Book of Abstracts

30th Annual Conference
of the
International Association
for Forensic Phonetics and Acoustics

Charles University, Prague, Czech Republic
July 10 – 13, 2022

FACULTY OF ARTS
Charles University

Institute
of Phonetics

Welcome

We are pleased to welcome you to the 30th Annual Conference of the International Association for Forensic Phonetics and Acoustics at the Charles University in Prague, Czech Republic. Our department, the Institute of Phonetics, hosts the conference at the university's Faculty of Arts – an almost 100-year-old historical building at Staroměstská, with a stunning view of the Vltava river and the Prague Castle.

We are going to welcome three keynote speakers – **Francis Nolan**, from the University of Cambridge, talking on the future of forensic phonetic experts; **Susanne Fuchs**, from the Leibniz-Centre General Linguistics in Berlin, speaking about the flexibility and stability of respiration in human actions; and **Petr Schwarz**, from Brno University of Technology, discussing the current trends in voice biometry research. Besides the plenary talks, the conference will feature 20 oral presentations and 30 posters by authors coming from 16 countries.

To learn more about the scientific content of the conference, please, see the programme below. In the programme, presentations marked with an "S" are student submissions eligible for **student paper awards**. After listening to all of them, you will be provided with a QR code and chance to vote for the best student paper award. This year, two students (of one oral presentation and of one poster) will be awarded this prize. The recipients will have the registration fee for the next IAFPA conference waived.

As for the conference social events, we start with a welcome drink reception in the 14th-century New Town Hall located at Karlovo náměstí (['naːmɲɛsciː] = *square*) on Sunday; students are eagerly expecting a beer night out at Zahrádky Letná on Monday evening; and on Tuesday evening we hope to see all participants on a boat tour during which we will sail through Prague and enjoy Czech food and beer together with live music.

We would like to thank the Faculty of Arts for their assistance with the organization of the conference, and to the KREAS project for support.

We are looking forward to greeting you in Prague!

on behalf of the local organizing committee

Radek Skarnitzl

The local organizing committee:

Tomáš Bořil

Alžběta Houzar

Tomáš Nechanský

Naty Nudga

Míša Svatošová

Jan Volín

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

30th annual conference

of the

International Association for Forensic Phonetics and Acoustics

Monday, July 11, 2022

9:00	Registration	
9:30	Conference opening	
	Chair: Tomáš Bořil	
9:40	**Vincent Hughes, Carmen Llamas and Thomas Kettig** A game-based approach to eliciting and evaluating likelihood ratios for speaker recognition	
10:00	**Fernanda Lopez-Escobedo and N. Sofía Huerta-Pacheco** Web application that generates reference values of acoustic parameters for forensic studies in Mexican Spanish	
10:20	**Elliot Holmes** Recognising socio-phonetically comparable speakers using phonetic approaches to automatic speaker recognition	**S**
10:40	**Finnian Kelly, Harry Swanson, Kirsty McDougall and Anil Alexander** Classifying non-speech vocalisations for speaker recognition	
	11:00 - 11:30 Coffee break	
	Chair: Radek Skarnitzl	
11:30	PLENARY TALK **Francis Nolan** Will forensic speech scientists still need ears?	
	12:15 - 13:45 Lunch	
	Chair: Alice Paver	
13:45	**Kristina Tomić and Peter French** Voice quality and voice similarity in cross-language forensic speaker comparison – Perception experiments	**S**
14:05	**Valeriia Perepelytsia, Nathalie Giroud, Tugce Aras, Martin Meyer and Volker Dellwo** Neural underpinnings of familiar talker advantage: an EEG study	**S**

14:25	**Willemijn Heeren, Laura Smorenburg and Erica Gold** Optimizing the strength of evidence: Combining segmental speech features	
	14:45 – 15:15 Coffee break	
	Chair: Alice Paver	
15:15	**Jim Hoskin and Paul Foulkes** Shifting the Burden: towards more robust and transparent procedures for LADO	**S**
15:35	**James Tompkinson, Kate Haworth, Emma Richardson, Felicity Deamer and Magnus Hamann** For the Record: Improving standards in the production of non-expert police interview transcripts	
16:00 – 17:30 Poster session I		
18:00	Students' meeting	

Tuesday, July 12, 2022

	Chair: Richard Rhodes	
9:00	**Ben Gibb-Reid, Paul Foulkes and Vincent Hughes** Just the way you are: The potential of the word just as a speaker discriminant	**S**
9:20	**Laura Smorenburg and Willemijn Heeren** The effect of linguistic contexts on the acoustics and strength-of-evidence of /s/	**S**
9:40	**Alice Paver, Natalie Braber and David Wright** Listener judgements for social traits and criminal behaviours as a function of speaker pitch and articulation rate	
10:00	**Nikita Suthar and Peter French** Role of within-vowel formants in forensic speaker comparison	**S**
	10:20 – 10:50 Coffee break	
	Chair: Jan Volín	
10:50	PLENARY TALK **Susanne Fuchs** Flexibility and stability of respiration in human actions	
	11:50 – 13:20 Lunch	
	Chair: Jan Volín	
13:20	**Leah Bradshaw and Volker Dellwo** Speech variability in telephone openings and its implications for speaker discrimination	**S**
13:40	**Luke Carroll and Georgina Brown** Towards a perceptual rhythm framework for forensic analysis: methodological developments	**S**

14:00 – 15:30 Poster session II	
	15:30 – 16:00 Coffee break
16:00	AGM
	19:00 – 23:00 Conference dinner

Wednesday, July 13, 2022

Chair: Radek Skarnitzl	
10:00	PLENARY TALK **Petr Schwarz** Current trends in voice biometry research and industrial efforts
Chair: Tomáš Bořil	
10:50	**Simon Gonzalez** An acoustic-phonetic description of diphthongs in Venezuelan Spanish
	11:10 – 11:40 Coffee break
11:40	**Honglin Cao, Chuyi Pan and Lei He** Speech length threshold in forensic voice comparison by using long-term fundamental frequency in Chinese Mandarin
Chair: Volker Dellwo	
12:00	**Arjan van Dijke** Case report: Forensic analysis of a ticking clock in a recording
12:20	**Vincent van Heuven and Sandra Ferrari Disner** Utility of length-normalization for predicting trademark sound-alikes from Levenshtein string edit distance
12:40	**Francis Nolan, Nikolas Pautz, Kirsty McDougall, Katrin Müller-Johnson, Harriet Smith and Alice Paver** The impact of reflection and retention intervals on earwitness accuracy: Two experiments
13:00	Conference closing

Poster session I

Monday, July 11, 16:00 – 17:30

PB = poster board

PB 1	**Jakub Bortlík** The performance of two ASR systems in language mismatch, foreign accent, and channel mismatch conditions	
PB 2	**Leah Bradshaw, Chiara Tschirner, Lena Jäger and Volker Dellwo** Using eye-tracking as a method to explore decision making in voice recognition tasks	**S**

PB 3	**Linda Gerlach, Kirsty McDougall, Finnian Kelly and Anil Alexander** Seeking voice twins – an exploration of VoxCeleb using automatic speaker recognition and two clustering methods	**S**
PB 4	**Lauren Harrington** The effect of listener accent background on the transcription of Standard Southern British English	**S**
PB 5	**Alžběta Houzar, Tomáš Nechanský and Radek Skarnitzl** Impact of vocal tract resonance modifications on LTF and f0	**S**
PB 6	**Vincent Hughes and Bruce Wang** Forensic experts should focus on uncertainty rather than discriminability	
PB 7	**Katharina Klug** Assessing the specificity of creaky voice quality for forensic speaker comparisons	**S**
PB 8	**Carolina Lins Machado and Lei He** Inter-speaker variability in the American English /æ/ and /ɑ/: a dynamic view from both tongue articulation and the first two formants	
PB 9	**Sarah Melker** Salient cues to age identity in an LX: a longitudinal pilot study on a female L1 Hungarian speaking English	
PB 10	**Sophie Möller and Gea de Jong-Lendle** When singing becomes illegal	**S**
PB 11	**Bryony Nuttall, Phillip Harrison and Vincent Hughes** Automatic Speaker Recognition performance with (mis)matched bilingual speech material	
PB 12	**Elisa Pellegrino, Homa Asadi and Volker Dellwo** Voice discrimination across speaking styles	
PB 13	**Sascha Schäfer and Paul Foulkes** Voice memory as an estimator variable in lay speaker identification tasks	**S**
PB 14	**Ravina Toppo and Sweta Sinha** Unmasking identity through acoustic analysis: A case study of Indian English	
PB 15	**Raphael Werner, Juergen Trouvain and Bernd Möbius** Speaker discrimination and classification in breath noises by human listeners	**S**

Poster session II

Tuesday, July 12, 14:00 – 15:30

PB 1	**Homa Asadi, Maral Asiaee and Volker Dellwo** Acoustic variation within Persian-English bilingual speakers	
PB 2	**Meike de Boer and Willemijn Heeren** Language-dependency of /s/ in L1 Dutch and L2 English	**S**
PB 3	**Ricky K.W. Chan and Bruce Xiao Wang** Evidential value of long-term laryngeal voice quality acoustics	
PB 4	**Lois Fairclough** Exploring covariation as a marker of speaker specificity	**S**

PB 5	**Andrea Fröhlich, Volker Dellwo and Meike Ramon** The quest to find auditory 'Super-Recognizers' - Results from a pilot study	
PB 6	**Bence Halpern and Finnian Kelly** Can DeepFake voices steal high-profile identities?	**S**
PB 7	**Lauren Harrington, Robbie Love and David Wright** Analysing the performance of automated transcription tools for covert audio recordings	**S**
PB 8	**Vincent Hughes, Paul Foulkes, Philip Harrison, David van der Vloed and Finnian Kelly** Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR	
PB 9	**Jacek Kudera and Bernd Möbius** Auditory and machine-based identification of closely related languages: A comparison of methods for LADO procedure	
PB 10	**Beeke Muhlack, Jürgen Trouvain and Michael Jessen** Acoustic characteristics of filler particles in German	**S**
PB 11	**Amanda Muscat** Speaker/author profiling in Maltese	**S**
PB 12	**Tomáš Nechanský, Alžběta Houzar and Radek Skarnitzl** The effect of free voice-disguise methods on ASR performance	**S**
PB 13	**Vojtěch Skořepa and Radek Skarnitzl** Notorious and new voice: How does a professional imitator fare?	**S**
PB 14	**Bruce Wang and Vincent Hughes** Reducing the degree of uncertainty within automatic speaker recognition systems using a Bayesian calibration model	
PB 15	**Samantha Williams** Analysis of Forced Aligner Performance on Non-native (L2) English Speech	**S**

Prague, July 10–13, 2022

A game-based approach to eliciting and evaluating likelihood ratios for speaker recognition

Vincent Hughes[1], *Carmen Llamas*[1], *and Thomas Kettig*[1]
[1]*Department of Language and Linguistic Science, University of York, UK*
{vincent.hughes|carmen.llamas|thomas.kettig}@york.ac.uk

This presentation describes a bespoke computer game which doubles as a sociolinguistic experiment that elicits and evaluates likelihood ratio (LR)-like scores from human non-expert listeners in a speaker recognition task. Previous work has examined human speaker recognition performance with unfamiliar voices; however, very little research has attempted to compare and combine such results with those of automatic speaker recognition (ASR) systems due to the considerable challenges in extracting judgements from humans that are both logically and empirically comparable with outputs of data-driven systems. Our project, *Humans and Machines: Novel Methods for Assessing Speaker Recognition Performance*, aims to provide a framework for combining LR-like judgements from human listeners with the output of an ASR system. We also explore sources of cognitive bias on human responses. We focus here on our methodology and experimental design.

Over the course of gameplay, participants encounter a series of voice comparisons. In each comparison, participants first listen to one stimulus (the nominal 'criminal' or 'unknown' sample) and rate on a 0–100 scale how typical they consider the voice to be relative to other speakers of the same accent. They are then presented with a second stimulus (the nominal 'suspect' or 'known' sample) and asked to provide a judgement of the similarity between this and the first sample on a 0–100 scale. Finally, participants indicate on a 0–100 scale whether they think the two voices belong to the same speaker.

The stimuli used in the game are 10-second audio samples of the speech of male British English speakers extracted from two corpora: the *Dynamic Variability in Speech* corpus (Nolan et al. 2009) and *The Use and Utility of Localised Speech* corpus (Llamas, French & Watt 2016-19). In addition to demographic information, participants initially provide judgements on a 0–100 scale to indicate how familiar they are with the three accents represented by the stimuli: Newcastle, Middlesbrough, and Standard Southern British English (SSBE).

The first stimulus in each pair is a far-end landline telephone recording while the second stimulus is a high-quality studio recording; this channel-mismatch replicates common conditions within forensic voice comparison casework. Participants encounter both same-speaker (SS) and different-speaker (DS) pairs; DS pairs are matched for regional accent except for a set of cross-accent Middlesbrough-Newcastle pairs. We can thus explore the effect of self-identified familiarity with an accent on speaker recognition performance.

Furthermore, in order to probe how listeners' LR scores might be affected by situating the task in a legal context, the game is comprised of several levels in which: 1) no legal context is supplied; 2) participants are immersed in their role on a 'jury of the future'; 3) participants are primed with extralinguistic evidence; 4) participants are given advice from an 'expert phonetician'.

LR-like scores are calculated by dividing average listener similarity and typicality judgements. Tests of initial prototypes have confirmed that listener judgements about similarity and typicality do produce LR-like scores that can be calibrated and evaluated like any other speaker recognition system.

References

Llamas, C., Watt, D. and French, P., 2016-2019, *The use and utility of localised speech forms in determining identity: forensic and sociophonetic perspectives*. UK Economic and Social Research Council (ES/M010783/1)

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

Web application that generates reference values of acoustic parameters for forensic studies in Mexican Spanish

Lopez-Escobedo Fernanda[1], *Huerta-Pacheco N. Sofía*[2]
[1]*UNAM, CdMx, México*
flopeze@unam.mx
[2] *Cátedra-CONACYT, UNAM, CdMx, México*
nshuerta@cienciaforense.facmed.unam.mx

This paper presents a web application to obtain statistics of acoustic properties in a Mexican Spanish Oral Corpus designed as a forensic application. An international survey about practices in forensic speaker comparison considers four main approaches to compare recordings (Morrison & Enzinger 2019): auditory, spectrographic, acoustic-phonetic, and automatic. The acoustic-phonetic approach consists in making quantitative measurements of acoustic properties such as fundamental frequency and formant frequencies. The acoustic information is used to compare an unknown recording to a known recording, but often technical conditions are not the same. It is common, in forensic speaker comparison, that the unknown recording is a telephone call, and the known recording is a direct microphone recording of a police interview. Many studies have shown that telephone transmissions have several effects on measuring fundamental frequency and formant frequencies (Künzel, 2001; Byrne & Foulkes, 2004; Guillemin & Watson, 2008; Lawrence, Nolan, & McDougall, 2008; Zhang, Morrison, Enzinger & Ochoa, 2013) that make an important contrast to other sources such as high-quality recordings.

The web application is meant to be a support for experts using the acoustic-phonetic approach providing them with reference values of the acoustic parameters in recordings with high-quality versus telephone transmission conditions. Data was obtained using recordings from 133 voluntary speakers (48 male and 65 female), Mexican Spanish speakers from Mexico City aged 20 years or more. Participants were recorded reading a phonetic balance text, among other tasks, in high-quality conditions at 44100 sampling frequency in a noise-cancelling room. For the telephone transmission condition, each recording was edited to simulate the telephone bandpass: from 300 Hz to 3300 Hz. Then, each vowel and diphthong were segmented and tagged with different phonetic characteristics such as word stress and syllable structure. Fundamental frequency, the first four formants and duration were extracted using the code of Parselmouth library in Python of Praat software. Figure 1 displays an idea of the functionality of the web application where experts can filter the data selecting technical conditions; age, level of education, and genre of speakers; vowel/diphthong; word stress, and syllable structure. Descriptive statistics like minimum, maximum, mean, median, variance, and standard deviation are display for each acoustic parameter. Figure 2 shows an example of distributional graphics (boxplot and density plot) generated by the application.

Finally, in this application, the frequency distribution of two technical characteristics (telephone call and interview) is presented comparatively, thus showing the reference values expected in each one, according to the filters selected by the expert.

It is important to note that this tool could be continually updated with new data in the future.

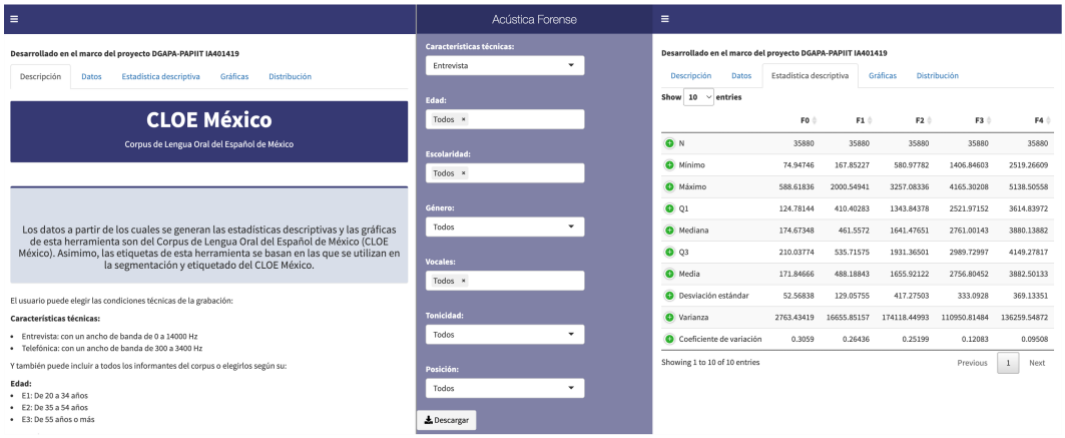


Figure 1. Appearance of the web application.

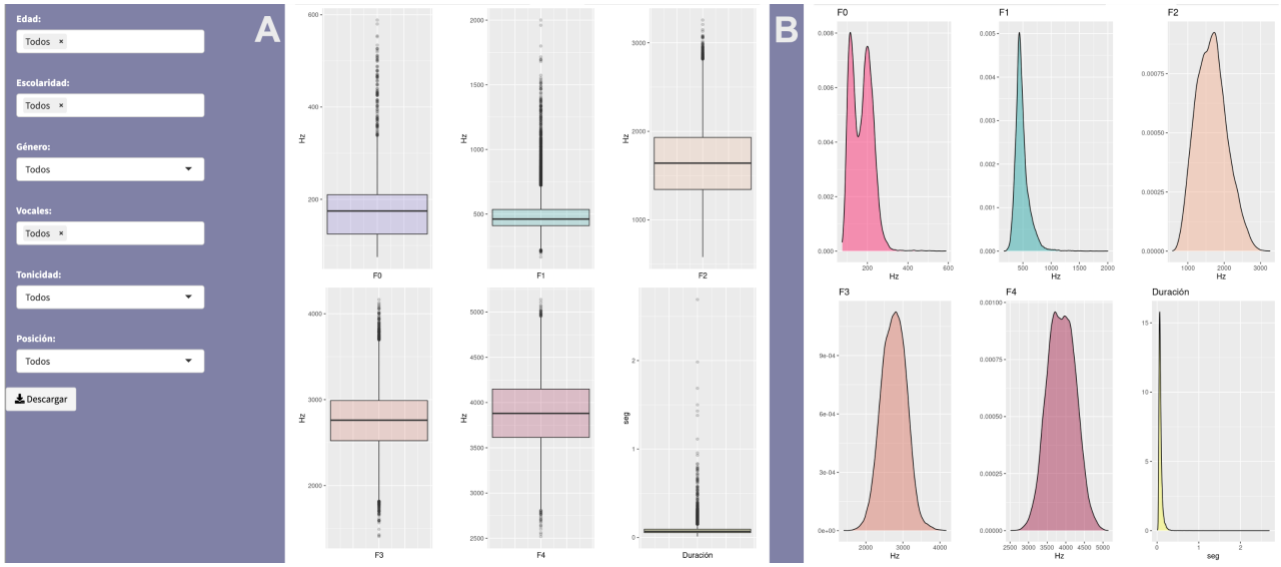


Figure 2. Example of distributional graphics generated by the web application.

References

Byrne, C., & Foulkes, P. (2004). The 'Mobile Phone Effect' on Vowel Formants. *International Journal of Speech Language and the Law*, 11 (1): 83-102.

Guillemin, B. J., & Watson, C. (2008). Impact of the GSM mobile phone network on the speech signal: some preliminary findings. *International Journal of Speech, Language & the Law*, 15(2).

Künzel, H. J. (2001). Beware of the 'telephone effect': the influence of telephone transmission on the measurement of formant frequencies. *Forensic linguistics*, 8(1), 80-99.

Lawrence, S., Nolan, F., & McDougall, K. (2008). Acoustic and perceptual effects of telephone transmission on vowel quality. *International Journal of Speech, Language & the Law*, 15(2).

Morrison, G. S., & 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.

Zhang, C., Morrison, G. S., Enzinger, E., & Ochoa, F. (2013). Effects of telephone transmission on the performance of formant-trajectory-based forensic voice comparison–female voices. *Speech Communication*, 55(6), 796-813.

Prague, July 10–13, 2022
Student Paper

Recognising Socio-Phonetically Comparable Speakers Using Phonetic Approaches to Automatic Speaker Recognition

Elliot Holmes
Department of Language and Linguistic Science, University of York, York, UK
Aculab PLC, Milton Keynes, UK
elliot.holmes@york.ac.uk

Modern approaches to Automatic Speaker Recognition (ASR) are undeniably powerful though they are undermined by their lack of interpretability. Such systems, like Mokgonyane et al.'s (2019), employ machine learning algorithms to achieve accuracy rates as high as 96%; however, these algorithms are 'black boxes.' As Rudin (2019) explains, this means that their processes are uninterpretable to their creators; consequently, when these systems fail, their problems cannot be diagnosed nor rectified. However, Phonetic Theory offers opportunities to re-integrate interpretability into these modern approaches whilst also improving performance: for example, Gonzalez-Rodriguez (2014) analysed the false rejections of modern uninterpretable ASR systems and found that voice creakiness, a perceptual feature of the voice that is quantifiable using interpretable phonetic features like Jitter, was found in all false rejections. Thus, if these interpretable phonetic features were integrated into the ASR system, its performance and interpretability would improve.

Motivated by the evident power of phonetic approaches to improve ASR performance and interpretability, Holmes (2021) designed a novel methodology that can identify the optimum combination of phonetic features to measure on a given phoneme to recognise socio-phonetically comparable speakers. These phonetic features include Pitch, Intensity, Formants, Autocorrelation, Harmonics-To-Noise Ratio, Periods, Jitter, and Shimmer. Phonemes have been employed as phonetically-interpretable segments of speech and socio-phonetically comparable speakers are focused on because it allows for optimum combinations of features on phonemes to be tailored to different socio-phonetic groups of speakers; by controlling social categories like age, gender, and accent, the differences identified will be idiosyncratic and therefore not attributable to broader social differences.

