis. In *Proc. Interspeech*, Brno, 411–415.

Klug, K., Jessen, M., Solewicz, Y.A. & Wagner, I. (2021). Collection and analysis of multi-condition audio recordings for forensic automatic speaker recognition. In C. Bernardasci et al. (Eds.), *Speaker individuality in phonetics and speech sciences: Speech technology and forensic applications* (pp. 57–76). Publ. by Associazione Italiana Scienze della Voce, Series Studi AISV, Vol. 8.

Solewicz, Y.A., Jessen, M. & van der Vloed, D. (2017). Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition. In *Proc. Interspeech*, Stockholm, 2849–2853.

The future of evidential voice analysis in the UK: 'Self-employed' is not a dirty word

Christin Kirchhübel[1], *Georgina Brown*[1,2] *and Luke Carroll*[1,2]
[1] *Soundscape Voice Evidence, Lancaster, UK*
ck@soundscapevoice.com
[2] *Department of Linguistics and English Language, Lancaster University, Lancaster, UK*
{g.brown5|l.a.carroll}@lancaster.ac.uk

It is a truism that the provision of evidential voice analysis is a niche area of forensic science. There are only a small number of providers that offer this service when taking a global view, let alone when focusing on individual countries. To the best of our knowledge, the UK provision at this moment in time comprises around five sole practitioners, some of whom work with an assistant, and one micro company with four full-time analytical staff. In the last few years, the UK has already seen a reduction in provision because a) trained practitioners have left casework to move on to pastures new, and b) fewer and fewer academics get involved in casework. It is to be expected that the current provision is going to further diminish in the not too distant future due to practitioners either retiring or pursuing other careers. While casework capacity appears to be declining, demand for evidential voice analysis is not. As such, a new generation of practitioners needs to step in.

Following a recent forensic voice comparison short-course (see Gerlach et. al., 2023), it is clear that there is enthusiasm among junior researchers for a career in casework. However, what routes into casework practice are available to those aspiring practitioners? It is acknowledged that the options are limited, particularly in the UK context where voice comparison casework for evidential purposes has so far been solely undertaken by private providers, rather than government organisations. Having said that, waiting for that rare job offer from one of the existing private providers is not the only option available to those who hope to practise in the UK.

In this talk, we will share our experience of setting up a forensic voice analysis practice, *Soundscape Voice Evidence*, which is based on a self-employment model. But, hold on: "What about having your work checked?", "What if the work dries up?", "How would you get started as someone without previous casework experience?". In this paper, we address some of the preconceived obstacles around working as a self-employed practitioner, but also highlight some of the under-recognised benefits. The aim is to demonstrate that working as a self-employed practitioner is a *real,* rather than a *fanciful,* route into casework, and that it can be done responsibly. More than that, it is to show that this is one of the ways in which we can ensure the continued availability of adequate evidential casework provision in the UK.

References

Gerlach, L., Carroll, L., Fairclough, L., Gibb-Reid, B., Harrington, L., Lee, D.D., Lieb, A., Möller, S., Patman, C., Paver, A., Schäfer, S., Siewert, M., Suthar, N., Valenzuela, M.G., Williams, S., Brown, G. and Kirchhübel, C. (2023). "Learning by doing: An example of casework-relevant training in forensic speech science". Poster presented at the 31st conference of the International Association for Forensic Phonetics and Acoustics (IAFPA). Zürich, Switzerland. 9th – 13th July.

Casework Procedures & Information Management Strategies

Richard Rhodes[1,2], Katherine Earnshaw[1,2], Bryony Nuttall[1], Edie Murray[3] and Peter French[2,3]

[1]*The Forensic Voice Centre, York, UK*

richard.rhodes@forensicvoicecentre.com

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

[3]*J P French Associates, York, UK*

The purpose of this presentation is to explain and discuss the procedures used by The Forensic Voice Centre (FVC) and J P French Associates (JPFA) to manage forensic cases and mitigate the risks presented by cognitive biases.

Many forensic science activities involve subjective decisions because their processes rely on choices and evaluations made by practitioners. These can potentially be influenced by cognitive biases - these biases might come from, for example, contextual information about the case, the way analyses are carried out, pressures or motivations induced by instructing parties, role effects or by the practitioners' own organisations. (For a wider overview see Kassin, Dror and Kukucka (2013) and Cooper and Meterko (2019)). Forensic practitioners/laboratories should take reasonable and practicable steps to avoid bias; however, some biases are unavoidable as some potentially biasing information is relevant to the forensic task: the way this task-relevant information is managed becomes the key priority. Laboratories must also function within realistic time and cost constraints, and so bias management strategies must be practical.

When presenting bias management issues in the past (Rhodes, 2016), one of the main responses found by the authors was that people did not know where to look for guidelines or practical ways to manage bias risks. Outside of the academic literature, and relating to the UK, the Forensic Science Regulator has published guidance on this subject, and there are chapters (in press) in a forthcoming OUP Handbook of Forensic Phonetics which outline different bias mitigation strategies at a basic level.

This presentation will provide an insight into the practical steps taken in every case at the author's organisations to mitigate the risks from cognitive biases effects. We will outline our casework procedures and 'information management' strategies with reference to idealised case examples generated from our experience in 'the wild' in real cases, including forensic speaker comparison, transcription/disputed utterance analysis, and other types of forensic speech and audio casework. As part of the this, we will outline the different roles that staff members take on in different types of cases, as well as how we deal with different types of information, audio material and instructions.

Information management - key strategy

- The general approach is that staff at FVC and JPFA take responsibility for how information provided by the instructing party is handled.
- This is normally done by isolating the reporting analyst from this information; another expert or staff member (with relevant forensic training) will act as an 'information manager' for the case.
- The information manager will decide 1) whether information is task-relevant and 2) if it is task-relevant, when it should be introduced into the analysis process. These decisions are normally checked at the time with another practitioner, and/or as part of a formalised checking process.
- Generally, key work is carried out initially without access to outside information; this information is then introduced at designated stages.
- The timing and staging of revelation of key information is recorded so this process is transparent.
- The information manager might also take part in early case consultations with the instructing party to decide what information is task-relevant; at this stage they might explain how the information is controlled.
- If the main, reporting analyst for a case is exposed to information which has a significant risk of biasing their conclusions, the case will be reallocated to another expert.

We are not arguing that everyone should adopt these specific strategies, or that they will provide a good 'fit' for every practitioner or every organisation across the range of jurisdictions represented in IAFPA; however, we think they are at least a good starter for a discussion on what a proportionate response should be. We aim to spend part of the Q&A discussion exploring possible solutions to bias management for sole practitioners or more isolated practitioners in larger organisations. Through the casework examples and discussion, we will also introduce some problems helpfully raised in a review of this abstract, including: what happens when case instructions, materials, or relevant questions are unclear or not well defined, or what to do when this is only discovered only later during analysis, and how to proceed when the recordings themselves transmit a lot of information.

In reality, a formalised policy cannot cater for all scenarios, and so it is important that all analysts involved in a case understand the underlying principles and the risks associated with manging information. Often, there is a balance between achieving a rigorous standard of safeguarding and doing what is practicable. Sometimes, it is simply not possible to remove sources of potential bias because they are in the recordings themselves. However, it is an increasingly indefensible position to do nothing about the risks of cognitive bias; partly as a sensible form of self-protection because courts and customers are more aware of these issues, but the principal reason is to preserve the integrity of the analysis and the impartiality of expert evidence.

References

Forensic Science Regulator: Cognitive Bias Effects Relevant to Forensic Science Examinations - FSR-G-217 (Issue 2).

Cooper, G. S., and Meterko, V. (2019) Cognitive bias research in forensic science: A systematic review. Forensic Science International, 297, 35-46. https://doi.org/10.1016/j.forsciint.2019.01.016

Kassin, S. M., I.E. Dror and J. Kukucka (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. *Journal of applied research in memory and cognition*, 2(1), 42-52. https://doi.org/10.1016/j.jarmac.2013.01.001

Rhodes, R. (2016) Cognitive bias in forensic speech science: risks and proposed safeguards. *IAFPA conference presentation,* York, UK.

Forensic voice comparison in Canada

Colleen Kavanagh, Peter Milne and Emily Lawrie-Munro
Audio Video Analysis Unit, Royal Canadian Mounted Police, Ottawa, Canada
{colleen.kavanagh|peter.milne|emily.lawrie-munro}@rcmp-grc.gc.ca

As a follow-up to the "year in the life" case study presented at the IAFPA meeting in Istanbul (Kavanagh, Milne, van der Vloed & Dellwo, 2019), we will present an updated overview of voice comparison casework trends within the Royal Canadian Mounted Police's Audio & Video Analysis Unit. We will examine trends across several years, focusing on the types of mismatches encountered, the analysis methods actually used, and any patterns in the conclusions reached.

The 2019 survey covered various types of voice-related cases including speaker profiling, disputed utterances, and forensic transcription. In this iteration, we will focus only on the voice comparison requests, examining more closely the types of mismatches, the methods chosen (and whether these align with our own ideas of best practices), and the conclusions reached in each case. Some interesting case reports will be highlighted to illustrate the particular challenges we have encountered. This will include a discussion of the first case of forensic ASR being presented and accepted as expert evidence in a Canadian court.

Request Type	Speaker Properties	Mismatches	Analysis Methods	Conclusion
Voice comparison	Gender	Channel	ASR	SS
	Language(s)	Language	Au-Ac Phon	DS
	Variety/-ies	Language variety	Combination	Strength of evidence
	Suspected disguise	Recording situation		
	Physiological/ emotional state	Speaking style		
	Age range	Gender		
	Vocal effort (Primary & Secondary)	Age		
		Vocal effort		
		File format		
		Codec		
		Time delay		

Table 1. Case characteristics for which data trends will be reported.

Reference

Kavanagh, C., Milne, P., Van der Vloed, D., & Dellwo, V. A survey of voice-related cases in three forensic speech laboratories. *International Association for Forensic Phonetics and Acoustics 28th Annual Meeting*. Istanbul, Turkey. 14-17 July 2019.

Best Practice Manual for the Methodology of Forensic Speaker Comparison – A Framework Document developed within ENFSI

Isolde Wagner[1], *Dagmar Boss*[2] *and Vincent Hughes*[3]
[1]*Department of Text, Speech & Audio, Federal Criminal Police Office, Germany*
isolde.wagner@bka.bund.de
[2]*Department of Forensic Phonetics, Bavarian State Criminal Police Office, Germany*
dagmar.boss@polizei.bayern.de
[3]*Department of Language and Linguistic Science, University of York, UK*
vincent.hughes@york.ac.uk

Introduction

The Best Practice Manual (BPM) for the Methodology of Forensic Speaker Comparison (ENFSI-FSA-BPM-003, 2022; https://enfsi.eu/wp-content/uploads/2022/12/5.-FSA-BPM-003_BPM-for-the-Methodology-1.pdf) was developed as part of the 'Accreditation of Forensic Laboratories in Europe (AFORE)' project, supported by the 'European Union's Internal Security Fund - Police'. As indicated by the title, it addresses one of the most important tasks in the field of forensic speech and audio analysis, i.e. examining audio recordings containing the voices of unknown and known speakers in order to help answer the question of whether these voices belong to the same speaker or different speakers. The document describes the traditional methodology of a combined procedure of phonetic-linguistic auditory and acoustic analyses of a range of speech features. It does not address automatic and semiautomatic speaker recognition, which is described in the ENFSI document 'Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition' (Drygajlo et al. 2015).

Figure 1. Best Practice Manual of the Methodology of Forensic Speaker Comparison

Results

Following ISO 17020 and 17025 standards the BPM outlines the entire process from receipt of recordings to conclusion, report and presentation at court as well as aspects of quality assurance, such as qualification, training & assessment of personnel, facilities & environmental conditions of the laboratory, equipment used, health & safety aspects, validation, handling items, and quality controls, like e.g. proficiency testing and peer review.

The core of the document is dedicated to analytical methods: Within the tradition of forensic speaker comparison (Jessen, 2018), a range of speech features is analysed to capture the many dimensions on which speakers can be distinguished. To reach a high degree of speaker-discriminatory power speech

features should be as independent of each other as possible. The relevant discriminatory information is determined by the relationship between intra- and inter-individual speaker variation. After the comparison and evaluation process on the basis of the (dis-)similarity and the typicality of speaker-specific characteristics a conclusion statement is given.

There is a wide range of discriminatory speech features that could in principle be analysed within the traditional methodology of forensic speaker comparison. In the BPM some of the most common ones are highlighted, i.e. language, dialect & foreign accent, fundamental frequency & variation, voice quality, formant frequencies, speech tempo, hesitation phenomena & other non-pathological speech disfluencies, and speech pathologies. As detailed information about these speech features is already available in the literature, only overview information is mentioned. In addition, an extended bibliography is appended to the BPM.

Discussion and Conclusion

Major points of discussion concern the combination of different methodologies, validation procedures (QCC-VAL-002, 2014; QCC-PT-001, 2014) and especially the process of evaluation and interpretation of results. Evaluative reporting in terms of the likelihood ratio approach is discussed in particular (ENFSI Guideline for Evaluative Reporting in Forensic Science', 2015). But at present, there is no universally used scale for reporting conclusions in the traditional methodology of forensic speaker comparison. The type and range of scales differ widely between different laboratories and jurisdictions (Gold & French 2011, 2019, Morrison et al., 2016). For this reason, there is no recommendation for one specific scaling in the BPM. Instead, we suggest that irrespective of the conclusion scale used, the whole examination shall undergo validation and quality assurance processes, and that statements shall be considered and expressed with thoroughness and care.

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*. Frankfurt: Verlag für Polizeiwissenschaft.

ENFSI-FSA-BPM-003, Best Practice Manual for the Methodology of Forensic Speaker Comparison, version 01, 22/12/2022.

ENFSI guideline for evaluative reporting in forensic science: Strengthening the Evaluation of Forensic Results across Europe, version 3.0, 08/03/2015.

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

Gold, E. & French, P. (2019) International practices in forensic speaker comparison: second survey. *Journal of Speech, Language and the Law* 26(1): 1-20.

Jessen, M. 2018. Forensic voice comparison. In: Visconti, J. (ed), *Handbook of Communication in the Legal Sphere*. Berlin: Mouton de Gruyter, 219-255.

Morrison, G.S., Sahito, F., Jardine, G., Djokic, D., Clavet, S., Berghs, S. & Dorny, C. (2016) INTERPOL survey of the use of speaker identification by law enforcement agencies. *Forensic Science International* 263: 92-100.

QCC-PT-001, Guidance on the Conduct of Proficiency Tests and Collaborative Exercises within ENFSI, version 001, 27/06/2014.

QCC-VAL-002, Guidelines for the single laboratory Validation of Instrumental and Human Based Methods in Forensic Science, version 001, 10/11/2014.

VocalHUM: real-time whisper-to-speech enhancement

Sonia Cenceschi, Francesco Roberto Dani and Alessandro Trivilini
Digital forensic Service, Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland
{sonia.cenceschi|Francesco.dani|Alessandro.trivilini}@supsi.ch

The proposal focuses on the algorithmic components of VocalHUM (Innosuisse project 52779.1 IP-ICT), a smart system embedding a real-time language-independent whisper-to-speech algorithm. Despite VocalHUM focuses on clinical applications, its algorithmic component (hereinafter cited as HUM) could be exploited in the forensic field too, where hot topics are, for example, speech enhancement in sensitive contexts (Schiavoni et al, 2011; Fan & Hansen, 2008; Rekimoto, 2023), or whispered speech intelligibility (Morris, 2003; Smith, 2015; Bartle & Dellwo, 2015). We refer here to the so-called soft whisper group (Weitzman,1976; Tsunoda et al.,1997; Lim,2011), whose reconstruction needs to recover the missing language-related segmental and suprasegmental components of the speech (French & al., 1947; Amano-Kusumoto et al., 2011; Jovičić, 1998; Gao, 2003; Sharifzadeh et al., 2010; Morris & Clements, 2002). As underlined by Loizou & Kim (2010), despite progress that has been made in the development of speech quality enhancement algorithms, little progress has been made in improving speech intelligibility in critical conditions such as unphonated speech. Recent signal-processing techniques demonstrate the possibility of enhancing whisper's intelligibility and perform voice conversion (e.g. Stylianou, 1996; Toda & Shikano, 2005; Tran et al., 2009), but they need to be strengthened to be embedded in real-time applications. These reasons led us to exclude machine learning techniques, opting for deterministic and generative approaches based on the audio streaming only, trying to avoid the necessity for annotated databases (extremely rare for whispered speech).
The work has been prototyped in Python and rewritten and optimized in C++ programming language in a second step. The whisper is acquired at chunks of 1024 samples, 44100Hz sample rate, 50% overlap. Each chunk is passed through a simple anti-Larsen filter, preprocessed with a first order high-pass filter (100Hz) and then further split into overlapping frames windowed with two different methods: (1) Bartlett window for further usage; (2) Hamming window for FFT and LPC analysis. The windowed frames are then stored into stacks along with the previously computed frames, so that a 1024 samples chunk can always be reconstructed with inverse overlap-add technique. For each new frame RMS, formant frequencies, bandwidths and amplitudes are calculated through LPC analysis. Formants are rounded into the canonical approximation of frequency ranges (Kent & Vorperian, 2018), and F0 is computed from the filtered RMS value (Morris & Clements, 2002). If a frame is considered to contain whispered speech, a corresponding synthesized frame is generated: the synthesizer uses a mixed technique of additive synthesis and filtered noise. Overlapping frames are then reconstructed with inverse overlap-add from the data stacks. The FFT is computed on the whispered frames, and two magnitude masks are calculated by normalizing the average of the magnitude spectrum. Each synthesized frame is transformed by FFT, each bin is multiplied by the two corresponding bins of the previous masks, then it is translated back to the time domain by IFFT. The synthesized frames are finally windowed and the output is calculated by another inverse overlap-add.
Three sample audios for a male voice are provided in Italian, English and Spanish[2]: in this preliminary phase the voice is still metallic, but we will focus on improving its naturalness and pleasantness once the first prototype is patented (currently pending final acceptance) to be tested basing on previous methodologies (e.g. Finizia et al., 1998, Lagerberg, et al., 2014; Springett, 2009, Sharifzadeh et al., 2010). Also, we plan to base on the experience-focused evaluations described in Shahina (2007) in order to take into account of the emotional aspects involving user's voice but using open comments and over at least a month for each user. It should be noted that VocalHUM applies a deterministic approach improving speech intelligibility but not maintaining the timbrical features of the speaker.

[2] https://drive.google.com/drive/folders/1zHRg5u_fRTtgLe65LXhAlw_WqEeb4PMl

References

Amano-Kusumoto, A., & Hosom, J. P. (2011). A review of research on speech intelligibility and correlations with acoustic features. *Center for Spoken Language Understanding, Oregon Health and Science University (Technical Report CSLU-011-001)*.

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language & the Law*, *22*(2).

Fan, X., & Hansen, J. H. (2008). Speaker identification for whispered speech based on frequency warping and score competition. In *Ninth annual conference of the international speech communication association*.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, *19*(1), 90-119.

Finizia, C., Lindström, J., & Dotevall, H. (1998). Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy. *The Laryngoscope*, *108*(1), 138-143.

Jovičić, S. T. (1998). Formant feature differences between whispered and voiced sustained vowels. *Acta Acustica united with Acustica*, *84*(4), 739-743.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of communication disorders*, *74*, 74-97.

Lagerberg, T. B., Åsberg, J., Hartelius, L., & Persson, C. (2014). Assessment of intelligibility using children's spontaneous speech: Methodological aspects. *International Journal of Language & Communication Disorders, 49(*2), 228-239.

Lim, B. P. (2011). *Computational differences between whispered and non-whispered speech*. University of Illinois at Urbana-Champaign.

Loizou, P. C., & Kim, G. (2010). Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. *IEEE transactions on audio, speech, and language processing*, *19*(1), 47-56.

Morris, R. W., & Clements, M. A. (2002). Reconstruction of speech from whispers. *Medical Engineering & Physics*, *24*(7-8), 515-520.

Morris, R. W. (2003). *Enhancement and recognition of whispered speech*. Georgia Institute of Technology.

Rekimoto, J. (2023, April). WESPER: Zero-shot and Realtime Whisper to Normal Voice Conversion for Whisper-based Speech Interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1-12).

Schiavoni, V., Riviere, E., & Felber, P. (2011, June). Whisper: Middleware for confidential communication in large-scale networks. In *2011 31st International Conference on Distributed Computing Systems* (pp. 456-466). IEEE.

Shahina, A., & Yegnanarayana, B. (2007). Mapping speech spectra from throat microphone to close-speaking microphone: A neural network approach. *EURASIP Journal on Advances in Signal Processing*, *2007*, 1-10.

Sharifzadeh, H. R., McLoughlin, I. V., & Russell, M. J. (2010, November). Toward a comprehensive vowel space for whispered speech. In *2010 7th International Symposium on Chinese Spoken Language Processing* (pp. 65-68). IEEE.

Springett, M. (2009). Evaluating cause and effect in user experience. Digital Creativity, 20(3), 197–204. https://doi.org/10.1080/14626260903083637

Smith, K. (2015). A forensic study of whisper and recall. *Working Papers of the Linguistics Circle*, *25*(1), 1-9.

Stylianou, Y. (1996). Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification. *Ph. D thesis, Ecole Nationale Superieure des Telecommunications*.

Weitzman, R. S., Sawashima, M., Hirose, H., & Ushijima, T. (1976). Devoiced and whispered vowels in Japanese. *Annual Bulletin, Research Institute of Logopedics and Phoniatrics*, *10*(61-79), 29-31.

Toda, T., & Shikano, K. (2005). NAM-to-speech conversion with Gaussian mixture models.

Tran, V. A., Bailly, G., Loevenbruck, H., & Toda, T. (2009, September). Multimodal HMM-based NAM-to-speech conversion. In *Interspeech 2009-10th Annual Conference of the International Speech Communication Association* (pp. 656-659).

Tsunoda, K., Ohta, Y., Niimi, S., Soda, Y., & Hirose, H. (1997). Laryngeal adjustment in whispering: magnetic resonance imaging study. *Annals of Otology, Rhinology & Laryngology*, *106*(1), 41-43.

Smile with your eyes! The impact of face coverings on speech intelligibility and perceptions of speaker attributes.

Chloe Patman[1], *Paul Foulkes*[2] *and Vincent Hughes*[2]
[1]*Phonetics Laboratory, University of Cambridge, UK*
cep72@cam.ac.uk
[2]*Department of Language and Linguistic Science, University of York, UK*
{paul.foulkes|vincent.hughes}@york.ac.uk

This study explores the extent to which face coverings create difficulties in understanding speech and judging speakers' social attributes. Previous research has mainly focused on the acoustic impact of face coverings (Llamas et al., 2008; Fecher, 2014). This project addresses the social implications of face coverings, asking the following questions. Do different face-coverings affect speech intelligibility differently? Do they elicit different perceptions of speaker attributes? Do auditory and visual stimuli interact to affect listener judgements? The intelligibility results are insightful for speaker comparison cases and the healthcare sector (where face coverings are still worn). The speaker attribute results shed light on the potential prejudice faced by veil wearers.