With this established rationale, Holmes' (2021) methodology takes a database of speakers, segments the entailed speech into its component phonemes, takes measurements of the selected features across the chosen phonemes, and finally calculates the optimum combination of phonetic features per phoneme based on whether the removal of that feature increases upon the baseline Log-Likelihood Ratio Cost (C_{llr}) of having all features considered. This indicates that the feature is necessary; without it, performance declines. The present study demonstrates the effectiveness of this interpretable phonetic approach to ASR in recognising the speakers of Nolan et al.'s (2009) DyVis Corpus who are all male, aged 18-24, and speak Southern Standard British English (SSBE). It focuses on three phonemes identified by Paliwal (1984) as particularly useful for speaker recognition, namely /i/, /a/, and /u/, and through three replications found the optimum combination of features for each phoneme for this socio-phonetically comparable group of speakers. The results can be found in Table 1 below.

Phoneme	*Baseline C_{llr}*	*Features that increase on Baseline C_{llr} when removed*	*Optimum C_{llr}*
/i/	0.77	Intensity, Formant 2, Formant 4, Bandwidth 2, Bandwidth 4	0.65
/a/	0.81	Formants 1-5, Bandwidths 1-5, Shimmer (Local)	0.66
/u/	0.83	Intensity, Formant 1, Bandwidth 1, Mean Autocorrelation, Jitter (Local, Absolute), Shimmer (Local)	0.74

Table 1. Optimum combinations of features per phoneme for the chosen socio-phonetically comparable speakers as realized by C_{llr}.

Overall, this study marks movement towards phonetically-informed ASR, having identified the best combination of interpretable phonetic features per phoneme for the recognition of speakers who can be defined as male, aged 18-24, and speaking SSBE. It has also provided a novel methodology that can allow more phonemes and more speaker groups to be tested so that the performance and interpretability of ASR can be improved using Phonetic Theory. Now, these optimum combinations should cross-validated on other data before ultimately being included alongside modern ASR to bolster their interpretability and, hypothetically, their performance.

References

Gonzalez-Rodriguez, J., Gil, J., Pérez, R. and Franco-Pedroso, R. (2014). *What are we missing with i-vectors? A perceptual analysis of i-vector based falsely accepted trials.* Speaker Odyssey 2014, Joensuu, Finland.

Holmes, E. J. (2021, February 4-5). Using Phonetic Theory to Improve Automatic Speaker Recognition. [Conference Presentation]. AISV, Zurich. https://elliotjholmes.wordpress.com/.

Mokgonyane, T. B., Sefara, T. J., Modipa, T. I., Mogale, M. M., Manamela, M. J., and Manamela, P. J. (2019). *Automatic Speaker Recognition System based on Machine Learning Algorithms.* [Paper]. 2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa, Bloemfontein, South Africa. https://ieeexplore.ieee.org/.

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. Forensic Linguistics, 16(1). https://www.researchgate.net/.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1, 206-215. https://www.nature.com/.

Classifying non-speech vocalisations for speaker recognition

Finnian Kelly[1], *Harry Swanson*[2], *Kirsty McDougall*[2], *and Anil Alexander*[1]
[1]*Oxford Wave Research Ltd., Oxford, U.K.*
{finnian|anil}@oxfordwaveresearch.com
[2]*University of Cambridge, Cambridge, UK.*
{hs686|kem37}@cam.ac.uk

Non-speech vocalisations (NSVs) are sounds speakers can produce with their vocal organs that do not have linguistic content, and that may or may not contribute meaning to a communication. Among such sounds are laughter, screams, yawns, moans, groans, sighs, throat clearings, hiccups, sneezes, and paralinguistic clicks.

There is little existing research on the relevance of NSVs to forensic speaker recognition, and in automatic speaker recognition they are typically discarded by the voice activity detection process, which occurs prior to speaker modeling and comparison. However, there have been some promising findings, e.g. Bachorowski et al. (2001) used laughter to classify speakers at above-chance levels using an automatic approach, and Engelberg et al. (2019) found participants were able to discriminate between speakers at above-chance levels from scream stimuli.

Despite the very limited research base, forensic practitioners do examine and sometimes use NSVs in real casework. Gold and French's survey of 36 FSC practitioners noted that 94% of respondents reported "examining non-linguistic features at least some of the time" (2011:302). This study explores whether real examples of NSVs can be classified automatically, with the aim of assessing their speaker-characterising properties, and ultimately of informing their use in speaker recognition.

Anikin & Persson's (2017) corpus of spontaneous NSVs (N = 603) was used as a source of data. The corpus comprises audio clips extracted from YouTube videos, containing either a single syllable or a bout (series of syllables) produced in a single emotional state. Each clip is labelled with one of nine emotional categories, and one of eight call types (grunt, laugh, moan, roar, scream, sigh, tone, whimper). Anikin & Persson's corpus is favourable to other NSV corpora (Belin et al. 2008, Sauter et al. 2010, Lima et al. 2013, Holz et al. 2021) as it contains spontaneous, rather than acted vocalisations.

A pilot classification study was conducted using VOCALISE x-vectors (Kelly et al., 2019) within a speaker profiling framework. Audio clips from four classes of NSV call types were selected: roar (N=84), scream (N=91), laugh (N=109), and moan (N=38). Additionally, a 'normal' speech class (N=100) was created by extracting short audio clips of spontaneous speech from YouTube videos. All NSV and speech clips came from different speakers. A two-class experiment was conducted, whereby a classifier was trained and tested for every possible pairwise combination of the five classes (4 NSV, 1 speech). In each case, recordings were randomly divided into training and testing sets (ratio 3:1). A linear support vector machine (SVM) was trained using spectral (MFCC) x-vectors extracted from the training set, and applied to classify x-vectors extracted from the testing set. This was repeated 10 times with a different random train-test partition. The resulting average classification EERs (Equal Error Rates) were <1% for all combinations of NSV vs speech, and between 5.6% (laugh vs roar) and 11% (scream vs roar) within NSVs. This promising discrimination performance supports the use of spectral x-vectors for NSV classification, which will enable a systematic assessment of the effects of NSVs on speaker recognition.

References

Anikin, A., & Persson, T. (2017). Nonlinguistic vocalizations from online amateur videos for emotion research: A validated corpus. Behavior research methods, 49(2), 758-771.

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. The Journal of the Acoustical Society of America, 110(3), 1581-1597.

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal affective voices: A validated set of nonverbal affect bursts for research on auditory affective processing. Behavior Research Methods, 40, 531–539. doi:10.3758/BRM.40.2.531

Engelberg, J. W., Schwartz, J. W., & Gouzoules, H. (2019). Do human screams permit individual recognition? PeerJ, 7, e7087.

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

Holz, N., Larrouy-Maestri, P., & Poeppel, D. (2021). The paradoxical role of emotional intensity in the perception of vocal affect. Scientific reports, 11(1), 1-10.

Kelly, F., Forth, O., Kent, S., Gerlach, L., & Alexander, A. (2019). Deep Neural Network Based Forensic Automatic Speaker Recognition in VOCALISE using x-Vectors, 2019 AES International Conference on Audio Forensics.

Lima, C. F., Castro, S. L., & Scott, S. K. (2013). When voices get emotional: A corpus of nonverbal vocalizations for research on emotion processing. Behavior Research Methods, 45, 1234–1245.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. Quarterly Journal of Experimental Psychology, 63(11), 2251-2272.

Voice Quality and Voice Similarity in Cross-Language Forensic Speaker Comparison – Perception Experiments

Kristina Tomić[1] *and Peter French*[2]
[1]*Faculty of Philosophy, University of Niš, 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@jpfrench.com

Voice quality is the cumulative effect of laryngeal and supralaryngeal features of a speaker's voice present most of the time while the person is talking and can, thus, be perceived as 'characteristic auditory colouring of an individual speaker's voice' (Laver, 1980, p.1).
Laver (1980) provided a descriptive framework for articulatory, phonatory and muscular tension settings that was later developed into a full-fledged protocol known as the Vocal Profile Analysis Scheme (Laver et al., 1991; Wirz & Mackenzie Beck, 1995). The protocol has been modified and updated throughout the years and is now used in forensic speaker comparison research and casework (Mackenzie Beck, 2005; San Segundo & Mompeán, 2017; San Segundo et al., 2019; San Segundo, 2021). Despite potential practical difficulties in application, VPA is relatively robust rater-wise, provided that the terminology is clearly defined (San Segundo et al., 2019).
Cross-language forensic speaker comparison has not been widely undertaken in forensic casework, and its application was strongly discouraged at the beginning of the century (Rose, 2002, p.342). Furthermore, clause 3.10 of the IAFPA Code of Practice (2020) reminds forensic practitioners to 'exercise particular caution with cross-language comparisons'[1]. However, even though some previous research has found that aspects of voice quality are heavily language-dependent (Wagner & Braun, 2003), and that languages can differ in the articulatory settings (Cruttenden, 2014, p. 302), there is reason to believe that some habitual voice quality features are preserved even when the person speaks a different language (*see* Heeren et al.,2014; Meuwly et al., 2015; Krebs & Braun, 2015; Asiaee et al., 2019; Tomić & French, 2019). With an increasing casework need (*see* Künzel 2013; Milne et al., 2019), the present research has focused on identifying voice features that 'persist' across language switches by multilingual speakers and are more speaker-dependent than language-dependent, and therefore potentially useable in cross-language forensic speaker comparison.

Methodology

The study consists of two perceptual experiments. In the first experiment, three expert listeners evaluate the voices of 20 female native speakers of Serbian in telephone conversations according to the modified VPA scheme (San Segundo et al. 2019).[2] The experts are presented with two sets of recordings, in Serbian and English, respectively. To avoid rater bias, this is a blind listening task, meaning the listeners are not explicitly told which two recordings originate from the same speaker.
In the second experiment, 30 lay listeners evaluate the similarity of pairs of voices in telephone conversations on a 10-point Likert scale. The variables studied in this experiment are the speaker and the language; therefore, the listeners are presented with four sets of recordings: (1) Serbian – Serbian, same speaker; (2) Serbian – Serbian, different speakers; (3) Serbian – English, same speaker; (4) Serbian – English, different speakers. The pairs are presented in a randomized order to mitigate any listener fatigue effects. After each evaluation, the listeners are asked to describe the specific characteristics on which they have based their answer.
The results of the tests are proposed as an initial basis for identifying the nature and extent of voice quality features that might reliably be used in cross-language forensic speaker comparison cases.

[1] http://www.iafpa.net/about/code-of-practice/

[2] The modified version of the VPA scheme is used by JP French Associates Forensic Speech and Acoustics Laboratory, York, UK

References

Asiaee, M., Nourbakhsh, M., & Skarnitzl, R. (2019). Can LTF discriminate bilingual speakers? *Proceedings of the 28th Annual Conference of the Internatonal Assocaton for Forensc Phonetcs and Acoustcs.* Istanbul, July 14th-17th, 2019.

Cruttenden, A. (2014). *Gimson's Pronunciation of English* (Eigth ed.). London & New York: Routledge.

Heeren, W., van der Vloed, D., & Vermeulen, J. (2014). Exploring long-term formants in bilingual speakers. *Proceedings of the International Association for Forensic Phonetics and Acoustics conference, Zurich, Switzerland*, (pp. 39-40).

IAFPA Code of Practice. (2020). Retrieved February 2022, from IAFPA: http://www.iafpa.net/about/code-of-practice/

Keating, P., Esposito, C. M., Garellec, M., Khan, S., & Kuang, J. (2010). Phonation Contrasts Across Languages. *UCLA Working Papers in Phonetics, 108*, 188-202. Retrieved from https://escholarship.org/uc/item/9xx930j1

Krebs, P., & Braun, A. (2015). Long Term Formant measurements in bilingual speakers. *Proceedings of the IAFPA Annual Conference, 8 - 10 July 2015.* Leiden.

Künzel, H. (2013). Automatic speaker recognition with crosslanguage speech material. *International Journal of Speech Language and the Law, 20*(1), 21-44. doi:10.1558/ijsll.v20i1.21

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

Laver, J., Wirz, S., Mackenzie, J., & Hiller, S. (1991). A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress, 14*, pp. 139-155.

Mackenzie Beck, J. (2005). Perceptual Analysis of Voice Quality The Place of Vocal Profile Analysis. In W. J. Hardcastle, & J. Mackenzie Beck (Eds.), *A Figure of Speech: A Festschrift for John Laver* (pp. 285-322). Mahwah, New Jersey & London: Lawrence Erlbaum Associates.

Meuwly, D., Heeren, W., & Bolck, A. (2015). Exploring the strength of evidence of long-term formants in bilingual speakers. *Proceedings of Annual Conference of the International Association for Forensic Phonetics and Acoustics*, (pp. 75-76).

Milne, P., Cavanagh, C., van der Vloed, D., & Dellwo, V. (2019). A survey of voice-related cases in three forensic speech laboratories. *A paper presented at the 28th Annual Conference of the Internatonal Assocaton for Forensc Phonetcs and Acoustcs.* Istanbul, 14th -17th July 2019.

Rose, P. (2002). *Forensic Speaker Identification.* London and New York: Taylor & Francis.

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

San Segundo, E., & Mompeán, J. A. (2017). A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity. *Journal of Voice, 31*(5), 644.e11–644.e27. doi:10.1016/j.jvoice.2017.01.005

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. doi:10.1017/S0025100318000130

Tomić, K., & French, P. (2019). Long-term Formant Frequencies in Cross-language Forensic Voice Comparison under Likelihood Ratio Framework. *A paper Ppesented at The 28th Annual Conference of the International Association for Forensic Phonetics and Acoustics in Istanbul, July 13th to 17th.*

Wagner, A., & Braun, A. (2003). Is voice quality language-dependent? Acoustic analyses based on speakers of three different languages. *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 651-654). Adelaide: Causal Productions.

Wirz, S., & Mackenzie Beck, J. (1995). Assessment of voice quality: The Vocal Profiles Analysis Scheme. In S. Wirz (Ed.), *Perceptual approaches to communication disorders* (pp. 39-55). London: Whurr Publishers Ltd.

Prague, July 10–13, 2022
Student Paper

Neural underpinnings of familiar talker advantage: an EEG study

Valeriia Perepelytsia[1], *Nathalie Giroud*[1], *Tugce Aras*[2], *Martin Meyer*[3], *and Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
valeriia.perepelytsia@uzh.ch
[2]*Department of Psychology, University of Zurich, Zurich, Switzerland*
[3]*Department of Comparative Language Science, University of Zurich, Zurich, Switzerland*
{tugce.aras|nathalie.giroud|martin.meyer|volker.dellwo}@uzh.ch

Introduction

It has been previously shown that words spoken by familiar talkers are more intelligible compared to words spoken by unfamiliar talkers, especially in adverse listening conditions such as background noise or competing talkers (Nygaard, Sommers & Pisoni, 1994; Nygaard & Pisoni, 1998; Souza et al., 2013, Levi, 2015). However, the cognitive and neural mechanisms of this effect, termed familiar talker advantage, are not known. The main aim of this study was to explore the processing of familiar and unfamiliar voices at the neural level with neural speech tracking as a possible underpinning of familiar talker advantage. To investigate this, we trained our participants to recognize four female talkers and then obtained scalp EEG while the participants listened to sentences produced by familiar and unfamiliar talkers in quiet and in multitalker babble noise. The study has implications for general understanding of familiar voice perception and processing, which is relevant for earwitness performance in voice parades.

Materials and Methods

Participants. A sample of 39 right-handed young adults, all native speakers of Swiss German (Mean age = 24.2 years, Range = 18–32 years, SD = 3.3, 10 male) participated in the study.
Stimuli and experimental paradigm. We used speech recordings from four female native speakers of Zurich Swiss German for voice recognition training. Recordings were of variable quality (i.e., studio quality and recordings from Zoom videoconferencing application (Zoom Video Communications, San Jose, CA) and contained read speech, semi-spontaneous speech, and charismatic speech. Read speech comprised sentences and text passages; semi-spontaneous speech included description of a cartoon, retelling a short story, describing what speakers did on a previous day, and a dialogue with other speakers on a selected topic; charismatic speech consisted of a speech on a topic of the speakers' choice to a virtual audience as if they were delivering a TED-talk. Each training session consisted of three blocks: familiarization (i.e., participants could play audio samples of each speaker as many times as they wanted), practice (i.e., participants heard an audio sample and were asked to indicate which of four speakers it is; feedback was given whether the response is correct), and a voice recognition test (i.e., participants heard an audio sample and were asked to indicate which of four speakers it is; no feedback was provided). To ensure that participants are well familiarized with the voices and can generalize them, the speaking styles in the familiarization/practice and the test phases in each training session were different. Each training session lasted approximately 15 minutes, and participants were required to pass ten training sessions. The sessions were distributed across three days. By the end of the training, participants had to reach at least 95% correct responses in the voice recognition test.
EEG recording. EEG was recorded after voice recognition training, on the fourth day. During EEG, participants listened to sentences produced by four familiar voices and four unfamiliar voices in quiet and in multitalker babble noise (SNR = 0). After each sentence, they answered a multiple-choice comprehension question about the content of the sentence.

Neural speech tracking. Neural speech tracking is the process of synchronization between the low-frequency activity in the brain and temporal regularities in the speech signal (Luo & Poeppel, 2007). It reflects neural encoding and processing of acoustic and linguistic speech features (Giraud & Poeppel, 2012; Poeppel & Assaneo, 2020). Neural speech tracking was quantified by cross-correlation between EEG signals and speech temporal envelopes was used.

Results

Voice recognition training. Participants differed in their learning curves during voice recognition training with some listeners being quicker in reaching ceiling levels of correct responses (see Figure 1). The first half of the training was most variable in terms of percent correct responses, while in the second half the response rates leveled out. Two talkers were more recognizable than others, while the other two talkers elicited less correct responses.

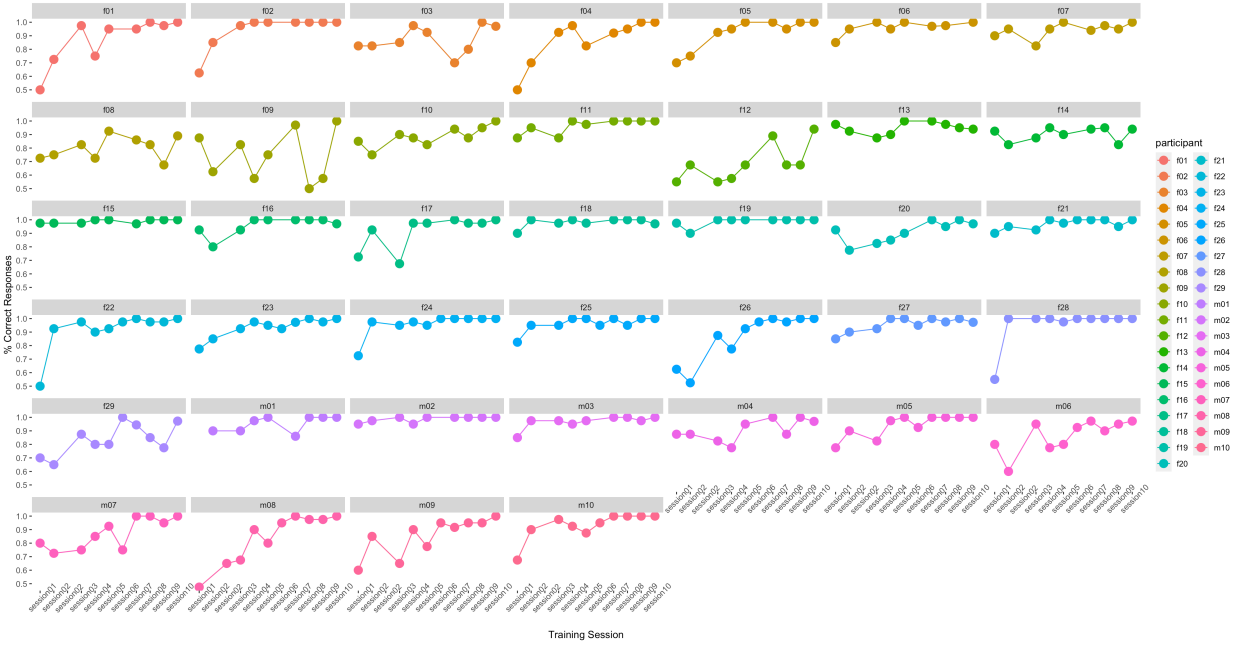


Figure 1. Percent correct responses during voice recognition training across training sessions for all participants.

EEG recordings. First results of the behavioural responses (comprehension questionnaire) during EEG indicated that participants' speech comprehension was significantly worse for speech embedded in multitalker babble noise compared to speech in quiet, but familiar voices did not result in significantly higher speech comprehension compared to unfamiliar voices. Neural speech tracking analyses are in progress and will be finalized soon.

References

Giraud, A. L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. In *Nature Neuroscience* (Vol. 15, Issue 4, pp. 511–517).

Levi, S. V. (2015). Talker familiarity and spoken word recognition in school-age children. *Journal of Child Language, 42*(4), 843–872.

Luo, H., & Poeppel, D. (2007). Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech Perception as a Talker-Contingent Process. *Psychological Science*, *5*(1), 42–46.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376.

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The Advantage of Knowing the Talker. The Advantage of Knowing the Talker. *Journal of the American Academy of Audiology*, *24*(08), 689–700.

Optimizing the strength of evidence: Combining segmental speech features

Willemijn Heeren[1], *Laura Smorenburg*[1], *and Erica Gold*[2]
[1]*Leiden University Centre for Linguistics, Leiden University, The Netherlands*
{w.f.l.heeren, b.j.l.smorenburg}@hum.leidenuniv.nl
[2]*California State University San Marcos, California, USA.*
egold@csusm.edu

In research, speaker specificity is often investigated at the level of individual speech sounds. In casework, however, conclusions are drawn by evaluating multiple features (e.g., Gold & French, 2011). Gold & Hughes (2015) compared several ways of combining different acoustic-phonetic features into one overall likelihood ratio (LR). They argued for the evaluation of correlations between speech features prior to combining evidence from various phonetic features. The current study considers segmental correlations prior to combining various Dutch speech sounds into a joint strength of evidence. It is expected that the combination of different speech sounds will support stronger conclusions by an LR system with higher validity.

For the current study, spontaneous conversational telephone speech from adult male speakers was used. In the first phase of the study correlations between speech features from the same sounds and across all six speech sounds were computed. In the second phase, an overall LR was computed, taking the correlations into account.

Method

Landline telephone data (300-3400 Hz) were taken from Heeren (2020, [a:, e:]), Smorenburg & Heeren (2020, [s, x]), and Smorenburg & Heeren (2021, [m, n]). Per speech segment, two well-performing acoustic-phonetic features from each segment were selected: F2 and F3 for the vowels, N2 and N3 for the nasals, and CoG and spectral standard deviation for the fricatives. The same set of 60 speakers contributed 20 tokens per speech sound[1].

Correlations between features within a speech sound and across speech sounds were computed using distance correlations implemented in the *R* package *Energy* (https://cran.r-project.org/web/packages/energy/) to assess non-linear relationships and Pearson's *r* to assess a linear relationship.

An overall LR was computed by developing separate MVKD LR systems (Aitken & Lucy, 2004) for non-correlating features using the MATLAB implementation by Morrison (2007), and by then multiplying the resulting LRs per system. The 60 speakers were randomly divided into equally-sized groups for development, reference, and test data, which was repeated 10 times per system (using fixed grouping per repetition). After score-to-LR conversion, ELUB limiting with 1 CMLR (Vergeer et al., 2016) was applied to each of the intermediate results, and also to the overall LRs after multiplication. Each system's validity was evaluated by computing Cllr, Cllr_{min} and EER (Brümmer & Du Preez, 2006, in Van Leeuwen, 2008).

Results

Within speech sounds and within speech sound classes (vowels, nasals, fricatives), significant correlations between speech features (e.g. F2, F3) were found ($r >= .33$, $p <= .016$). The only significant correlations between speech sounds from different classes were found between N2 of nasal [m] and the vowel [a:]'s formants F2 ($r = .42$, $p = .001$) and F3 ($r = .45$, $p = .002$). Given this result, two LR systems were built: one for vowels+nasals and one for fricatives.

The LR results are summarized in Table 1, presenting medians and interquartile ranges across 10 repetitions per system.

[1] One speaker had only 17 [s] tokens.

	[a:] + [e:] + [m] + [n]	[s] + [x]	*combined*
$LLR_{same\text{-}speaker}$	1.15 [1.0, 1.3]	0.55 [0.44, 0.55]	1.45 [1.40, 1.45]
$LLR_{different\text{-}speaker}$	−1.25 [−1.25, −0.95]	−0.50 [−0.61, −0.35]	−1.10 [−1.25, −1.10]
Cllr	0.53 [0.49,0.56]	0.84 [0.82, 0.86]	0.51 [0.47, 0.53]
$Cllr_{min}$	0.28 [0.24, 0.36]	0.71 [0.66, 0.79]	0.22 [0.18, 0.27]
EER (%)	9.11 [7.44, 9.52]	25.08 [23.31, 26.55]	6.28 [5.06, 7.70]

Table 1. Results of the LR analysis, per system and for the combined result.

Discussion

Results show that an acoustic-phonetic system for Dutch may perform well using features from just six, carefully-selected segments. Without limiting, comparable results were obtained for combined features in English (formants, F0, articulation rate, Gold, 2014). Even though Dutch [s, x] may not appear to be a strong system on its own, when combined with [a:, e:, n, m] it is very helpful in increasing strength of evidence and validity. Accounting for the correlations between and within features allows us to avoid miscarriages of justice that would traditionally over-estimate strength of evidence. The results in this study show that accounting for correlations within and between just six phonetic parameters provides appropriate same-speaker and different-speaker strengths of evidence. We also have a respectable EER for the system and the overall Cllr is not too high.

References

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

Brümmer, N., a& Du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3), 230–275.

Gold, E. (2014). *Calculating likelihood ratios for forensic speaker comparisons using phonetic and linguistic parameters*. PhD Dissertation, University of York.

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

Gold, E., & Hughes, V. (2015) Frontend approaches to the issue of correlations in forensic speaker comparison. In: *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow.

Heeren, W.F.L. (2020). The Effect of Word Class on Speaker-dependent Information in the Standard Dutch Vowel /aː/. *Journal of the Acoustical Society of America*, 148(4), 2028–2039.

Smorenburg, B.J.L., & Heeren, W.F.L. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America*, 147(2), 949–960.

Smorenburg, B.J.L. & Heeren, W.F.L. (2021). Acoustic and speaker variation in Dutch /n/ and /m/ as a function of phonetic context and syllabic position. *Journal of the Acoustical Society of America*, 150(2), 979–989.

Morrison, G.S. (2007). *Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation*. [software]

Van Leeuwen, D. A. (2008). SRE-tools, a software package for calculating performance metrics for NIST speaker recognition evaluations. Downloaded from http://sretools.googlepages.com/. [software]

Vergeer, P., Van Es, A., de Jongh, A., Alberink, I., & Stoel, R. (2016). Numerical likelihood ratios outputted by LR systems are often based on extrapolation: when to stop extrapolating? *Science & Justice*, 56(6), 482–491.

Prague, July 10–13, 2022
Student Paper

Shifting the Burden: towards more robust and transparent procedures for LADO

Jim Hoskin[1], *Paul Foulkes*[1]
[1]*Department of Language and Linguistic Science, University of York, England*
jah638@york.ac.uk; ja.hoskin@yahoo.com
Paul.Foulkes@york.ac.uk

Since the mid-1990s, when governments commenced the use of Language Analysis for the Determination of Origin (LADO), little experimental work relevant to the field has been conducted (Hoskin, Cambier-Langeveld & Foulkes 2020). Most published work is case-based critique and in-principle polemic on the involvement of native-speaker non-linguists (NSNLs) as analysts (Foulkes, French & Wilson 2019; Hoskin 2018).

Expressed simply, the question in LADO is: How likely is it that this person is an authentic speaker of the variety they claim to speak? (Cambier-Langeveld 2012) At present, asylum applicants must demonstrate the authenticity of their language use in a one-shot interview (Matras 2018; Patrick 2012). The field is therefore open to the development of supplementary tests which recruit asylum applicants' perceptions, as well as their production, of language.

The experimental work discussed here constitutes one among a handful of attempts (see also Wilson 2009, Hedegard 2015, Shen & Watt 2015) to investigate empirically issues central to LADO. The ultimate objective is to develop new tests of perception, and perhaps of production, to augment extant LADO procedure.

This presentation details a comparison of the perceptions of Syrian listeners with those of other speakers of Arabic, both NSNLs and linguists trained to postgraduate level. In total, 79 listeners responded: 21 Syrian and 22 non-Syrian NSNLs, and 10 Syrian and 21 non-Syrian linguists; five non-native speaker linguists were also (fortuitously) recruited. There were 22 stimuli; 10 featured speakers of Syrian and 12 speakers of non-Syrian Arabic, all reading extracts of an Arabic folktale in their native dialect. The question asked was, 'Is this a Syrian accent?'

Using the glmer function in *R*, a series of model comparisons was performed on the five groups' accuracy in accepting or rejecting stimuli as Syrian. Results demonstrate that, on Syrian stimuli, Syrian listeners' responses were significantly more accurate than those of non-Syrian listeners: $\chi 2$ (2, N = 79) = 59.248, p = <0.0001. The variable of education showed no significant effect on accuracy.

From these results two inferences can be made. First, Syrians' greater accuracy on Syrian stimuli indicates that the administration of a similar, real-life test of perception might reliably separate genuine from non-genuine Syrian asylum applicants. Second, academic training in linguistics may have no effect on the reliability of primary-phase LADO analysis, even allowing for differences between the highly-controlled speech material used here and that encountered in LADO.

Also discussed are listener comments on the cues that guided their acceptance or rejection of speakers as Syrian. All five groups appear to depend more on cues not codified in the available dialectological surveys than on those that are. This tendency is most marked among non-linguists and least among non-native linguists. From these findings two conclusions are drawn. First, there may be crucial omissions from the available surveys—perhaps, as argued by Nolan (2012), of the 'below-consciousness' type—of shibboleths separating Arabic dialects from one another. Second, recourse to the existing literature does not appear to be correlated with accuracy in the perception task. It is suggested that careful analysis of the cues mentioned by listeners may furnish an empirical basis for developing a further series of supplementary production tests for LADO.

References

Cambier-Langeveld, T. (2012) Clarification of the issues in language analysis: a rejoinder to Fraser and Verrips. *International Journal of Speech, Language and the Law,* 19(1), 95-108.

Foulkes, P., French, P. & Wilson, K. (2019) LADO as forensic speaker profiling. In P. Patrick, M. Schmid, & K. Zwaan (eds.) *Language Analysis for the Determination of Origin*. Cham: Springer, 91-116.

Hedegard, H. (2015) *Language Analysis for the Determination of Origin (LADO): The Native Speaker Dimension*. Unpublished MSc dissertation, University of York.

Hoskin, J.A. (2018) Native speaker non-linguists in LADO: an insider perspective. In I.M. Nick (ed) *Forensic Linguistics: Asylum-seekers, Refugees and Immigrants*. Malaga: Vernon Press, 23-40.

Hoskin, J., Cambier-Langeveld, T. & Foulkes, P. (2020) Improving objectivity, balance and forensic fitness in LAAP: a response to Matras. *International Journal of Speech, Language and the Law*, 26(2), 257-277.

Matras, Y. (2018) Duly verified? Language analysis in UK asylum applications of Syrian refugees. *International Journal of Speech, Language and the Law*, 25(1), 53-78.

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(2), 263-289.

Patrick, P. (2012) Language analysis for determination of origin: objective evidence for refugee status determination. In P. Tiersma and L. Solan (eds.) *The Oxford Handbook of Language and Law*. Oxford University Press, 533-546.

Shen, C. & Watt, D. Accent Categorisation by Lay Listeners: Which Type of “Native Ear” Works Better? *York Papers in Linguistics*, 2(14), 106-131. https://www.york.ac.uk/language/ypl/ypl2/14/YPL2-14-05-Shen-Watt.pdf

Wilson, K. (2009) *Language Analysis for the Determination of Origin: Native Speakers vs. Trained Linguists*. Unpublished MSc dissertation, University of York.

For the Record: Improving standards in the production of non-expert police interview transcripts

James Tompkinson[1], *Kate Haworth*[1], *Emma Richardson*[1], *Felicity Deamer*[1] *and Magnus Hamann*[1 2]

[1]*Institute for Forensic Linguistics, Aston University, UK*

{j.tompkinson|k.haworth|e.richardson4|f.deamer}@aston.ac.uk

[2]*Loughborough University, UK.*

m.hamann@lboro.ac.uk

Following the completion of a police-suspect interview in England and Wales, a written record is usually produced of the interview recording. These ROTI (Record of Taped Interview) transcripts are evidential documents which are used in courtrooms. However, in contrast to the kinds of expert transcripts produced by qualified phoneticians (Fraser 2003, 2017; French and Fraser, 2018; Love and Wright 2021), ROTI transcribers receive no phonetic or linguistic training, and there is no standardised guidance for how ROTI transcripts should be produced. The production of ROTI transcripts is also largely an 'in-house' process conducted within police forces, despite concerns having been raised (French and Fraser, 2018) about the suitability of police personnel to produce certain types of transcripts.

A concern highlighted by Haworth (2018) is that significant alterations can be made to the interview evidence as it is converted from an audio recording to a written transcript, especially as these ROTI transcripts are treated "an unproblematic copy" of the interview recording. However, several problems arise through the process, including differences in the way that certain features are represented, inaccurate or incomplete summaries of evidence, inconsistent representation of features such as emotion and pausing, and the subsequent presentation of this amended evidence in courtrooms when transcripts are converted back into spoken format by being read out loud by legal representatives.

This paper highlights a series of results from a project aimed at improving transcription practices within one English police force as a pilot. The project takes a mixed-methods approach, with three specific methodological strands combined for the overall findings. These methods are 1) conversation analysis of police interview recordings and their corresponding ROTI transcripts; 2) detailed focus group discussions with current ROTI transcribers and subsequent qualitative analysis of the response data, and 3) psycholinguistic experiments to test how people's perceptions of interviewees are affected by the representation of linguistic features in transcripts.

From this combined analysis, our project has developed a set of criteria which we believe should apply to all transcripts – consistency, accuracy and neutrality (the "CAN" model). The model proposes the three areas as the foundational features that any transcript should uphold. Focusing on these issues, key findings from our ROTI transcript analysis include:

- Information is routinely omitted from ROTI transcripts, either due to incomplete or inadequate summaries of the audio contents.
- There is uncertainty among transcribers as to how certain types of information should be represented, leading to inconsistencies in the final transcripts.
- Whether an interview is presented in audio or written format can significantly affect people's perceptions of interviewees.

Along with highlighting a range of concerns with the production of ROTI transcripts, the paper will also focus on how linguists, and specifically linguists with an interest in the production and perception of speech, can work to improve current processes. Drawing on existing guidelines and research for the production of transcripts by expert phoneticians, the paper will provide a series of recommendations which can be applied to the production of non-expert transcripts of police interviews.

References

Fraser, H. (2003). Issues in transcription: factors affecting the reliability of transcripts as evidence in legal cases. *International Journal of Speech, Language and the Law*, *10*, 203-226.

Fraser, H. (2017). Transcription of indistinct forensic recordings: Problems and solutions from the perspective of phonetic science. *Language and Law/Linguagem e Direito*, *1*(2).

French, P., & Fraser, H. (2018). Why" Ad Hoc Experts" should not Provide Transcripts of Indistinct Audio, and a Better Approach. *Criminal Law Journal*, 298-302.

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

Love, R., & Wright, D. (2021). Specifying challenges in transcribing covert recordings: Implications for forensic transcription. *Frontiers in Communication*, *6*.

Prague, July 10–13, 2022
Student Paper

Just the way you are: The potential of the word *just* as a speaker discriminant

Ben Gibb-Reid Holmes, Paul Foulkes and Vincent Hughes
Department of Language and Linguistic Science, University of York, York, UK
{ben.gibb-reid|paul.foulkes|vincent.hughes}@york.ac.uk

Background

A particularly important issue in forensic voice comparison (FVC) is the lack of direct correspondence in the content of different recordings. That is, recordings are unlikely to share many of the same words. Therefore, a frequently used word (or other feature) in naturally occurring speech is of value because it permits direct comparison. To examine the forensic value of any linguistic feature, it is necessary to understand how variable it is between and within speakers, and the factors that affect it in different discourse positions or prosodic contexts. In the present study, the short discourse-pragmatic marker (DPM) *just* is analysed in this way for suitability as a diagnostic feature in FVC.

In previous research, filled pauses (*uh, um*), also defined as DPMs, have been analysed as FVC features with promising results (Tschäpe et al., 2005). In Hughes et al. (2016)'s study, the best speaker comparison models were based on all three formants. For the present study, formants and durations from the vowel portions of *just*, STRUT and filled pauses were analysed for 100 male Southern Standard British English speakers (DyViS corpus, Nolan et al., 2009). The polyfunctional word *just* was selected because of its high frequency in spontaneous speech. *Just* is the 27th most frequent word in the British National Corpus (2014) at 0.75 per 100 words (Love et al., 2015). Research also shows that *just* is increasing in frequency over time, as demonstrated for younger speakers in Toronto (Tagliamonte, 2016) and Tyneside (Woolford, 2021). It is also of interest to FVC whether speakers use *just* in different ways, and therefore the various different functions of *just* (as discussed by Woolford, 2021) are also analyzed to aid speaker comparison.

Results

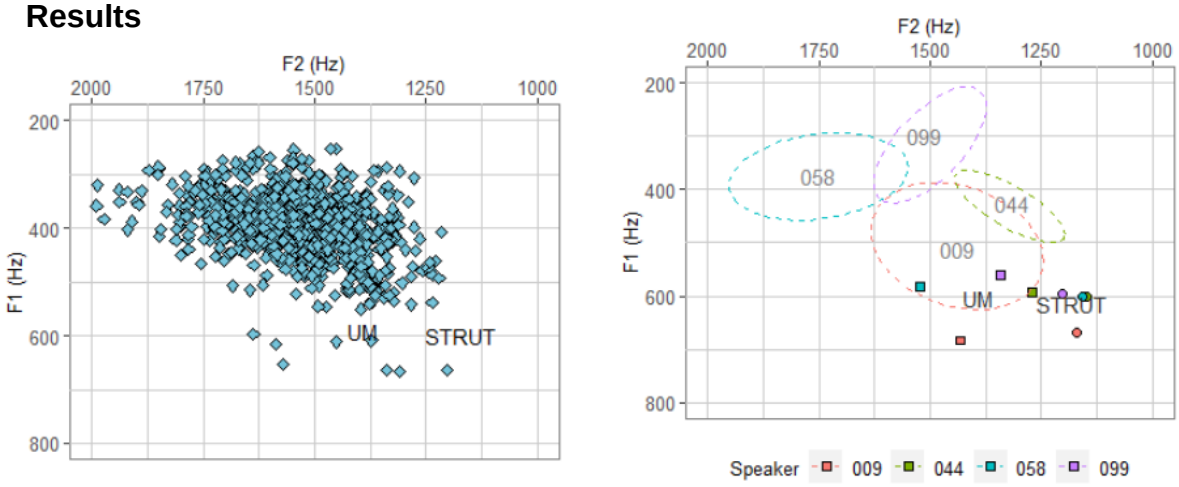


Figure 1. F1-F2 plot of *just* vowel midpoints alongside means for STRUT and the vowel of *um* (left). F1-F2 of *just* for 4 speakers compared with ellipses showing standard deviation and mean F1-F2 values for STRUT and the vowel in *um* (right).

1,276 tokens of *just* were extracted for analysis, transcribed to show function, segment elision and to allow for formant readings to be taken from the vowel. As expected, *just* was highly frequent, occurring overall 0.88 times per 100 words. In total, 1,019 *just* vowels were found

suitable for formant analysis. Midpoint formant measures for STRUT and the vowel of *um/uh* were also extracted as points of comparison for likelihood ratio-based testing across 76 speakers. Vowel midpoints for all tokens are displayed in Figure 1 along with the mean readings for STRUT and *um* vowels. Generally, the vowel in *just* is considerably raised and/or fronted compared to STRUT or *um*. Figure 1 also displays four speakers who had mean F1 and F2 values at the upper and lower extremes.

In likelihood ratio-based testing, various tests were run comparing acoustic measures of *just. Just* was also compared with STRUT and *um* in its discriminatory capacity. Figure 2 shows the validity measures for these tests, where lower log LR cost (C_{llr}) and equal error rates (EER) correspond to a better-performing system. The left panel shows that F1-F3 of *just* outperforms the formants of STRUT. It has a lower C_{llr} than *um* but a very slightly higher EER. The right panel of Figure 2 displays the effect of adding discourse functions of *just* to speaker comparison models. *Just* F1-F3 without any function information performs best, whereas adding restrictive or discourse *just* information reduces model validity. It is possible that speaker comparison models which do not include tokens of discourse *just* perform better – and therefore discourse *just* is a slightly less good feature than say adverb or restrictive *just*. Overall, *just* shows some promise for FVC application, performing better than *um* or STRUT. Adding information about *just* functions, however, may only aid the task of FVC a little. This is positive, as FVC analysts can treat all tokens of *just* similarly, regardless of the word's function – making *just* a broad idiosyncratic feature of the voice.

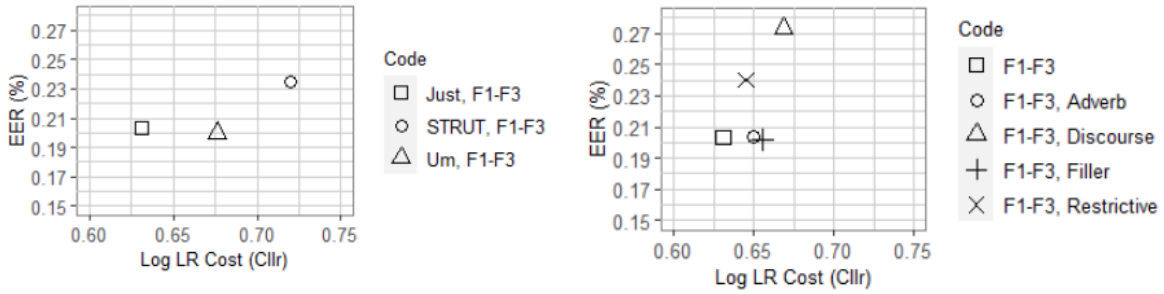


Figure 2. Plot of log LR cost (C_{llr}) and equal error rate (EER%) for *just*, STRUT and *um* F1-F3 vowel midpoints (left) and for *just* F1-F3 across various functions.

References

The British National Corpus, version 3 (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

Love, R., Dembry, C., Hardie, A., Brezina, V. and McEnery, T. (2017). The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3): 319-344. DOI: 10.1075/ijcl.22.3.02lov

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.

Nolan, F., McDougall, K., De Jong, G. & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech Language and the Law 16(1): 31-57.

Tagliamonte, S. (2016). Teen talk: The language of adolescents: Cambridge University Press.

Tschäpe, N., Trouvain, J., Bauer, D., & Jessen, M. (2005). Idiosyncratic patterns of filled pauses. Paper presented at the 14th Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA), Marrakesh, Morocco.

Woolford, K. (2021). Just in Tyneside English. World Englishes. doi:10.1111/weng.12542

Prague, July 10–13, 2022
Student Paper

The effects of linguistic contexts on the acoustics and strength-of-evidence of /s/

Laura Smorenburg and Willemijn Heeren
Leiden University Centre for Linguistics, Leiden, The Netherlands
{b.j.l.smorenburg|w.f.l.heeren}@hum.leidenuniv.nl

Previous research has shown that linguistic structure and phonetic contexts can affect the acoustics and consequently the strength-of-evidence in speaker comparisons. For example, stressed vowels seem to perform better than unstressed vowels [1] and vowels from content words seem to perform slightly better than vowels from function words [although only in multinomial regression, not in likelihood-ratio analysis: 2].

Fricative /s/ is a relatively speaker-specific consonant, but is reported to be strongly affected by coarticulatory labialization, which lengthens the anterior cavity and lowers the resonance frequencies in /s/ [e.g. 3]. Data from Dutch spontaneous telephone speech has shown that slightly more speaker information is available when fricatives /s/ and /x/ occur in these labial contexts [4], which was attributed to between-speaker variation in the degree and timing of the co-articulatory movement. We now investigate the effects of phonetic context and syllabic position on British English /s/, also considering speech channel effects.

Method

Materials consisted of mock telephone conversations with an accomplice taken from Task 2 in WYRED [5]. One 15-min conversation per speaker (*N*=60, all adult males from Wakefield, Yorkshire) was analysed. Per speaker, ~100 /s/ tokens along with their immediate phonetic neighbours were manually segmented and labelled on syllabic position. Spectral moments (M1: centre of gravity, M2: variance, L3: skewness, L4: kurtosis), duration, and spectral tilt were measured for each /s/ in the simultaneously recorded studio and landline telephone channel. M1 was also measured dynamically in 5 non-overlapping windows and captured with a quadratic polynomial fit. Effects of contextual labialization (non-labial, labial) and syllabic position (onset, coda) were assessed with linear mixed-effects (LME) modelling for the acoustics and with multinomial logistic regression (MLR) and MVKD [6] likelihood ratio analysis (LR) for the speaker discrimination. Only speakers with at least 10 tokens per factor level were included in the analysis (*N*=55).

Results

For M1 measured in the studio channel, it can be seen in Table 1 that acoustic results are mostly congruent with the literature [e.g. 3, 4]. There are significant effects of labialization, although anticipatory (i.e. right context) effects are relatively small compared to carry-over (i.e. left context) effects. Coda reduction is also observed. Generally, these effects are not maintained in the telephone channel, rather, they sometimes go in the opposite direction and seem random.

	Studio (550-8000 Hz)			Telephone (550-3400 Hz)		
Effects	*Est.*	*SE*	*t*	*Est.*	*SE*	*t*
(intercept)	5190	77	67.3	2075	32	64.2
Left context = LABIAL	–365	20	–18.7	112	10	10.6
Right context = LABIAL	–94	22	–4.3	–31	12	–2.6
Syll. Position = CODA	–200	15	–13.2	–1	8	–0.1
Left x Syll. Position						
Right x Syll. Position	–118	37	–3.2	68	20	3.4

Table 1. Best-fitting LME model for M1 in the studio and telephone channels (*N*=55, *n*=6634). Hertz ranges refer to the measurement ranges per channel, not the available signal.

Regarding the speaker discrimination, both the MLR and LR analyses showed small contextual sampling effects in the studio, but not (in MLR), or to a lesser extent (in LR), in the telephone channel. However, even in the studio channel, the effect of syllabic position is negligeable and the effect of context labialization not consistent for preceding versus following context (see Figure 1 and Table 2). The effect of speech channel, on the other hand, is much larger for both the acoustics and speaker discrimination. To conclude, contextual sampling effects are present in broadband, but not so much in narrowband signals. It is rather the speech channel that has the largest effects on the acoustics and strength-of-evidence of English /s/.