Method

Intelligibility perceptions were addressed via an orthographic transcription task, testing participants' ability to identify minimal pair words. The stimuli comprised a set of Harvard sentences, adapted to contain minimal pair words for /f/ and /s/, /p/ and /k/, and /f/ and /θ/, where either minimal pair word is semantically acceptable. For example: "They saw the FOX/SOCKS in the dirty water". These pairs were chosen as they are known to be perceptually difficult to distinguish (Miller & Nicely 1955; Wang & Bilger 1973). Speaker attributes were investigated by presenting listeners with speakers reading a set of short anecdotal stories about face coverings and miscommunications. Participants then rated the voice on five attributes (friendliness, nationality, trustworthiness, intelligence, intelligibility) using a 7-point Likert scale. Four SSBE speakers produced the stimuli in a recording studio, each wearing different face coverings (surgical mask, niqāb and no mask). The stimuli were then presented to listeners using a matched-guise method, such that for half the stimuli there was a mismatch in the visual and audio presentation. For example, a photograph of a veil wearer accompanied mask-free speech. This method was implemented to assess the relative contribution on intelligibility/attribute scores of the acoustic signal and listeners' prior expectations about face coverings.

Results

The study found that wearing a face covering did not significantly affect speech intelligibility (word identification), and neither did the different mask types. However, for [s], [k] and [f] (when in a minimal pair with /θ/), the probability of obtaining a correct transcription significantly decreased when listening to the face-mask speech (Figure 1). For the attributes, face-mask auditory stimuli were rated significantly less friendly than mask-free auditory stimuli. Additionally, face-mask visual stimuli (specifically a niqāb) were rated as significantly less British than mask-free visual stimuli (Figure 2). Overall, face coverings did not affect speech intelligibility. However, speakers wearing them were perceived as sounding less friendly and looking less British.

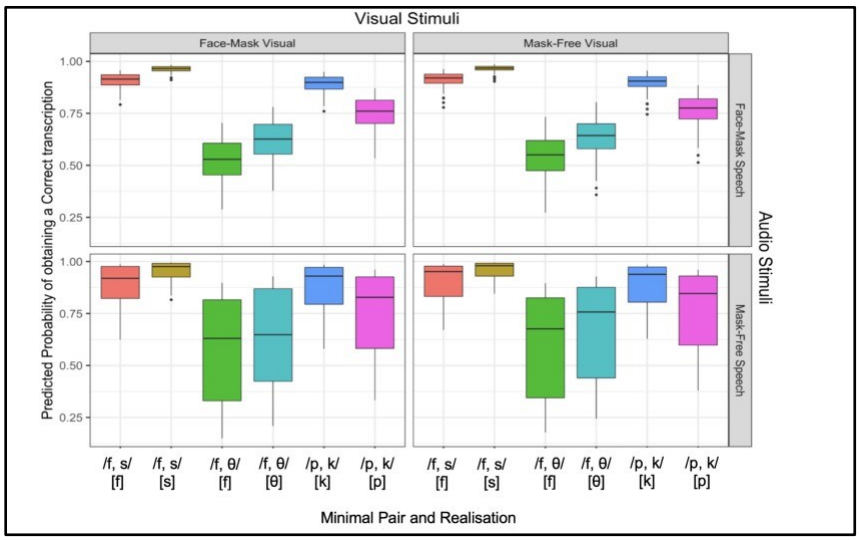


Figure 1. Box plot illustrating the probability of obtaining a correct transcription as predicted by the statistical model, by minimal pair category, and audio and visual stimuli. /f, s/ = [f] shows 'fox' responses to ambiguous 'fox/socks' stimuli.

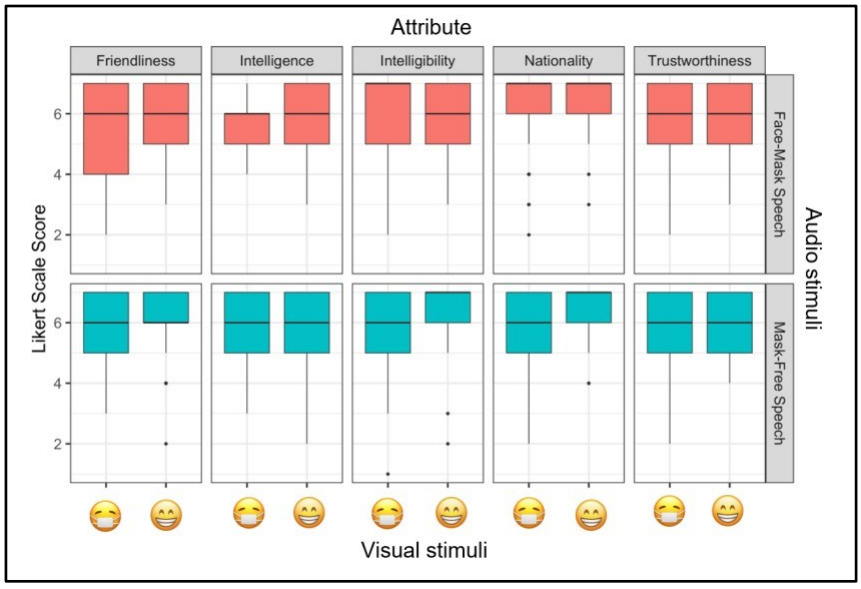


Figure 2. Box plot illustrating the distribution of Likert scale scores for the speaker attributes according to attribute, visual and auditory stimuli. 1 refers to foreign as well as the lowest scores for friendliness, intelligence, intelligibility, and trustworthiness.

References

Fecher, N. (2014). *Effects of forensically-relevant facial concealment on acoustic and perceptual properties of consonants.* PhD Thesis, University of York, UK. https://etheses.whiterose.ac.uk/7397/1/Natalie_Fecher_PhD_2014.pdf

Llamas, C., Harrison, P., Donnelly, D., & Watt, D. (2008). Effects of different types of face coverings on speech acoustics and intelligibility. *York Papers in Linguistics, 2*(9), 80-104.

Miller, G. A., & Nicely, P. A. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America*, *27(1)*, 338–352.

Wang, M. D., & Bilger, R. C. (1973). Consonant confusions in noise: A study of perceptual features. *The Journal of the Acoustical Society of America*, *54*(5), 1248–1266.

To what extent can expert listeners distinguish between speakers based on speech rhythm?

Luke Carroll[1,2] *and Georgina Brown*[1,2]
[1]*Department of Linguistics and English Language, Lancaster University, Lancaster, UK*
[2]*Soundscape Voice Evidence, Lancaster, UK*
{l.a.carroll|g.brown5}@lancaster.ac.uk

Within the auditory-phonetic and acoustic approach to forensic voice comparison, there is currently no structured framework analysts can use to effectively account for speakers' speech rhythm patterns. Previous research has sought to assess the discriminatory potential of speech rhythm parameters in different ways using read (content-controlled) speech data. Leemann, Kolly and Dellwo (2014) used measures of relative syllable durations to characterise speech rhythm across utterances, whilst He and Dellwo (2016) used measures of relative syllabic intensity values within utterances.

More recent research by the first author of the present work has investigated the feasibility of applying such measures to spontaneous (content-mismatched) speech. Across spontaneous utterances, syllabic measurements of intensity, f0 and duration yielded very little speaker discriminatory power, however, more promising results were obtained when the rhythmic characteristics of specific frequently occurring speech units (*erm*, *er*, *yeah* and *no*) were analysed. Despite such results indicating there is some value in pursuing rhythm for speaker identification, it is suspected that some rhythmic information will likely be missed using these methods. Furthermore, comparing the acoustics of speakers' rhythm patterns is reliant upon 'enough' adequate speech data being available to the forensic analyst – a privilege that cannot be guaranteed within the forensic context.

The current study examines the contribution of holistic assessments of rhythm grounded in perception. To do this, expert listeners were invited to discriminate between speakers and evaluate the similarity of speech samples based on primarily rhythmic attributes of speech. Speech samples from the WYRED corpus (Gold et al. 2018) were subjected to delexicalisation, whereby syllables were represented by schwa-like tones, creating 30-second samples which foregrounded rhythmic characteristics. These delexicalised samples were presented to 32 expert listeners (forensic caseworkers, phonetics researchers, etc.) and 13 non-expert listeners in an online perception experiment. The experiment consisted of three sections. In sections one and two, participants were required to make a binary decision as to which delexicalised samples contained the same speaker as the original (non-delexicalised) samples whilst also providing qualitative feedback. In section three, listeners had to rate the similarity of pairs of delexicalised speech samples on a nine-point Likert scale from very similar (1) to very different (9).

Results revealed that expert listeners were better than non-expert listeners at making correct speaker identification assessments across all sections of the experiment. Amongst the expert listeners, those who had expertise in forensic phonetics generally performed better than those who did not. For all participant groups, section three was the most challenging. Within section three, certain sample pairs had substantially more "correct" similarity ratings than others. In addition to presenting these quantitative findings, we review the qualitative observations to determine whether it is possible to develop meaningful descriptors of speech rhythm which could feed into a perceptual rhythm framework for forensic speech analysis.

References

Gold, E., Ross, S. and 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, September 2-6, 2018, Hyderabad, pp. 2748-2752.

He, L. and Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. International Journal of Speech, Language and the Law, 23(2), 243–273.

Leemann, A., Kolly, M.-J. and Dellwo, V. (2014). Speech-individuality in suprasegmental temporal features:

implications for forensic voice comparison. Forensic Science International 238: 59–67.

The perception and interpretation of additional information in legally relevant transcripts

James Tompkinson[1] *and Kate Haworth*[1]
[1]*Institute for Forensic Linguistics, Aston University, UK*
{j.tompkinson|k.haworth}@aston.ac.uk

A key consideration in transcription tasks involves knowing how much information to include in a transcript in addition to the words that are spoken. Should, for example, a transcriber denote when pauses, crying, or overlap occurs? If these are to be included, how should this be done? This is a key consideration for a range of legally relevant transcription applications, from transcripts of poor-quality audio (Fraser, 2022) to transcripts of police interviews (Haworth, 2018; Tompkinson et al., 2023) and courtroom proceedings (Walker, 1986).

This study assessed how non-linguists understood a range of notation conventions for the representation of additional information in transcripts. We designed an experiment with multiple versions of a police interview transcript, each containing different representations of six additional aspects of speech: *pauses*, *overlapping talk*, *inaudible speech*, *emphasis*, *crying* and *sniffing*. These are displayed in Table 1. 150 participants (50 per transcript) read one transcript. Participants were then asked to state what they thought each representation meant, and whether the inclusion of additional information made the transcript easier or harder to read and understand. No additional key was provided at this stage.

Linguistic feature	***Transcript 1***	***Transcript 2***	***Transcript 3***
Pauses	(…)	(>0.6 sec) / (.)	Both representations used in Transcripts 1 and 2 for each feature
Overlap	[speech] [speech]	Speech… … Speech	
Inaudible speech	(XXXX)	\\\ (-)	
Emphasis	<u>Underlined</u> speech	CAPITAL LETTERS	
Crying	HHHHuh	((crying))	
Sniffing	.shih	((sniffs))	

Table 1. Representations used in the experiment

The results showed a high level of variation in interpretations. For example, 91% of participants who read Transcript 2 said that the "(0.6 sec)" notation represented a pause. Contrastingly, the representations for crying and sniffing in Transcript 1, and overlap and inaudible speech in both Transcripts 1 and 2 were predominantly misunderstood, with less than 10% of participants providing correct interpretations. As expected, the inclusion of multiple representations for each feature in Transcript 3 lowered accuracy rates. A high proportion of participants (82% for Transcript 1, 72% for Transcript 2 and 80% for Transcript 3) also said the inclusion of the additional information made the transcript harder to read and understand.

In a follow-up stage, we created two further transcripts and provided participants with a transcription key which detailed what each feature represented. We opted to use the representations which were not accurately identified in Transcripts 1-3 for Transcript 4. In Transcript 5, we chose features that

were more accurately identified in Transcripts 1-3, but assigned different meanings to assess whether participants would be 'led' by the information in the key. Our expectation was that the inclusion of a key would improve accuracy. A further 100 people (50 per transcript) took part in the same experimental process described above, and accuracy rates of between 63% and 88% were observed. Furthermore, 48% of participants who read Transcript 4, and 44% of participants who read Transcript 5, said that the use of additional notation conventions made the transcript harder to read and understand. This represented an improvement on the results from Stage 1, but perhaps not as much as would have been expected.

Overall, this paper highlights some of the issues with representing additional information in transcript. The results should serve to promote caution around assuming that a) non-linguists can correctly interpret transcription notation conventions, and b) that the use of such conventions improves the understandability of transcripts.

References

Fraser, H. (2022). A Framework for Deciding How to Create and Evaluate Transcripts for Forensic and Other Purposes. *Frontiers in Communication*, *7*, 898410.

Haworth, K. (2018). Tapes, transcripts and trials: The routine contamination of police interview evidence. *The International Journal of Evidence & Proof*, *22*(4), 428-450.

Tompkinson, J., Haworth, K., Deamer, F., and Richardson, E. (2023, in press). Perceptual instability in police interview records: Examining the effect of pauses and modality on perceptions of an interviewee. *International Journal of Speech, Language and the Law.*

Walker, A. G. (1986). Context, transcripts and appellate readers. *Justice Quarterly*, *3*(4), 409-427.

Taking a closer look at formants for the purpose of voice comparison: a meta-analysis of research literature

Lois Fairclough[1], *Georgina Brown*[1,2] *and Christin Kirchhübel*[2]
[1]*Department of Linguistics and English Language, Lancaster University, Lancaster, UK*
l.fairclough@lancaster.ac.uk, g.brown5@lancaster.ac.uk
[2]*Soundscape Voice Evidence, Lancaster, UK*
ck@soundscapevoice.com

Formant measurements have become a widely accepted feature in forensic speech science research and casework. In fact, courts in Northern Ireland even go as far as to insist that a formant analysis must be carried out as part of a forensic voice comparison analysis (R v O'Doherty [2002] NICA 20). Recent research in the broader phonetics literature has raised concerns about the reliability of formants and the weight that should be given to them in related sub-disciplines (Whalen et al, 2022). Given the increased scrutiny of formants across the phonetic sciences, it is fitting to take stock and review the use of formants within forensic speech science, ultimately reconsidering their actual contribution in casework settings.

The current study starts to do this by carrying out a meta-analysis of the use of formants in forensic phonetics research studies. The meta-analysis covers 25 years, 82 unique research papers, and close to 400 quantitative and qualitative results. The research papers are divided into two sub-categories depending on the motivation behind the examination of formants: a) speaker discrimination performance and b) robustness of formants across conditions (i.e., within-speaker variation owing to speaker-internal or speaker-external factors). The speaker discrimination papers were quantitatively analysed according to several factors which capture the nature of the dataset, the formant measurement technique, and the presentation of results. The robustness papers were qualitatively assessed.

Results

In the speaker discrimination part of the meta-analysis, results were classified according to whether the study reported Equal Error Rates (EERs) or Classification Rates (CRs). Formant measurement type (i.e., midpoint, dynamic, LTF) produced highly variable EER results (Figure 1 Left). When applying a linear regression analysis to the EER results, no significant results effects were found for measurement type or linguistic variety. Interestingly however, a significant effect was found for number of speakers ($p = 0.04$). There appears to be a weak tendency for studies which include a larger number of speakers to display higher EERs. This could be connected to the idea that a larger pool of speakers creates more room for different speakers to sound more similar to one another - therefore there is opportunity for more speaker discrimination errors. The CR results are more difficult to compare because there are insufficient studies which share similar dataset sizes. As expected, CRs are highly dependent on number of speakers included ($p = 0.01$), which makes these results difficult to extrapolate to casework settings (Figure 1 Right).

Of the robustness papers included so far, the following factors have been investigated: telephone transmission (landline and mobile), compression, measurement software/settings, vocal effort, non-contemporaneity, speech style (spontaneous vs. scripted) and the speakers themselves. There is a tendency for studies to focus on just one factor at a time, rather than multiple factors at once as would be reflective of casework.

It is hoped that this meta-analysis provides a comprehensive review of current research practices and their alignment with casework. In doing so, the status of formants in the forensic speech domain can be more clearly understood.

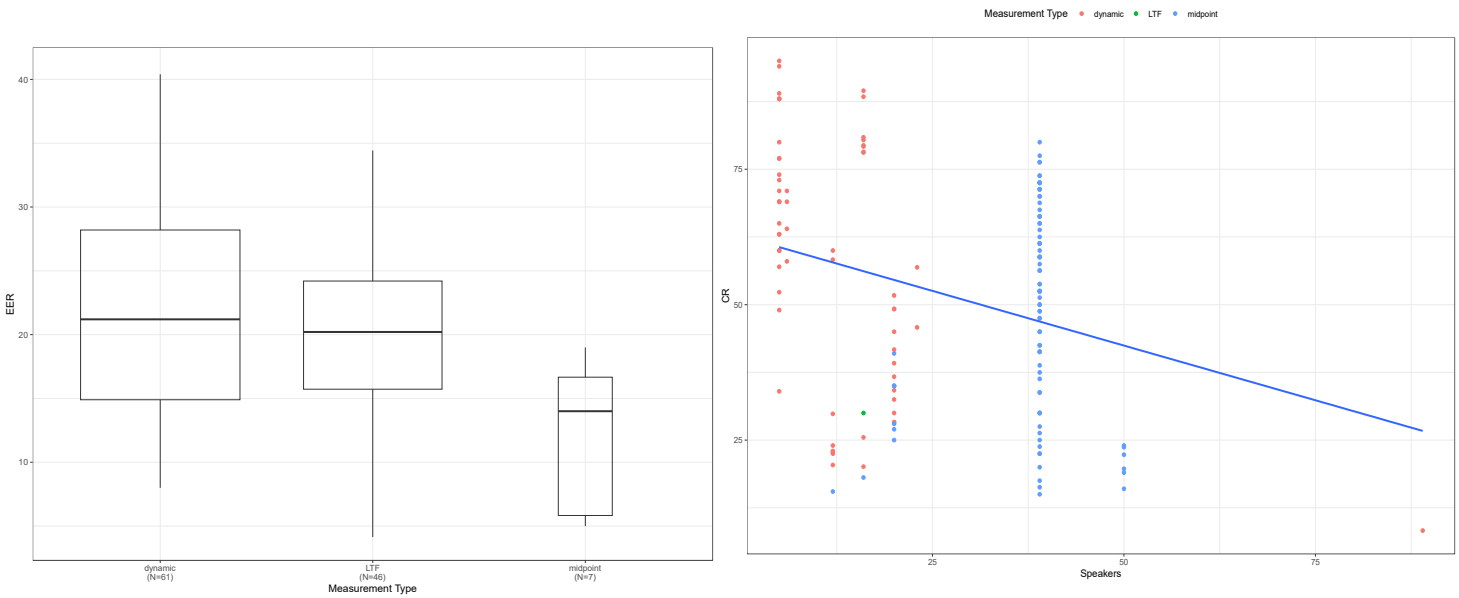


Figure 1. Results of the speaker discrimination analyses. Left: EER results showing EER vs measurement type, right: CR results showing number of speakers vs CR coloured by measurement type.

References

Whalen, D. H., Chen, W. R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). *The Journal of the Acoustical Society of America*, *152*(2), 933-941. doi: 10.1121/10.0013410.

Personality ratings for male and female speakers of different age groups

Paula Rinke[1], Nadine Lavan[2] and Mathias Scharinger[1]
[1]*Research Group »Phonetics«, Institute of German Linguistics, Philipps-Universität Marburg, Germany*
paula.rinke@uni-marburg.de, mathias.scharinger@staff.uni-marburg.de
[2]*School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom*
n.lavan@qmul.ac.uk

The human voice carries information about the linguistic content of a message but also about the speaker. Previous studies have shown that physical information, such as the speaker's age or gender, are prominently perceived during early stages of speech processing (Latinus & Taylor, 2012). In addition, there is evidence that trait perception from voice (e.g., regarding trustworthiness or dominance) proceeds rapidly (Mileva & Lavan, 2023). Moreover, these authors found differences in the time course for male and female voices, with male voices requiring less exposure for consistent ratings of attractiveness and dominance than female voices.

The present study aims to extend these findings by investigating whether person characteristics such as trustworthiness, dominance, attractiveness, professionalism, and education are modulated by physical speaker characteristics, such as gender or age.

The present behavioral study was carried out as part of an EEG study investigating the temporal aspects of speech and speaker processing. The rating task was based on 96 recordings of the German 400ms-long vowels [a], [i], and [u] from 16 male and 16 female speakers, which were extracted from the "Saarbrücker Stimmdatenbank". The vowels were produced in a sustained manner at the speakers' normal pitch and trimmed to 400ms with a 25ms fade-in and fade-out.

Male and female speakers were sampled to fall into an older (55-74 years) or younger (20-34 years) age group. Consequently, each condition in this 2x2 design included eight speakers. The five characteristics selected for this experiment were among the most frequently mentioned person characteristics in a previous study (Lavan, 2023).

A rating task was carried out using a 9-point rating scale, with 9 indicating that a voice was perceived as scoring high on a specific characteristic (e.g., 'very educated', 'very trustworthy'). All participants were presented with all 96 stimuli for each characteristic and were asked to rate each stimulus via mouse click on the scale.

30 participants took part in the study (12 male, 18 female; mean age 25.4±4.42 years, range 19–36). One participant was excluded as they did not complete the full experiment. All participants were native German speakers and had no reported hearing or neurological disorders.

For statistical analysis, Linear Mixed Models were performed in Jamovi (2022) and were based on the ratings for each characteristic, with speakers' gender and age as fixed effects and subject as a random intercept. For all characteristics, ratings differed significantly ($p<.001$) between the speakers' gender and age. Female speakers were perceived as more attractive, educated, professional and trustworthy, while male speakers were rated as more dominant (Fig. 1). In addition, for all characteristics, the younger speakers reached higher rating scores than the older speakers.
Overall, these results confirm that personal speaker characteristics are modulated by physical characteristics with significant differences regarding the speakers' age and gender.
For forensic settings, these findings can be relevant in the context of earwitness testimony, especially regarding how testimonies might be influenced by physical speaker characteristics.