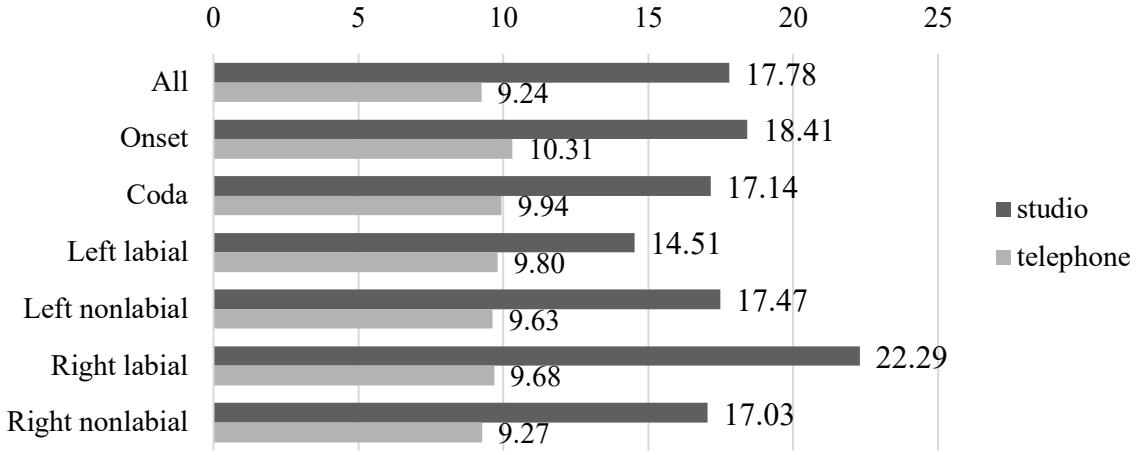


Figure 1. MLR speaker-classification accuracies (in %) using duration, M2, L3, L4, spectral tilt, and linear and quadratic M1 coefficients as predictors. Chance level = 1.82%.

	Studio (550-8000 Hz)					Telephone (550-3400 Hz)				
	LLR^{SS}	LLR^{DS}	Cllr	$Cllr^{min}$	EER	LLR^{SS}	LLR^{DS}	Cllr	$Cllr^{min}$	EER
All	1.77	–2.73	0.52	0.46	16.04	0.76	–0.52	0.82	0.71	24.02
Onset	1.83	–2.86	0.51	0.45	14.09	0.85	–1.02	0.72	0.63	23.12
Coda	1.96	–3.00	0.49	0.45	14.08	1.05	–0.84	0.72	0.64	23.50
Left labial	1.20	–1.16	0.69	0.61	19.31	0.64	–0.30	0.85	0.77	28.43
Left nonlabial	1.75	–2.58	0.53	0.48	16.17	0.67	–0.48	0.80	0.74	24.97
Right labial	1.88	–3.82	0.50	0.36	10.68	0.90	–0.87	0.72	0.66	25.05
Right nonlabial	1.52	–2.27	0.58	0.49	15.47	0.74	–0.51	0.81	0.73	25.36

Table 2. Calibrated LLRs, Cllr, $Cllr^{min}$ and EER. Max. sample size (n=18) per speaker per condition.

References

[1] McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *Int. J. of Speech, Lang. and the Law, 13*(1), 89–125.

[2] Heeren, W. F. L. (2020). The effect of word class on speaker-dependent information in the Standard Dutch vowel /aː/. *J. of the Acoust. Soc. of Am., 148*(4), 2028–2039.

[3] Koenig, L. L., Shadle, C. H., Preston, J. L., & Mooshammer, C. R. (2013). Toward Improved Spectral Measures of /s/: Results from Adolescents. *J. of Speech Lang. and Hearing Res., 56*(4), 1175.

[4] Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *J. of the Acoust. Soc. of Am.*, *147*(2), 949–960.

[5] Gold, E., Ross, S., & Earnshaw, K. (2018). The "West Yorkshire Regional English Database": Investigations into the generalizability of reference populations for forensic speaker comparison casework. In Proceedings of *INTERSPEECH* (Vol. 2018–Sept., pp. 2748–2752).

[6] Morrison, G.S. (2007). Matlab implementation of Aitken & Lucy's (2004) forensic likelihood-ratio software using multivariate-kernel-density estimation.

Listener judgements for social traits and criminal behaviours as a function of speaker pitch and articulation rate.

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

The perceptions and prejudices that people hold about voices are brought with them when they enter the legal system. Previous research has found that people consider some voices to sound 'more guilty' (Axer 2019) and more likely to commit certain crimes (Dixon et al. 2002; Paver et al. 2021). This can have implications in forensic contexts, such as jury perceptions of witness credibility (Cantone et al. 2019) and potentially biasing earwitness evidence (Nolan and Grabe 1996).

This paper reports on the latest experiments in 'Improving Voice Identification Procedures' (IVIP) funded by the Economic and Social Research Council (ESRC). In line with previous sociolinguistic work (e.g. Coupland & Bishop 2007) our earlier experiments have found that listeners associate certain accents with particular traits and behaviours, including certain crimes. This paper extends the focus to listener judgements of pitch and articulation rate (AR) which have been found to influence perceptions relating to competency, attractiveness, and threat (e.g. Street and Brady 1982; Jones et al. 2008; Tompkinson 2018).

We ran a voice rating task in which three British English accents were selected as stimuli based on our previous experiments – Belfast, Liverpool, and SSBE. Three 15s samples of each accent were manipulated for 'high', 'average' and 'low' versions of both pitch (Experiment 1) and AR (Experiment 2). In each experiment, 180 participants were asked to listen to each of the nine voices (and four distractor voices) rate them using a 7-point rating scale. Half answered questions on ten social traits and half on ten behaviours, including criminal offences.

Mixed-effects ordinal regression models confirmed that listeners made judgements based on pitch for some of the social traits, but not behaviours. However, results suggest that speaker accent is more important for listeners than speaker pitch (Figure 1). The models also verified an effect of AR on listener judgements of most social traits, as well as some behaviours.

Low pitched voices were rated lower for solidarity-based traits, whereas high pitched voices rated lower for status-based traits. Low AR resulted in lower ratings for status, solidarity, and dynamism traits. There was a statistically significant effect of accent on participants' judgements of traits in both experiments, both when the traits were grouped and observed individually. These findings were in line with previous research which finds that non-standard accents are rated less favourably for status dimensions, but more favourably for solidarity dimensions.

For the behavioural questions, there was a significant main effect of speaker accent and AR, but no main effect of speaker pitch. In line with our previous findings, across both experiments the Belfast and Liverpool speakers were both rated more likely to commit crimes than the SSBE speaker. The Belfast speaker was rated less likely to perform morally bad behaviours, SSBE most likely to be morally ambiguous, and Liverpool less likely to be morally good. Low AR voices had higher ratings for criminal behaviours, and lower ratings for morally good behaviours.

The results reveal that voices are subject to listener perceptions, potentially with serious implications in forensic contexts.

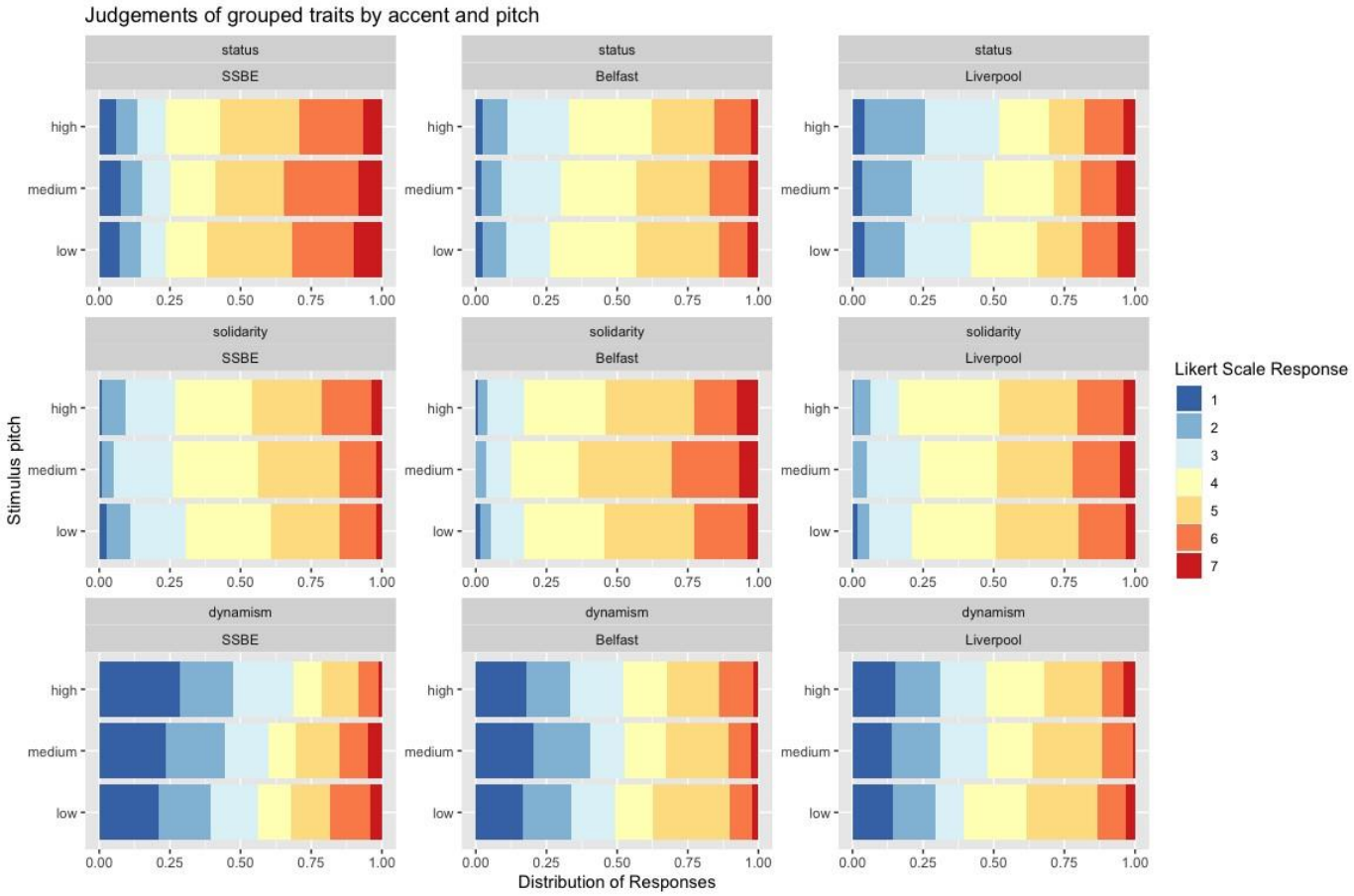


Figure 1. Stacked barplots showing the distribution of responses from participants (Experiment 1) for each group of traits, separated by speaker accent. The y-axis shows the speaker pitch. For the Likert scale, 1 = strongly disagree, 7= strongly agree.

References

Axer, G. (2019). British accent perceptions and attributions of guilt by native and non-native speakers. Journal of Language and Discrimination, 3(2), 195–217.

Cantone, J.A., L.N. Martinez, C. Willis-Esqueda and T. Millerd. (2019). Sounding guilty: How accent bias affects juror judgments of culpability. Journal of Ethnicity in Criminal Justice, 17(3), 228–253.

Coupland, N. and H, Bishop. (2007). Ideologised values for British accents. Journal of Sociolinguistics 11(1): 74–93.

Dixon, J.A., B. Mahoney, and R. Cocks. (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.

Jones, B. C., Feinberg, D. R., Debruine, L. M., Little, A. C., and Vukovic, J. (2008). Integrating cues of social interest and voice pitch in men's preferences for women's voices. Biology letters, 4(2), 192–194.

Nolan, G. and E. Grabe. (1996). Preparing a voice lineup. The International Journal of Speech, Language and the Law (Forensic Linguistics), 3(1), 74–94.

Paver, A., Wright ,D., and Braber, N. (2021) 'Accent judgements for social traits and criminal behaviours: ratings and implications.' Paper presented at the International Association for Forensic Phonetics and Acoustics Annual Conference (online), Marburg, 22-25 August 2021.

Street, R. L. and Brady, R. M. (1982)' Speech rate acceptance ranges as a function of evaluative domain, listener speech rate, and communication context. Communication Monographs, 49(4), 290–308.

Tompkinson, J. (2018). Assessing the influence of phonetic variation on the perception of spoken threats. PhD dissertation, University of York.

Role of Within-Vowel Formants in Forensic Speaker Comparison

Nikita Suthar[1], Peter French[2],
[1,2]Department of Linguistics, University of York, UK
{nikita.suthar|peter.french} @york.ac.uk

Formant analysis has been used as one of several methods for speaker discriminant studies (Cao & Dellwo, 2019; McDougall, 2006; McDougall & Nolan, 2007). Most studies have focused only on formant centre frequencies, trajectories and/or, to a more limited extent, bandwidths (Fleischer et al., 2015; Gonzalez-Rodriguez, 2011; Kent & Vorperian, 2018). The current work takes the potential role of formants as individual speaker discriminants further by investigating a range of *within*-formant measures. It reports on work conducted on the centre of gravity, relative amplitude, spectral bandwidth, LPC bandwidth, spectral peaks, skewness, kurtosis, and standard deviation of spectral moments. The Marwari language was used as a testbed for the work; in principle, the analysis could be conducted on any other language. Marwari belongs to the Indo-Aryan language family and is spoken in the north-western state of Rajasthan (India).

A total of forty-five female Marwari monolingual speakers from the Bikaner district were recruited for the study. Recordings were collected from spontaneous and non-spontaneous speech and focused on eight different vowels. Three modes of data collection were employed. The first was a list of 80 words (10 tokens per vowel) that the participants were asked to read aloud. The second was a picture description task, i.e., participants were shown a picture of local deities and were asked to narrate a story associated with them. The third mode was a conversation where participants were paired and asked to converse on a topic of their choice, or to choose a topic from a provided list.

The current analysis is conducted on the wordlist and story data. An ANOVA conducted in R showed a significant vowel difference between the two types of data. Once these differences had been established, the goal was to look at individual speaker discrimination. Eight spectral measures were taken from the first four formants. Manually assisted and corrected automatic extractions were conducted using a Praat script. As a next step, a linear discriminant analysis (lda) was conducted on the features extracted from every formant to determine the classification rate of these measures in identifying individual participants.

	Wordlist		*Story*	
Acoustic Measure	*Classification Rate*	*Times greater than chance*	*Classification Rate*	*Times greater than chance*
F1+F2+F3+F4	15%	6.5 times	11%	4.5 times
Centre of Gravity: F1-F4	15%	6.5 times	12%	5 times
Spectral Peak F1-F4	14%	6 times	12%	5 times
Spectral Amplitude: F1-F4	13%	5.5 times	13%	5.5 times
LPC Bandwidth: F1-F4	11%	4.5 times	10%	4 times
Spectral Bandwidth: F1-F4	9%	3.5 times	8%	3 times
Standard Deviation F1-F4	9%	3.5 times	9%	3.5 times
Kurtosis F1-F4	8%	3 times	9%	3.5 times
Skewness F1-F4	7%	3 times	7%	2.5 times

Table 1. A classification rate of centre frequencies F1-F4 and within-formant features for wordlist and story data.

The initial results showed that all the spectral measures and the centre formant frequencies increased the classification rates by a factor of at least 2.5. Some features performed better at classifying participants than others. Table 1 presents the initial results of the discriminant analysis and the classification rates of the individual features for the wordlist and story data. The results show that the centre of gravity, amplitude and spectral peaks were the best performing features for both datasets.

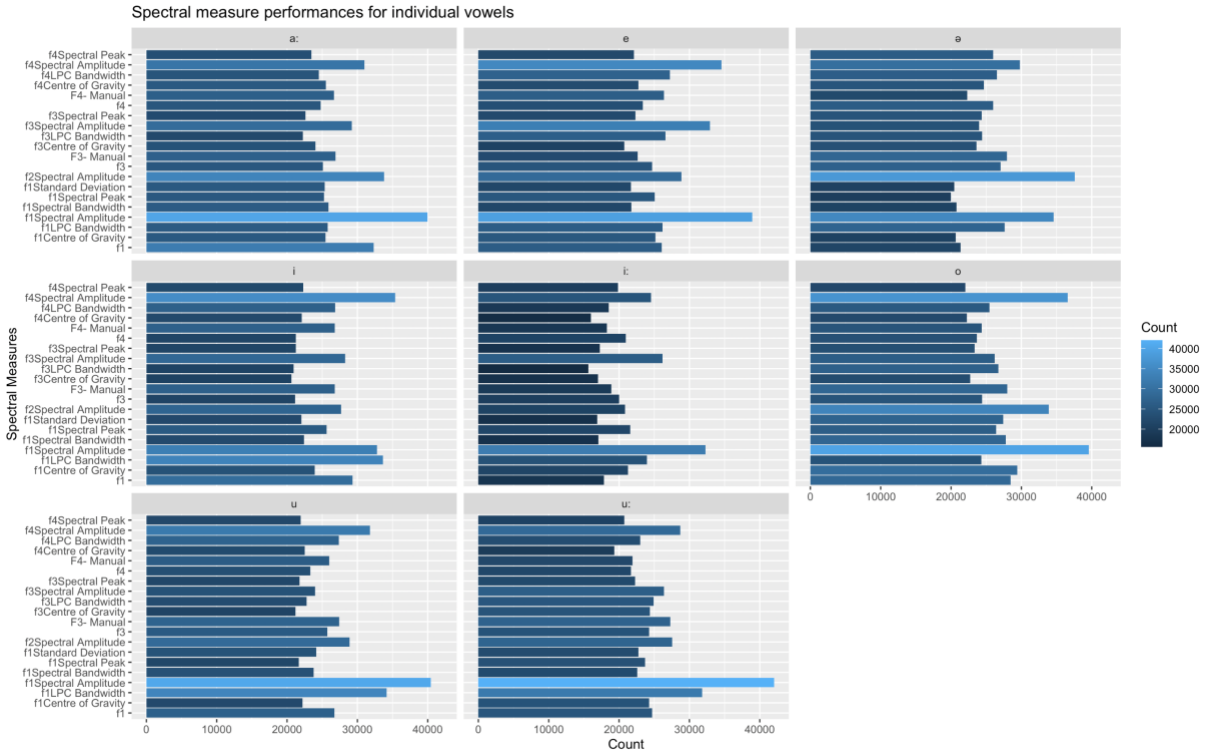


Figure 1. Number of times individual vowels, as they occurred in feature combinations for lda

Figure 1 shows the performance of individual features for every vowel. The x-axis represents the number of times an individual feature occurred for the lda analysis. (The analysis was conducted multiple times with up to five feature combinations.) The figure shows that long vowels provide more information than short vowels for speaker classification and that some features (f1/f4 spectral amplitude and bandwidth) are performing exceptionally better than the others.

Further analysis of these features will be conducted on the conversation data. As mentioned earlier, the Marwari language was used as a testbed, and in principle, this analysis could be performed on any other language. The performance of the measures will be tested on spoken corpora of other languages.

References

Cao, H., & Dellwo, V. (2019). The role of the first five formants in three vowels of mandarin for forensic voice analysis. *International Congress of Phonetic Sciences*. https://doi.org/10.5167/uzh-177494

Fleischer, M., Pinkert, S., Mattheus, W., Mainka, A., & Mürbe, D. (2015). Formant frequencies and bandwidths of the vocal tract transfer function are affected by the mechanical impedance of the vocal tract wall. *Biomechanics and modeling in mechanobiology*, *14*(4), 719-733.

Gonzalez-Rodriguez, J. (2011). Speaker recognition using temporal trajectories in linguistic units: the case of formant and formant-bandwidth contours. *INTERSPEECH.*

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

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: towards a new approach using formant frequencies. *The International Journal of Speech, Language, and the Law*, *13*(1), 89–126. https://doi.org/10.1558/sll.2006.13.1.89

McDougall, K., & Nolan, F. (2007). Discrimination of Speakers using the Formant Dynamics of /uː/ in British English. In Proceedings of *the International Congress of Phonetic Sciences*, 1825–1828.

Prague, July 10–13, 2022
Student Paper

Speech variability in telephone openings and its implications for speaker discrimination.

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

Speakers frequently modify their speech according to interlocutor-specific communicative goals, e.g., improving intelligibility with computer-directed speech (e.g., Mayo et al., 2012), regulating attention with child-directed speech (e.g., Fernald et al., 1984), or easier recognition for a speaker-recognition system (Dellwo et al., 2019).

Explorations of contexts in which within-speaker variability is prevalent are typically focused on different interlocutors, and little is known about discourse-specific contexts which may influence speech variability. For instance, telephone openings are marked by a high number of communicative goals, including topic establishment, identity confirmation and attention grabbing. Further, given the lack of visual cues, vocal adaptations are more necessary for successfully establishing these goals and more generally ensuring communicative efficiency. Therefore, it is likely speech in these contexts may be more variable due to the employment of speech modifications to achieve these communicative goals. Indeed, findings from a brief acoustic analysis conducted prior to this study, showed that f_0 standard deviation, used as an indicator of variability, was higher in initial-utterances compared to mid-conversation utterances in telephone calls (Figure 1). Although the explanation for this trend may be a result of one, or multiple factors, not limited to those mentioned above, we see preliminary evidence that speech in telephone openings may be more variable.

Greater within-speaker variability poses challenges for speaker discrimination and identification, with lay-listener performance frequently shown to be worse where more variability is present (e.g., Lavan et al., 2016; Lavan et al., 2019; Afshan et al., 2020). Therefore, it is plausible that should speech modifications occur in telephone openings, this speech will also be more variable and may equally pose challenges for listeners.

The following presents findings from a pilot study which aimed to explore if prevalent within-speaker variability is occurring in telephone openings, compared to later in the calls. A speaker discrimination task was used to see if performance differs when listeners are presented with samples taken from call openings, compared to mid-conversation. We predicted that if openings contain greater within-speaker variability, performance would be worse in these conditions.

Methods

12 American-English listeners conducted a speaker discrimination task containing three conditions; initial vs. initial, where both samples were from the call opening, mid vs. mid, where both samples were taken from the middle of the call, and initial vs. mid, a mismatched condition. Stimuli consisted of 0.75ms samples extracted from telephone calls between unfamiliar speakers obtained from the Switchboard corpus (Godfrey and Holliman, 1993). Samples were taken either from the opening of a telephone conversation (the first 10 utterances) or the middle of a call (10^{th} utterance onwards). Participants were tested on 24 same and different stimuli sets for each condition, resulting in 144 trials (6 x 24). Signal detection statistics accounting for listener sensitivity and bias were used to assess performance, along with linear mixed effects models with *A'* as the dependent variable, *condition* and *gender* as predictors, and by-*participant* random effect.

Results

Findings show slightly worse performance in the initial vs. initial condition (*A'*: 0.70) compared to the initial vs. mid (*A'*: 0.77) and mid vs. mid (*A'*: 0.76) conditions (Figure 2). Although, model outputs show this only to be a slight trend; initial vs. mid ($p < 0.05$), mid vs. mid ($p < 0.1$). However, given the small sample size, we can tentatively interpret these findings as suggesting that the initial utterances contain greater variability, making unfamiliar speaker discrimination more difficult. Further, if telephone openings contain greater within-speaker variability, this could have potential implications for forensics tasks which utilize telephone speech, including forensic speaker comparison and speaker profiling, as well as unfamiliar naïve voice recognition. Depending on the extent of this variability, it is plausible that within-style variability poses similar mismatch conditions to between-style, and they should be approached with the same kind of caution when completing forensic analyses. Overall, contexts where additional within-speaker variability can be observed within a single speech style should be considered further.