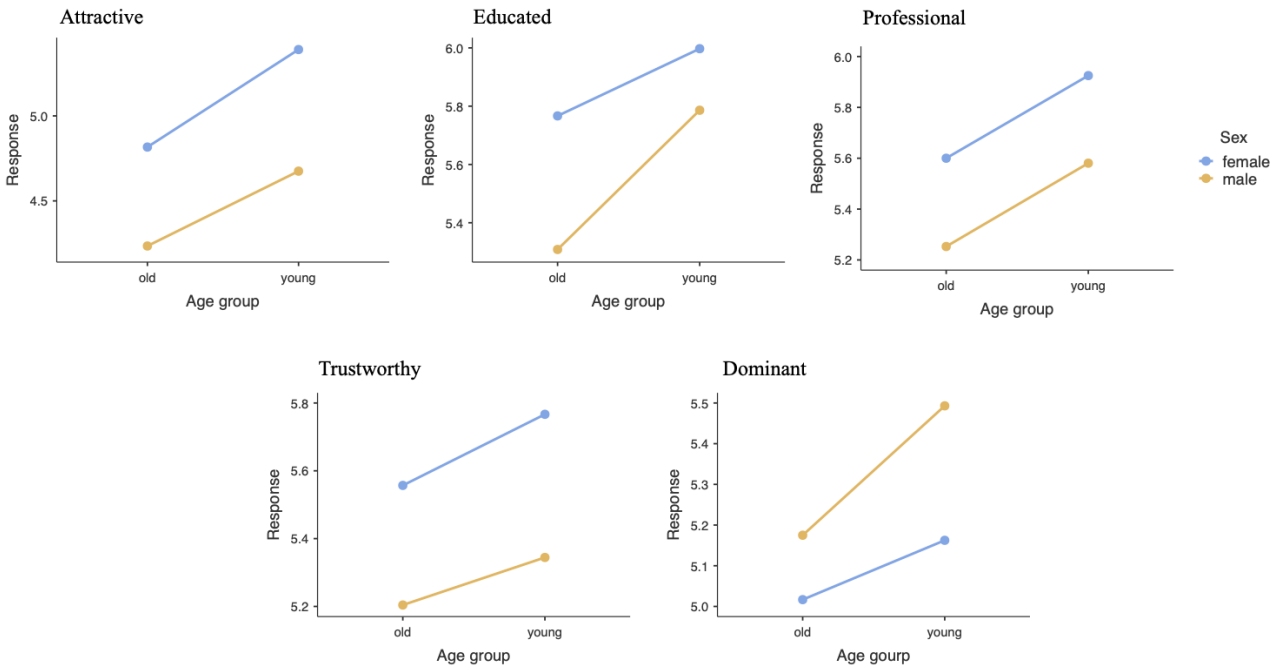


Figure 1. Mean rating scores for male and female speakers of older and younger age group for the five character traits ‘attractive’, ‘educated’, ‘professional’, ‘trustworthy’ and ‘dominant’.

References

Latinus, M., & Taylor, M. J. (2012). Discriminating Male and Female Voices: Differentiating Pitch and Gender. *Brain Topography*, *25*(2), 194–204. https://doi.org/10.1007/s10548-011-0207-9

Lavan, N. (2023). How do we describe other people from voices and faces? *Cognition*, *230*, 105253. https://doi.org/10.1016/j.cognition.2022.105253

Mileva, M., & Lavan, N. (2023). Trait impressions from voices are formed rapidly within 400 ms of exposure. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001325

The Jamovi Project. (2022). *Jamovi (Version 2.3)*. https://www.jamovi.org/

The Vocal Parameters of Dissociative Identity Disorder

Alexandra Lieb and Gea de Jong-Lendle
Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany
lieba@students.uni-marburg.de, dejong@staff.uni-marburg.de

Dissociative Identity Disorder (aka multiple/split personality disorder) is defined in the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) published by the American Psychiatric Association (2013) as a disruption of identity characterized by two or more distinct personality states. DID is caused by trauma, the vast majority of cases by prolonged abuse during childhood (Dimitrova et al. 2020, DSM-5 2013, Merck Manual). The estimated prevalence of DID in the general population has been reported to range from 1 to 3% (Pietkiewicz et al. 2021:1). The three main symptoms include identity confusion, identity alteration and amnesia. Patients present with two or more personalities. Each has their own distinct identity and perception of the environment and of themself (Sinnott-Armstrong et al. 2000) - gender, sexuality, status, likes and perceptive appearance can all differ. Memories may not be shared between personalities. The actual person/body is called the *host*. Additional personalities are *alters*. As these can be so different, could their speech be different as well? If so, could their speaking characteristics be so different, that they are perceived as different individuals?

The aim of this study is to find out to what extent these personalities can be different from a phonetic perspective; substantial differences may be forensically relevant. It consists of a phonetic comparative analysis of personalities (part 1) and an estimation experiment testing how different personalities are perceived in terms of age (part 2).

Methodology:

The speakers consisted of two women with DID: a 42-year-old German and a 27-year-old US-American woman. For this pilot-study two personalities per speaker were chosen that perceptively differed the most: for the German speaker the 4-year-old alter and the 42-year-old host, for the American speaker the 16- and the 50-year-old alter. Per personality spontaneous speech samples (1.5-4min) were created that were subsequently analysed in terms of articulation rate, mean F0 and dialect/pronunciation.

Results – Part 1:

	DE-4 (4y-alter)	DE-42 (42y-host)	USA-16 (16y-alter)	USA-50 (50y-alter)
AR (syll/sec)	4.39	5.34	6.30	6.19
F0 (Hz)	309	188	263	242
Dialect/Pron.	weak articulation/ deletion of /t/ (pre-treatment?)	aspirated /t/ (post-treatment?)	my [a͡i]	my [a͡:] velar fronting [ing-in] (Southern-American)

Table 1. Results of articulation rate, F0 and dialect/pronunciation comparing two identities for two female (1 German, 1 US-American) speakers with DID.

Most striking was the difference in F0 between the two German personalities. The analysis also showed some noticeable differences in dialect and pronunciation. The DE-4 exhibited a non-standard version of /t/, whereas DE-42 consistently produced a heavily aspirated /t/. The difference could be an indication that the speaker received speech-therapy treatment in her youth. USA-50 showed typical features of a Southern-American accent (Allbritten 2011), whereas USA-16 did not.

Part 2 consisted of two age-estimation surveys carried out by native speakers of German or English aged 18-55 years. 4-6s sound clips were extracted from all personalities and accompanied by five additional distractor speakers covering a similar set of ages between 4 and 70 years.

Results – Part 2:

The German survey results show a fairly large age estimation difference for the 4-year-old alter and the 42-year-old host. Although the age judgements showed a wide variety for the younger alter (4 – 80), around one third of the listeners perceived her indeed as child of 10 years or younger. Regarding DE-42: 50% of the listeners estimated her to be within a ±5 range of her real age. The age estimates for the American alters showed a much smaller gap between both alters and a tendency towards the age of the host (see Figure 1 and 2 below).

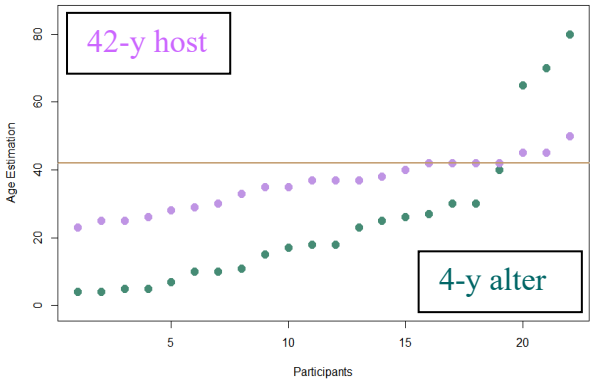


Figure 1. Distributions of all age estimations for DE-4 (alter, green) and DE-42 (host, purple) with the orange line indicating the real age of the speaker (42-y host).

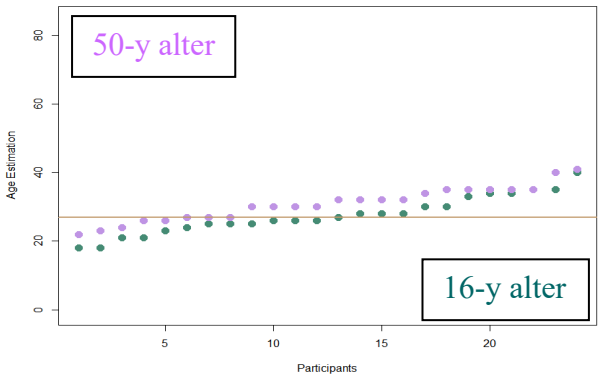


Figure 2. Distribution of all age estimations for USA-16 (alter, green) and USA-50 (alter, purple) with the orange line indicating the real age of the speaker (27-y host).

Conclusion

Both the phonetic measurements and the age-perception results show that a person with DID can exhibit personalities that are capable of sounding considerably and consistently different. These results may indicate that 1) vocal parameters can serve as an additional diagnostic tool for DID, and 2) forensic speaker comparison may be more complicated when it concerns a speaker with DID. Following the age estimation study, a follow-up study will be carried out to test, whether similar results are obtained using a speaker discrimination paradigm.

References

Allbritten, R. M. (2011). Sounding Southern: Phonetic features and dialect perceptions. Georgetown University.

American Psychiatric Association. (2013). Diagnostic and statistical manual of mental disorders (5th ed.)

Berry-Dee, C. (2017). Talking With Psychopaths and Savages-A journey into the evil mind: A chilling study of the most cold-blooded, manipulative people on planet earth. Kings Road Publishing.

Dimitrova, L., Fernando, V., Vissia, E. M., Nijenhuis, E. R., Draijer, N., & Reinders, A. A. (2020). Sleep, trauma,fantasy and cognition in dissociative identity disorder, post-traumatic stress disorder and healthy controls: a replication and extension study. European journal of psychotraumatology, 11(1), 1705599.

Reinders, A. A., & Veltman, D. J. (2021). Dissociative identity disorder: out of the shadows at last?. The British Journal of Psychiatry, 219(2), 413-414.

Sinnot-Armstrong, W., Behnke. S. (2000) “Responsibility in cases of multiple personality disorder.” Philosophical Perspectives 14: 301-323

Automatic Speaker Recognition: does dialect switching matter?

Marlon Siewert[1], *Linda Gerlach*[2, 3], *Anil Alexander*[2], *Gea de Jong-Lendle*[1], *Alfred Lameli*[1] *and Roland Kehrein*[1]

[1]*Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany*
Siewert4@students.uni-marburg.de, lameli@uni-marburg.de,
{dejong|kehreinr}@staff.uni-marburg.de

[2]*Oxford Wave Research, Oxford, UK.*
{linda|anil}@oxfordwaveresearch.com

[3]*Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, UK*
lg589@cam.ac.uk

Introduction

In forensic speaker comparisons, the conditions of the disputed and reference recordings (e.g. recording quality, speaking setting and style, dialect) should be as similar as possible to obtain reliable results. In reality however, a mismatch at some level between forensic recordings is almost always the case. For example, a speaker may speak a standard variety of a particular language in the (formal) police interview and a regional variety in the disputed recording. Research involving German speakers has shown (Kehrein 2012) that speakers of regional/dialectal varieties of German tend to vary between their regional base dialect and the German standard variety, displaying differing degrees of linguistic proximity to standard German depending on communication setting and conversation partner. As automatic speaker recognition systems are increasingly being integrated into forensicphonetic casework, this study was designed to test the effect of intraspeaker dialectal variation on ASR performance. The ASR software used in this case is the VOCALISE system produced by Oxford Wave Research (VOCALISE 2021). The study consisted of two parts: in the first experiment, withincondition and across-condition comparisons of multiple samples per speaker with varying degrees of phonetic proximity to standard German were conducted. In the second part, these ASR recognition results were then compared with the phonetic distance-from-standard-German measures calculated for both recordings of each speaker as part of the project Regionalsprache.de (Schmidt et al. 2020ff).

Methodology

Audio recordings of 30 speakers (45-55 years old) from the main dialectal areas of Germany (Figure 1) were selected from the Regionalsprache.de Database (Schmidt et al. 2020ff) produced by the Deutscher Sprachatlas in Marburg. These speakers had been recorded during a formal interview with the experimenter to evoke a near-standard speaking style, and during a casual conversation with a close friend born and raised in the same area to evoke a dialectal speaking style. For each speaker three 2-minute extracts were selected from each recording. These six extracts were subsequently analyzed in VOCALISE (Oxford Wave Research 2023b). From each of these extract files MFCC features were extracted to generate an x-vector based on a pre-trained deep neural network. In the next step, all these vectors were compared with each other resulting in comparison scores. Further, equal error rates (EERs) were calculated using Bio-Metrics 2021 (Oxford Wave Research 2023a) in order to compare the following conditions: dialect vs. standard, dialect vs. dialect, and standard vs. standard. In the final step, these data were then correlated with the phonetic distance measures and interpreted within a dialectal context.

Results

The analysis of the first experiment showed a higher EER for the across-condition comparison than for the within-condition comparison: 0.00% for the dialect vs. dialect comparison, 0.66% for the standard vs. standard comparison and 2.68% for the standard vs. dialect comparison. Note however,

that these EER values cannot be directly compared; the audio selections used for the within-condition comparisons were extracted from the same audio-recording, whereas the selections for the across-condition comparisons came from different recordings. Nevertheless, the across-condition EER proves that VOCALISE is robust against intra-speaker dialect-variation. Results of the second experiment indicated no significant correlation between comparison scores and phonetic distance measurements, which implies that intra-speaker dialectal mismatch did not affect the software's performance.

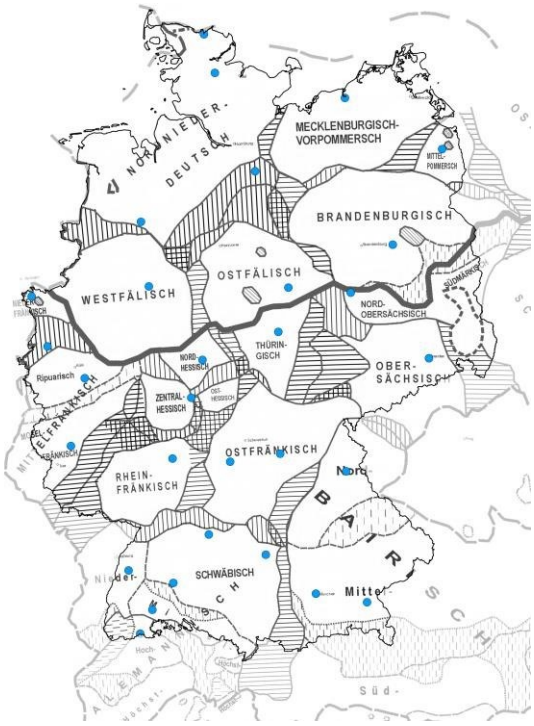


Figure 1. Places of origin of the informants (blue) and main regional dialectal areas of German (created with <www.regionalsprache.de> using a dialect classification map by Wiesinger 1983 (Schmidt et al. 2008ff)).

References

Kehrein, R. (2012). *Regionalsprachliche Spektren im Raum. Zur linguistischen Struktur der Vertikale*. Stuttgart: Steiner.

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

Oxford Wave Research (2023a): Bio-Metrics. Performance Metrics Software. [https://oxfordwaveresearch.com/products/bio-metrics/].

Oxford Wave Research (2023b): VOCALISE. Automatic Speaker Recognition Software. [https://oxfordwaveresearch.com/products/vocalise/].

Schmidt, J. E., Herrgen, J., Kehrein, R., Lameli, A., Fischer, H. (eds.) (2008ff). Dialekteinteilung nach Wiesinger (1983). Abzeichnung der Karte 47,4 in Wiesinger, Peter (1983). Die Einteilung der deutschen Dialekte. In: Besch, Werner [u.a.] (eds.*): Dialektologie. Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. 2. Halbband (= Handbücher zur Sprach- und Kommunikationswissenschaft. 1.2). Berlin/New York. [Published in URL: <www.regionalsprache.de>].

Schmidt, J. E., Herrgen, J., Kehrein, R., Lameli, A., Fischer, H. (2020ff). Regionalsprache.de (REDE III). Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Bearbeitet von Robert Engsterhold, Heiko Girnth, Simon Kasper, Juliane Limper, Georg Oberdorfer, Tillmann Pistor, Anna Wolańska. Unter Mitarbeit von Dennis Beitel, Milena Gropp, Maria Luisa Krapp, Vanessa Lang, Salome Lipfert, Jeffrey Pheiff, Bernd Vielsmeier. Studentische Hilfskräfte. Marburg: Forschungszentrum Deutscher Sprachatlas.

Voice parade guidelines: what happened since?

Gea de Jong-Lendle
Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg,Germany
gea.dejong@staff.uni-marburg.de

Background: variables affecting voice parade reliability

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 knowledge and can be a time consuming and costly process. 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 improves earwitness accuracy and minimises witness error. The reliability of voice parades depends on a number of factors. Wells (1978) in his overview of eyewitness-testimony research distinguished two types of variables: estimator and system variables. The first type concerns variables that affect eyewitness accuracy but are not under the control of the criminal justice system. Examples of estimator variables are severity of the crime, exposure duration, characteristics of the victim like memory retrieval skills or identification ability, or characteristics of the defendant like distinctiveness. System variables are under the direct control of the criminal justice system. Examples are retention interval, lineup structure (e.g. functional size versus nominal size), instructions to the witness, etc. It is these guidelines that may lead to optimal settings of these system variables being used, so that the witness is offered the best possible chance of being accurate and the innocent suspect the best chance of not being selected from the parade. Guidelines also provide the scientists or police officers involved with useful instructions; there are many issues to consider and in the case of an unfamiliar voice, time is an important factor (deJong-Lendle et al. 2015). Thirdly, guidelines are crucial in the judicial process: they help to increase the number of lineups that are appropriately constructed and conducted and by doing so they may reduce the number of costly appeals.

A comparison of voice lineup guidelines

In the past police forces have relied primarily on guidelines for visual lineups. Although the purpose of both types is the same, the requirements partially differ. In the 90s the first guidelines for voice parades were published (Broeders and Van Amelsfoort 1999) or papers with suggestions for guidelines (Hollien et al. 1995, Hollien 1996, Nolan and Grabe 1996). Guidelines for international organisations like IAFPA or ENFSI however, have not yet materialised, although attempts to construct guidelines for IAFPA were initiated by appointing special committees (Hollien et al. 1995). In the meantime however, guided by research and by legal requirements the early guidelines have made place for guidelines that are significantly more detailed and advanced. This talk provides a general review of the different guidelines and recommendations for voice parades worldwide, focusing in more detail on those that are most detailed and established – the guidelines used in The Netherlands, Germany and The United Kingdom.

Although the general idea behind these three guidelines is the same - improving earwitness accuracy and minimising witness error – they disagree on quite a number of points. Table 1 shows the most important differences using, whenever possible, the original guideline text.

	NL-G	**DE-G**	**UK-G**
Number of foils	At least 5 foils + suspect	5 foils + suspect	8 foils + suspect
Duration of the stimulus	Ca. 20 sec.	Ca. 60 sec.	Ca. 60 sec.
Playing the samples	The samples are played *sequentially*: one by one. They are played *only once*. As soon as the witness hears the perpetrator,	The samples are played *sequentially*: one by one. Each sample can be heard *as often as the witness*	The samples are played *serially*. The witness is instructed to listen to each tape [i.e. sample] at least once

	he/she will have to say that and the recording is briefly halted. After that the recording continues until all samples have been placed.	*requires.* A decision has to be made after each sample.	before he/she makes a selection. The witness is allowed to listen to any or all the samples *as many times* as they wish.
Instructions to the witness	If you recognize the voice of *the person you had in mind*, you should inform the police officer who is playing the voices to you at once. Afterwards you hear some more voices. *The voice you heard at the time* may not be included in the parade. If you have doubts about whether you recognize a person, you should not point out anyone. (Furthermore, the witness: 1) is not informed about the number of samples, 2) is not put under pressure by stating that this parade is important, and 3) is not encouraged).	The *criminal* may or may not be part of the parade. Select a sample only then when you believe you have heard the criminal. Each time you have heard a voice, please make your decision.	The witness must be instructed that: 1.) the voice of the *suspect* may or may not be in one of the samples played during the procedure. 2.) he/she must listen to each tape [i.e. sample] at least once before he/she makes a selection.

Table 1. Showing the most important differences between voice line guidelines used in The Netherlands, Germany and The United Kingdom.

The consequences of these differences in terms of voice parade reliability will be discussed and their motivation explained. Finally, where possible, suggestions for standardization will be provided.

References

van Amelsvoort, A.G. (2018). *Handleiding confrontatie* (10e herziene druk). Den Haag: Sdu.

Gfroerer, S. & Jessen, M. (2021). Sprechererkennung und Tonträgerauswertung. In: Eckhart Müller/Reinhold Schlothauer/Christoph Knauer (Hrsg.) *Münchener Anwaltshandbuch Strafverteidigung* (3. überarbeitete Auflage). München, Beck, 2862-2890.

Hollien, H. (1996). 'Consideration of guidelines for earwitness line-ups', Forensic Linguistics 3(1): 14–23.

Hollien, H., R. Huntley, H. Künzel, and P. A. Hollien (1995). 'Criteria for earwitness line-ups', Forensic Linguistics 2(2): 143–53.

Home Office, UK (2003) Advice on the Use of Voice Identification Parades. Circular 057/2003. London: Crime Reduction and Community Safety Group, Police Leadership and Powers Unit.

de Jong-Lendle G., Nolan F., McDougall K. & Hudson T. (2015) Voice lineups: a practical guide. *Conference-paper at the 18th International Congress of Phonetic Sciences*. Glasgow, Scotland, 10–14 August 2015.

Nolan, F., and E. Grabe (1996). 'Preparing a voice lineup', Forensic Linguistics 3(1): 74-94.

Wells, G. L. (1978). Applied eyewitness testimony research: System variables and estimator variables. *Journal of Personality and Social Psychology*, 36, 1546-1557.

Reaction time predicts accuracy in the estimation of speaker origin

Adrian Leemann[1], *Carina Steiner*[1], *Péter Jeszenszky and Yara Miescher*[1]
[1]*Department of German and Center for the Study of Language and Society, University of Bern*
adrian.leemann|carina.steiner|peter.jeszenszky|yara.miescher@unibe.ch

Introduction

Swiss German speakers talk in dialect on a daily basis. They only "switch" to standard in formal situations like official statements or teaching – this variety is typically referred to as Swiss Standard German. Guntern (2011) found that listeners can identify a speaker's regional origin even when they speak Swiss Standard German. One aspect that has been under-researched and which bears relevance for forensics, however, is the question of how reaction time in guessing a speaker's origin is linked with identification accuracy. In the same way that more offender material improves voice line-up accuracy (Künzel 1990), we expected slower reaction times (i.e., listening to more material) to improve accuracy in the identification of a speaker's origin.

Methods

For this study, 25 speakers from across German-speaking Switzerland were selected from the SDATS database (Leemann et al. 2020). On average, they were 69 years old and read off a text in dialect ($\bar{X}$:165s) and in Swiss Standard German ($\bar{X}$: 38s). Twelve listeners from Solothurn (six between 20-30yo and six 70+yo) listened to the stimuli, tasked with guessing the speaker's origin. Listeners took part in two conditions, with a 2-month break in between: in the first condition, dialect stimuli were rated, whereby participants placed a pin on a Google Map using an iPad. Participants were asked to place the pin as soon as they knew where the speaker was from. In the second condition, the same procedure was applied to the Swiss Standard German stimuli. Latitude and longitude were retrieved using Google Maps. Accuracy was determined by calculating a difference score in structural linguistic similarity between actual guessed place of origin (cf. Scherrer 2021).