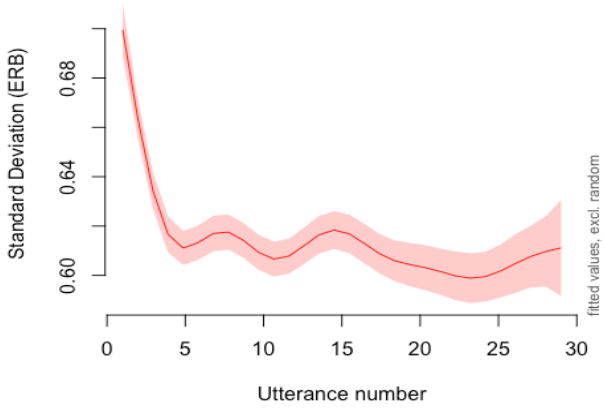


Figure 1. Nonlinear smooth fitted for f_0 standard deviation in each utterance across a call. Shaded bands represent the pointwise 95%-confidence interval.

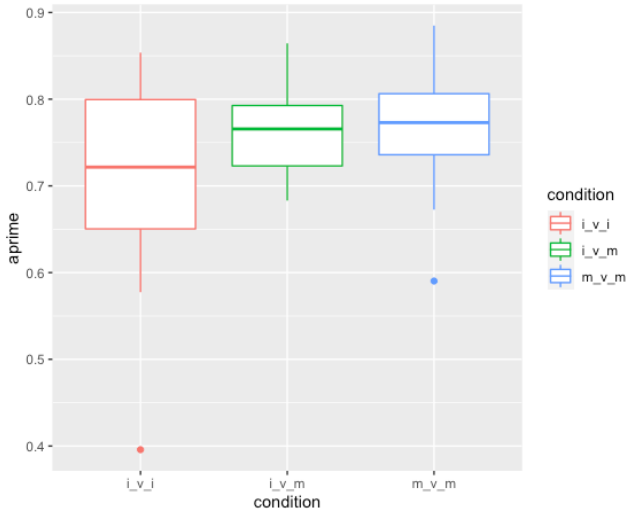


Figure 2. Boxplot showing *A'* differences across the three experimental conditions; initial vs. initial, initial vs. mid, mid vs. mid.

References

Afshan, A., Kreiman, J., & Alwan, A. (2020). Speaker discrimination in humans and machines: Effects of speaking style variability. *arXiv preprint arXiv:2008.03617*.

Dellwo, V., Pellegrino, E., He, L., & Kathiresan, T. (2019). The dynamics of indexical information in speech: Can recognizability be controlled by the speaker? *AUC PHILOLOGICA*, *2019*(2), 57–75.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501.

Godfrey, J. J., & Holliman, E. (1993). *Switchboard-1 Release 2 (LDC97S62)*. Web Download. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/sw3h-rw02

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). 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

Lavan, N., Scott, S. K., McGettigan, C. (2016). Impaired generalisation of speaker identity in the perception of familiar and unfamiliar voices. Journal of Experimental Psychology: General, 145, 1604–1614. doi:10.1037/xge0000223

Mayo, C., Aubanel, V., & Cooke, M. (2012). Effect of prosodic changes on speech intelligibility. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Prague, July 10–13, 2022
Student Paper

Towards a perceptual rhythm framework for forensic analysis: methodological developments

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

Although it is suspected that the rhythm of speakers' speech has something to offer forensic speech analysis, it is not clear how it could be best integrated into these analyses. Previous studies have looked into possible ways and variables to characterise individual speakers' speech rhythm and their speaker discriminatory power. Leemann, Kolly and Dellwo (2014) characterised speech rhythm using measures of relative syllable durations within utterances, and He and Dellwo (2016) reported more promising speaker discrimination results by using measures of relative intensity values of syllables within utterances. While these studies made use of content-controlled speech data to characterise rhythmic patterning, applying these measures to spontaneous speech is of greater relevance to forensic analysis.

An initial investigation by the first author preceding the current study looked at the discriminatory power of measures of intensity, f0 and duration across spontaneous (content-mismatched) utterances. It revealed that applying these measures to these data is largely unproductive. Utterances from 20 male speakers from the WYRED corpus (Gold et al. 2018) were analysed in relation to syllabic intensity and syllabic f0 values, as well as syllabic durations. Measurements were subjected to linear discriminant analysis which produced rather weak speaker discrimination results for all measures (at best, only a small margin above chance level). A follow-up study, focussed on the same speech data (same speakers within the mock police interview task), analysed the rhythmic characteristics of frequently occurring speech units (*erm*, *er*, *yeah* and *no*) as a means to measuring speakers' rhythm patterns. Speaker discrimination rates were markedly improved (e.g., *erm* = 81.3% correct - dynamic measurement of intensity, f0 and duration combined).

Results from these production experiments have shown that there is some value in pursuing rhythm for speaker identification. However, relying on the acoustics of rhythmic information can only capture so much. The current work acts as a natural next step as it aims to strengthen the auditory analytical potential of rhythm as a speech analysis feature. The long-term goal is to develop a perceptual rhythm framework which can be used in the context of forensic casework. The present work is a pilot study that looks into a perception test methodology that could feed into this goal.

Using speech samples from the WYRED corpus, the present work makes use of a "delexicalisation" tool to create samples that express only the rhythmic attributes of the speech sample. These samples were then presented to listeners to establish whether listeners can make meaningful identification assessments based only on the delexicalised samples. It is also hoped that meaningful descriptors of speech rhythm that could contribute towards a future auditory analysis framework for speech rhythm will be developed.

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.

An acoustic-phonetic description of diphthongs in Venezuelan Spanish

Simon Gonzalez
The Australian National University
u1037706@anu.edu.au

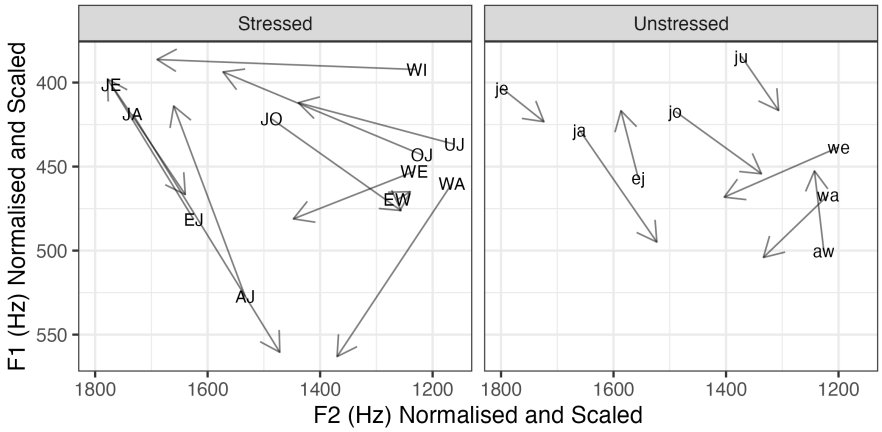
Venezuelan Spanish has been described in terms of its phonetic features. Sstudies include acoustics of [voice] correlation (Lain, 2012), consonantal deletion (Días-Campos & Killam, 2012; D'lntrono & Sosa, 1986), and sociolinguistic studies (Bentivoglio & Sedano, 1993; D'lntrono & Sosa, 1986; Obediente, 1999). These have been of great value for the understanding of Venezuelan Spanish. However, there remains one main gap in the research of Venezuelan Spanish and it is in the area of vowel acoustics within the context of regional variation. In several studies, like in Scrivener (2014), vowels are analysed in relation to consonantal phenomena but not on vowels directly. The current study aims to fill this research gap in this Spanish variety. Therefore, the main purpose of this work is to carry out the first in-depth acoustic analysis on the vocalic space of Venezuelan Spanish based on speakers from all dialectal regions for both monophthongs and diphthongs in conversational speech.

Results in the current work show the phonetic analysis of the observed diphthongs [je], [ja], [ej], [aj], [wi], [uj], [ju], [oj], [jo], [we], [ew], [wa], and [aw] in Venezuelan Spanish, in both stressed and unstressed contexts. An important contribution of this study is that it analyses data from conversational speech, which differs from other studies that only analyse controlled data, such as read sentences and isolated words. The data was collected from nine speakers, who came from five of the seven different regions in the country: *Andino, Central, Guayanés, Llanero,* and *Zuliano*. Each speaker was part of an interview run by professional journalists, and we selected five minutes of interrupted conversation. These were obtained from official YouTube channels and time-stamped closed-captions were available for each video. Closed-caption texts facilitate the audio-text alignment needed for the acoustic forced-alignment, which was done using the Montreal Forced Aligner (McAuliffe et al., 2017). Duration and formant measurements were analyzed, which were extracted in Praat (Boersma, 2001). In particular, we investigated the amount of spectral change in the formant trajectories by obtaining F1 and F2 values at 11 equally spaced time points from the beginning to the end of the duration for each vowel. Formant values were normalized using the Lobanov (Lobanov, 1971) normalization method available in the vowels package (Kendall & Thomas, 2018) in R (R Core Team, 2021).

Formant and duration measurements indicate unambiguous separation of stressed and stressed vowels. These results are expected, especially in conversational speech. However, when these results are compared to the current literature, they bring much more depth to the understanding of this variety. First of all, vowel durations are now understood in the light of more naturalistic speech, with unstressed vowels having a mean duration of 50 ms and stressed vowels with 80 ms. This is different from other studies where they find durations of approximately 110 and 120 ms. Another relevant finding is that we have a clear view of the phonetic compression in conversational speech in unstressed vowels.

Future research on this data will explore more acoustic features in this Spanish variety as well as developing acoustic models than can be used for forensic research.

Figure 1. Trajectories for phonetic diphthongs, with initial point at 20% and end points at 80% of the trajectories. Unstressed trajectories are prominently shorter than the stressed ones.



References

Bentivoglio, P. & Sedano, M. (1993). Estudio sociolingüístico del habla de Caracas: el corpus de 1987. Universidad central de Venezuela

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. Glot International 5:9/10, 341-345.

D'lntrono, F., & Sosa, J. M. (1986). La elisión de la /d/ en el espafiol de Caracas: Aspectos sociolingüísticos e implicaciones teóricas. En Rafael Iraset Páez & Jorge Guitart (Eds.). Estudios sobre la fonología del español del Caribe, Caracas: La Casa de Bello, pp. 135-63

Díaz-Campos, M. & Killam, J. (2012). Assessing Language Attitudes through a Matched-guise Experiment: The Case of Consonantal Deletion in Venezuelan Spanish. Hispania, 95(1), pp. 83-102

Kendall, T. & Thomas, E.-R. (2018). vowels: Vowel Manipulation, Normalization. Web Application Framework for R (Version 1.2-2) [R package]. Retrieved from http://blogs.uoregon.edu/vowels/

Lain, S. (2012). Acoustic [voice] correlate variation by dialect: Data from Venezuelan Spanish. Sociophonetics at the crossroads of speech variation, processing and communication, pp. 37-40

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different listeners. Journal of the Acoustical Society of America 49:606-08.)

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In Proceedings of the 18th Conference of the International Speech Communication Association.

Obediente, E. (1999). Identidad y dialecto: el caso de los Andes venezolanos. In Perl, Matthias & Pörtl, Klaus (Eds.). Identidad cultural y lingüística en Colombia, Venezuela y en el Caribe hispánico. Tübingen: Niemeyer, pp. 213-219

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

Scrivener, O. (2014). Vowel Variation in the Context of /s/: A Study of a Caracas Corpus. In Rafael Orozco (Ed.). New Directions in Hispanic Linguistics, pp. 162-183

Speech Length Threshold in Forensic Voice Comparison by Using Long-Term Fundamental Frequency in Chinese Mandarin

Honglin Cao[1], Chuyi Pan[1], Lei He[2]

[1] Key Laboratory of Evidence Science (China University of Political Science and Law), China.
[2] Department of Computational Linguistics, University of Zürich, Switzerland
caohonglin@cupl.edu.cn, bobocharisma@163.com, lei.he@uzh.ch

Introduction

Long-Term Fundamental Frequency (LTF0) is one of the most frequently-used acoustic features in Forensic Voice Comparison (FVC), although generally its discriminating power is limited (Rose 2002, Gold and French 2011). The mean value is the most common measure of LTF0 in FVC, followed by standard deviation (SD), median, baseline, etc. (Gold and French 2011). For all Long-Term features in FVC, one fundamental question is that how long the minimal duration of a speech sample (speech length threshold) is enough. This question for LTF0 has been investigated by a few studies, of which the results are inconsistent. For example, around 60s is recommended by (Nolan 1983), about 20s of voiced speech is enough for Wu Chinese dialect (Rose 1991, Rose 2002), and some other smaller time values of the speech length thresholds are given in (Arantes and Eriksson 2014, Arantes, Eriksson et al. 2017) for many tone and non-tone languages. Because of the limitations of the previous studies: duration of speech samples is short; number of speakers are few, etc., in this study, we investigate the speech length threshold of mean LTF0 based on a large database in Chinese Mandarin.

Method

400 speakers between 18 and 30 years old are involved, balanced by both genders and two speaking styles (read and spontaneous): 100 speakers for male-reading (MaR), male-spontaneous (MaS), female-reading (FeR) and female-spontaneous (FeS) each. The reading materials contain 25 sentences (262 Chinese characters in total) and 200 speakers read the materials five times (mean 337.4s/326.2s, and SD 50.5s/43.4s for male and female speakers, respectively.). Spontaneous speech samples are elicited by asking the other 200 speakers to describe a series of pictures about Beijing subway. The duration of all spontaneous speech samples is more than 3 minutes (mean 278.8s/271.4s and SD 55.2s/51.1s for male and female, respectively). WaveSurfer is used to extract the F0 raw data (frame interval 0.01s) and Matlab is used to do further statistics.

In order to minimize the impact of linguistic information (contents, tones, intonation, stress, etc.), we select every successive voiced speech segment of 0.5s (50 frames of F0 data) as a subunit (i.e., for a 300-second-long recording, the voiced speech is 150s, and there are 300 (=150s/0.5s) subunits included). The sequence of the subunits is then randomized another 9 times and the LTF0 is re-calculated the same number of times, consequently. To our knowledge, the minimal intra-speaker variation of stabile mean LTF0 for FVC is not clear and maybe speaker-specific. First of all, we assume that more than 100s of voiced speech is adequate to obtain the stable LTF0 value. Then, we set three dynamic levels for the threshold of intra-speaker variation of mean LTF0, which are "±1% of stable value", "±2% of stable value" and "±3% of stable value". An example is shown in figure 1. The stable value of the mean LTF0 of a male speaker's reading speech is 117.5Hz. At the "±2% of stable value" level (range: 115.15Hz-119.85Hz, differences: 4.7Hz), for random 0 (the original sequence of the subunits), the mean LTF0 can be regarded as forensically stable when the duration of voiced speech is more than 50s (i.e., the located stabilization point for random 0 is 50s). We use the average duration of the ten located stabilization points for random 0 to 9 as the final minimum speech length threshold.

Results

The mean/SD of stable values of MaR speech, MaS speech, FeR speech and FeS speech are 131.1/18.8Hz, 121.8/17.0Hz, 222.3/17.3Hz, 202.2/20.6Hz, respectively. The specific speech length thresholds of mean LTF0 for 4 gender-style combinations at three dynamic levels are shown in figure

2. Obviously, the stricter the dynamic level for threshold of intra-speaker variation is, the longer the duration of speech samples is required to obtain a characterization of speaker's LTF0. For FVC, on average, at least 50.0s, 20.2s and 10.6s of voiced speech is required to approach a stable mean LTF0 value at the ±1%, ±2% and ±3% of "stable value" levels, respectively.

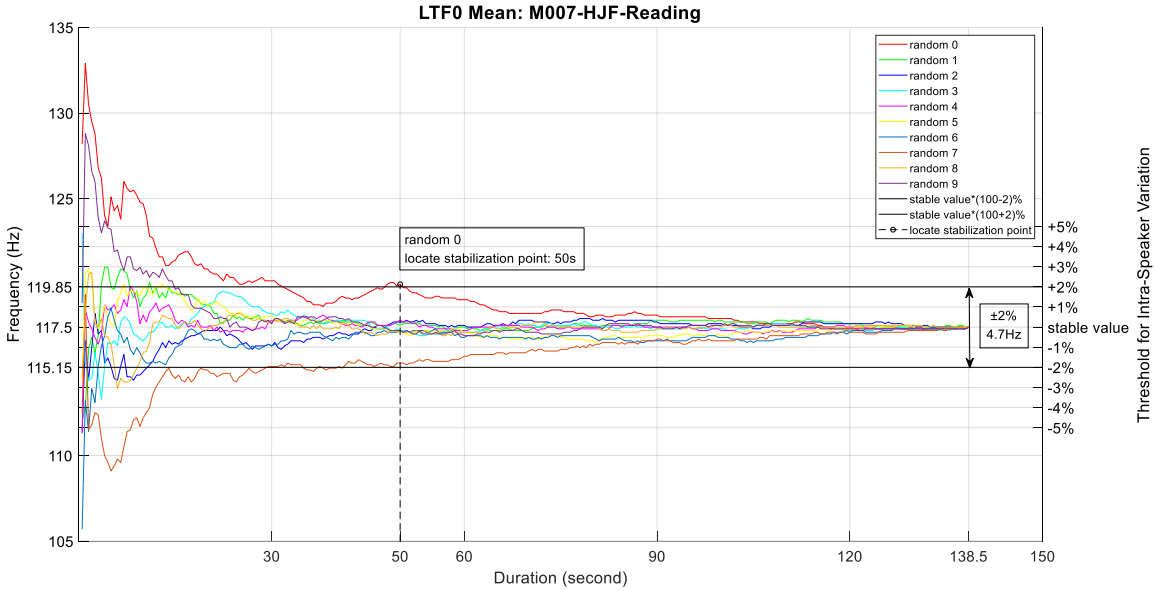


Figure 1. The 10 random distribution curves of the mean LTF0 of one male speaker's reading speech

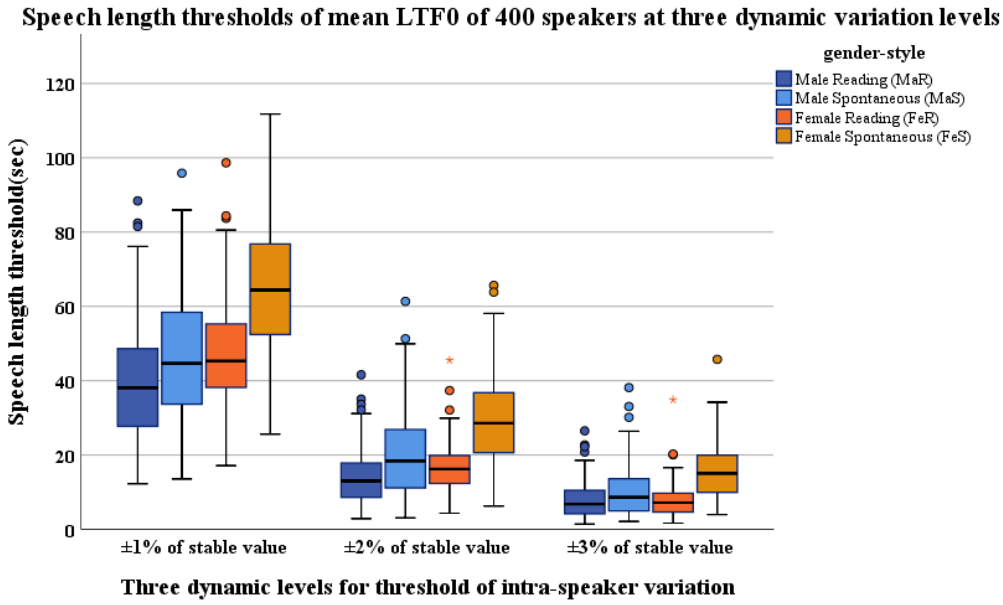


Figure 2. Box plots for speech length thresholds of mean LTF0 at three dynamic variation levels

References

Arantes, P. and A. Eriksson (2014). Temporal stability of long-term measures of fundamental frequency. *Proc. Speech Prosody*. 1149-1152.

Arantes, P., A. Eriksson and S. Gutzeit (2017). Effect of Language, Speaking Style and Speaker on Long-Term F0 Estimation. *Proc. INTERSPEECH*, Stockholm, Sweden. 3897-3901.

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

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

Rose, P. (1991). How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication* 10(3): 229-247.

Rose, P. (2002). *Forensic Speaker Identification*. London and New York, CRC Press.

Case report: Forensic analysis of a ticking clock in a recording

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

Recently the Netherlands Forensic Institute investigated a case about authenticity of audio recordings. The recordings were claimed to be made with a voice recorder app on a mobile phone hidden in a coat. These recordings were extremely incriminating for the suspect. The defense did not contest speaker identity, but claimed that the recordings were manipulated by cutting and pasting audio from several conversations to make the suspect say things that were never said. The judge ordered NFI to investigate the recordings and the claims of the defense.

Research question

While listening to the audio recordings, it became clear that the recorder had picked up a ticking clock in a room where the conversations were held. This made it possible to check the recordings for cut and paste manipulations by analyzing the timing of the clock ticks. This clock tick analysis was part of a larger authenticity research, which included linguistic conversation analysis and inspection of recording artefacts.

To gain more insight in the internal working of mechanical clocks, several publicly available resources were consulted, such as the YouTube channel from the National Watch & Clock Museum (National Watch & Clock Museum, n.d.). Mechanical clocks work by letting a gear, the escapement wheel, move forward in small steps of the same length. The periodic halting of the escapement wheel creates a ticking sound, which was audible in parts of the recordings. At some points, speech and noise from movement and clothing drowned out part of the clock ticks, as can be seen in figure 1.

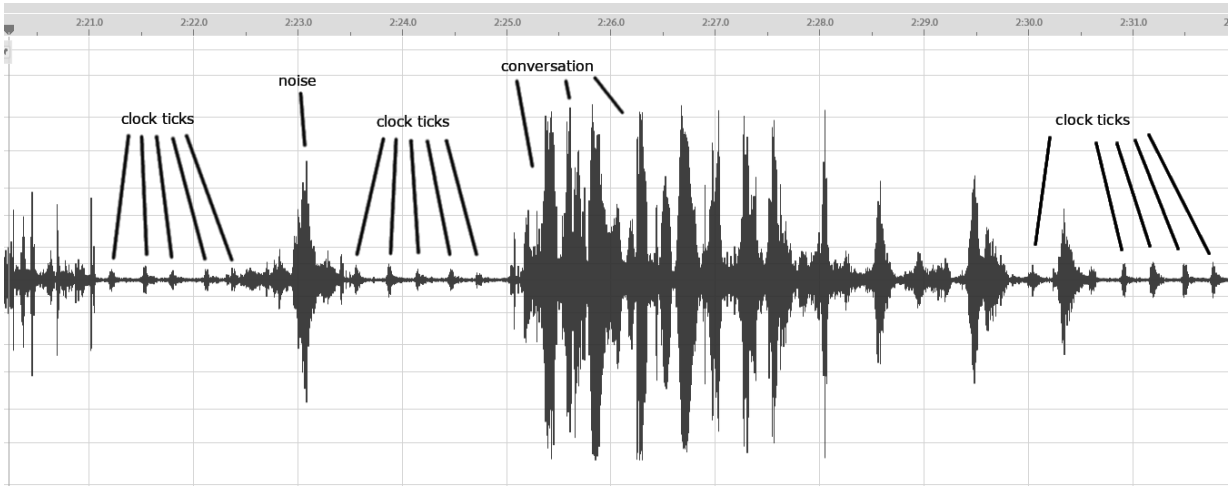


Figure 1. Part of the waveform of one of the recordings, with the clocks ticks clearly visible.

Method

A two-part method for analysis of the clock ticks was developed, within the guidelines from the SWGDE Best Practices for Forensic Audio (Scientific Working Group on Digital Evidence, 2016).

First was checked if extra clock ticks would fit in the areas where they were inaudible in the recording. If it was not possible to fit a whole number of clock ticks, this would be an indication that there was audio removed or added in these areas.

Next, a 'perfect clock' was created, based on the measured period of the clock ticks in the recordings. This perfect clock was then placed over the recording and the time difference between every perfect clock tick and its corresponding actual clock tick in the recording was calculated, producing an error. The summed error was then minimized to make the best possible fit of the perfect clock on the real recording.

A sudden jump of this timing error would indicate that a part of the recording was shifted in time and thus manipulated. Some drift in the timing analysis was expected, because when placing the timing markers in the recordings, it was not possible to be more precise than the duration of the clock tick. The results from the developed method showed no large deviations of the clock signal and showed no support for the hypothesis that the audio was manipulated.

The make and model of the recorded mechanical clock was unknown, so typical characteristics of this clock (such as jitter) were not taken into account. Because mechanical clocks are normally designed to keep accurate time for a long period of time (usually weeks), it was assumed that there were no large changes in the timing of the clock during the relatively short recordings.

The research method was validated by simulating several manipulations on the clock timing data. At randomly selected points in the clock tick timing data, extra time was added, simulating cut and paste operations. These manipulations showed up as sudden and permanent deviations from the clock signal. Figure 2 shows the analysis of a part of one of the recordings without (a) and with (b) artificial manipulation. In the latter case the timing error between real clock and the perfect clock shows a large jump.

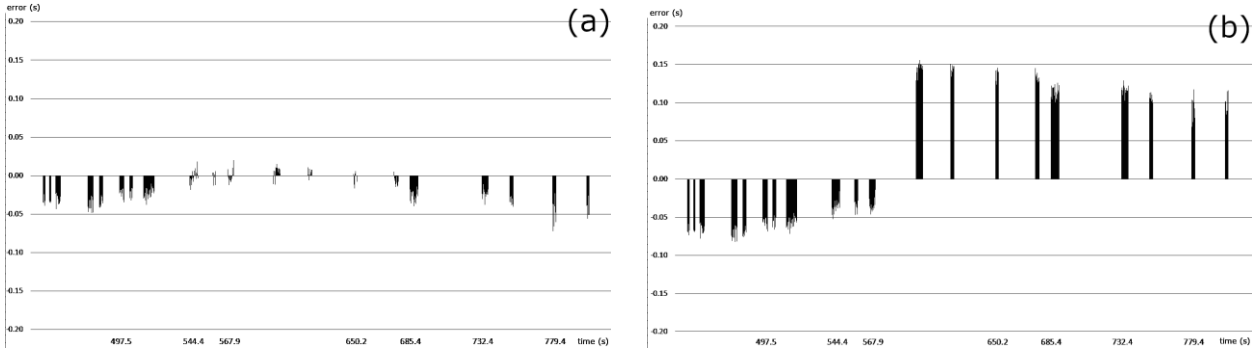


Figure 2. Timing errors in a part of the original recording (a) and with a simulated cut and paste operation (b) in that same part.

Conclusion

In a case where the sound of a mechanical clock was picked up in an audio recording, it was possible to use these clock ticks to check the authenticity of the recording. A method for checking the timing was developed and validated at NFI for this case and for future use.

References

National Watch & Clock Museum. (n.d.). *Home*. [YouTube channel]. YouTube. Retrieved March 17, 2022, https://www.youtube.com/c/NationalWatchClockMuseum

Scientific Working Group on Digital Evidence. (October 8, 2016). *SWGDE Best Practices for Forensic Audio Version 2.2*. https://www.swgde.org/documents/published

Utility of length-normalization for predicting trademark sound-alikes from Levenshtein string edit distance

Vincent J. van Heuven[1, 2, 3]
Sandra F. Disner[4]
[1] *Department of Hungarian and Applied Linguistics, University of Pannonia, Hungary*
[2] *Leiden University Centre for Linguistics, The Netherlands*
[3] *Frisian Academy, Leeuwarden, The Netherlands*
[4] *University of Southern California, Los Angeles, USA*
v.j.j.p.van.heuven@hum.leidenuniv.nl
sdisner@usc.edu

Abstract

Background. We are interested in providing a phonetic underpinning for hitherto impressionistic/intuitive court decisions in trademark infringement cases, where an existing ('senior') mark argues that the name of a newcomer ('junior mark') on the market is so similar to the existing mark that consumers may get confused.

Lambert (1997) showed that perceptual confusion between brand names of pharmaceutical products, such as *Diphenatol ~ Diphenidol* /daɪ'fɛn**ət**ɑl ~ daɪ'fɛn**ɪd**ɑl/, could be better predicted from the Levenshtein distance (LD) between the names than from either the number of phone bigrams or phone trigrams shared between names. Van Heuven et al. (2021) examined ways to optimally separate between pairs of trademarks in the USA that were judged by the court to sound too similar to compete on the market (e.g., mortgage lenders *Ameriquest ~ Americrest* /ə'mɛrɪk**w**ɛst ~ ə'mɛrɪk**r**ɛst/) and pairs that were sufficiently different to be allowed to co-exist (e.g., cholesterol-lowering drugs *Advicor ~ Altocor* /'æ**dvɪ**kɔr ~ 'æ**lto**kɔr/). They, too, found that LD outperformed the proportion of shared bigrams and trigrams – although a combination of LD and bigram frequencies yielded even better results.

Van Heuven et al. computed LD as the smallest number of string operations needed to convert the phonetic transcription of a trademark to the transcription of its competitor. Possible string operations are deletion, insertion and substitution of a symbol. Following Heeringa (2004: 130), Van Heuven et al. length normalized their LD, arguing that longer symbol strings offer more opportunities for mismatches than shorter strings. Lambert (1997), however, did not length-normalize.

Goal of the present study. The present paper reports on a pilot study exploring the utility of raw versus length-normalized LD. Moreover, the comparison was done once with a plain LD, in which all string operations contribute equally to the LD, and a second time with feature-weighted differences between segments. The latter is a more sophisticated approach which takes into account that some sounds (e.g., /m, n/) are more easily confused than others (e.g., /m, s/). Our feature weighting is based on Almeida and Braun (1986), adapted for LD measurement (Heeringa & Braun 2003), and recently implemented as an option in the LED-A app (Heeringa 2021, Heeringa et al. 2022).

The hypothesis is tested that the length-normalized LD is the better predictor of past decisions on allowable trademark pairs in the (rather small) database of documented court cases in Van Heuven et al. (2021). Similarly, a similar test is reported on the ability of length-normalized LD

to differentiate between confused and non-confused (generic) product names in (a representative sample from) Lambert's (1997) data.[1] Finally, we report on the potential advantage of the feature-weighted LD relative to the plain LD, in both comparisons, testing whether the benefits of length normalization and feature weighting are additive.

References

Almeida, Antonio & Angelika Braun (1986). "Richtig" und "falsch" in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, 53(2), 158–172. Retrieved from https://www.jstor.org/stable/40502947

Heeringa, Wilbert J. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen: Groningen Dissertations in Linguistics, 46. Retrieved from https://pure.rug.nl/ws/portalfiles/portal/9800656/thesis.pdf

Heeringa, Wilbert J. (2021). Levenshtein Edit Distance App (LED-A). Computer program. https://fryske-akademy.nl/fa-apps/led-a/#run

Heeringa, Wilbert J. & Angelika Braun (2003). The use of the Almeida-Braun system in the measurement of Dutch dialect distances. *Computers and the Humanities*, 37(3), 257–271. Retrieved from http://wjheeringa.nl/papers/cath02a.pdf

Heeringa, Wilbert J., Vincent J. van Heuven & Hans Van de Velde (2022). LED-A: a web app for measuring distances in the sound components among local dialects. Poster to be presented at the 17th New Methods in Dialectology Conference, Mainz.

Heuven, Vincent J. van, Sandra F. Disner & Wilbert J. Heeringa (2021). What's in a name? On the phonetics of trademark infringement. Paper presented at the IAFPA 2021 Conference, Marburg.

Lambert, Bruce L. (1997). Predicting look-alike and sound-alike medication errors. *American Journal of Health-System Pharmacy*, 54(10), 1161–1171. Doi: 10.1093/ajhp/54.10.1161

[1] We are most grateful to professor Bruce Lambert, director of the Center for Communication and Health at Northwestern University (Evanston, IL), for making his (updated) list of 1,250 confused medication brand names available to us.

The impact of reflection and retention intervals on earwitness accuracy: two experiments

Francis Nolan[1], *Nikolas Pautz*[2] *Kirsty McDougall*[1], *Katrin Müller-Johnson*[3], *Harriet Smith*[2], *and Alice Paver*[1]
[1]*University of Cambridge, UK,* [2]*Nottingham Trent University, UK,* [3]*University of Oxford, UK*
[1]{fjn1|kem37|aep58}@cam.ac.uk
{nikolas.pautz|harriet.smith02}@ntu.ac.uk
[3]katrin.mueller-johnson@crim.ox.ac.uk

In most lab-based experiments on earwitness performance, participants hear a target ('perpetrator') voice then either return later for a voice parade (McDougall *et al*. 2015), or complete a 'filler' task (Smith *et al.* 2020) simulating the delay between encoding and retrieval. Yet one might predict that if the witness realises they have heard a perpetrator's voice, they are likely to think back over the event, i.e. post-encoding reflection may occur. The present study investigates whether such reflection improves voice recognition accuracy, through two experiments which manipulate the role of post-encoding reflection in a voice parade task.

Experiment 1 used a 2×2 factorial (target presence: present, absent; reflection: reflection, no reflection) design. Three target speakers of SSBE were selected from *DyViS* (Nolan *et al.* 2009). 9-speaker parades with 15-second samples were constructed for each target. *DyViS* speakers were selected as foils using multi-dimensional scaling of listener similarity ratings (McDougall 2013) so as to approach the lower bound of earwitness performance by using voices highly similar in accent and personal voice quality. 80 listeners were randomly assigned to the targets and exposed to a 60-second encoding sample. Half of the participants were instructed: "*Imagine that the voice you have just heard is that of a criminal. You may be asked by the police to make an identification some time in the future. Take a few moments now to reflect on the voice."* (the reflection condition). The other half were in the control condition. In both the instruction and the no-instruction condition, a 20-second interval elapsed before listeners completed a 5-minute word-search task with lobby noise. The listeners then undertook a target-absent or target-present voice parade. Listeners with post-encoding reflection showed no meaningful differences in identification performance from those without; nor was there an interaction between target presence and reflection. Responses to target-present parades were more likely to be accurate than those to target-absent parades, consistent with previous findings (Smith *et al.* 2020). Listeners had above-chance accuracy in target-absent parades, but only in the reflection condition (see Figure 1).

Experiment 2 echoed Experiment 1, except listeners (N=181) had a 20–28-hour retention interval between exposure and parade, instead of a filler task. Listeners in the reflection condition showed no meaningful differences in accuracy from those in the control condition, nor was there an interaction between target presence and reflection. There was no effect of target presence on accuracy, surprisingly, since target presence is a relatively robust effect. Contrary to Experiment 1, target-absent parades did not demonstrate above-chance levels in the reflection condition.

The results provide some evidence that including a reflection period reduces the likelihood of making a positive identification for a target-absent parade. This effect, however, was present only in parades which used a five-minute filler task (Experiment 1) and not when using the longer retention interval (Experiment 2). The results suggest that experiments using a filler task can contribute to our understanding of voice parades, but also suggest that outcomes need to be tested using a more ecologically valid retention interval.

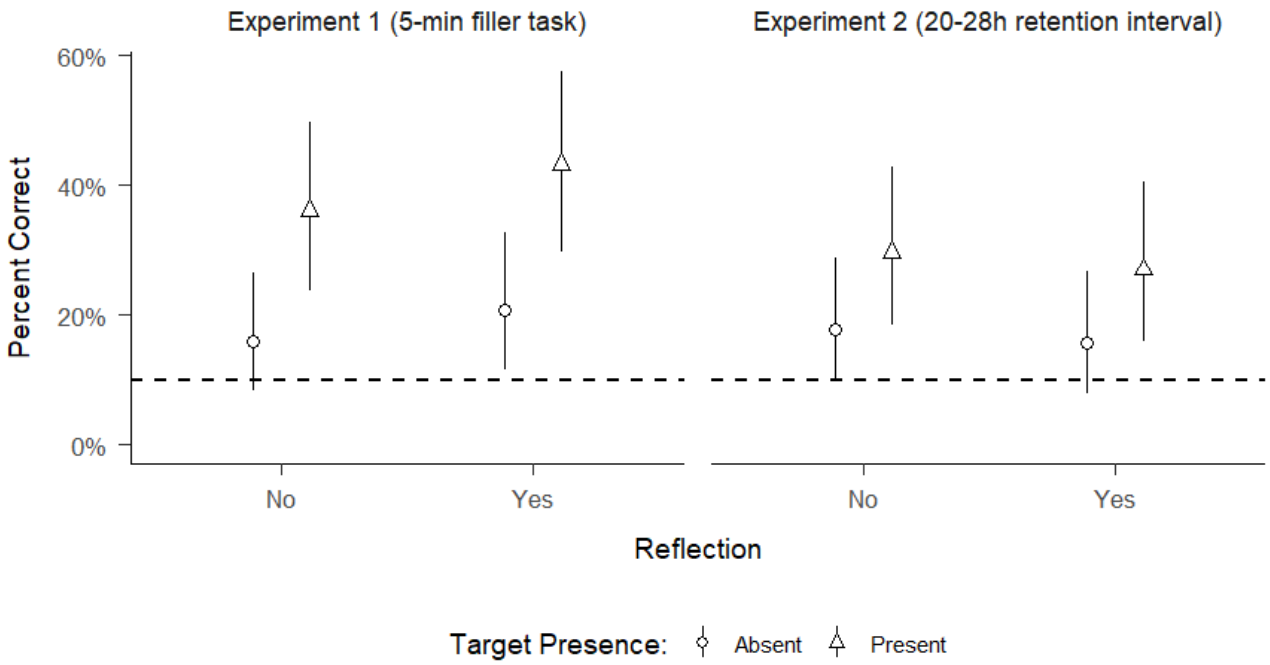


Figure 1. Point estimates and corresponding 95% Highest Density Interval (HDI) of accuracy extracted from the Bayesian logistic models.

References

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

McDougall, K., Nolan, F., & Hudson T. (2015). Telephone transmission and earwitnesses: performance on voice parades controlled for voice similarity. *Phonetica*. 72(4): 257-72.

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009) The *DyViS* database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech Language and the Law*. 16(1): 31-57.

Smith, H.M., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P.C. (2020). Voice parade procedures: optimising witness performance. *Memory*. 28(1): 2-17.

The performance of two ASR systems in language mismatch, foreign accent, and channel mismatch conditions

Jakub Bortlík
Phonexia s.r.o., Brno, Czech Republic
jakub.bortlik@phonexia.com

In my contribution I will present part of my dissertation in which I tested the performance of two Automatic Speaker Recognition (ASR) systems in the conditions of language mismatch, foreign accent, and channel mismatch. Channel mismatches have been shown to have a negative effect on ASR performance (Morrison et al. 2012). In personal communication, several forensic speaker comparison experts have attested that language mismatches are a serious issue and often a reason for performing no forensic speaker comparison of the data at all.

I collected a dataset of recordings of 31 Czech native speakers reading short Czech and English texts. Four English and four Czech phrases were selected from each speaker. The original recordings (44.1 kHz sampling rate) were transformed to simulate the quality of telephone calls (8 kHz and other modifications) to test channel effects (Enzinger et al. 2016).

Both original and "phone" versions of the English recordings were rated for foreign accent. The average ratings ranged from very accented to nearly native speaker levels and so could be used to look for the effect of foreign accent on the performance of ASR systems.

Two ASR systems, commercial Phonexia *SID4-XL4* and open-source SpeechBrain *spkrec-ecapa-voxceleb* (Ravanelli et al. 2021), were tested with the dataset (496 recordings). Scores from the ASR systems were prepared for each unique combination of the recordings.[1]

The effect of foreign-accent strength was analyzed by calculating Spearman coefficients of the correlation between the ASR scores from the cross-language trials and the accent ratings. There turned out to be a weak positive correlation (Spearman 0.096-0.263) in the case of same-speaker trials, i.e., cross-language trials received somewhat higher ASR scores if the English recording in the pair was more foreign-accented. This suggests the ASR systems benefitted from the presence of foreign accent to provide more accurate identifications, while speakers with a more native-like pronunciation could more easily pass for a different person in the cross-language condition.

Equal Error Rates (EER) were measured for several subsets of the ASR scores to analyze effects of language and channel mismatch. Figure 1 shows that when both original and simulated "phone" call recordings were considered together, there were only small differences between the cross-language and matched-language trials in terms of EER. The *spkrec* system even had the highest EER (26.2%) in the Czech matched-language condition which was against our prediction that the cross-language condition would have the highest EER.

Figure 2 shows that channel mismatches were responsible for the highest error rates and that both ASR systems performed the best in the matched-channel condition with original recordings. Further analysis showed that after factoring out channel, both ASR systems performed significantly better in the matched-language than in the cross-language condition. It turned out that it was the combination of mismatches and, more importantly, the inclusion of both matched and mismatched trials which caused the highest error rates.

[1] Only same-sex comparisons were considered (Doddington et al. 2000) and recordings were never compared to their alternative versions (original vs "phone").

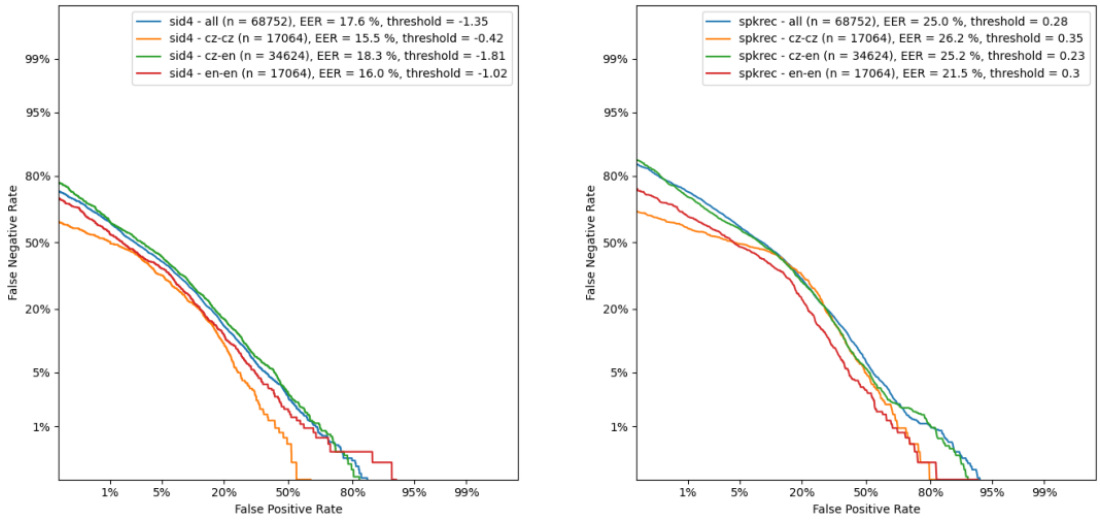


Figure 1: Error rates and EER values of SID4-XL4 and spkrec-ecapa-voxceleb for different combinations of language of the samples, cz = Czech, en = English, all = all samples taken together (both original and simulated "phone" recordings are included). The *n* denotes the number of trials in each condition.

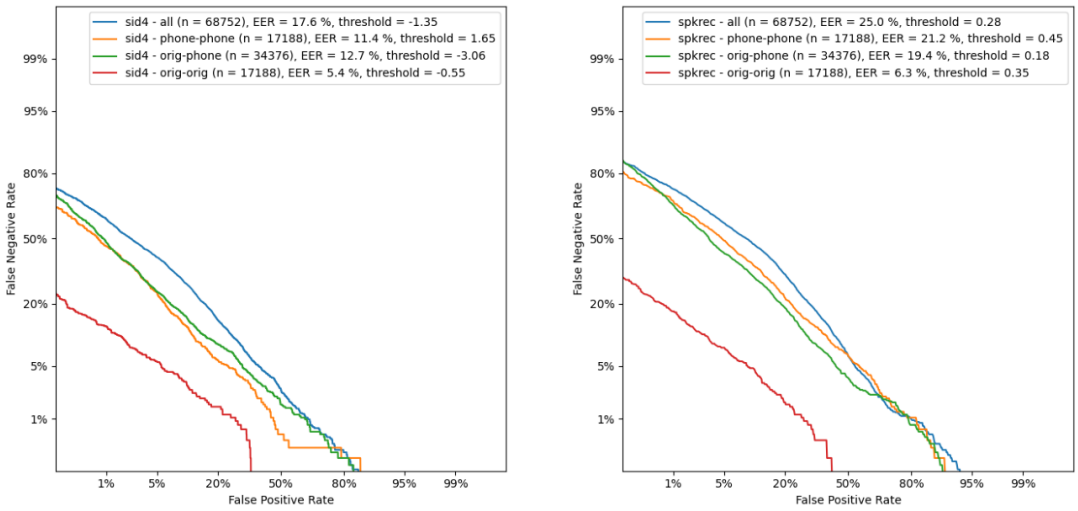


Figure 2: Error rates and EER values of SID4-XL4 and spkrec-ecapa-voxceleb in the channel (mis)match conditions. Note that the "all" condition combines both cross-channel and matched-channel conditions.

References

Doddington, G.R., Przybocki M.A., Martin A.F. & Reynolds D.A. (2000). The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. *Speech Communication*, no. 31: 225–254.

Enzinger, E., Morrison G.S. & Ochoa F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science and Justice*, no. 56: 42–57.

Morrison, G.S., Ochoa F., and Thiruvaran T. (2012). Database selection for forensic voice comparison. In *Proceedings of Odyssey 2012: The Speaker and Language Recognition Workshop*, 62–77. Singapore.

Ravanelli, M., Parcollet T., Plantinga P., Rouhe A., Cornell S., Lugosch L., Subakan C., et al. (2021). SpeechBrain: A General-Purpose Speech Toolkit. ArXiv:2106.04624. arXiv: 2106.04624 [eess.AS].

Stoet, G. (2010). PsyToolkit – A software package for programming psychological experiments using Linux. *Behavior Research Methods,* no. 42 (4): 1096–1104.

Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology,* no. 44 (1): 24–31.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Using eye-tracking as a method to explore decision making in voice recognition tasks

Leah Bradshaw[1], *Chiara Tschirner*[1], *Lena Jäger*[1] *and Volker Dellwo*[1]
[1]*Department of Computational Linguistics, University of Zurich, Zürich, Switzerland*
leah.bradshaw@uzh.ch

Background

Voice recognition and identification tasks are an experimental technique used regularly in forensic phonetic research to explore lay-listener identification capabilities of personally familiar or trained-to-familiar voices. Multiple studies have used them to explore the influence of missing acoustic information on naïve speaker recognition abilities, e.g. glottal-waveform, fundamental frequency and formant modifications (Lavner, Gath & Rosenhouse, 1999), or noisy/degraded signals such as telephone speech (Foulkes & Barron, 2000), as well as the influence of acoustic feature adaptations, e.g. whisper voice (Foulkes & Sóskuthy, 2017) and sweeping harmonics (Dellwo et al., 2018).

Despite their frequent occurrence in forensic phonetic research, little is known about how participants complete these tasks, on account of performance and sensitivity typically being the only measures assessed in analyses. Indeed, a greater understanding of decision-making and the time course of these tasks would provide valuable insight for evaluating the credibility of earwitness evidence for court admission.

The Visual World Paradigm (VWP – Allopenna et al., 1998) is a popular eye-tracking experimental technique used in psycholinguistic and phonetic research to explore online processing of various linguistic information. Typically, it involves participants being presented with a visual scene while listening to speech. Where and when participants' visual attention shifts to a given object in the visual world is taken to reflect their current interpretation of the audio stimulus. Although typically used to explore online speech processing, there are limited but promising findings showing its utility for assessing online processing of speaker identification (Schindler & Reinisch, 2015).

Given its capacity to assess online processing, the VWP represents a viable technique for exploring the timing of voice recognition, namely how long it takes for the target to be selected following stimulus onset and the sequence in which voices are considered. Further, it could also be beneficial for assessing exactly what role voice similarity plays in decision-making. This project presents a first-of-its-kind experimental technique, combining a VWP and voice recognition task, for exploring decision making in naïve familiar voice recognition.

Research Questions

For this experiment, we propose the following questions:

- Can proportion of looks/fixations be used to assess confusion or difficulties regarding similar voice competitors?
- Can fixation sequences be used to examine how participants complete tasks? How frequently do participants revisit competitors in decision-making?
- Is the timing to the decision of target speaker generalisable or individual to listeners?

Methods

This experiment design represents a pilot study which will test the validity of this experimental concept as a method to explore decision making in voice recognition tasks. For this pilot, participants will be presented with four voices and four corresponding images which they must learn in an initial familiarization stage. Participants will then complete a training phase and test phase, where they are presented one of the four voices and required to select the image corresponding to that voice while their eye movements are being recorded. In the training phase only, participants will be given feedback on their accuracy. Voice stimuli will be selected based on a lay-listener voice similarity judgements collected using a human perception questionnaire to ensure that what we are classifying as similar in the experiment correlates with lay-listener judgements.

References

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. *Journal of Memory and Language*, *38*(4), 419–439. https://doi.org/10.1006/jmla.1997.2558

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

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

Foulkes, P., Smith, I., & Sóskuthy, M. (2017). Speaker Identification in Whisper. *Letras de Hoje*, *52*(1), 5. https://doi.org/10.15448/1984-7726.2017.1.26659

Prague, July 10–13, 2022

Work-in-Progress, Student paper

Seeking voice twins – an exploration of VoxCeleb using automatic speaker recognition and two clustering methods

Linda Gerlach[1,2]*, Kirsty McDougall*[1]*, Finnian Kelly*[2]*, and Anil Alexander*[2]

[1]*Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, Cambridge, UK.*

{lg589|kem37}@cam.ac.uk

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

{linda|finnian|anil}@oxfordwaveresearch.com

Speaker similarity is a highly relevant concept in forensic phonetics, be it for constructing a voice parade fair to all involved parties or to assess the theoretical impact of relevant populations similar to a suspect speaker in forensic speaker recognition. Taking similarity to the extreme, the question arises whether it is possible to find voice twins, i.e. speech recordings originating from different, unrelated speakers that sound extremely similar to one another. Applications of voice twins may be found in earwitness assessment tasks (see Schäfer & Foulkes, 2022) or in medical voice banking when the available audio material of a voice-impaired person is insufficient for personalising a speech-generating device (see e.g. Yamagishi et al., 2012).

Previous studies have indicated that automatically obtained similarity scores based on perceptually relevant acoustic features (i.e. LTF1 to LTF4) are able to approximate, to a certain extent, ratings of perceived voice similarity as judged by listeners (Gerlach et al., 2020, 2021). Using automatically obtained similarity scores, recent research has tried to further concentrate a selection of speakers into more similar subgroups using agglomerative hierarchical clustering (AHC), gaining some general insights regarding clustering of speaker sex and the potential to find very similar sounding speakers in clusters where AHC branches merge early on (Gerlach et al., 2022). However, AHC has the disadvantage of forcing items into clusters and hence may form them even when speakers do not sound particularly similar. Additionally, the study relied on a small selection of 180 speakers with one recording each, thus, increasing the number of speakers as well as the diversity of recordings may improve the chances of discovering voice twins.

The aim of the present study is to expand on Gerlach et al. (2022) using a subset of good quality recordings (n=831, thereof 348 female; 30s minimum net speech, 24dB SNR, 0% clipping) of VoxCeleb (Nagrani et al., 2020), a large, diverse speaker database encompassing multiple recordings per speaker. Two clustering approaches for detecting voice twins will be explored: the previously used AHC, as well as the clustering method DBSCAN (density-based spatial clustering of applications with noise), which allows for items to not belong to a cluster ("noise"). Similarity ratings between the recordings will be obtained using VOCALISE automatic speaker recognition software relying on x-vectors and automatically-extracted phonetic features (LTF1 to LTF4). It is hypothesised that dense clusters containing more than one speaker and multiple files per speaker are possible voice twin candidates. An initial auditory and acoustic assessment of potential voice twins will be conducted, and challenges pertaining to voice similarity assessment and constructing a listener experiment will be discussed.

References

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2021, August). How do automatic speaker recognition systems 'perceive' voice similarity ? Further exploration of the relationship between human and machine voice similarity ratings. *Proceedings of the Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*.

Gerlach, L., McDougall, K., Kelly, F., & Alexander, A. (2022, April). Selecting similar-sounding speakers for

forensic phonetic applications: an exploration of cluster analysis using automatic speaker recognition. *Presented at British Association of Academic Phoneticians (BAAP) Colloquium*.

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

Nagrani, A., Chung, J. S., Xie, W., & Zisserman, A. (2020). Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, *60*, 101027. https://doi.org/10.1016/j.csl.2019.101027

Schäfer, S., & Foulkes, P. (2022, April). Assessing the Individual Voice Recognition Skills of Earwitnesses. *Presented at British Association of Academic Phoneticians (BAAP) Colloquium*.

Yamagishi, J., Veaux, C., King, S., & Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, *33*(1), 1–5. https://doi.org/10.1250/ast.33.1

Prague, July 10–13, 2022
Work-in-Progress, Student paper

The effect of listener accent background on the transcription of Standard Southern British English

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

In England and Wales, orthographic transcripts are often provided alongside speech evidence in courts of law so that members of the jury can be provided with a written copy of what was said. The accuracy of these transcripts is extremely important as a result of 'priming' effects following the exposure to a transcript; listeners may be influenced to hear something that is contained within the transcript but not within the speech signal and, even after compelling evidence of the transcript's implausibility, remain confident in their interpretation (Fraser et al., 2011). Current work on the transcription of forensic or legal audio has not considered how the regional accent of both the speaker and listener may affect transcription accuracy.

'Familiarity' with an accent has been shown to affect performance in a range of speech processing and transcription tasks (Sumner & Samuel, 2009; Floccia et al., 2006; Adank & McQueen, 2007). A significant drop in performance tends to be observed for 'unfamiliar' accents while accents which are judged to be familiar, such as the speaker's home accent and their country's standard variety, both elicit a higher and similar level of performance. For transcription tasks, this effect is particularly prevalent in poorer listening conditions (Smith et al., 2014). Many studies focus solely on within-group behaviour rather than comparing the performance for one particular accent across different listener groups (e.g., Adank et al., 2009). This study investigates how listener accent background may affect transcription accuracy of the standard variety, which all listeners are judged to be familiar with due to its dominance in education, public life and international media (Lindsey, 2019).

24 short utterances were extracted from the mock police interview task from the DyViS database (Nolan et al., 2009). Each extract contains 3-6 seconds of speech from one of four speakers of Standard Southern British English (SSBE). Dynamic compression was applied to the recordings to reduce the difference in amplitude between the loudest and quietest sections of each utterance to ensure that the signal-to-noise ratio was relatively consistent within each utterance. The recordings were then mixed with speech-shaped noise derived from the experimental recordings in Praat (Boersma & Weenink, 2022) and the signal-to-noise ratio (SNR) was manipulated to create three listening conditions: +6 dB SNR, 0 dB SNR and -3 dB SNR representing fair, moderate and poor intelligibility respectively. 120 participants were divided evenly into two listener groups such that listeners' accents matched or mismatched with the accent of the speech in the stimuli. Participants were therefore either speakers of SSBE or speakers of nonstandard regional varieties of British English. Participants were presented with the set of 24 unique utterances at a mixture of audio qualities and instructed to transcribe all of the speech that they could hear within the audio file, listening as many times as they wished. Participant transcripts were automatically aligned with a reference transcript on a word-level basis using a custom-built online tool, and word pairs were assigned an error category (no error, substitution, insertion or deletion).

This paper will present preliminary findings of the study, comparing transcription performance across listener groups and listening conditions by taking into account the types and frequencies of errors made. The study's design will allow direct comparisons to be made for each utterance across different accent backgrounds and different audio qualities. This work forms part of a larger doctoral research project which aims to identify the ways in which regional accent may affect transcription performance and as a result assist police, security and forensic agencies to provide better, more accurate evidence in criminal cases.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. Journal of Experimental Psychology: Human Perception and Performance, 35(2), 520-529.

Adank, P., & McQueen, J. M. (2007). The effect of an unfamiliar regional accent on spoken-word comprehension. 16th International Congress of Phonetic Sciences (ICPhS 2007).

Boersma, P. & Weenink, D. (2022) “Doing phonetics by computer,” [Computer program]. Version 6.2.13, retrieved 20th March 2022 from http://www.praat.org/.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? Journal of Experimental Psychology: Human Perception and Performance, 32(5), 1276-1293.

Fraser, H., Stevenson, B., & Marks, T. (2011). Interpretation of a crisis call: Persistence of a primed perception of a disputed utterance. International Journal of Speech, Language and the Law, 18(2), 261-292.

Lindsey, G. (2019). *English after RP: Standard British pronunciation today*. Springer.

F. Nolan, K. McDougall, G. de Jong and T. Hudson, “The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research.,” International Journal of Speech Language and the Law, vol. 16(1), pp. 31-57, 2009.

Smith, R., Holmes-Elliott, S., Pettinato, M., & Knight, R. (2014). Cross-accent intelligibility of speech in noise: Long-term familiarity and short-term familiarisation. The Quarterly Journal of Experimental Psychology, 67(3), 590-608.

Sumner, M. & Samuel, A. G. (2009). The effect of experience on the perception and representation of dialect variants. Journal of Memory and Language 60, 487-501.

Prague, July 10–13, 2022
Student Paper

Impact of vocal tract resonance modifications on LTF and f0

Alžběta Houzar, Tomáš Nechanský[1]*, and Radek Skarnitzl*[1]

[1]*Institute of Phonetics, Faculty of Arts, Charles University, Czech Republic*
alzbeta.houzar@ff.cuni.cz, tomas.nechansky@seznam.cz,
radek.skarnitzl@ff.cuni.cz

In everyday life, a speaker's voice characteristics change due to a range of factors such as speech style, affective states, daytime, or sickness; these shifts in the speaker's voice are possible due to extensive vocal tract plasticity and they often happen without the speaker's intention or even knowledge. This intra-speaker variability is, of course, of vital importance in the forensic phonetic context.

Another crucial aspect which can come into play in forensic voice comparison is intentional voice disguise, i.e., the speaker's deliberate attempt to conceal their identity by changing their voice characteristics. Some of the techniques used include placing a foreign object in front of or into their mouth (such as holding a tin can in front of their mouth as a resonator, covering their mouth with a cloth, or holding a pen between their front teeth; see Figueiredo & Britto, 1996), imitating a regional dialect or foreign accent, changing rhythmic characteristics of their speech, or trying to change their voice by articulatory or phonatory settings modifications.

Růžičková & Skarnitzl (2017) observed ways in which 100 Czech male speakers modified their voice when instructed to conceal their voice identity as much as they could in a manner of their own choice. The employed strategies differed among speakers, but in most cases, they were not very sophisticated, mostly including a change of a single parameter (predominantly speaking fundamental frequency, whose changes appeared in 70% of the speakers) or a combination of two. This study presents our research focused on targeted voice disguise in five male speakers of Common Czech. All of them are experienced voice users, trained in phonetics, and they were generally able to perform targeted voice manipulations. They were instructed to read a short text (the Czech translation of the Rainbow Passage) in their habitual voice, and several more times, each time performing a different resonance modification; the modifications were chosen mostly based on the SVPA scheme (San Segundo & Mompean, 2017): (1) strong lip spreading, (2) lip rounding, (3) closed jaw, (4) open jaw, (5) palatalization, and (6) pharyngealization. Speakers could repeat any passage of the text multiple times in case they did not maintain the targeted voice manipulation. The recordings were performed in a sound-treated studio; they were later edited so as to contain the best realization of each sentence vis-à-vis the intended modifications.

The aim of this study is to describe in what exact ways the individual resonance modifications affect LTF and f_0. This knowledge might be of help in cases where a speaker on a recording is apparently disguising their voice in a specific way and a prediction of the speech signal properties without the disguise is needed. This is a pilot study of our broader research on targeted voice disguise strategies' influence on instrumental and automatic speaker identification.

References

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

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

San Segundo, E. & Mompean, J. A. (2017). A Simplified Vocal Profile Analysis protocol for the assessment of voice quality and speaker similarity. *Journal of Voice*, 31(5), 644.e11–644.e27.

Forensic experts should focus on uncertainty rather than discriminability

Vincent Hughes, Bruce Xiao Wang
Department of Language and Linguistic Science, University of York
vincent.hughes@york.ac.uk/bruce.wang@alumni.york.ac.uk

Saks & Koehler (2005) described a *paradigm shift* within what they called the forensic identification sciences (now better referred to as forensic comparison sciences, of which forensic voice comparison is one). This involved a move away from unscientific methods, founded on the principle of *discernible uniqueness* – the notion that patterns can be compared to determine a match or mismatch. Since then, the paradigm shift has been extended to include (i) expression of expert conclusions using likelihood ratios, (ii) data-driven estimation of typicality, and (iii) validation of methods in line with international standards. The issue of validation, in particular, has received considerable attention and experts are now under considerable pressure from policy makers and regulators to validate their methods and systems in order to demonstrate that they work. However, forensic validation (and the same is true of forensic research) tends to focus on the overall performance of methods under casework conditions as evaluated by metrics such as EER or C_{llr}. This implicitly focuses the expert's attention on discriminability with different methods chosen and decisions made based on low values (or low assumed values) for the validity metric used. For example, an expert might choose to analyse F3 in a forensic voice comparison case and attach additional weight to the difference between known and unknown samples on the assumption that F3 is generally a good speaker discriminant. The view that discriminability should be the expert's primary focus has been proposed in Smith & Neal (2021).

We disagree with this view. Rather, we believe that the expert's primary concern should be to reduce uncertainty, rather than maximising potential discriminability (i.e. the possibility that a method could produce a low validity value). This is because reducing uncertainty is directly related to reducing the probability of a miscarriage of justice, which is the ultimate aim of the judicial process. Uncertainty here is defined broadly as variability in the specific conclusion (the LR) or validity value that a method produces (also referred to as reliability). Forensic voice comparison, of any kind, involves a series of decisions (be the principled or pragmatic) that potentially introduce uncertainty – in other fields this is referred to as *researcher degrees of freedom* (Roettger 2019).

In this paper, we discuss the issue of uncertainty in forensic voice comparison and demonstrate how it may be reduced via techniques such as Bayesian calibration (Brümmer & Swart 2014). We also present a series of recommendations for forensic experts. Specifically, experts should:

1. Recognise that forensic comparison is a process involving numerous decisions which introduce uncertainty via both systematic and random factors
2. Be explicit about the decisions made at each stage of the process and the implications of such decisions for uncertainty in terms of the results LRs **and** overall method validity
3. Take steps to measure and minimise uncertainty

The focus on uncertainty also directly relates to issues of reproducibility and replicability. In this paper, we also consider the specific challenges these concepts pose for forensic voice comparison.

References

Brümmer, N. and Swart, A. (2014) Bayesian calibration for forensic evidence reporting. *Proc. Interspeech*, 388-392.

Roettger, T. B. (2019) Researcher degrees of freedom in phonetic research. *Journal of the Association for Laboratory Phonology*, 10, 1.

Saks, M. J., & Koehler, J. J. (2005) The coming paradigm shift in forensic identification science. Science, 309, 892–895.

Smith, A. M. & Neal, T. M. S. (2021) The distinction between discriminability and reliability in forensic science. *Science & Justice*, 61, 319–331.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Assessing the specificity of creaky voice quality for forensic speaker comparisons

Katharina Klug
Department of Language and Linguistic Science, University of York, UK
kk667@york.ac.uk

Voice quality in forensic speaker comparisons is mainly judged auditorily. However, as with many other parameters, an attempt should be made to increase the objectivity and replicability of the analysis by establishing measures and procedures for conducting a thorough acoustic analysis. Klug et al. (2019) demonstrated the potential to assess voice quality quantitatively by using acoustic information from spontaneous speech samples of speakers who were rated as dominantly breathy. Repeating the same approach with speakers rated to be dominantly creaky proved problematic. In contrast to breathy voice, creaky voice (CV) is used as an umbrella term for different modes of glottal pulses that cannot easily be compared. What they have in common, however, is the perception of distinct glottal pulses (Laver, 1980: 124) due to amplitude damping between glottal excitation (Coleman, 1963). The present study attempts to refine the auditory assessment of CV modes to allow testing for acoustic correlates for each CV mode. In this way, speakers who are auditorily rated to be dominantly creaky could also be acoustically distinguished from non-creaky speakers.

There are two extremes of how CV is treated in the literature: over-simplification or over-specification. Over-simplified studies usually define one CV mode as the 'typical' CV phonation and ignore the multi-faceted nature of that phonation type which may have speaker-specific power (e.g. Dallaston & Docherty, 2019). Over-specified studies introduce too many CV modes which are difficult to apply to spontaneous speech samples and/or the degraded audio recordings typical in forensics (e.g. Keating et al., 2015; Batliner et al., 1993). Neither approach meets the requirements of forensic application. Therefore, I am looking for a new approach to assess specific CV phonation and hope to encourage forensic phoneticians to implement it in casework.

In searching for suitable approaches in the literature, I finally came back to the concept of 'compound phonation types' introduced by Laver (1980: 135). Laver describes the potential co-occurrence of CV with whispery voice, harsh voice and falsetto, and various combinations of the four, e.g. harsh whispery creaky voice (Laver, 1980: 161). I suggest excluding falsetto for classification purposes here as it rarely occurs as a long-term voice quality feature among non-pathological speakers (San Segundo et al., 2018: 368). This leaves us with three CV modes: clean CV, whispery CV and harsh CV. Each mode is assumed to be distinguished by different acoustic characteristics. Table 1 suggests a simplified classification based on auditory and acoustic information, which can be corroborated by spectrographic and spectral information.

	Clean CV	**Whispery CV**	**Harsh CV**
Acoustics	Single, periodic pulses	Single, aperiodic pulses	Multiple, a/periodic pulses
Auditory	Tension	High friction noise	Resonant hum

Table 1. Creaky voice modes characterised by the specified acoustic and auditory features (Laver, 1980; Esling et al., 2019).

I encourage discussion and feedback at my poster to refine the conception for the intended study. Interested conference participants are invited to group various spontaneous speech samples according to CV mode and test the proposed terminology during the poster presentation.

References

Batliner, A., Berger, S., Johne, B., Kießling, A. (1993). MÜSLI: A classification scheme for laryngealizations. In *Proceedings of the ESCA Workshop on Prosody,* Lund, 176-179.

Coleman, R.F. (1963). Decay characteristics of vocal fry. *Folia Phoniatrica et Logopaedica*, 15(4), 256-263.

Dallaston, K., & Docherty, G. (2019). Estimating the prevalence of creaky voice: A fundamental frequency-based approach. In *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, 532-536.

Esling, J.H., Moisik, S.R., Benner, A., & Crevier-Buchman, L. (2019). *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice, In *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 0821.1-0821.5.

Klug, K., Kirchhübel, C., Foulkes, P., & French, J.P. (2019). Analysing breathy voice in forensic speaker comparison Using acoustics to confirm perception. In *Proceedings of the 18th International Congress of Phonetic Sciences,* Melbourne, 795-799.

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

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

Inter-speaker variability in the American English /æ/ and /ɑ/: a dynamic view from both tongue articulation and the first two formants

Carolina Lins Machado[1] *and Lei He*[12]
[1]*Department of Computational Linguistics, Zurich University, Zurich, Switzerland*
cmachado@ifi.uzh.ch
[2]*Department of Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland.*
lei.he@uzh.ch

Formant dynamics is believed to be a rich source of individual information, reflecting the characteristic articulatory behavior of a speaker (Yang et al., 1996; McDougall, 2006). Primarily, formants carry information about vowel-phoneme identity in a property denominated "vowel-inherent spectral change", or VISC (Nearey & Assmann, 1986). The degree of VISC varies between the American English vowels /æ/ and /ɑ/. While for /æ/ formant changes pertain to vowel identity, for /ɑ/ less intrinsic spectral change is required due to its relatively stable acoustic goal (Stevens, 1989). Therefore, /ɑ/ may be produced with more variability in the underlying articulatory strategy allowing the speaker to reduce the need for precision in articulatory movements (Perkell et al., 1997). Consequently, this would result in more speaker-dependent information in /ɑ/ VISC, since the linguistic constraints imposing articulatory control may be less strong in this vowels allowing room for speakers' preferred articulatory strategies. Results from Lins Machado et al. (submitted) indicated that when speakers produced the vowel /ɑ/, there was more variability in the articulatory strategies employed than in the production of /æ/. Furthermore, the results suggested that speakers with similar VISC contours also seemed to have similar kinematic profiles. Although this point analysis was a first step into understanding individual articulatory behavior and its acoustic outcome, it did not account for actual formant dynamics, entailing values of direction and slopes. Thus, the present study builds on previous work and explores inter-speaker variability at a higher level of dynamics, namely in the relationship between articulatory velocities and rate of formant change.

The first two formants and the x, y coordinates of the tongue blade and dorsum of the vowels /æ/ and /ɑ/ in isolated monosyllabic words of twenty native speakers of U.S. English (10 F 10 M) were selected from the EMA-MAE corpus (Ji et al., 2014). These acoustic and articulatory measures were taken between vowel onset and offset in Praat (Boersma & Weenink, 2021). Articulatory velocities and rate of formant change were calculated from four intervals with equal duration within each vowel. Speaker-specific articulatory behaviors and its relationships to F1 and F2 outcomes are visualized as network graphs (Newman, 2018) similar to the ones in Figure 1. Dynamic networks contain nodes as the acoustic and articulatory variables and edges representing the partial correlations between them. Similarity between speaker networks was calculated using cosine similarity.