Results

Overall, results revealed that accuracy was better in the dialect condition ($0.12(\pm 0.01)$, $t=9.3$, $p<0.001$). When looking at the reaction times, an intriguing pattern emerged, see Figure 1 (dialect left panel, Swiss Standard German right panel): the error on the Y-axis denotes linguistic difference between actual and guessed place of origin; the higher the value, the greater the linguistic difference, the greater the error. In the dialect condition, the error is smallest with little material, e.g., 1-5s. When following the trajectory of reaction time vs. error (red line), more material typically means lower accuracy. This error peaks at ~80s. The trend for the Swiss Standard German stimuli was even more pronounced: the longer participants listened to the stimulus, the poorer the accuracy in the identification of the speaker's origin. A 'sweet spot' is achieved, again, with very little material at around 1-5s of stimulus exposure.

Discussion

Results corroborate Guntern's (2011) findings and point towards a link between accuracy and 'gut feeling': when listeners trust their intuitive response, accuracy is highest. In fact, when listeners overthink the regional origin of a speaker, accuracy decreases, especially in Swiss Standard German. This bears implications for forensics, where offender material is often short: lay listeners' guesses of origin may be high even with only a few seconds of speech and a listener's guess does typically not improve the longer the voice was heard.

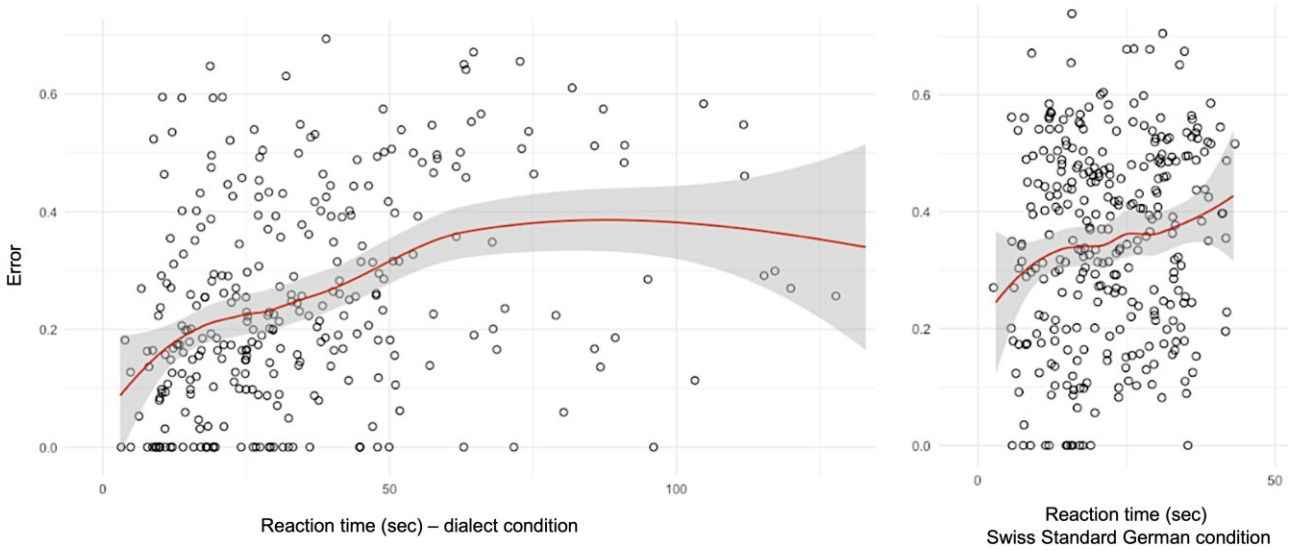


Figure 1. Reaction time (sec.) vs. guessing error: dialect stimuli (left panel) versus Swiss High German (right panel).

References

Guntern, M. (2011). Erkennen von Dialekten anhand von gesprochenem Schweizerhochdeutsch. Zeitschrift für Dialektologie und Linguistik, 155-187.

Künzel, H. J. (1990). Sprechererkennung durch linguistisch naive Personen, ZDL Monographs No. 69, Stuttgart, Steiner-Verlag.

Leemann, A., Jeszenszky, P., Steiner, C., Studerus, M., & Messerli, J. (2020). Linguistic fieldwork in a pandemic: Supervised data collection combining smartphone recordings and videoconferencing. Linguistics Vanguard, 6(s3).

Scherrer, Yves (2021): dialektkarten.ch - Interactive dialect maps for German-spea- king Switzerland and other European dialect areas. In T. Krefeld, S. Lücke, & C. Mutter (Eds.), Berichte aus der digitalen Geolinguistik (II): Akten der zweiten Arbeitstagung des DFG-Langfristvorhabens VerbaAlpina und seiner Kooperationspartner am 18.06.2019. Munich: Korpus im Text, University of Munich.

Predicting a face from a voice and a voice from a face: the effect of expressive audio-visual information on cross-modal identity matching

Elisa Pellegrino[1], *Enrico Varano*[2], *Alexis Hervais-Adelman*[2], *Nadine Lavan*[3] *and Volker Dellwo*[1]

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

[2] *Department of Psychology*, *University of Zurich, Switzerland, CH*
{enrico.varano | alexis.hervais-adelman}@psychologie.uzh.ch

[3] *School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom, UK*
n.lavan@qmul.ac.uk

Previous research demonstrated that next to audio information visual aspects of the speech articulators (in particular the lips) play a key role in voice recognition. More intriguingly, individuals can match the identity of an unfamiliar voice across these modalities (henceforth: cross-modal identity matching), however, their performance is far from perfect. Surprisingly, cross-modal identity matching has only been tested in rather non-expressive adult communication. It seems plausible that the presentation of more expressive facial and voice information might enhance cross-modal identity matching. The present project is designed around testing this hypothesis.

We will collect an audio-visual corpus of infant-directed speech, a speaking style characterized by exaggerated prosody and exaggerated facial expressions. We will video-tape mothers reading to their own baby (expressive speech) and to an adult in a formal situation (less-expressive speech). To isolate visible facial movements for cross-modal identity matching, for each video clip we will produce point-light video and fully illuminated displays. We will present observers either a face together with the corresponding and a non-corresponding voice or a voice together with two faces (corresponding and non-corresponding). Their task will be to pick the correct match across the modalities. We will run three perception experiments. In Experiment I we will test the influence of expressive cues on voice face matching, in Experiment II we will use point-light faces to test the effect of the lack of static facial information and in Experiment III we will test the importance of rhythmic cues of speech (using sine-wave speech) on matching a face to a voice.

The results will explain the role of expressive information and the mechanisms behind cross-modal identity matching. We will also determine whether visible or audible articulatory information alone (point-light faces vs. sine-wave speech) contain sufficient information. Cross modal identity matching represents one of the challenging targets for automatic speech recognition systems. We expect that our results will contribute to the theoretical understanding of such human abilities, which can be used to create more robust speaker recognition algorithms. Such applications are particularly relevant in the field of Forensic Phonetics when evidence for criminal investigations is available either in the visual or in the auditory modality

References

Green, J. R., Nip, I. S. B., Wilson, E. M., Mefferd, A. S., & Yunusova, Y. (2010). Lip movement exaggerations during infant-directed speech. *Journal of Speech, Language, and Hearing Research : JSLHR*, *53*(December 2010), 1529–1542.

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). “Putting the Face to the Voice”: Matching Identity across Modality. *Current Biology*, *13*(19), 1709–1714. Journal Article. http://doi.org/10.1016/j.cub.2003.09.005

Lander, K., Hill, H., Kamachi, M., & Vatikiotis-Bateson, E. (2007). It's not what you say but the way you say it: matching faces and voices. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 905–914.

Vocal Profile Analysis as a Tool in Cross-Language Forensic Speaker Comparison

Kristina Tomić[1] *and Peter French*[2]
[1]*Faculty of Philosophy, University of Niš, Serbia*
kristinatomic89@hotmail.com
[2]*Department of Language and Linguistic Science, University of York & JP French Associates, Forensic Speech and Acoustics Laboratory, York, UK.*
peter.french@york.ac.uk

The "identifiability" of a person's voice derives from a combination of individual biology and learning. A major contribution to identifiability is made by the physical dimensions and configuration of the organs within and around the larynx and the vocal tract. Another is made by the phonatory and supralaryngeal settings of those organs the speaker habitually adopts, either by dint of individual preference or as a result of their socialisation within a particular sociolinguistic community (Laver, 1980). Whilst we cannot entirely separate or quantify the relative contributions of nature and nurture, we can nevertheless capitalise on their combined product, i.e., voice quality, when undertaking forensic speaker comparison (FSC) casework. A tool for so doing is the Vocal Profile Analysis (VPA) scheme (Laver et al., 1981), which was designed to capture "long-term-average" phonatory and vocal tract adjustments that persist throughout utterances.

According to two international surveys of forensic practices (Gold & French, 2011; San Segundo, 2021), many forensic practitioners routinely perform perceptual analysis of voice quality in their casework following the VPA or a similar protocol. Recently, San Segundo et al. (2019) proposed a methodological framework for the successful application of the VPA protocol in forensic speaker characterisation using a modified 32-feature version of the original scheme. In addition, some contemporary studies have confirmed that voice quality can corroborate other forensic analyses, including MFCC-based ASR (see Cardoso et al., 2019; French et al., 2015; Gonzalez-Rodriguez et al., 2014; Hughes et al., 2017).

The Present Study

The present research explores VPA reliability in speaker characterisation across languages. Twenty female native speakers of Serbian were recorded over a mobile phone in a spontaneous-speech task in Serbian and English. The 40-second long recordings were rated using a truncated version of VPA that included 27 articulatory and phonatory settings (cf. San Segundo et al., 2019) by three trained experts[3]. In addition, to explore the relationship between voice quality and foreign language proficiency, all participants took a mock language proficiency test by British Council[4], and the recordings were subject to an IELTS-based proficiency scoring by an ESL expert (Author 1).

Preliminary Results

The preliminary results based on 10 speakers and two raters suggest a strong inter-rater agreement in the VPA scores (Gwet's AC2 = .726, SE = .015). "True scores" were obtained by calculating the median of the scores by individual raters, and Euclidean distances and cosine similarity were calculated for same-speaker (cross-language) and different-speaker (same-language) pairs. Paired t-test comparisons of averaged distances suggest that between-speaker distances are higher in the foreign language than in the mother tongue (t = 4.079, p - .003), whereas between-speaker and within-speaker (cross-language) similarities and distances do not exhibit a significant difference.

Correlation statistics revealed a negative association between VPA similarity and language proficiency estimated via both methods (VPA similarity-test: r - .699, p - .025; VPA similarity-

[3] We would like to express our deepest gratitude to Katharina Klug, Dr Jessica Wormwald, and doc. Dr Radek Skarnitzl for undertaking the comprehensive task of scoring the participants on the VPA protocol.

[4] Online English Level Test (British Council) https://learnenglish.britishcouncil.org/english-levels/online-english-level-test

IELTS-based score: r - .669, p – .034), indicating that the less proficient the speaker is, the higher vocal profile similarity across two languages.

Preliminary Conclusion

A preliminary conclusion is that the VPA scheme may have limited application in cross-language forensic speaker comparison, provided that the speaker displays a "stronger" foreign accent and lower proficiency. Additionally, the accent-specific articulatory adjustments may outweigh the individual differences. Namely, there are indications that a non-neutral adjustment in the specific cluster of settings (backed tongue body, raised tongue body, extensive tongue range, pharynx constriction, raised larynx and creaky voice) is assumed by most of the analysed speakers. No correlation was found between averaged between-speaker distances/similarities across Serbian and English, which suggests that speakers whose vocal profile deviates the most from the population in the mother tongue do not necessarily exhibit equal deviation in the foreign language. Such a result confirms the hypothesis that the degree of foreign accent is crucial in maintaining vocal features across languages. The hypothesis remains to be tested on a larger dataset.

References

Cardoso, A., Foulkes, P., French, P. J., Gully, A. J., Harrison, P. T., & Hughes, V. (2019). Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS), Melbourne, Australia.*

French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E., & Stevens, L. (2015). The vocal tract as a biometric: output measures, interrelationships, and efficacy. *Proceedings of the 18th International Conference of Phonetic Sciences (ICPhS), Glasgow, Scotland.*

Gold, E., & French, P. (2011). International Practices in Forensic Speaker Comparison. *The International Journal of Speech Language and the Law, 18*(2), 293-307. https://doi.org/10.1558/ijsll.v18i2.293

Gonzalez-Rodriguez, J., Gil, J., Perez, R., & Franco-Pedroso, J. (2014). What are we missing with ivectors? A perceptual analysis of i-vector-based falsely accepted trials. *Proceedings of Odyssey 2014: The Speaker and Language Recognition Workshop. Joensuu, Finland*, (pp. 33-40). https://doi.org/10.21437/Odyssey.2014-6

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017). The complementarity of automatic, semi-automatic, and phonetic measures of vocal tract output in forensic voice comparison. *A paper presented at the 26th Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Split, Croatia.*

Laver, J. (1980). *The Phonetic Description of Voice Quality.* Cambridge University Press.

Laver, J., Wirz, S., Mackenzie, J., & Hillier, S. M. (1981). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress, 14*, 139-155.

San Segundo, E. (2021). International survey on voice quality: Forensic practitioners versus voice therapists. *Estudios de Fonética Experimental, 29*, 8-34.

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2019). The use of the Vocal Profile Analysis for speaker characterization: Methodological proposal. *Journal of the International Phonetic Association, 49*(3), 353-380. https://doi.org/10.1017/S0025100318000130

Filler particles and pausing behaviour in Egyptian Arabic

Beeke Muhlack and Omnia Ibrahim
Department of Language Science and Technology, Saarland University (Germany)
[muhlack|omnia]@lst.uni-saarland.de

The research area of disfluencies and filler particles has gained interest across many languages in the last decades. However, languages that are not part of the Indo-European language family are considerably under-researched. Disfluency research in healthy adults in Arabic is, to the best of our knowledge, a research gap. This study delivers a first puzzle piece that advances the way into filling this research gap, starting out with a closer look into filler particles and pausing behaviour in the Egyptian dialect. As the Arabic language is the fourth most spoken language in the world, with an estimated number of 400 million speakers distributed over 23 countries (Bateson 2003), disfluency research in this language is highly necessary. As previous research for British English suggests, disfluency patterns may be speaker-specific and thus, could aid as a feature in forensic phonetic casework (McDougall & Duckworth 2018).

Data: The data used for this study is a subset taken from Ibrahim et al. (2020)'s corpus on speech rhythm in Arabic. In the current study, we extracted the spontaneous data of 7 Egyptian Arabic speakers, 4 males and 3 females. Each speaker performed 2 tasks: In the first task, speakers talked freely for more than one minute about their daily life, while in the second task, the speakers were asked to describe the directions to go to the university from a famous nearby location using a map as visual aid. Both tasks by all speakers amounted to a total of 19 minutes of speech.

Lexical (*well, you know*) as well as non-lexical filler particles (*uh, um, hm*), lengthenings, and repetitions have been annotated in the files by a native speaker of Egyptian Arabic (second author) with the aim to present the inventory of disfluencies used in the Arabic dialect and investigate speaker-specific patterns. Furthermore, formant measurements are taken from the non-lexical vocalic filler particle (*uh*) at the midpoint of the vowel to assess the vowel quality.

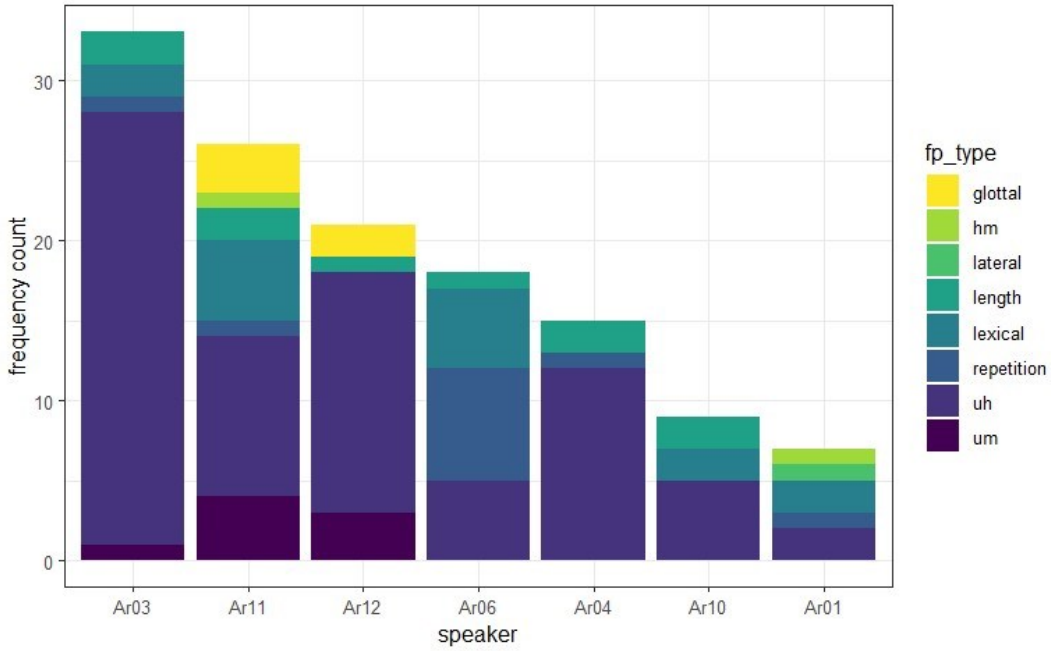


Figure 1: Disfluency patterns of 7 Arabic speakers pooled over both speech tasks.

Results: 129 filler particles and 331 silent pauses have been found in the dataset which suggests that Arabic speakers of this dialect prefer non-vocalised pauses to vocal filler particles. The preferred filler particle is the vocalic *uh*, other particles or other disfluencies are used to a lesser extent. However, speakers seem to show individual preferences (see Figure 1). A comparison of both tasks may be fruitful to explore speaker-specific patterns and the influence of cognitive load.
Cross-language comparisons of the vowel qualities of Arabic filler particles with those produced in L1 German, English, and Spanish (Muhlack et al. 2023, Muhlack forthcoming), show that Arabic hesitation vowels overlap with those produced in German to a higher degree than English and Spanish in a two-dimensional vowel space (Figure 2).

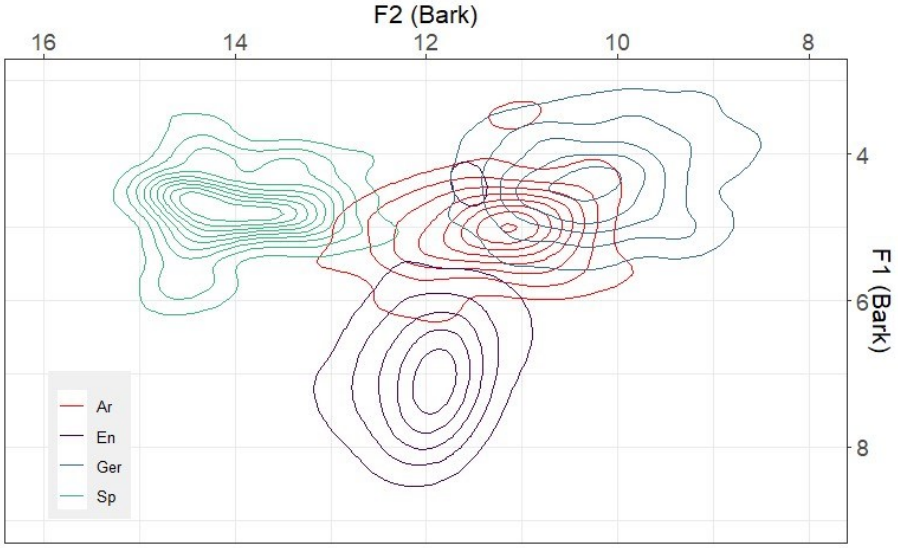


Figure 2: Vowel quality of Arabic (red) filler particles compared to filler particles produced in English (purple), German (blue), and Spanish (green) spontaneous speech.

References

Bateson, M. C. (2003). *Arabic Language Handbook*. Washington, Georgetown University Press.

Ibrahim, O., Asadi, H. Kassem, E., & Dellwo, V. (2020). Arabic Speech Rhythm Corpus: Read and Spontaneous Speaking Styles. *Proc. of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, pp. 5337–5342.

McDougall, K. & Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *Journal of Speech, Language and the Law*, 25(2), pp. 205-230.

Muhlack, B. (forthcoming). Filler particles in English and Spanish L1 and L2 speech. *Proc. 20th International Congress of Phonetic Sciences (ICPhS '23)*, Prague.

Muhlack, B., Trouvain, J., Jessen, M. (2023). Distributional and Acoustic Characteristics of Filler Particles in German with Consideration of Forensic-Phonetic Aspects. *Languages*, 8(2), 100.

The possibilities that come with using whole voice comparison processes in voice comparison research

Georgina Brown[1,2] *and Christin Kirchhübel*[2]
[1]*Department of Linguistics and English Language, Lancaster University, Lancaster, UK*
g.brown5@lancaster.ac.uk
[2]*Soundscape Voice Evidence, Lancaster, UK*
ck@soundscapevoice.com

The auditory-phonetic and acoustic approach (AuPhA) to forensic voice comparison casework involves an analyst making qualitative and quantitative observations about a range of different voice and speech features. In reaching a voice comparison conclusion, the analyst takes into account the whole collection of qualitative and quantitative findings. Both the analysis and the conclusion are a product of the decisions the analyst has made throughout the process. Often, forensic phonetics research focuses on a small number of features in isolation and remarkably little research replicates the more complex process that is AuPhA. This project aims to produce results and findings that are derived from a whole analysis pipeline that could more realistically be applied to forensic voice comparison casework.

60 voice comparison trials were carefully constructed using the NIST-SRE 2006 (Przybocki et. al., 2007) and Pool 2010 (Jessen et. al., 2005) datasets which present a range of telephone-style recordings containing spontaneous speech. Specifically, pairs of male same-speaker and different-speaker recordings were chosen by an experienced forensic practitioner to ensure that they reflected the types of recordings that surface in casework. The practitioner purposefully selected recordings that featured mismatches of different types (e.g., vocal effort, distance, recording quality). Importantly, the pairs of recordings did not feature obviously similar or obviously different voices that made any of the trials "too easy".

20 of the trials feature English speech, 20 feature Mandarin speech and 20 feature German speech. The intention of incorporating three languages was to enable more research into "foreign-language cases" where the language spoken in the recordings is not one that is spoken by the forensic analyst. Foreign-language case enquiries often arise in the UK, but providers generally decline to take on these cases because they are not equipped to handle them. This has resulted in interpreters (who were not voice comparison experts) giving voice comparison opinions (R v Tamiz [2010] EWCA Crim 2638). In many other cases, the consequence is that the speech material does not get forensically analysed at all.

The 60 voice comparison trials are in the process of being analysed using an AuPhA protocol as well as an automatic speaker recognition system (developed using Kaldi). The analyst carrying out these trials is blind to the ground-truth. The results of these analyses are anticipated to:

1) Develop competence in applying AuPhA alongside an automatic speaker recognition system, ultimately arriving at an analysis protocol that combines these methods;
2) Test a modified conclusion framework for voice comparison casework (FSR-C-118);
3) Test analysis protocols on foreign-language recordings;
4) Create training and testing material.

References

Forensic Science Regulator Codes of Practice and Conduct. Development of Evaluative Opinions. FSR-C-118. Issue 1. 2021.

Jessen, M., Köster, O. and Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech, Language and the Law. 12. pp 174-213.

Przybocki, M., Martin, A. and Le, A. (2007). NIST Speaker Recognition Evaluations Utilizing the Mixer Corpora – 2004, 2005, 2006. IEEE Transactions on Audio, Speech and Language Processing. 15. pp 1951-1959.

Within-speaker fundamental frequency variations in bi-dialectal speakers: The case of Mandarin and Danyang dialect

Yu Zhang, Lei He and Volker Dellwo
Department of Computational Linguistics, University of Zurich, Zurich, Switzerland
{yu.zhang|lei.he|volker.dellwo}@uzh.ch

Fundamental frequency (F0) differences have been the focus of interest in a number of voice studies to explore the effect of language on the acoustic aspects of speech acoss different languages and speech groups (Altenberg and Ferrand, 2006; Andrianopoulos et al., 2001; Awan and Mueller, 1996; Baken and Orlikoff, 2000; Gelfer and Denor, 2014; Hanley et al., 1996; Jarvinenet al., 2013; Keating and Kuo, 2012; Lee and Sidtid, 2017; Ng et al.,2012; Ordin and Mennen, 2017; Sapienza, 1997; Todaka, 1993; van Bezooijen, 1995). However, these studies failed to reach a consensus in their findings of language differences in F0 statistics across different languages and language groups. Disparities might stem from methodological differences and limitations (difference languages involved in comparison, isolated vowels v.s connected speech; different proficiency levels of languages among speakers, etc.). As a special case in point, it is worthnoting that previous studies generally approbated the notion of a normative higer F0 and larger F0 range in tonal languages than in non-tonal languages(Andrianopoulos et al., 2001; Eady, 1981; Keating and Kuo, 2012). In tonal languages, Fo variations are important at the phonemic level for realising lexical contrast, and thus the Fo patterns are determined mainly by the tone contours of all the lexical items in a sentence. The F0 patterns of tonal language speakers display a greater dynamic fluctuations as a function of time. However, contrasting results were still obtained in different studies and it seems that language proficiency is a factor at play (Lee and Sidtid, 2017; Ng et al., 2010; Ng et al., 2012). Inconsistencies in the findings suggest that it is still unclear whether observed differences in F0 measurements can be attributable to anatomical factors or cross-linguistic differences, and studies examining the voice profiles of balanced bilingual speakers when speaking two phonologically different but culturally intertwined languages are lacking.

The current study looked at the F0 variations within a group of bi-dialectal speakers who are native speakers of both Mandarin and Danyang dialect (henceforth “Dialect”). Danyang dialect is a northern Wu dialect which contrast six lexical tones in it’s spoken form and is widely used daily in the city of Danyang(Lü, 1991). Mandarin has four tones and is the standard Chinese variety that is used natively in both formal and informal settings in Danyang area. A total of 14 speakers (9 females and 5 males) were recruited to participated in the present study. All of the selected participants are native speakers of both Mandarin and Danyang dialect and use Mandarin and Danyang dialect equally extensively at home. All participants performed a sentence-reading task and a passage reading task in each of their spoken languages. In the present study only the sentence-reading materials are used. The present study builds on a relatively new approach to measuring holistic F0 variations that computing how fast the direction of F0 contour changes utterance-wise, i.e. the wigglyness of the F0 contour. F0 contours were extracted from the recorded speech samples using Praat (Boersma, 2021) with a standard range setting of 75–600 Hz and interpolation was further implemented to smooth the F0 contours acquired. Z-score transformation was applied to each utterance to mitigate the influence of massive outliers. We are interested in how cross-dialect variability can be captured by the wigglyness of the F0 contour as a function of time, which can be quantified by calculating the integral of the squared second derivative($\int ((d^2\ y)/(d\ x^2))^2$) of pitch per utterance (hearafter integral_sqr_der). To test the significance of cross-dialect variability captured by the integral_sqr_der, mixed-effects models were employed using the R package lmer (ref.). Language was modeled as the fixed factor, language and speaker as a random slope, and sentences as a random intercept. Results indicated a significant global effect of language ($p=0.034$). Danyang dialect exhibited a significantly greater F0 variations compared with Mandarin, which is in line with previous studies on the effect of tones on

fundamental frequency (Chen, 2005). To further assess possible language effect on holistic F0 variations within each speaker, we performed paired sample t-test (Bonferroni corrected). Results show that 11 out of 14 speakers didn't show significant difference across these two dialects in terms of holistic F0 variations.

Results suggest that the intrinsic acoustic features of each language may have influenced the vocal parameters in overall speech, but individual anatomy still serves as a physiological foundation for acoustic outcomes. Implications for FSC is that acoustic norms should be selected carefully for speakers when different languages are involved.

References

Altenberg, E.P., & Ferrand, C.T. (2006). Fundamental frequency inmonolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice*, 20:89–96

Andrianopoulos, M.V., Darrow, N.K., & Chen, J. (2001). Multimodalstandardization of voice among four multicultural populations: fundamental frequency and spectral characteristics. *Journal of Voice*, 15: 259–272.

Awan, S.N.,& Mueller, P.B.(1996). Speaking fundamental frequency characteristics of white, American, and Hispanic kindergarten-ers. *Journal of Speech,Hearing, and Research*, 39: 573–577.

Baken, R., & Orlikoff, R. (2000). *Clinical measurement of speech andvoice.* Singular Thomson Learning.

Eady, S.(1982). Differences in the F0 patterns of speech: tone language versus stress language. *Language and Speech*, 25:29–42.

Gelfer, M., & Denor, S. (2014). Speaking fundamental frequency andindividual variability in Caucasian and African-Americanschool-aged children. *American Journal of Speech-Language Pathology*, 23(3): 395–406.

Hanley, T.D., & Snidecor, J.C. & Ringel, R.L.(1966). Some acoustic differences among languages. *Phonetica*, 14: 97–107.

Jarvinen, K., Laukkanen, A., & Aaltonen, O. (2013). Speaking a foreign language and its effect on F0. *Logopedics Phoniatrics Vocology*,38: 47–51

Keating, P., & Kuo, P. (2012).Comparison of speaking fundamental fre-quency in English and Mandarin. *Journal of Acoustic Society of America*, 132: 1050–1060.

Lee, B., & Sidtis, D. (2017). The bilingual voice: Vocal characteristics when speaking two languages across speech tasks. *Speech, Language and Hearing*, 20:3, 174-185.

Lü, S. X. (1991). The Phonology of Danyang Dialect. Yu Wen Press.

Ng, M., Chen, Y., & Chan, Y.K. (2012). Differences in vocal characteristics between Cantonese and English produced by Cantonese-English bilingual speakers–a long-term spectral analysis. *Journal of Voice*, 26: e171–e176.

Ordin M, & Mennen I. (2017). Cross-Linguistic Differences in Bilinguals' Fundamental Frequency Ranges. J Speech Lang Hear Res., 60(6), 1493-1506.

Boersma, P., & Weenink, D. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.54, retrieved 3 May 2023 from http://www.praat.org/

Todaka, Y.1993. A cross-language study of voice quality: bilingual Japanese and American speakers. Los Angeles: Doctoral dissertation, University of California.

Sapienza, C.M. (1997). Aerodynamic and acoustic characteristics of the adult African American voice. J*ournal of Voice*, 11: 410–416.

van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch women. *Language and Speech*,38(3): 253–256.

Speaker Diarization Systems in the Context of Forensic Audio Analysis

David Grünert[1], *Alexandre de Spindler*[1] *and Volker Dellwo*[2]
[1]*ZHAW Zurich University of Applied Sciences, Switzerland*
{grud,desa}@zhaw.ch
[2]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
volker.dellwo@uzh.ch

Forensic audio does not seldom consist of long recordings of multiple speakers engaged in a dialogue. An important task for forensic phoneticians is then to say (a) how many speakers are present in the recording and (b) who speaks when. This can also lead to the identification of audio segments that are relevant for detailed inspection. Speaker diarization provides the fundamental ability to automatically split audio streams into segments assigned to speakers (Tranter, 2003). Since current diarization systems do not require audio profiles of speakers and do not assume any given number of speakers, they support a wide range of applications. In this paper, we present our approach to leverages speaker diarization to support forensic audio analysis. This also implies the identification of challenges for diarization systems in this context and the necessity of a novel evaluation metric.

Applications

A common task in forensic audio analysis is to find audio segments that are relevant for detailed inspection. Nowadays, speaker identification, domain recognition, and physiological state analysis of speakers support the identification of such segments (Pathak 2020). In our work we try to show that detecting communication structures and their development can help to identify additional segments of interest. A segment of interest may be where a communication structure changes, e.g., from the structure in sequence 1 to the structure in sequence 2 (c.f., figure below). Furthermore, a structural change found in this way can serve as a template to search for similar occurrences in an audio collection.

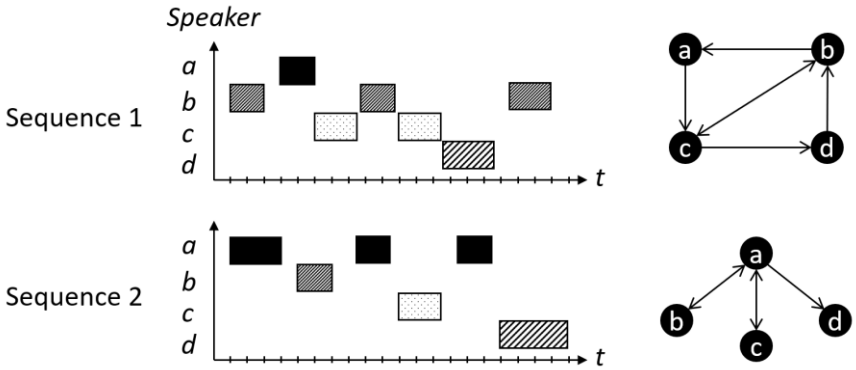
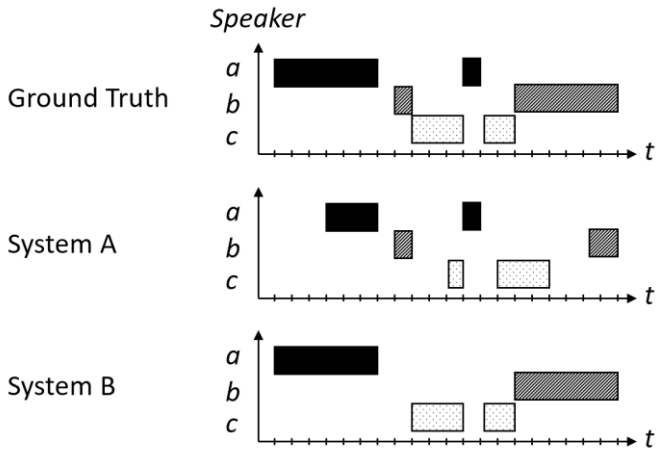


Figure 1. Diarization output (left) and derived communication structures (right) for two sequences involving four speakers.

Evaluation

It is unclear whether the performance of existing diarization systems is sufficient for such forensic applications. While the challenge of audio being recorded in a variety of environments with different acoustic properties and background noises has been addressed in recent evaluations (Ryant, 2021) the detection of changes in communication structure requires evaluation data containing such changes. In our work, we analyze exiting data sets regarding their test coverage for changes in communication structures.

Another challenge for the evaluation is caused by the metrics typically used such as the Diarization error rate (DER) (Fiscus, 2006) and the Jaccard error rate (JER) (Ryant, 2019). DER measures the overall performance by calculating the ratio between the sum of all errors to the total duration. In contrast, JER computes the average of a per-speaker error rate. In our work, we have observed that these metrics are only appropriate if the system is used to analyze the overall participation of all speakers. In contrast, if the communication structure is analyzed, we noticed that many of the errors detected by DER and JER can be tolerated if the sequence of contributions is recognized correctly. Figure 2 illustrates this with a simplified scenario including the output of two diarization systems. The output of System A contains many errors leading to high DER and JER, but it correctly recognizes the sequence. Output of System B contains only two very short errors leading to a much lower DER and JER, but it fails to detect the sequence.



System	*DER*	*JER*
A	52.63	57.14
B	10.53	9.52

Figure 2. Simplified scenario with two different diarization systems in a use case with three speakers (left) and corresponding DER and JER values computed with dscore (Ryant, 2019).

Next Steps

The proposed application of speaker diarization systems and their evaluation needs further investigation. We will analyze the existing evaluation data for its communication structure, and we will present novel evaluation metrics. Once we have a reliable system, we can test its usefulness for finding relevant segments in forensic audio analysis.

References

Fiscus, J. G. et al. (2006). The rich transcription 2006 spring meeting recognition evaluation. Proceedings of International Workshop on Machine Learning and Multimodal Interaction, pp. 309–322.

Pathak V. et al (2020). A survey of Audio Forensic Techniques Applied on various Encoders. PROTEUS JOURNAL. ISSN/eISSN: 0889-6348

Ryant, N. et al. (2019). The second DIHARD diarization challenge. Proceedings of the Annual Conference of the International Speech Communication Association, pp. 978–982.

Ryant, N. (2019). Dscore (Accessed on May 15, 2023) [Source code]. https://github.com/nryant/dscore

Ryant, N. et al. (2021). The Third DIHARD Diarization Challenge

Tranter, S. E. et al (2003). An investigation into the interactions between speaker diarisation systems and automatic speech transcription, CUED/FINFENG/TR-464.

Language Analysis in the Swiss Asylum System: Towards Inclusive Collaboration and Best Practice

Hannah Hedegard[1], Priska Hubbuch[2] and Simonette Favaro-Buschor[2]
[1]*Department of English, Universität Bern, Switzerland*
hannah.hedegard@unibe.ch
[2]*Lingua, Staatssekretariat für Migration, Switzerland*
{priska.hubbuch|simonette.favaro-buschor}@sem.admin.ch

The procedure known as Language Analysis for the Determination of Origin (LADO), i.e., the corroboration or contradiction of the claimed regional origins of asylum seekers based on systematic linguistic examination of their speech, is a common element of asylum protocol in unclear cases across Europe. LADO is mostly undertaken by either independent commercial entities or government agencies, with linguists and analysts of various educational backgrounds involved.

Amid disagreement on several issues in the field, there is a clear consensus that more LADO-related theoretical and methodological enquiry is urgently required. Fraser notes that "virtually all recent commentators have called for more research to assist analysts in their task of providing LADO evidence", but also highlights the inconsistent manner of the communication between researchers and practitioners: "at another level, written advice and recommendations from linguists…are interpreted with minimal interaction with the authors, sometimes resulting in non-optimal responses" (2019, p.85). How and whether empirical findings are taken up in practise is determined by, amongst other things, their contextual applicability and ongoing communication between the parties involved. In the Swiss context, knowledge transfer has been most productive when specialized linguists were locally engaged for a research task specific to the agency (e.g., McNamara & Schüpbach, 2019). Part of the issue is that, to date, academic exchange has primarily focused on theoretical issues in LADO alone, rather than practical and professional ones that are shared across the forensic sciences e.g., point of engagement in the legal process or navigating time constraints. Furthermore, due to security reasons, discussion and co-operation between LADO agencies and sociolinguists has tended to occur without significant input from the analysts themselves, despite the latter's critical role in, and insight into, the procedure.

In this paper, we present an overview of preliminary work to address these remaining shortcomings and build on past successful inter-disciplinary dialogue in Switzerland: the start of an extended collaborative research project at Universität Bern with *Lingua* (the unit responsible for LADO in the Swiss State Secretariat for Migration). Recounting the organisation and output of a round table connecting *Lingua* analysts, academic experts, and other forensic specialists in 2022, as well as outlining planned in-house research assignments that target agency-specific challenges, we set out how these kinds of activities and close cooperation between practitioners and scholars can further best practice in the LADO context.

References

Fraser, H. (2019). The Role of Native Speakers in LADO: Are We Missing a More Important Question? In: Patrick, P.L., Schmid, M.S., Zwaan, K. (eds) *Language Analysis for the Determination of Origin: Current Perspectives and New Directions*. Language Policy, vol 16. (pp. 71-90) Springer.

McNamara, T., Schüpbach, D. (2019). Quality Assurance in LADO: Issues of Validity. In: Patrick, P.L., Schmid, M.S., Zwaan, K. (eds) *Language Analysis for the Determination of Origin: Current Perspectives and New Directions.* Language Policy, vol 16. (pp. 253-271) Springer.

Machine Assisted Voice Evaluation (MAVE)

Timo Becker[1], *Herbert Masthoff*[2]
[1]*Phonam, Germany*
becker@phonam.de
[2]*Phonam, Germany.*
masthoff@phonam.de

Forensic audio experts are continuously being challenged by the specific constraints of the actual case at hand: voice comparisons, transcriptions, voice profiling et cetera generally require a case by case custom analysis. Even though the analysis may follow some procedural routines, it is by no means a matter of "click the button" to come forward with results. Such routines include the repetition of tasks like annotation, file analysis (file structure analysis and audio metadata analysis), acoustic analysis and audio editing (filtering, cutting, editing). Available off-the-shelf software solutions usually do not fully serve the needs of the forensic experts who often have to resort to more than one program dealing with separate and/or overlapping tasks. When for example applying automatic voice comparison systems, the forensic expert is required to pre-process, edit and analyze the audio files in question manually – usually by employing different programs.

While various such dedicated automatic voice comparison systems to support the forensic expert are available, in the authors' opinion there currently does not exist a comprehensive software solution for forensic audio analysis. Hence, we are in the process of developing a self-contained software solution "Machine Assisted Voice Evaluation (MAVE)" in order to support forensic audio analysis on the basis of "one (hopefully) serves all". MAVE is geared to the practical implications of daily casework and its ongoing challenges. It is based on the R language for statistical computing (R Core Team (2023)) which provides hundreds of algorithms for forensic audio analysis.
Currently, MAVE provides several useful automatic routines:
- conversion and manipulation of annotations (e.g. Praat TextGrid)
- audio file container and general file header analysis
- voice activity detection
- speaker separation, cluster analysis of speaker groups, speaker sex classification
- audio quality assessment by acoustic analysis
- extraction of acoustic voice parameters (formants, MFCCs) - automatic voice comparison

The goal of the MAVE project is to successively determine and automize those tasks in forensic audio analysis which are suitable for the least intervention by human experts. While the above mentioned routines have already been implemented, future challenges in forensic audio will demand new routines which need to be attentively designed to serve forensic experts in their daily case work. Audio recordings from messenger services such as Whatsapp pose new challenges to forensic casework: channel impact, coding algorithms, data reduction/audio compression, speaking style variations and the resulting mismatch conditions can be dealt with in a concise and timesaving manner when repetitive tasks are automized and handled by one single toolbox, especially when comparisons of messenger recordings to telephone intercepts are required. While current offthe-shelf products generally behave like a black box, MAVE is fully transparent and can also be fitted to the specific conditions typically found in practical forensic casework. This transparency is instrumental to providing valid and reproducible results as well as quality assurance.

References

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.09, retrieved 2 March 2023 from http://www.praat.org/

Questioning the authorship of the voice of a famous Mexican painter. An acoustic-phonetic approach to the case

Fernanda López-Escobedo[1], *N. Sofía Huerta-Pacheco*[1, 2] *and Iván Vladimir Meza Ruiz*[2]

[1]*Escuela Nacional de Ciencias Forenses, UNAM.*
flopeze@unam.mx

[2]*Consejo Nacional de Ciencia y Tecnología de México.*
nshuerta@enacif.unam.mx

[3]*Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas, UNAM.*
ivanvladimir@turing.iimas.unam.mx

Frida Kahlo (1907-1954) was an early 20th-century Mexican painter. She was famous not only for her artwork but also for championing indigenous Mexican culture and influencing feminist movements. Although she was an important artist and public intellectual during her lifetime, it was not until 2019 when a recording of her voice was discovered in a clip from a radio show where its host expressly attributed to Kahlo the reading of a short poem dedicated to Diego Rivera (her husband).

Many national and international media picked up the news due to people's curiosity to know what the voice of such a famous artist sounded like. However, after the release of the recording, a well-known voice actress Amparo Garrido—active in the period when the recording was made—told the national media she remembered performing that reading. Before this declaration, there were no doubts about the authorship of the recording.

In this work, a voice comparison to determine the similarity between the voice attributed to Frida Kahlo and that of the actress Amparo Garrido was performed. If there is no similarity between the voices of both artists, the hypothesis that it is Frida Kahlo's voice should not be questioned by the declaration of Amparo Garrido. We collected a corpus comprising audio recordings from 27 female voice actresses active in Mexico City's radio broadcasting in the mid-fifties when the questioned recording was made, including one of Amparo Garrido. Two multivariate models developed by Rose et al. (2004) and Morrison (2011) were implemented for data analysis from an acoustic-phonetic approach. Recordings were segmented manually and different metrics of the first four formants—F1, F2, F3, and F4—of the five Spanish vowel sounds—/a/, /e/, /i/, /o/, and /u/— were analyzed. Approximately 100 tests were carried out, from which the maximum value of the vowel formants F1, F2, F3, and F4 was found to be the most representative metric and, consequently, it was selected as the basis for the comparisons of voice samples. The evidence obtained by comparing the voice sample attributed to Frida Kahlo against the corpus of female speakers did not support the hypothesis of the same origin (LR_Rose < 3.91E-102 and LR_Morrison < 1.41E-127). This means that the voice attributed to Frida Kahlo can not be attributed to any of the 27 female speakers analyzed. Since the focus was on the voice sample of the actress Amparo Garrido, the result of the comparison showed no evidence to support the hypothesis of the same origin.

References

G.S. Morrison, A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM), Speech Commun. 53 (2011) 242–256.

P. Rose, D. Lucy, T. Osanai, others, Linguistic-Acoustic Forensic Speaker Identification with Likelihood Ratios from a Multivariate Hierarchical Random Effects Model-A Non-Idiot's Bayes' Approach, (2004).Provine, R. R. (2001). *Laughter: A scientific investigation*. Penguin.

Are hesitation patterns individual?

Angelika Braun and Nathalie Elsässer
Department of Phonetics, University, of Trier, Germany
brauna@uni-trier.de, s2naelsa@uni-trier.de

This contribution takes a look at hesitations from the forensic practitioner's point of view. It focuses on individuality in the use of hesitation markers. There are observations in various previous studies pointing to the fact that patterns in the use of hesitation markers may be individual, (Belz and Trouvain 2019; Goldman-Eisler 1961, 1968; Finlayson and Corley 2012; Duez 1982; Kjellmer 2003; Clark and Fox Tree 2002; Belz 2021; Eklund 2001; Fant et al. 2003). This is consistent with the notion that hesitation behavior reflects the cognitive planning process of a specific individual. In the early literature on hesitations, individuality is stressed much more than in more recent publications.

A dedicated forensic approach to disfluencies was developed by McDougall and colleagues (McDougall and Duckworth 2018; McDougall et al. 2019). They list a number of parameters which describe the behavioral profile of a given speaker and are to be used in forensic casework. This is more comprehensive than any other framework, but it still falls short of being exhaustive, and the intraspeaker consistency of the features is assumed but not tested. The present contribution seeks to establish a more comprehensive concept of hesitation than has been done previously. It makes use of the "classical" fillers, which have been studied for decades, but it also proposes new elements which have so far hardly, if ever, been considered, such as the nasal filler and verbal fillers. Verbal fillers are multifunctional lexical items (Stenström 2012) which may either carry propositional meaning or serve as fillers. Examples from German are *ja* or *und*, but also phrases which make the search for the appropriate word explicit, such as *wie sagt man* ('how do you put it') or *mir fällt gerade das Wort nicht ein* ('I can't think of the word right now').

The key questions to be explored by this research are thus
(a) Are there speaker characteristic features in the hesitation behavior which have so far not been exploited?
(b) Are speakers at all consistent in their hesitation behavior?
(c) Are there features which are suitable for distinguishing between speakers?

The materials analyzed consist of several minutes of spontaneous speech largely consisting of monologues by eight female middle-aged speakers from the larger Frankfurt area at three different points in time. Analyses cover fillers including two elements which have not received much attention in previous research: the nasal filler and verbal fillers. Within- and between-speaker differences are assessed. In order to shed light on speaker individuality, results are presented separately for each speaker and also by session. Statistical analysis shows that hesitation markers will distinguish speakers at a level well above chance. At the same time, results show that it is impossible to pin down a single measure which will characterize the hesitation behavior of individual speakers. Rather, a combination of parameters is needed. The forensic implications of these findings are discussed.

References

Belz, M. (2021). Die Phonetik von *äh* und *ähm*. Akustische Variation von Füllpartikeln im Deutschen. In: Eklund, R. (ed.) *DiSS. The 7th Workshop on Disfluency in Spontaneous Speech*.. Berlin, 1–4.

Belz, M. and J. Trouvain. (2019). Are 'silent' pauses always silent? *Proc. 19th ICPhS,* Melbourne.

Clark, H. H., Fox Tree, J.E. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73-111.

Duez, D. 1(982). Silent and non-silent pauses in three speech styles. *Language and Speech* 25, 11-28.

Eklund, R. (2001). Prologations: a dark horse in the disfluency stable. *Proceedings of DISS '01*, Edinburgh, 5-8.

Fant, G,, Kruckenberg, A. and Barbosa Ferreira, J. (2003). Individual variation in pausing. A study in read speech. *PHONUM* 9, 193-196.

Finlayson, I.R, Corley, M. (2012). Disfluency in Dialogue: An Intentional Signal from the Speaker? *Psychon Bull Rev* 19, 921-928.

Goldman-Eisler, F. (1961). A comparative study of two hesitation phenomena. *Language and Speech* 4: 18-26.

Goldman Eisler, F. (1968). *Psycholinguistics. Experiments in Spontaneous Speech*. London and New York: Academic Press.

Harrington, L., Richard Rhodes, and Vincent Hughes. (2021). Style variability in disfluency analysis for forensic speaker comparison. *International Journal of Speech Language and the Law*, 28(1): 31–58.

Kjellmer, G. 2003. Hesitation. In defence of Er and ERM. *English Studies* 84(2): 170-198.

McDougall, K., Duckworth, M. (2018). Individual patterns of disfluency across speaking styles: A forensic phonetic investigation of Standard Southern British English. *International Journal of Speech, Language and the Law* 25(2): 205–230. https://doi.org/10.1558/IJSLL.37241

McDougall, K., Rhodes, R., Duckworth, M., French, P., Kirchhübel, Chr. (2019). Application of the 'Toffa' Framework to the Analysis of Disfluencies in Forensic Phonetic Casework. In: Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren (eds), *Proceedings of the 19th International Congress of Phonetic Sciences,* Melbourne, Australia. Canberra, Australia: Australasian Speech Science and Technology Association Inc, pp. 731- 735.

Stenström, A-B. (2012). Pauses and hesitations. In G. Andersen and K.Aijmer (eds.), *Pragmatics of Society.* Berlin/Boston: De Gruyter Mouton, 537-567.

Towards automatic speech recognition in police operations: the difference that real case resources can make

Ellen Grand[1,2]*, Georgina Brown*[2] *and Anonymous*[1]
[1]*National Crime Agency, UK*
[2]*Department of Linguistics and English Language, Lancaster University, UK*
{e.grand|g.brown5}@lancaster.ac.uk

A lot of police time and money is devoted to transcribing covertly recorded material collected to inform investigations. Automating this task could make substantial savings. Research, such as Loakes (2022), has considered the performance of automatic speech recognition (ASR) technology when applied to covertly recorded material. Because of the usual security and data protection restrictions, such research tends to make use of a) test data that only simulates the covert recording conditions, and b) open-source or commercially-available ASR systems which would not necessarily be employed by police departments. These methodological compromises mean that the research falls short of replicating the real use-case. There is an acceptance within the policing community of the need to train and test a bespoke system on recordings that are as reflective as possible of actual case data. This paper therefore presents ASR experiments that have been carried out on real covert operational data using a system that was specifically developed to entertain its use for investigative purposes.

In this work, a collaborating police force provided many hours of audio data that were captured as part of a past covert operation. The recordings were taken from within a moving vehicle; as a result, the speech is overlaid with varying levels of background noise. There is also "clipping" within the recording. There is one "main" speaker within the vehicle, but he has a number of telephone conversations where the device is set to loudspeaker (meaning other speakers have been captured within the recording).

114 minutes of the covert material were transcribed by a specialist transcriber who has extensive experience of transcription and the forensic speech analysis field. The same material was passed through the ASR system. Word Error Rate (WER) was used as a means to compare the outputs of the ASR system with the human-transcribed material. For this part of the work, only the stretches of speech where the transcriber did not indicate uncertainty were included in this comparison. A WER was calculated for 495 speaking turns (some turns only consisted of one word, while others consisted of many words). This comparison resulted in 86.33% WER overall. When taking a closer look at individual turns that yielded WERs of 0%, they are all very short utterances, most only consisting of one or two words (e.g. 'YEAH', 'YOU OKAY' and 'HELLO'). This observation resonates with some of the findings in Loakes (2022). However, in contrast to Loakes (2022), there are also longer stretches of speech containing numerous content words that achieved a relatively low WER.

Although an overall WER of 86.33% is still very high, a closer inspection of the results suggests that there are some performance gains when a system is specifically trained for the investigative purpose. The reported differences between research that does not use real use-case data and research that does use such data demonstrates the value of accessing real use-case resources.

References

Loakes, D. (2022). Does automatic speech recognition have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*. DOI: 10.3389/fcomm.2022.803452.

Interactive Visualisation of Speech Data in Virtual Reality

Philip Harrison, Paul Foulkes, Vincent Hughes, Poppy Welch, Jessica Wormald and Chenzi Xu

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

{philip.harrison|paul.foulkes|vincent.hughes|poppy.welch|jessica.wormald|chenzi.xu}@york.ac.uk

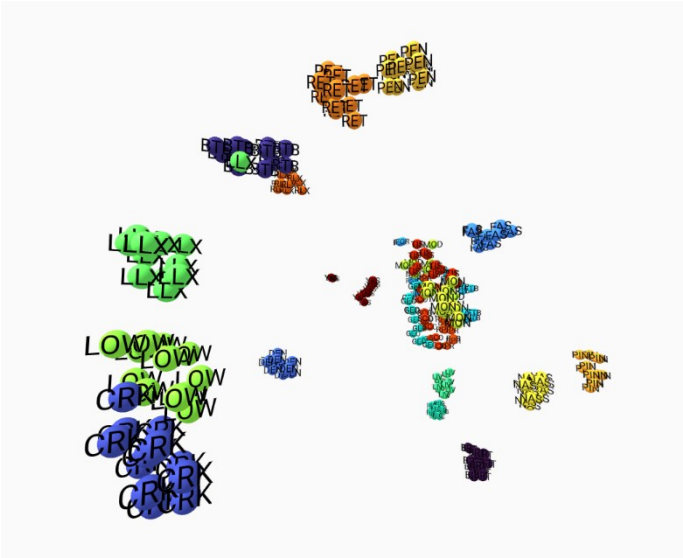
The complexity and variation found in human speech combined with the many ways in which it can be examined demonstrates the richness of speech as a source of data. Visualising this data is an integral part of the analysis pipeline. Visualisation serves multiple purposes including providing an overview of the data, allowing the comparison of data from multiple sources or with different attributes, and assists with identifying groups, trends, patterns or outliers. Exploring complex sets of data with traditional tools, such as two-dimensional scatter plots, can be limiting, especially when working with multidimensional data. Such exploration often involves listening to the source speech recordings to supplement or inform interpretation, but this must usually be done using separate software.

In this paper we present an innovative tool that displays multi-dimensional speech data as a three-dimensional scatter plot within a virtual reality environment. The user can explore the data by moving around and within the data points. The user can also rotate the data points around two axes, which provides additional perspectives. This immersive and interactive approach allows the user to obtain a more comprehensive view of the data compared with using static plots. One key feature is that the user can simply click on a data point to replay a clip of the source recording associated with it, without having to leave the visualisation. The tool is agnostic to the type of data being displayed so it can be used for many different purposes. The data is provided via a single CSV file which minimally contains the values of three variables for each data point, which are used as the three-dimensional coordinates in space, and a link to an audio clip. Other information can be provided such as additional variables, group or attribute information. These can be incorporated into the visualisation as colour information, a text label, point size or point shape, to allow groups to be more easily identified and distinguished.

The tool has been used to plot the output of t-SNE dimension reduction (van der Maaten & Hinton 2008) applied to x-vectors generated by the automatic speaker recognition system VOCALISE (Kelly et al. 2019). This allowed the examination of the clustering of different voice qualities within the x-vector speaker space (Wormald et al. 2023) - see Figure 1. Similar examinations can also be undertaken directly in VOCALISE as it includes an effective in-built interactive x-vector visualisation tool with audio file replay. The current tool extends this functionality by placing the visualisations within virtual reality and provides the ability to examine other types of speech data. For example, it has also been used to explore the distribution of speakers in a three-dimensional formant space based on their mean long-term formants - see Figure 2. It can easily be used with other dimension reduction techniques such as PCA (principal component analysis) or UMAP (Uniform Manifold Approximation and Projection) (McInnes & Healy 2018), or other combinations of acoustic speech measures.

The tool has been built using open web technologies including JavaScript, A-Frame (2023) and A-Frame P5 (2021) so that the technological barriers to using it are as low as possible. The most immersive way of using the tool is with a VR headset with hand controllers, but the tool can also be used effectively in a web browser on a standard computer using a mouse and keyboard.

The tool will be demonstrated using both a VR headset and a web browser on a standard computer.



BRT - Breathy
BTB - Backed tongue body
CRK - Creaky
DEN - Denasal
FAS - Fast
FOR - French accented
GEO - Geordie
HIG - High pitch
LIV - Lively
LLX - Lowered larynx
LOW - Low pitch
MOD - Modal voice
MON - Monotone
NAS - Nasal
PEN - Pen between the teeth
PIN - Pinched nose
RET - Retroflex
RLX - Raised larynx
RPR - Received Pronunciation
SCO - Scottish
TIS - Tissue over the phone mouthpiece
YKS - Yorkshire

Figure 1. Visualisation of different voice qualities represented in x-vectors after the application of t-SNE dimension reduction. Voice qualities are labelled and grouped by colour.

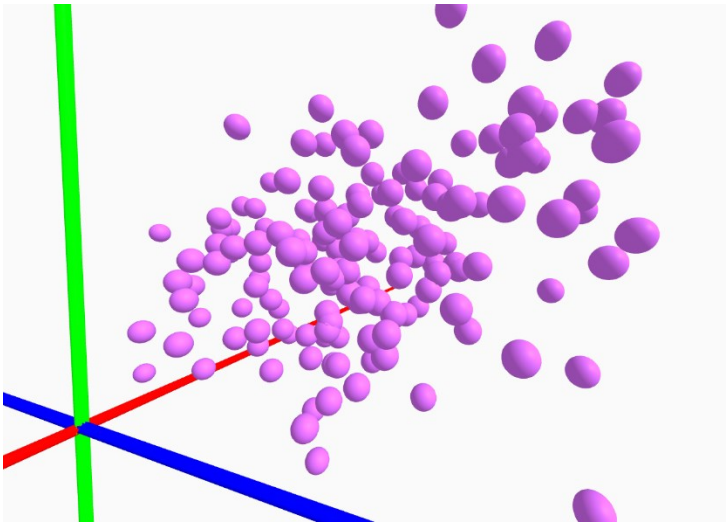


Figure 2. Visualisation of long term mean F1, F2 and F3 formant values of 160 speakers - red axis = F1, green axis = F2, blue axis = F3.

References

A-Frame [Computer software] (2023). Retrieved from https://aframe.io/.

A-Frame P5 [Computer software] (2021). Retrieved from https://cims.nyu.edu/~kapp/aframep5/.

Kelly, F., Forth, O., Kent, S., Gerlach, L. & Alexander, A. (2019) Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. *Proceedings of the Audio Engineering Conference: 2019 AES International Conference on Audio Forensics*.

van der Maaten, L., & Hinton, G. 2008. Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

McInnes, L, & Healy, J, 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426.

Wormald, J., Foulkes, P., Harrison, P., Hughes, V., Kelly, F., van der Vloed, D., Welch, P. & Xu, C. 2023. Sensitivity of x-vectors and automatic speaker recognition scores to vocal variation. Proceedings of ICPhS 2023, Prague.

Learning by doing: an example of casework-relevant training in forensic speech science

Linda Gerlach[1,2], Luke Carroll[3], Lois Fairclough[3], Ben Gibb-Reid[4], Lauren Harrington[4], Daniel Denian Lee[1], Alexandra Lieb[5], Sophie Möller[5], Chloe Patman[1], Alice Paver[1], Sascha Schäfer[4], Marlon Siewert[5], Nikita Suthar[4], M.Gabriela Valenzuela Farías[4], Samantha Williams[4], Georgina Brown[3,6] and Christin Kirchhübel[6]

[1]*Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, UK*
[2]*Oxford Wave Research, Oxford, UK*
[3]*Department of Linguistics and English Language, Lancaster University, UK*
[4]*Department of Language and Linguistic Science, University of York, UK*
[5]*Department of German Studies and Arts, Philipps Universität Marburg, Germany*
[6]*Soundscape Voice Evidence, Lancaster, UK*

There are plenty of reasons to increase the provision of casework-relevant training in forensic speech science. Firstly, training is now more prominent in the revised 2021 version of the IAFPA Code of Practice (Section 2.2). Secondly, the detachment between casework practice and academic research was raised as a concern at the IAFPA Annual General Meeting in 2022. Thirdly, an IAFPA student representative requested further casework-relevant training. All of these reasons prompted two of the authors to design and deliver a Forensic Voice Comparison (FVC) short-course. Simultaneously, this short-course allowed them to trial Problem-Based Learning (PBL) in the forensic speech science context (Brown, 2022). This paper presents how the course took place, challenges that arose and what the authors (the organisers and attendees) learned from the experience.

The centrepiece of the short-course was a mock voice comparison case (the "Problem"). The recordings that made up the Problem were carefully selected in order to ensure that the Problem a) reflected real-life casework, and b) carried a number of challenges in relation to the analysis and interpretation of findings. The participants were given 9 weeks to work on the Problem before attending a one-day in-person workshop. The purpose of the workshop was to exchange ideas and to engage in detailed discussion on how to go about the Problem with an experienced practitioner present throughout the day. Because PBL is typically carried out in small groups, the original intention was to cap the number of participants at 10, but the organisers were surprised by the levels of interest and accepted 15 participants onto the course (and were put into the unfortunate position of saying "no" to a number of other registrants). The 15 participants were all postgraduate research students and early-career researchers from across four institutions.

To begin with, the workshop involved a detailed exchange of how different participants approached the Problem and what conclusions they reached. Interestingly, 8 of the participants had arrived at a same-speaker conclusion and 7 arrived at a different-speaker conclusion. The practitioner then provided on-the-spot feedback to some of the more common themes that emerged from the discussions, revealing some of her own analysis of the Problem. In the latter part of the workshop, participants were presented with an extension of the Problem which aimed to address engagement and communication with instructing parties and opposing voice experts. To spark ideas on this, participants were issued with mock responses from their instructing party together with a mock voice comparison report from an opposing expert. Not only did this prompt discussion about written communication of voice comparison conclusions, but it also triggered discussion about broader ethical issues attached to casework.

Following the course, all 15 participants completed a survey to document their experiences. Participants reflected on the analytical aspects of voice comparison tasks, the interpretive

responsibility placed on the analyst, and the wider pressures that come with voice comparison casework. This paper will expand on some of the key themes that emerged.

References

Brown, G. (2022). Proposing Problem-Based Learning for training future forensic speech scientists. *Science & Justice*, 62(6), 669-675.

The effects of voice stereotypes on voice parades

David Wright[1], *Alice Paver*[2] *and Natalie Braber*[1]
[1]*School of Arts and Humanities, Nottingham Trent University, UK.*
{david.wright|natalie.braber}@ntu.ac.uk
[2]*Theoretical and Applied Linguistics Section, University of Cambridge, UK.*
aep58@cam.ac.uk

This study is a follow-up experiment based on findings from the ESRC-funded project 'Improving Voice Identification Procedures' (IVIP). A series of studies in the project examined various voice parade parameters and how they affect earwitness identification accuracy (Pautz et al., 2023a; 2023b). In mock voice parade experiments, listeners were exposed to a 60 second sample of a target voice and, after a short distraction task, were asked to identify the target from a voice line-up of nine speakers. Experimental conditions in all of the studies included target-present and target-absent parades. Results showed that, in the target-absent parades, some foil voices were incorrectly selected by listeners and that the frequency with which different foils were selected varied.

The present study explores the possible motivations behind the incorrect selection of a foil voice in target-absent voice parades. Specifically, it investigates whether false alarm rates can be explained by stereotyped judgements that listeners make about the foil voices used in parades. Previous research has found that speakers judge some voices more negatively than others in forensic contexts, including rating some as sounding 'more guilty' of committing certain criminal offences (Dixon and Mahoney, 2004; Frumkin and Thompson, 2020; Paver et al., *in review*). Against this backdrop, the hypothesis to be tested in this study is that the foil voices that are frequently selected in target-absent parades are rated more negatively than those that are not frequently selected.

In an online listening experiment, 180 participants used a 7-point Likert scale to rate 12 voices from the previous experiments on ten traits (related to 'status', 'solidarity' and 'dynamism') and ten behaviours (broadly conceived as morally 'good', 'ambiguous' and 'bad', including a range of criminal offences). Participants heard four voices from each parade: the target speaker, and three foils, the one most frequently selected, one never selected, and one selected at a rate roughly at median value.

Preliminary descriptive results find that across the three parades combined, the most- and middleselected foils were both rated *lower* on status and solidarity traits when compared with the target speaker and foils never selected (Figure 1). They also both rated *higher* for criminal and morally 'bad' behaviours (Figure 2) meaning listeners judged them as more likely to behave in bad and illegal ways, while they were rated *lower* (i.e. less likely) for morally good and morally ambiguous behaviours. However, when examining results for individual parades, some variation is observed. For some parades, the rates of incorrect selection of a foil appear to pattern with its perceived voice similarity with the target (collected during pre-testing phases of the previous experiment), rather than stereotyped ratings of traits and behaviours. At this stage, the results provide partial support for the hypothesis, but this is not consistent across different parades or different voices. This paper will explore the results for individual parades and speakers in more detail and will discuss the implications of these findings for voice parades and the evidence that they elicit.

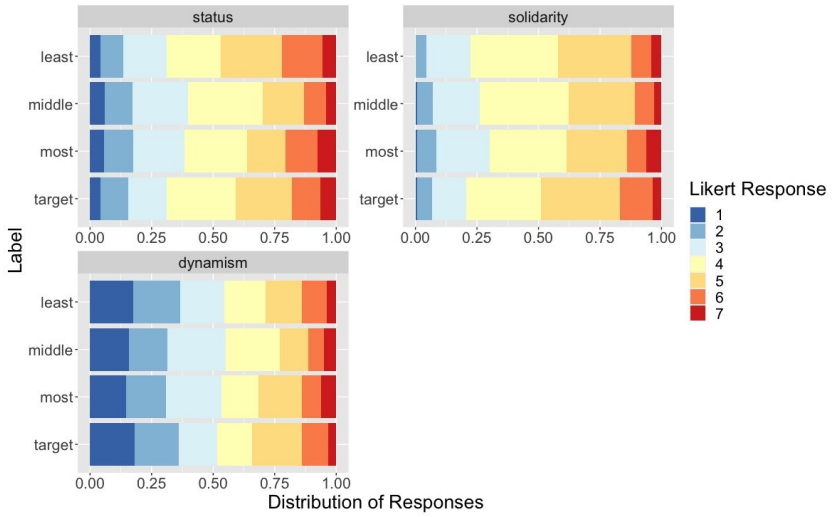


Figure 1. Judgements of social traits by stimulus incorrect selection rate (1= strongly disagree, 7= strongly agree)

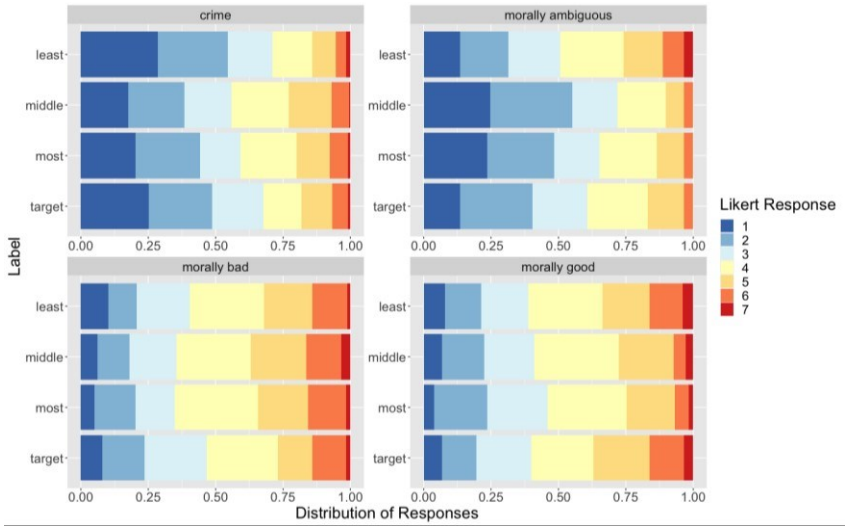


Figure 2. Judgements of behaviour types by stimulus incorrect selection rate (1= strongly disagree, 7= strongly agree)

References

Dixon, J. A., Mahoney, B. & Cocks, R. (2002). Accents of guilt? Effects of regional accent, race, and crime type on attributions of guilt. *Journal of Language and Social Psychology*, 21(2), 162–168, https://doi.org/10.1177%2F02627X02021002004

Frumkin, L. A. & Thompson, A. (2020). The impact of different British accents on perceptions of eyewitness statements. *Journal of Language and Discrimination,* 4(1), 119–138. https://doi.org/10.1558/jld.39368

Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A., and Smith, H.M.J. (2023a). Identifying unfamiliar voices: Examining the system variables of sample duration and parade size. *Quarterly Journal of Experimental Psychology*. https://doi.org/10.1177/17470218231155738.

Pautz, N., McDougall, K., Mueller-Johnson, K., Nolan, F., Paver, A., and Smith, H.M.J. (2023b). 'Improving Voice Identification Procedures'. Paper presented at the Strategic Research Themes Conferences, Nottingham Trent University, 29th March 2023.

Paver, A., Wright, D., and Braber, N. (in review) Criminal behaviour and trait judgements of accents in the UK. *Frontiers in Communication – Language Sciences.*

Regional accent identification by naïve listeners

Caroline Kleen[1] *and Angelika Braun*[2]
[1]*Research Center Deutscher Sprachatlas, Philipps University, Marburg, Germany*
caroline.kleen@uni-marburg.de
[2]*Department of Phonetics, University of Trier, Germany*
brauna@uni-trier.de

The present study investigates the ability of naïve listeners to identify regional accents. The aim is to compare the recognition performance of this listener group to that of forensic linguistic experts as studied in Köster et al. (2012).

A total of 40 native listeners of German were studied. The sample consisted of 20 participants (9 m, 11 f; mean age = 26.7 years, SD = 4.3 years) from the Moselle-Franconian language area and 20 from the North Lower Saxon language area (10 m, 10 f; mean age = 27.7 years, SD = 3.4 years). They were born and raised in the respective region. Their task was to listen to 20 stimuli and to locate them in the German language area. The stimuli were the same as those used by Köster et al. (2012). They are recordings of respondents to 911 calls and form part of the DIGS corpus. Each stimulus contains speech with a duration of 30-60 seconds. After listening to each stimulus participants were asked to mark the inferred region of the speaker's origin either by setting a check mark or by drawing a circle marking the inferred region. The responses were evaluated according to the point system developed by Köster et al. (2012).

Köster et al. (2012) found a recognition rate of about 85 % for the group of forensic experts. In the present study, a recognition rate of about 38 % is established. This differs highly significantly from that of the experts. The individual recognition rates range from 6.3 % to 63.8 %. This means that there is an overlap between the best performance among lay listeners and the worst performance by experts. Participants' sex and regional origin show no significant effect on the identification of regional accents, and neither is there an interaction.
In previous studies (e.g., Hundt et al. 2015), recognition rate has been analyzed with respect to distance between the listener's own regional accent and the accent to be judged. However, the concept of distance between listener's own dialect and the accent to be judged is difficult to grasp. Since present stimuli were unequally distributed within the German language area, the effect of distance could not be studied systematically. Therefore, only those samples showing a large geographical distance between listener and speaker accents were compared.

With the exception of the listener's own accent background (West Low German and West Middle German stimuli), no significant interaction or main effect was found. The wide dispersion of the individual recognition rates of the naïve listeners may be related to other factors, e.g., education, own interests, the profession practiced, and others, which may have influenced their sensitivity to language (Klein 2021). With the forensic perspective in mind, it is very clear that forensic expertise is indispensable in accent identification.

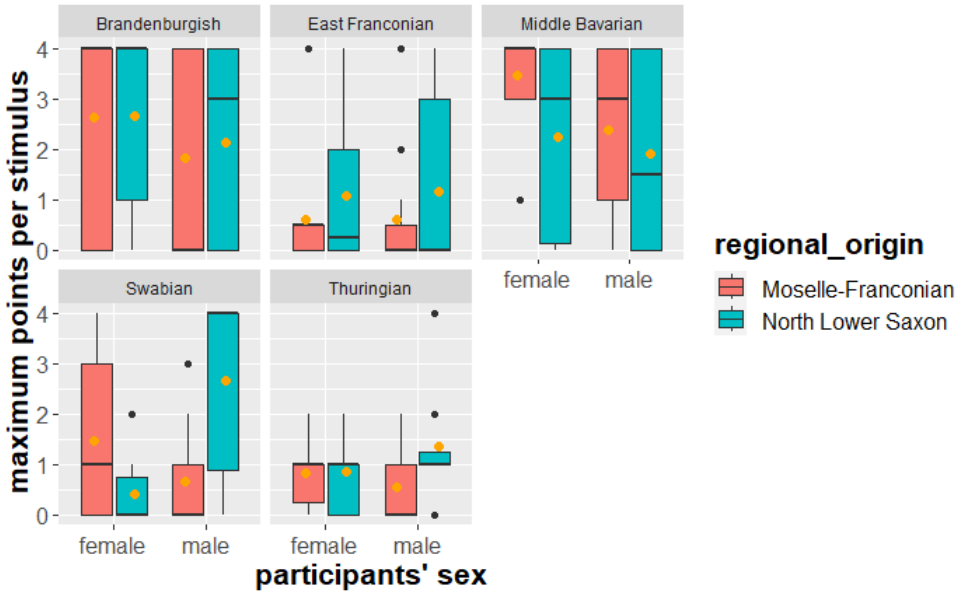


Figure 1. Distribution of recognition rates for stimuli from different dialect areas according to listener sex and regional origin.

References

Braun, A. (2015). Forensische Sprach- und Signalverarbeitung. In J. Bockemühl (Ed.), *Handbuch des Fachanwalts. Handbuch des Fachanwalts Strafrecht* (6. ed., pp. 1769–1790). Heymanns.

Hundt, M., Palliwoda, N. & Schröder, S. (2015). Wahrnehmungsdialektologie - Der deutsche Sprachraum aus der Sicht linguistischer Laien. In R. Kehrein, A. Lameli & S. Rabanus (Eds.), *Regionale Variation des Deutschen: Projekte und Perspektiven* (pp. 585–629). De Gruyter Mouton.

Klein, W. P. (2021). Was denken linguistische Laien über die (deutsche) Grammatik? In T. Hoffmeister, M. Hundt & S. Naths (Eds.), *Laien, Wissen, Sprache* (pp. 227–248). de Gruyter.

Köster, O., Kehrein, R., Masthoff, K. & Boubaker, Y. H. (2012). The tell-tale accent: Identification of regionally marked speech in German telephone conversations by forensic phoneticians. *International Journal of Speech Language and the Law, 19(1), 51-71.*

Schmidt, J. E., Herrgen, J. & Kehrein, R. (Eds.) (2020 ff.). Regionalsprache.de (REDE). Forschungsplattform zu den modernen Regionalsprachen des Deutschen. Edited by R. Engsterhold, H. Girnth, S. Kasper, J. Limper, G. Oberdorfer, T. Pistor, A. Wolańska. Unter Mitarbeit von D. Beitel, M. Gropp, M. L. Krapp, V. Lang, S. Lipfert, J. Pheiff, B. Vielsmeier. Research Center Deutscher Sprachatlas.

Multilingual voices database and COVID protection masks effect in Forensic Speaker Recognition

André Saraiva[1,2], *Attila Fejes*[3], *Jelena Devenson*[4] *and Vasile-Dan Sas* [5]
[1]*Forensic Science Laboratory, Lisbon, Portugal.*
[2]*Faculty of Engineering of the University of Porto, Porto, Portugal.*
andre.saraiva@pj.pt
[3]*Special Service for National Security, Institute for Experts Services, Budapest, Hungary.*
fejes.attila@nbsz.gov.hu
[4]*Forensic Science Centre of Lithuania, Vilnius, Lithuania.*
j.devenson@ltec.lt
[5]*National Institute of Forensic Expertise, Bucharest, Romania.*
vasile.sas@inec.ro

The variety of the languages spoken in Europe has proven to be a difficulty for forensic automatic and semi-automatic speaker recognition, given the absence of adequate reference populations of voices in languages other than native languages, or native languages spoken by foreigners. These reference populations play a key role in generating statistical models for this type of forensic examinations.

In addition, the COVID-19 pandemic represented a challenge for forensic speaker recognition given the mandatory use of protection masks in almost every daily situation, as these act as voice barriers attenuating speech signals. To the best of our knowledge, published research on the impact of facial coverage on forensic speaker recognition notes the need for larger and more diverse datasets, regardless of the significance of the conclusions reached (Bogdanel et al., 2022; Das & Li, 2020; Geng et al., 2023; Iszatt et al., 2021; Khan et al., 2022; Loukina et al., 2020; Mallol-Ragolta et al., 2021; Ristea & Ionescu, 2020; Saeidi et al., 2015, 2016).

Started in January of 2022, a work package, part of the EU-funded CERTAIN-FORS project, aims to tackle both issues by developing a voice samples database to be shared with ENFSI Forensic Speech and Audio Analysis Working Group (FSAAWG) members. It has been built with samples obtained from individuals speaking their native language, with and without protection masks, and speaking non-native languages.

The data collection has been performed by several collaborating FSAAWG members, according to a predefined protocol, including: reading a text in native language without mask, wearing a surgical mask and a FFP2 type mask; reading a text in non-native language(s); dialoguing in native language; and, when possible, dialoguing in non-native language(s).

The dataset is composed of samples collected from more than 650 volunteers from Croatia, Georgia, Portugal, Romania, Ukraine, Greece, Hungary, Lithuania and Spain. Each collaborating Institute was asked to collect samples from 80 volunteers (40 males and 40 females) minimum, according to the following age classes, in years (ten of each by gender): [18 - 30], [31 - 40], [41 - 50], [51 - +∞[.

In general, the samples were obtained from individuals with different origins within each country, allowing different accents to be represented. The vast majority was collected using microphones in a controlled environment, consecutively. Nevertheless, in some cases it was possible to obtain samples from mobile communications at the same time.

The characterization of the dataset will be presented, as well as the study of the effect of surgical and FFP2 type protection masks several acoustic parameters. The impact of the Covid protection masks in the performance of Forensic Automatic Speaker Recognition systems will also be evaluated.

References

Bogdanel, G., Belghazi-Mohamed, N., Gómez-Moreno, H., & Lafuente-Arroyo, S. (2022). Study on the Effect of Face Masks on Forensic Speaker Recognition. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13407 LNCS*, 608–621. https://doi.org/10.1007/978-3-031-15777-6_33

Das, R. K., & Li, H. (2020). Classification of Speech with and without Face Mask using Acoustic Features. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 747–752.

Geng, P., Lu, Q., Guo, H., & Zeng, J. (2023). The effects of face mask on speech production and its implication for forensic speaker identification-A cross-linguistic study. *PLoS ONE*, *18*(3 March). https://doi.org/10.1371/JOURNAL.PONE.0283724

Iszatt, T., Malkoc, E., Kelly, F., & Alexander, A. (2021). Exploring the impact of face coverings on x-vector speaker recognition using VOCALISE. *International Association of Forensic Phonetics and Acoustics*.

Khan, A., Javed, A., Malik, K. M., Raza, M. A., Ryan, J., Saudagar, A. K. J., & Malik, H. (2022). Toward Realigning Automatic Speaker Verification in the Era of COVID-19. *Sensors 2022, Vol. 22, Page 2638*, *22*(7), 2638. https://doi.org/10.3390/S22072638

Loukina, A., Evanini, K., Mulholland, M., Blood, I., & Zechner, K. (2020). Do face masks introduce bias in speech technologies? The case of automated scoring of speaking proficiency. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1942–1946. https://doi.org/10.21437/Interspeech.2020-1264

Mallol-Ragolta, A., Liu, S., & Schuller, B. W. (2021). The Filtering Effect of Face Masks in their Detection from Speech. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2079–2082. https://doi.org/10.1109/EMBC46164.2021.9630634

Ristea, N. C., & Ionescu, R. T. (2020). Are you wearing a mask? Improving mask detection from speech using augmentation by cycle-consistent GANs. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2102–2106. https://doi.org/10.21437/Interspeech.2020-1329

Saeidi, R., Huhtakallio, I., & Alku, P. (2016). Analysis of face mask effect on speaker recognition. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1800–1804. https://doi.org/10.21437/INTERSPEECH.2016-518

Saeidi, R., Niemi, T., Karppelin, H., Pohjalainen, J., Kinnunen, T., & Alku, P. (2015). Speaker Recognition For Speech Under Face Cover. *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 1012–1016.

Exploring the Articulatory Perspective of Mel-Frequency Cepstral Coefficients: Unravelling the Link between MFCCs and Vocal Tract Features

Bruce Xiao Wang[1] *and Lei He*[2]
[1]*Department of Chinese and Bilingual studies, Hong Kong Polytechnic University, HK*
brucex.wang@polyu.edu.hk
[2]*Department of Computational Linguistics - Phonetics, University of Zurich, Switzerland.*
lei.he@uzh.ch

In recent years, Mel-frequency cepstral coefficients (MFCCs; Davis and Mermelstein, 1980) have been widely used as the input features of semi-automatic (Nolan & Grigoras, 2005) forensic voice comparison (FVC) systems, and some studies have shown that MFCC features yield better speaker discriminatory performance (e.g., lower EER and/or C_{llr}) than traditional acoustic phonetic features (e.g., vowel formants).

MFCCs capture the spectral characteristics of speech signal, and spectral characteristics are a function of vocal tract (Fant, 1971), e.g., oral cavity. It is claimed that MFCCs capture shape and features of the human vocal tract; however, no studies have attempted to investigate how MFCCs and vocal tract features, if there are any, are related. A recent study (Hughes et al., 2023) has partially discussed the correlation between MFCCs and formant values; however, the interpretability of MFCCs has rarely been properly discussed or investigated as well as the question of *why* MFCCs, despite a higher dimensionality, outperform traditional acoustic phonetic features.

In the current work-in-progress paper, we aim to investigate MFCCs from an articulatory perspective. We extracted the first 12 MFCCs and articulatory kinematics from three vowels (i.e., FLEECE, TRAP, FOOT) in single words. The data, obtained from Ji et al. (2014), contained the raw recordings of single words produced by 20 Midwestern standard American English speakers (10 male and 10 female) as well as the articulatory kinematics. The articulatory kinematics data was measured using electromagnetic articulography (EMA containing the movement of tongue dorsum (TD), tongue lateral (TL), tongue blade (TB), upper lip (UL), lower lip (LL), lateral lip corner (LC) measured in three dimensions (Figure 1, i.e., x: front and back, y: height, z: left and right). We will perform principal component analysis (PCA) on the first 12 MFCCs as well as 12 articulatory kinematics data (i.e., 6 sensors * x-axis * y-axis) aiming to explore to what extent the MFFCs and articulatory kinematics contribute to the first three components.

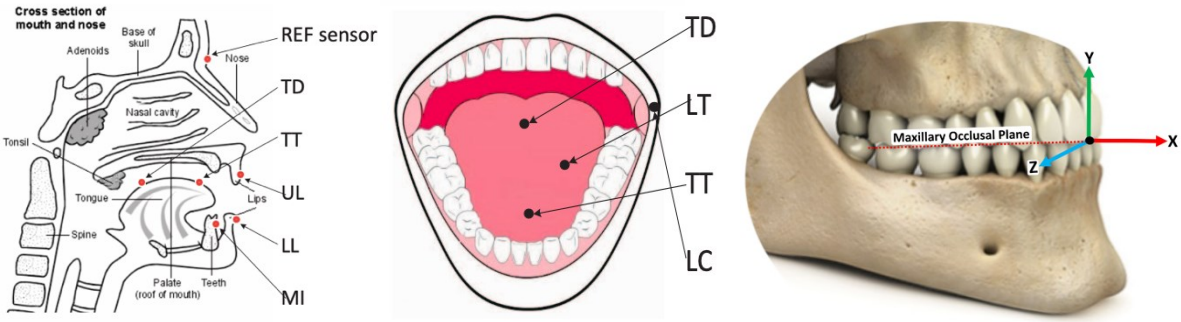


Figure 1. Leftmost and central panel: sensor placement (Figure 1 from Ji et al. 2014). Rightmost panel: Target anatomically-referenced coordinate system, Positive increases in sensor values denote forward, upward, and rightward movement along x, y, z, respectively (Figure 2 from Berry et al., 2016 EMA-MAE corpus User's Handbook).

References

Berry, J., Ji, A. & Johnson, T. (2016). *EMA-MAE Corpus User's Handbook* (Version 2.0). Marquette University, Milwaukee, WI, USA.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing, 28*(4), 357-366.

Fant, G. (1971). *Acoustic theory of speech production: with calculations based on X-ray studies of Russian articulations* (No. 2). Walter de Gruyter

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

Ji, A., Berry, J. J., & Johnson, M. T. (2014, May). The Electromagnetic Articulography Mandarin Accented English (EMA-MAE) corpus of acoustic and 3D articulatory kinematic data. In 2014 *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 7719-7723). IEEE.

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

It's all like yeah: Assessing the speaker discriminant potential of *yeah*

Ben Gibb-Reid[1], Vincent Hughes[1] and Paul Foulkes[1]
[1]*Department of Language and Linguistic Science, University of York, UK*
ben.gibb-reid@york.ac.uk

Background

Yeah is one of the most frequent words in spoken English. In the most recent edition of the spoken British National Corpus (Love et al., 2017), it is ranked 7th. Given its high rate of use, investigating its potential as a forensic voice comparison (FVC) feature has great merit. As previous literature (Nolan, 1983, p.11; Rose, 2003, p.52) has outlined, the more frequent a feature is, the greater its use in FVC because it is more likely to occur across disputed and known samples. The word *just* has been proven to perform well as a speaker discriminant (Gibb-Reid, accepted) as have filled pauses (*uh/um*) (Hughes et al., 2016). These previous studies utilised vowel formant and token duration measurements in their analysis within likelihood-ratio (LR) framework testing. The present paper replicates this by investigating the suitability of formant trajectory measurements taken from tokens of *yeah* in a homogenous group. This study will eventually also describe the interaction of *yeah* with related vowels DRESS and SQUARE to contribute to the understanding of the word's typical realisation.

Methods

The data used in this study is taken from the Quakebox corpus (Walsh et al., 2013): recordings of monologues made by speakers recounting their experiences of the 2010-2011 Canterbury earthquakes. It is studio-quality audio and is particularly suitable for FVC research as participants were invited to return to do a second recording seven years after the first. This subset of data (QB2) allows a comparison with the original data (QB1) that replicates the FVC context by having two distinct samples of audio for speaker comparison. All tokens of *yeah* were extracted from the database as.wav files resulting in 4,142 tokens. Then a Praat plug-in called Fast Track (Barreda, 2021) was utilised to extract five measures of F1, F2 and F3 formant trajectories. After data cleaning, there were 1,913 suitable tokens across 114 speakers. There were also thirty-six speakers who had enough tokens to allow a comparison across QB1 and QB2. Table 1 summarises the amount of data available in each subdivision.

Sex	*Tokens*	*Speakers*	*Across QB1 and QB2*
M	657	36	12
F	1247	77	24
NA	9	1	0
Sum:	1913	114	36

Table 1. *Yeah* token counts distributed across speaker sex and amounts in corpora.

Once the data was cleaned, quadratic polynomial equations were fitted to simplify the amount of datapoints from five measurements to three coefficients following similar methods employed by (Hughes et al. (2016) and Morrison (2009). Then, LR-based testing was undertaken achieved using the *fvclrr* package in R (Lo, 2022) which allows for replications of tests altering the speakers contained within background, training and test subsets.

Preliminary results

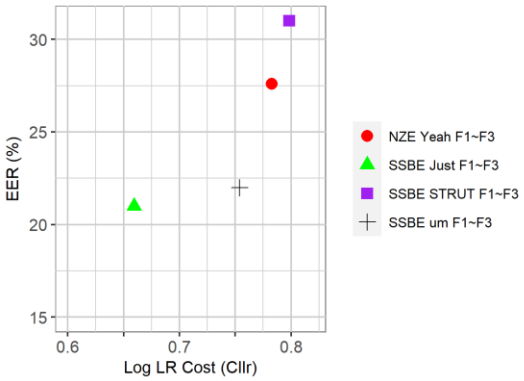


Figure 1. C_{llr} plotted against EER comparing the performace of the vowels in *just, um,* STRUT and quadratic regression taken from *yeah* formant trajectories.

Figure 1 displays a comparison with data from a previous study (Gibb-Reid, forthcoming). Preliminary results show that *yeah* performs comparably with STRUT. These are based on one single calibrated speaker comparison test. Overall this indicates some promise for the speaker discriminant ability of *yeah* which will be investigated with further data cleaning, subsetting and replicated LR testing.

References

Aitken, C. G., & Lucy, D. (2004). Evaluation of trace evidence in the form of multivariate data. Journal of the Royal Statistical Society: Series C (Applied Statistics), 53(1), 109–122.

Barreda, S. (2021). Fast Track: Fast (nearly) automatic formant-tracking using Praat. Linguistics Vanguard, 7(1). https://doi.org/10.1515/lingvan-2020-0051

Gibb-Reid, B. (forthcoming). Just one word: An analysis of just as a speaker discriminant using various acoustic measures. Proceedings of the 20th International Congress of Phonetic Sciences, Prague, Czech Republic, 2023. 20th International Congress of Phonetic Sciences (ICPhS), Prague, Czech Republic.

Hughes, V., Wood, S., & Foulkes, P. (2016). Strength of forensic voice comparison evidence from the acoustics of filled pauses. International Journal of Speech, Language & the Law, 23(1), 99–132.

Lo, J. (2022). fvclrr: Likelihood Ratio Calculation and Testing in Forensic Voice Comparison. https://github.com/justinjhlo/fvclrr#readme

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3), 319–344.

Morrison, G. S. (2009). Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs. The Journal of the Acoustical Society of America, 125(4), 2387–2397.

Morrison, G. S. (2013). Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio. Australian Journal of Forensic Sciences, 45(2), 173–197. https://doi.org/10.1080/00450618.2012.733025

Nolan, F. (1983). The phonetic bases of speaker recognition / Francis Nolan. Cambridge University Press.

Rose, P. (2003). Forensic speaker identification.

Walsh, L., Hay, J., Bent, D., King, J., Millar, P., Papp, V., & Watson, K. (2013). The UC QuakeBox Project: Creation of a community-focused research archive.

"The Impact of Vocal Variability on Voice Identification"

Alanna Tibbs
[1]*Department of Language and Linguistics, University of York, UK*
at1553@york.ac.uk

Considerable research has examined layperson ability to identify voices, crucial in legal contexts where earwitness testimonies can lead to convictions. However, multiple factors affect this ability and the construction of voice identities (Lavan et al., 2019a). To determine the strength of earwitness evidence, these factors must be investigated on a case-by-case basis.

Particularly in forensic settings, speakers can demonstrate wide linguistic variation resulting from physiological changes to the vocal tract. Intra-speaker variability can be exacerbated by different expressive speaking styles (Scherer, 2003), (Lavan et al., 2019b) making voice recognition more challenging. Research supports the greater difficulty identifying voices in a neutral quality if the witness only heard the suspect in another -i.e. whispering (Yarmey, 1995) or shouting (Blatchford & Foulkes, 2006). Unfortunately, mismatch may occur when voice line-ups only utilise neutral speech from police interviews (Nolan, 2003). However, evidence indicates that exposure to a greater variety of speech styles enhances accuracy in identifying speakers in unfamiliar styles (Lavan et al., 2018)

This paper describes a current Masters project to investigate how varied exposure to different voice qualities, through expressive speaking styles, effects recognition accuracy, and confidence ratings, in a modified voice parade.

The study records eight male speakers in their twenties from villages near Middlesbrough: firstly in mock police interviews resembling DyViS (Nolan et al. 2009), eliciting neutral spontaneous speech typical in real voice parades (Nolan, 2011). Participants will then be recorded shouting, whispering, and using creaky-voice in different expressive styles. The samples will be tested to ensure no voices stand out (Nolan, 2003) and will be in expressive speech styles as this will provide information to listens constructing a speaker identity (Lavan et al., 2019b), resembling real cases.

In an online survey, participants will listen to three ten-second samples of expressive speech, followed by ten seconds of neutral speech. They will rate their confidence (McDougall, 2013) in determining whether the samples come from the same or different speakers. Th tests will be randomised regarding speaker and style mismatch Performance will also be assessed regarding listener demographics -i.e age, familiarity with accent.

These results could strengthen the evidence provided by ear-witnesses, if they were exposed to a greater proportion of a speaker's possible variability. Without a clear person-specific representation of a voice, listeners exposed to high-variability input are more likely to expect speaker mismatch but should be more accurate at identifying voices in unfamiliar styles. Whereas those exposed to low-variability input feel more familiar with the voice, even if only in one variety (Lavan et al., 2019a). Those who heard only one variety received more consistent input and are likely to be more confident about whether or not a new speech sample fits their perceived speaker identity. However, their accuracy will depend on how similar the training stimulus was to the neutral test stimulus.

As well as supporting listener-specific effects on voice identification ability, these results will demonstrate the most reliable acoustic qualities when constructing speaker identities and how ear-witnesses will apply this representation in their evidence.

References

Blatchford, H., & Foulkes, P. (2006). Identification of voices in shouting. International Journal of Speech, Language and the Law, 13(2), 241-254. https://doi.org/10.1558/ijsll.2006.13.2.241

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2018). Flexible voices: Identity perception from variable vocal signals. Psychonomic Bulletin & Review, 26(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

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

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019 b). Breaking voice identity perception: Expressive voices are more confusable for listeners. Quarterly Journal of Experimental Psychology, 72(9), 2240–2248. https://doi.org/10.1177/1747021819836890

McDougall, K. (2013) Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. International Journal of Speech, Language and the Law 20(2): 163-172.

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

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(10).

Nolan, F. (2011). Dynamic Variability in Speech: a Forensic Phonetic Study of British English, 2006-2007. [data collection]. UK Data Service. SN: 6790, DOI: 10.5255/UKDA-SN-6790-1

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. Speech Communication, 40(1-2), 227–256. https://doi.org/10.1016/S0167-6393(02)00084-5

Yarmey, A. D. (1995). Earwitness speaker identification. Psychology, Public Policy, and Law, 1(4), 792–816. https://doi.org/10.1037/1076-8971.1.4.792

Yarmey, A. D., Yarmey, A. L., Yarmey, M & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. Applied Cognitive Psychology. 15. 283 - 299.

A guide on the exploration of the vocal identity space

Alessandro De Luca[1] *and Volker Dellwo*[2]
[1]*Linguistic Research Infrastructure, University of Zurich, Switzerland.*
alessandro.deluca@uzh.ch
[2]*Department of Computational Linguistics, University of Zurich, Switzerland.*
volker.dellwo@uzh.ch

Introduction

Vocal identity and within-speaker variability hold interest for researchers across various sectors of academia and industry. Recently, Dellwo et al. (2019) have highlighted how idiosyncratic properties of a speaker can be modulated to affect recognition. In addition, evidence supporting a norm-based coding mechanism of vocal recognition in humans (Latinus et al., 2013) suggests the possibility that listeners use a personal perceptive vocal identity map to discriminate between speakers.
Studying within-speaker variability from the perspective of psycholinguistics, evolutionary strategies, and human cooperation, it is often difficult to build hypotheses that regard specific features of speech. In these cases, it is more reasonable to make hypotheses about how speakers are placed in a norm-based vocal space under the conditions of interest. Despite this, there are no clear guides on the representation of such a vocal space from recorded utterances, and researchers choose methodology for this task arbitrarily. We would like to address this by exploring the different ways a norm-based coding vocal space (henceforth VS) can be built and analysed.

Proposal

We believe that the best starting point for feature extraction methods that can create a reliable VS can be found in speaker recognition (SR) software. Common examples of such methods are Mel-frequency cepstral coefficients (MFCCs) and deep neural network embeddings. In terms of vocal identity representation, pre-trained models using these methods excel. Pre-trained models are optimized to isolate different speakers; they process the input signal to represent a collection of a fixed number of features (embedding) that clusters same-speaker utterances. A VS representation is then a collection of embeddings of different speakers. On the other hand, pre-trained models fail to capture the dynamic nature of a speaker's speech, characterising within-speaker variability that may be of interest to researchers. In our opinion, this also fails to represent the nature of perception and SR in humans.
We will test several automatic feature extraction methods (including pre-trained, trained, and naïve models) on the task of building a reliable VS representation. In addition, we will also investigate which sets of measurements in this VS representation are the most useful in characterizing differences across samples in the representation. Examples of such metrics are Manhattan, Euclidean, and cosine distances, or statistical distribution divergence metrics.
Using the Matlab software TANDEM-STRAIGHT (Kawahara et al., 2009), we will create five morphs from two speakers using different proportions of each speaker's speech in each sample. We will then measure the correlation between the relative proportions of each speaker in the samples with the relative distances between the samples and the original speakers in the VS representation. This will be done for each method and metric combination, to find which best represents the true similarity between the morphed samples and the speakers. For example, applying the ideal method and metric combination to a morphed sample created by morphing 20% of speaker A with 80% of speaker B should result in a distance of the sample from each speaker in the vocal space representation that is respectively 80% and 20% of the total distance between A and B. Our goal is primarily to create a guide to vocal space exploration and analysis, describing which feature extraction methods and metrics best represent the norm-based coding vocal space and within speaker variation in vocal identity. If our resources allow it, we will also strive to publish a software tool to aid researchers in working in the framework described.

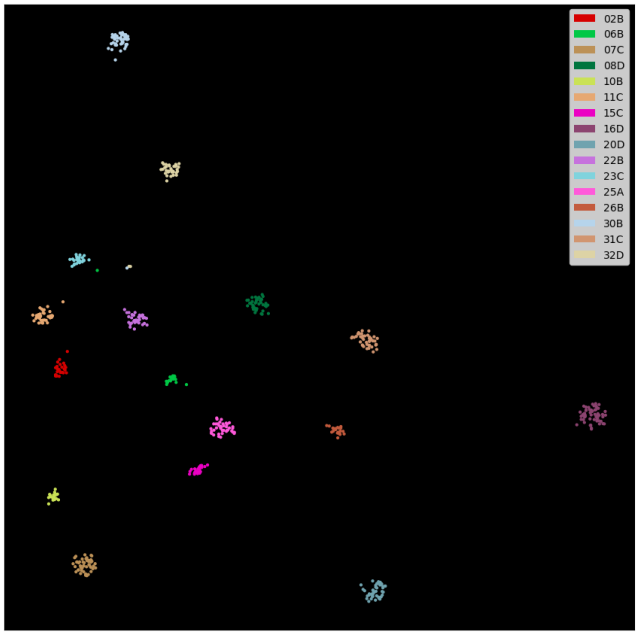


Figure 1. Example of a vocal identity space of 16 speakers built using ECAPA-TDNN speaker embeddings (Desplanques et al., 2020) and UMAP to reduce to 2D.

References

Dellwo, V., Pellegrino, E., Lei, H., & Kathiresan, T. (2019). The dynamics of indexical information in speech: can recognizability be controlled by the speaker?. *AUC PHILOGICA,* 2019(2), 57-75.

Desplanques, B., Thienpondt, J., & Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*. https://doi.org/10.48550/arXiv.2005.07143.

Kawahara, H., Takahashi, T., Morise, M., & Banno, H. (2009). Development of exploratory research tools based on TANDEM-STRAIGHT. In *Proceedings: APSIPA ASC 2009: Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference,* 111-120. Asia-Pacific Signal and Information Processing Association, 2009 Annual Summit and Conference, International Organizing Committee. http://hdl.handle.net/2115/39651.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology,* 23(12), 1075-1080. https://doi.org/10.1016/j.cub.2013.04.055.

The effects of sinusitis and voice core polyp surgery on forensic speaker diagnosis and recognition examination

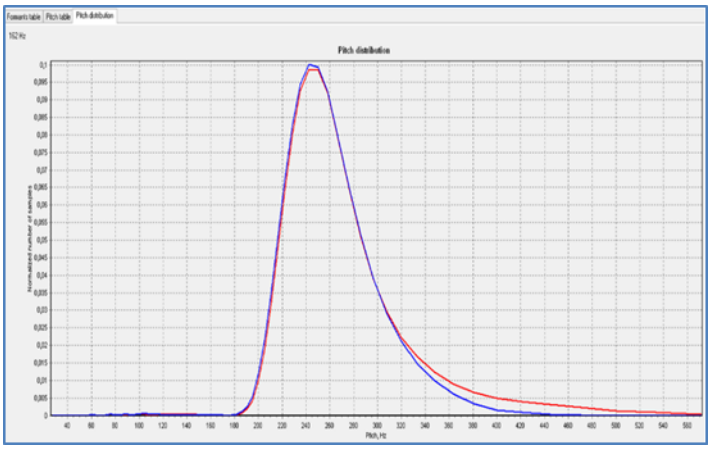
Bahar Akgün Okmuş[1] and Ayfer Batmaz[2]
Ankara Regional Criminal Police Laboratory Directorate, ANKARA, Türkiye
[1]Voice Idendification Expert, Inspector
bahar.akgunokumus@egm.gov.tr
[2]Voice Idendification Expert, Police Officer
ayfer.batmaz@egm.gov.tr

Crimes committed by physical tools and methods in the past are increasingly being replaced by crimes committed by electronic tools with the increase of digitalization. These crimes committed by electronic tools leave behind electronic evidence such as computers, portable devices, network devices and storage devices. The examination of audio evidence or audio forensic (audio forensic) is known as a sub-branch of Forensic Informatics discipline that contributes to the clarification of crimes by using electronic evidence and the data contained in them. There are many audio analysis software currently available for audio forensics. Using methods based on digital sound processing techniques, these softwares are designed for editing/assembly detection in audio recordings (verification of recording integrity), spectrum analysis, increasing the intelligibility of corrupted recordings (record improvement), speaker profile (age, gender, etc.) determination, voice analysis masked by ambient noises. It is expected to automatically perform voice forensics such as real-time voice analysis in forensic follow-ups and identifying language and accent elements related to the ethnicity of the speaker.

In this research; Medical sinusitis (Picture-1) and vocal cord polyp (Picture-2) surgery were evaluated in terms of Criminal Voice Examination. It was aimed to determine whether there was any change in the voices of people who had sinusitis and vocal cord polyp surgery, preoperative and postoperative auditory and visual parameters. Audio recordings were examined based on audio and visual methods and the forensic speaker was evaluated in terms of all parameters in the diagnosis and recognition method.

In the research carried out by Ankara University İbn-i Sina Hospital, Department of Otorhinolaryngology; Voice samples were taken before and after surgery from 19 volunteer patients who will have sinusitis and vocal cord polyp surgery and auditory and visual (similarity scores produced by software) comparisons were made on these recordings. A total of 38 recordings were obtained from the voice samples used before and after the sinusitis and vocal cord polyp surgery of the patients who had sinusitis and vocal cord polyp surgery.

After the auditory and visual analyzes, the F0, F1, F2, F3 and F4 formant frequency values of all speakers (Table-1) did not affect the examination of vowel phonemes. It has been determined that there are changes between 5/15 (+/-) Hz. Since these changes are at very low levels, it has been observed that they are not of value to affect speaker recognition and within the accepted limit.



Graphic-1 Correlation Chart of Nicknamed Male Speaker as "B.Y."

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (a)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
A-A-TAKAN	115	592	1236	2737	3437	117	578	1246	2758	3308
A-AT-A-KAN	120	542	1579	2483	3550	123	523	1571	2480	3473
A-ATAK-A-N	124	581	1275	2580	3598	124	559	1304	2678	3501
ORT	120	572	1363	2600	3528	121	553	1374	2639	3427

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
1%	-3%	1%	1%	-3%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (e)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
E-E-RKENDEN	119	516	1950	2586	3440	120	491	1992	2571	3464
E-ERK-E-NDEN	106	525	1819	2729	3562	108	574	1862	2734	3464
E-ERKEND-E-N	113	486	1513	3000	3502	114	501	1529	2996	3579
ORT	113	509	1761	2772	3501	114	522	1794	2767	3502

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
1%	3%	1%	0%	2%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (i)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
İ-İ-ZMİRLİ	115	360	1944	2764	3671	116	400	1985	2784	3665
İ-İZM-İ-RLİ	97	392	1980	3136	3763	97	418	1954	3103	3742
İ-İZMİRL-İ	110	344	1914	2724	3594	111	352	1941	2745	3452
ORT	107	365	1946	2875	3676	108	390	1960	2877	3620

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
1%	2%	1%	1%	-4%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (ı)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
I-I-LIMLI	397	405	1759	3208	3706	398	401	1617	3219	3641
I-IL-I-MLI	105	422	1055	3191	3678	106	423	1064	3109	3511
I-ILIML-I	120	370	1839	2925	3540	121	382	1824	2959	3442
ORT	207	399	1551	3108	3641	208	402	1502	1611	3531

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
1%	3%	-1%	1%	-3%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (o)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
O-O-NKOLOK	104	506	813	3089	3948	105	489	915	2966	3865
O-ONK-O-LOK	112	513	839	2779	3482	113	506	1004	2731	3303
O-O-NKOL-O-K	121	505	885	2759	3439	121	498	903	2770	3355
ORT	112	508	846	2876	3623	113	498	941	2822	3508

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
0%	-1%	2%	0%	-2%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (ö)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
Ö-AMAT-Ö-R	100	491	1516	2523	3515	109	512	1426	2530	3445
Ö-G-Ö-L	122	480	1729	2213	3444	122	497	1640	2235	3354
Ö-Ö-ZEL	446	445	1550	2552	3468	440	433	1548	2575	3389
ORT	223	472	1598	2429	3476	220	481	1538	2447	3396

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
-1%	-3%	0%	1%	-2%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (u)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
U-U-YUMLU	102	408	1337	2600	3744	103	405	1317	2665	3671
U-UY-U-MLU	113	402	981	3121	3595	113	435	1047	3076	3431
U-UYUML-U	131	424	1036	2971	3557	130	433	1040	2914	3435
ORT	115	411	1118	2897	3632	115	424	1135	2885	3512

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
-1%	2%	0%	-2%	-3%

Formant Değeri Alınan Söylem	B.Y. İSİMLİ ERKEK KONUŞMACIYA AİT FORMANT DEĞERLERİ (ü)									
	Ameliyat Öncesi					Ameliyat Sonrası				
	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)	F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
Ü-Y-Ü-ZÜNÜ	110	328	1821	2447	3524	110	388	1790	2434	3006
Ü-YÜZ-Ü-NÜ	106	415	1692	2500	3190	121	444	1665	2463	3037
Ü-YÜZÜN-Ü	129	442	1663	2473	3260	128	440	1658	2474	3134
ORT	115	395	1725	2473	3325	120	424	1704	2457	3059

F0(Hz)	F1(Hz)	F2(Hz)	F3(Hz)	F4(Hz)
-1%	0%	0%	0%	-4%

Table-1 Comparative Formant Values of "a,e,ı,i,o,ö,u,ü" Phoneme of Nicknamed Male Speaker as "B.Y."

Intensity dynamics variation across different speech rates

Homa Asadi
Department of Linguistics, University of Isfahan, Isfahan, Iran
h.asadi@fgn.ui.ac.ir

This study aims to investigate the speaker-specific temporal characteristics, with a particular focus on the variability of syllable intensity within a Persian-speaking population. Previous research has demonstrated that the analysis of speech rhythm using measurements derived from the duration of consonantal and vocalic intervals, along with syllable intensity, can effectively distinguish between speakers (Leemann et al., 2014; He & Dellwo, 2016; Asadi et al., 2018). Considerable evidence exists regarding the association between temporal aspects of speech and the articulatory movements of speech organs, specifically the lips, jaw, and tongue, which exhibit systematic variations corresponding to speech rate (DeNil & Abbs, 1991; Berry, 2011; ILLA & Ghosh, 2020). Speakers modify their speaking rate due to various factors, including voice style, voice disorders, aging, and transient emotional states such as anger, rage, or happiness. Variations in speech rate impose different articulatory demands, thereby influencing articulatory movements and affecting the acoustic properties of speech.

The current study seeks to determine if temporal features in terms of syllable intensity can successfully discriminate speakers when they employ different speech rates. To accomplish this, a group of ten male native Persian speakers, aged between 24 and 36, with a mean age of 31.3 and a standard deviation of 3.7, were instructed to read "*The North Wind and the Sun*" at three different speech rates (normal, slow, and fast). Data recordings were carried out in a soundproof booth using a sample rate of 44.1 kHz and a quantization of 16 bits. To make sure that the recording devices were not compressing or limiting the dynamic range of intensity measurements, a headset microphone was placed at a constant distance from the speaker's lips. The dataset comprises a total of 240 speech tokens (10 speakers × 8 sentences × 3 speech rates), leading to substantial syllable rate variation across reading passages. Following He and Dellwo (2017), mean, standard deviation and sequential variability (Pairwise Variability Index (PVI)) were measured for positive and negative intensity dynamics, as indicators of syllable intensity variability. Positive dynamics refer to the rate of intensity increase from an amplitude envelope trough point to a subsequent peak point, which corresponds to mouth-opening gestures. On the other hand, negative dynamics represent the speed of intensity decrease from a peak to a subsequent trough point, which correlates with mouth-closing gestures. Preliminary findings from the Multinomial Logistic Regression (MLR) analysis revealed that negative dynamics captured a significant portion of the variability between speakers at slow and normal speech rates. Conversely, between-speaker variability was primarily attributed to positive dynamics at the fast speech rate. This indicates that speakers exhibited greater variation in mouth-closing gestures during slow and normal rates, while differences in mouth-opening gestures were more pronounced during the fast rate. The individual contribution of each intensity measure is shown in Figure 1.

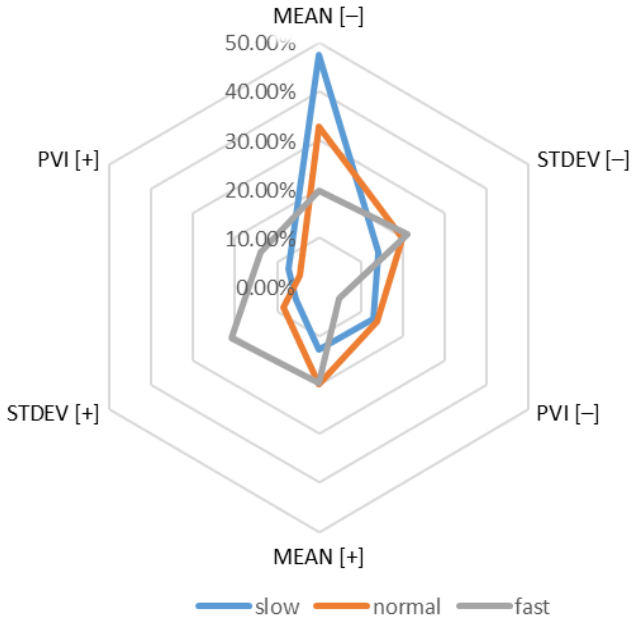


Figure 1. Radar chart visualizing the amount of individual contribution of each investigated parameter in the multinomial logistic regression model for speaker (slow rate= blue color, normal rate= red color, fast rate= gray color).

References

Asadi, H., Nourbakhsh, M., He, L., Pellegrino, E., & Dellwo, V. (2018). Between-speaker rhythmic variability is not dependent on language rhythm, as evidence from Persian reveals. *International Journal of Speech, Language and the* Law, 25(2), 151- 174.

Berry, J. (2011). Speaking Rate Effects on Normal Aspects of Articulation: Outcomes and Issues. *Perspectives on Speech Science and Orofacial Disorders, 21*, 15-26.

DeNil, LF., Abbs, JH. (1991). Influence of speaking rate on the upper lip, lower lip, and jaw peak velocity sequencing during bilabial closing movements. *Journal of the Acoustical Society of America*, 89(2): 845-9.doi:10.1121/1.1894645.

He, L., & Dellwo, V. (2016). The role of syllable intensity in between-speaker rhythmic variability. *The International Journal of Speech, Language and the Law*, Vol 23, 243-273.

He, L., & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *Journal of the Acoustical Society of America*, 141(5): EL488–EL494.

Illa, A., & Ghosh, P.K. (2020). The impact of speaking rate on acoustic-to-articulatory inversion. *Computer Speech & Language., 59*, 75-90.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: implications for forensic voice comparison. *Forensic Science International*, 238, 59-67.

Conference organization

The **31st International Association for Forensic Phonetics and Acoustics (IAFPA) Conference** is held in Zurich, Switzerland, July 9 – 12, 2023.

The conference is organized by
Centre for Forensic Phonetics and Acoustics (CFPA) at the University of Zurich (UZH) and the Zurich Forensic Science Institute (FOR).

In collaboration with
Linguistic Research Infrastructure (LiRI), Zurich, Switzerland.

Organizing committee

- Leah Bradshaw
- Volker Dellwo
- Alessandro De Luca
- Peter French
- Daniel Friedrichs
- Andrea Frölich
- Lei He
- Carolina Lins Machado
- Elisa Pellegrino
- Valeriia Perepelytsia
- Alejandra Pesantez

With the support of Agnes Kolmer and Raffaela Zaugg.

Venue
The conference will take place in rooms of the University of Zurich (UZH). The main program will be at building AFL of UZH directly situated at Zurich Oerlikon train station. Visiting address: Affolternstrasse 56, Zürich.

Contact Information
E-mail: iafpa2023@ifi.uzh.ch
Conference website: www.iafpa2023.uzh.ch

Acknowledgements

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

- UZH alumni, University of Zurich
- Swiss National Science Foundation (SNF)
- UZH Graduate Campus (GRC)