Preliminary results demonstrate that edges between acoustic and articulatory variables are not as salient in this analysis as compared to the previous point analysis. Nonetheless, dynamic networks still displayed less between-speaker variability in articulatory behaviors in the production of /æ/, suggesting that articulatory velocity may also be constrained by linguistic information related to the dynamic characteristics of this vowel.

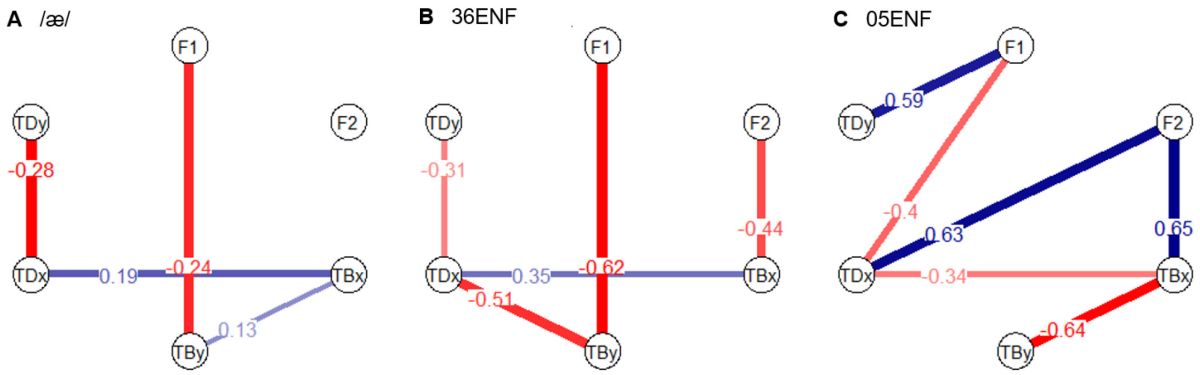


Figure 1. Networks displaying articulatory behaviors for the production of /æ/ at the third interval. Graphs indicate speakers with most similar (B) and most dissimilar (C) articulatory behaviors to the model network (A) , i.e. a vowel network built across all speakers capturing the most common articulatory strategies.

Acknowledgement

This work was supported by the Swiss National Science Foundation; Grant number PZ00P1_193328 to LH.

References

Boersma, P. and Weenink, D. (2021). Praat. Doing phonetics by computer [Computer program]. Version 6.2.04.

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

Lins Machado, C., Dellwo, V., & He, L. (submitted). Idiosyncratic lingual articulation of American English /æ/ and /ɑ/ using network analysis. *Interspeech 2022*.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *International Journal of Speech, Language and the Law, 13*(1), 89–126.

Nearey, T. M., & Assmann, P. F. (1986). Modeling the role of inherent spectral change in vowel identification. *The Journal of the Acoustical Society of America, 80*(5), 1297–1308.

Newman, M. (2018). *Networks* (Vol. 1). Oxford University Press.

Perkell, J., Matthies, M., Lane, H., Guenther, F., Wilhelms-Tricarico, R., Wozniak, J., & Guiod, P. (1997). Speech motor control: Acoustic goals, saturation effects, auditory feedback and internal models. *Speech Communication, 22*(2–3), 227–250.

Stevens, K. N. (1989). On the quantal nature of speech. *Journal of Phonetics, 17*(1–2), 3–45.

Yang, X., Millar, J. B., & Macleod, I. (1996). On the sources of inter- and intra- speaker variability in the acoustic dynamics of speech. *Proceeding of Fourth International Conference on Spoken Language Processing*. ICSLP '96, 3, 1792–1795 vol.3.

Salient cues to age identity in an LX: a longitudinal pilot study on a female L1 Hungarian speaking English

Sarah Melker[1]
[1]*Institute for English Studies, Karl-Franzens-Universität, Graz, Austria*
sarah.melker@uni-graz.at

In the past two decades much fruitful work has been done to unravel the complexities of the biological ageing process on the human voice (Harrington, Palethorpe, & Watson, 2007; Reubold, Harrington, & Kleber, 2010), including in the field of forensic phonetics (Künzel, 2007; Rhodes, 2017). At the same time, these findings have raised further questions about the universality of age-related changes in speech production and its perception, as well as discrepancies between individuals in the ageing process imputable to health or social network reasons. Hejná and Jespersen (2021) have recently drawn attention to this gap along with potential methods to remedy them by measuring both physiological and psychological factors.

A previously unexplored angle which may give insight into speaker design in choosing an age-identity is that of proficient LX adult immigrants. Studies on large immigrant groups have shown that LX speakers actively choose degrees of belonging based on indexical features (Kozminska, 2021). This study investigates: (1) how the ageing process affects LX speech; (2) how an LX's speech may incorporate features that are salient to her; (3) the extent to which she displays features identifying her with an age group when not having benefitted from as much exposure to different categories as compared to a typical L1.

For this study a set of televised interviews with the celebrity Zsa Zsa Gabor (born 1917 in Hungary; emigrated to the USA in 1941) was analysed and compared to similar recordings for two other female L1 English celebrities of about the same age (Lucille Ball, Ginger Rogers) as well as to her sisters, Eva and Magda. Hungarian was chosen due to its dissimilarity from English for several prosodic parameters. Moreover, Rácz and Papp (2016) report unexpected indexicalisation of pitch in Hungarian male speech, results which diverge from previous findings in other languages, which may inform findings in the present study. Female speakers were chosen for the comparison material because as listeners they have been found to be more accurate in age judgements (Kelly & Harte, 2015). If this is due to heightened awareness of salient ageing cues, then it might be expected that a female LX could also be attuned to relevant features and may attempt to match these in producing her own speech. Celebrities were chosen due to the ease of accessing material as well as the factors that pressure associated with their public profile might have on portraying their age identity.

This study follows previous studies in measuring fundamental frequency and vowel formants to track change over time, as well rhythm metrics (Pellegrino, 2019) which have shown varying results as an indicator of ageing, and are useful in gauging LX proficiency and transfer (White & Mattys, 2007). Results for these will be compared to both the L1 and sister recordings. A later stage of the project will involve perception testing among L1 English and L1 Hungarian listeners to determine which potentially age-identity related features are salient to different groups.

References

Harrington, J., Palethorpe, S., & Watson, C. I. (2007). Age-related changes in fundamental frequency and formants: a longitudinal study of four speakers. *Interspeech*, 2753-2756.

Hejná, M., & Jespersen, A. (2021). The coming of age: How do linguists tease apart chronological, biological and social age?. *Language and Linguistics Compass*, 15(1), e12404.

Kelly, F., & Harte, N. (2015). Forensic comparison of ageing voices from automatic and auditory perspectives. *International Journal of Speech, Language & the Law*, *22*(2), 167-202.

Kozminska, K. (2021). Scaling diasporic soundings in the globalised world: A study of Polish stops in the UK. Language & Communication, 77, 17-34.

Künzel, H. J. (2007). Non-contemporary speech samples: auditory detectability of an 11-year delay and its effects on automatic speaker identification. *International Journal of Speech, Language and the Law*, 14 (1), 109-136.

Pellegrino, E. (2019). The effect of healthy aging on within-speaker rhythmic variability: A case study on Noam Chomsky. *Loquens*, *6*(1), e060.

Rácz, P., & Papp, V. (2016). Percepts of Hungarian pitch-shifted male speech In E. Levon & R. B. Mendes (Eds.), *Language, Sexuality, and Power: Studies in Intersectional Sociolinguistics* (pp. 151-167). Oxford University Press.

Reubold, U., Harrington, J., & Kleber, F. (2010). Vocal aging effects on F0 and the first formant: a longitudinal analysis in adult speakers. *Speech Communication*, 52 (7-8), 638-651.

Rhodes, R. (2017). Aging effects on voice features used in forensic speaker comparison. *International Journal of Speech, Language & the Law*, 24(2), 177-199.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4), 501-522.

Prague, July 10–13, 2022

Work-in-Progress, Student paper

When singing becomes illegal

Sophie Möller & Gea de Jong-Lendle
Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg, Germany
{moeller9| dejong}@students|staff.uni-marburg.de

Introduction

Singing is not normally associated with an illegal act. However, recently a case was reported in which it was claimed that a song, that in Germany and Austria is considered illegal (§ 86a StGB), had been produced by a particular politician, well-known for his nationalist ideas. Supposedly, the song was recorded and part of it was published on the internet, causing a scandal. It concerned the so-called "Horst-Wessel-Lied", originally a battle song of the SA ("Sturmabteilung", paramilitary wing of the Nazi Party). In 1929 Horst Wessel, an SA-member himself, had rewritten the lyrics of the well-known Königsberg-Lied in order to produce a version that glorified the Nazi-regime. Thanks to Goebbels, it became an integral part of the nazi-propaganda and even received national anthem status. In 1945 the song was banned by the Allied Forces and in 1986 a higher regional court ruled that even singing the melody was illegal (Broderick 1995).

Ideally, the disputed and the reference recordings both contain material produced in a similar speaking mode. Here, both phonation modes were very different, exceeding the degree of mismatch normally accepted in casework. Although a positive identification was judged impossible in such a case, a negative identification may, under certain circumstances, still be possible: e.g. a speaker with a deep bass-voice probably sounds different when speaking or singing compared to a speaker/singer with a high tenor-voice.

While many studies exist on the vocal characteristics of speaking versus singing, studies on the comparison of both types of phonation from a forensic perspective seem rare. The aim of this study is to explore the discrimination ability of listeners confronted with singing-speaking comparison-stimuli.

Singing versus speaking

The phonatory mechanism of singing is different from speaking: In singing, the right posture, the correct breathing technique and the breath-support process are crucial (Nespital 2013, 49 and Friedrich et al. 2008, 30-33). The inhalation/exhalation duration ratio for singing can reach 1:50 (Hammer 2009, 20), whereas for speaking this ratio is approx.1:9 (Kreiman & Sidtis 2011, 30). The F0-range required for speaking is smaller and is found in the lower half of the total range (Wendler et al. 2005, 97). Spectral differences also exist: particular vowels may "suffer" in an effort to produce a resonant singing voice (Clermont 2002). In addition, spectral energy is concentrated in the 200-500Hz area of a speaking voice, slowly decreasing across higher frequencies. The proper singing voice may show another spectral peak, the so-called "singer´s formant" somewhere between 2000-3500Hz (Wendler et al. 2005).

Methodology

A pairwise design was selected in which listeners were tested in a discrimination test involving 3 conditions (speaking-speaking, speaking-singing, singing-singing). Speakers/singers consisted of 10 male and female speakers (18-35y) and were native speakers of German, who exhibited a standard variety of German. None of them were tone-deaf. Singing skills ranged from excellent to satisfactory.

The song “Muss I den zum Städtele hinaus” was selected, as the melody resembles that of the Horst-Wessel-Lied in terms of frequency-range, -pattern and songtype. In addition, most people who have grown up in Germany know this song and would be able to produce it.

References

Clermont, F. (2002). Systemic comparison of spoken and sung vowels in formant-frequency space. In C. Bow (Ed.), *Proceedings of 9th Australian International Conference on Speech Science & Technology* (p.124-129). Australian Speech Science and Technology Association.

Friedrich, G., Bigenzahn, W., & Zorowka P. (2008). *Phoniatrie und Pädaudiologie: Einführung in die medizinischen, psychologischen und linguistischen Grundlagen von Stimme, Sprache und Gehör* (4. Aufl.). Huber.

Geisler, M. E. (2005). In the Shadow of Exceptionalism: In M. E. Geisler (Ed.), *National Symbols, Fractured Identities: Contesting the National Narrative* (p. 71). UPNE.

Hammer, S.S (2009) Stimmtherapie mit Erwachsenen: Was Stimmtherapeuten wissen sollten (Praxiswissen Logopädie), 4. Edition. Springer.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies : an interdisciplinary approach to voice production and perception*. Wiley-Blackwell.

Miller, R. (1996). *On the art of singing*. Oxford University Press.

Nespital, U. (2013). Wirkungen des funktionellen Nachvollzugs physiologischer Gesangsstimmen auf die Qualität der Sprechstimme. *Hallesche Schriften zur Sprechwissenschaft und Phonetik*, 44.

Wendler, J., Eysholdt, U., & Seidner, W. (2005). *Lehrbuch der Phoniatrie und Pädaudiologie* (4. Aufl.). Thieme.

Automatic Speaker Recognition performance with (mis)matched bilingual speech material

Bryony Nuttall[1,2], *Philip Harrison*[2], *and Vincent Hughes*[2]
[1] *J P French Associates, York*
bryony.nuttall@jpfrench.com
[2] *Department of Language and Linguistic Science, University of York*
{philip.harrison|vincent.hughes}@york.ac.uk

To validate any forensic voice comparison system, it is necessary to test using samples that are reflective of the conditions of the case (Morrison et al. 2021). However, the extent to which certain speaker or technical factors affect system performance remains an empirical question (see Hansen and Hasan (2015) for an overview). This research, conducted as part of a MSc dissertation, contributes towards this area by considering the effect of language in automatic speaker recognition (ASR) systems used in forensic casework. Specifically, we examine the extent to which language mismatch either between the known and questioned samples, or between the evidential samples and the reference population (RP) used for calibration, affects overall system performance and the resulting strength of evidence (i.e., likelihood ratios for individual comparisons).

Testing was conducted using the state-of-the-art Phonexia Voice Inspector (v.4.0.0) x-vector system and speech samples from 88 Canadian English-French bilinguals from the Royal Canadian Mounted Police, Audio and Video Analysis Unit, Speech Research Database (AVAU_UO_data). There were three matched and mismatched language conditions that were examined across 16 different tests (see Table 1). The conditions were:

Condition 1 – Single language test data and different language RP data *Tests 1, 2, 5 & 6*
Single language RP data were compared with (mis)matched single language test data to test the effect of (mis)matched RP data in cases where a matched language reference database may be unavailable.

Condition 2 – Mixed language test data *Sets C and D*
Mixed language test data (where the known speaker sample is one language and questioned speaker sample is another language) were compared with single and mixed language RP data to assess ASR performance with bilingual material.

Condition 3 - Mixed language test and RP data *Tests 3, 4, 7 & 8 and Sets C & D*
Mixed language RP data were compared with single and mixed language test data to assess the effects of a (mis)matched RP to determine which combinations of language yield the lowest and least severe errors. These results form a basis for drawing evidential conclusions on appropriate reference populations in bilingual casework.

System performance was evaluated using the log LR cost function (C_{llr}) as well as its two constituent parts (C_{llr}^{cal} - calibration loss; C_{llr}^{min} - discrimination loss).

Results indicate that mixed language test comparison sets (C & D) pose a greater challenge to ASR systems than single language test sets (A & B), showing that the system's suitability for bilingual data still requires attention. More severe miscalibration was found in mixed language test and reference data (C & D) which makes drawing evidential conclusions based on this bilingual data challenging. Nonetheless, there are predictable patterns of directional shifts in

log LRs which are consistent with previous research. When combined with further empirical research, these shifts could provide a foundation on which to base expected calibration errors in real casework.

Table 1. Overall system performance across all tests. Languages are English (En) and French (Fr).

Test	Test set	Test language(s)		RP language(s)		Test & RP language match	Single or mixed language RP match	Cllr	$Cllr^{min}$	$Cllr^{cal}$
		KS	QS	KS	QS					
1	A	En	En	En	En	Match	Match	0.0016	0	0.0016
2				Fr	Fr	Mismatch	Match	0.0016	0	0.0016
3				En	Fr	Partial match	Mismatch	0.0540	0	0.0540
4				Fr	En	Partial match	Mismatch	0.1152	0	0.1152
5	B	Fr	Fr	En	En	Mismatch	Match	0.0074	0	0.0074
6				Fr	Fr	Match	Match	0.0071	0	0.0071
7				En	Fr	Partial match	Mismatch	0.2206	0	0.2206
8				Fr	En	Partial match	Mismatch	0.4066	0	0.4066
9	C	En	Fr	En	En	Partial match	Mismatch	6.28E-04	0	6.28E-04
10				Fr	Fr	Partial match	Mismatch	0.0023	0	0.0023
11				En	Fr	Match	Match	0.0738	0	0.0738
12				Fr	En	Partial match	Match	0.1487	0	0.1487
13	D	Fr	En	En	En	Partial match	Mismatch	2.2557	0.034	2.2217
14				Fr	Fr	Partial match	Mismatch	1.2312	0.034	1.1973
15				En	Fr	Partial match	Match	0.1103	0.034	0.0764
16				Fr	En	Match	Match	0.0731	0.034	0.0392

References

Hansen, J. H., and Hasan, T. (2015) Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Process. Mag.* 32(6): 74–99.

Morrison, G. S. et al. (2021) Consensus on validation of forensic voice comparison. *Science and Justice* 61(3): 299-309.

Prague, July 10–13, 2022
Work-in-Progress

Voice discrimination across speaking styles

Elisa Pellegrino[1], *Homa Asadi*[2], *and Volker Dellwo*[1]

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

elisa.pellegrino@uzh.ch

[2]*Department of Linguistics, University of Isfahan, Isfahan, Iran*

h.asadi@fgn.ui.ac.ir

Department of Computational Linguistics, University of Zurich, Zurich, Switzerland

volker.dellwo@uzh.ch

Human voices are individual but also extraordinary variable for the effect of numerous factors, such as speaking styles, background conditions, social contexts, physiologic states, etc. (Rose, 2002; Lavan et al., 2018). How human listeners can recognize individual speakers despite the enormous variability that individual voices reveal is far from fully being understood. This study reports ongoing research examining how voice recognition is affected by the variability introduced to voices through different speaking styles. We collected speech samples from 52 adult male Persian speakers in four different speaking styles that differ in the extent of observable within and between speaker acoustic variability: from the more homogenous read and clear speech to the more variable spontaneous and child-directed speech (henceforth CDS). Previous work has shown that situational voice alterations due to speaking styles can have an asymmetrical effect on automatic speaker recognition (Kathiresan et al. 2019). For example, learning a speaker under infant-directed speech (IDS) benefitted recognition in adult-directed conversational speech but not vice versa. To test how well human listeners can extract speaker-specific information from more or less variable vocalizations, we administered a voice discrimination test to 143 Persian listeners. The trials consisted of speech excerpts (approx. 2 sec. each) from the same and different speaking styles. Listeners' task was to decide whether the two voices in a trial come from the same or different speakers. To analyze listeners' performance, we calculated the bias-free sensitivity measure A' (henceforth aPrime) from signal detection theory (Grier, 1971). aPrime values range from 0.0 to 1.0, with 0.5 signifying chance level sensitivity and 1.0 indicating the highest sensitivity. aPrime sensitivity measure has been calculated per listener and speaking style. Data analysis is ongoing, but preliminary results based on 27 listeners showed that voice discrimination was carried out with a great deal of accuracy regardless of the speaking style (Fig. 1). The effect of speaking style on aPrime, tested with Mixed Effect Model (Speaking Style as fixed factor, aPrime as dependent variable, listeners as random factor) was indeed not significant [$\chi^2(3) = 6.7$, $p = 0.082$]. Further parameters will be examined to better understand the effect of speakers' acoustic variability on voice discrimination (e.g. listeners' bias towards responding same in less variable speaking styles, reaction times). Additionally, factors that in the current design (e.g. stimulus length, signal to noise ratio) might have masked a difference in performance across speaking styles will be also discussed.

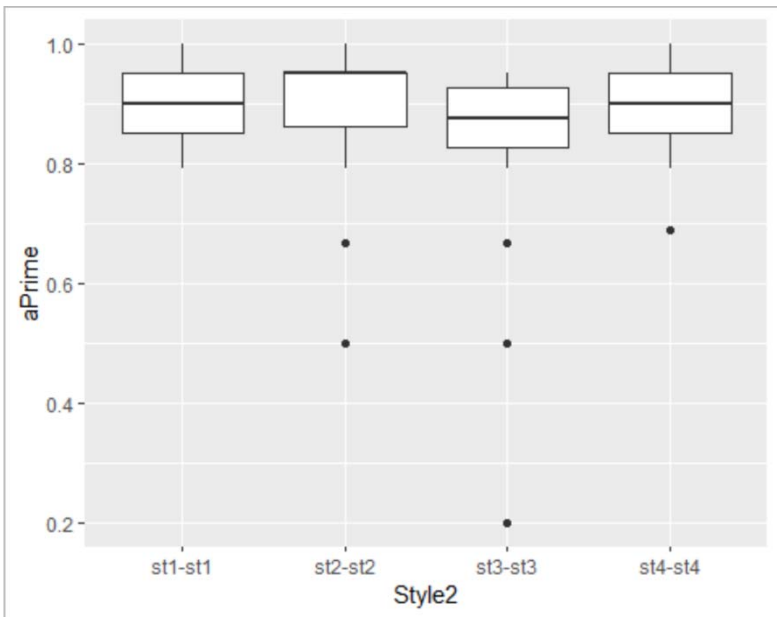


Figure 1. Boxplots showing median, range and inter-quartile range for aPrime by speaking style (st1 = read speech; st2=clear speech; st1= child-directed speech; st4= spontaneous speech).

References

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. Psychological Bulletin, 75(6), 424–429. https://doi.org/10.1037/h0031246.

Lavan, N., Burton, A., Scott, S.K. et al. (2019). Flexible voices: Identity perception from variable vocal signals. Psychon Bull Rev 26, 90–102. https://doi.org/10.3758/s13423-018-1497-7

Kathiresan, T., Dilley, L., Townsend, S., Shi, R., Daum, M., Arjmandi, M. & Dellwo, V. (2019). Infant-directed speech enhances recognizability of individual mothers' voices. Journal of the Acoustical Society of America, 145(3), 1766.

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

Prague, July 10–13, 2022
Student Paper

Voice Memory as an Estimator Variable in Lay Speaker Identification Tasks

Sascha Schäfer, and Paul Foulkes
Department of Language and Linguistic Science, University of York, York, UK
sascha.schaefer|paul.foulkes@york.ac.uk

Eliciting more reliable testimony from earwitnesses has been a long-standing endeavour in the forensic speech science community. Most recent efforts to do so have focused on the improvement of a particular procedure, the voice parade (VP), by finding optimal settings for the variables that can be controlled by the investigator ("system variables"), such as the quality (McDougall, 2021; Smith et al., 2019) and presentation (Smith et al., 2020) of the stimuli.

While optimising system variables may help establish credibility in the procedure, it does not necessarily establish credibility in the individual witness, whose general ability to identify voices might be questioned in court (Robson, 2017). The present experiment addresses this problem by exploring inter-listener differences in voice memorisation, which are beyond the control of the investigator ("estimator variables").

Hypotheses

A previous experiment (Schäfer & Foulkes, 2022) assessed the immediate voice recognition skills of listeners, i.e. excluding memory processes. Results showed that participants (n = 100, mean age = 36, SD = 13.8) differed markedly in their recognition accuracy (range 50 – 93.8%, mean = 75%, SD = 9.1%). The index *d prime* (d') also revealed high differences in listener discriminability (range 0 – 2.94, mean = 1.38, SD = 0.57).

Psychological studies showed even greater inter-listener differences when memory processes are involved. The *Glasgow Voice Memory Test* (*GVMT*, Aglieri et al., 2017), for instance, exhibited an accuracy range between 37.5 and 100% (mean = 78.15%, SD = 10.95%) and d' scores ranging from -0.67 to 3.07 (mean = 1.66, SD = 0.69). Whether the results of the *GVMT* apply to earwitnesses is unclear, however, as the stimuli were created from isolated vowels, rather than naturalistic speech. The present study therefore employs naturalistic stimuli in a voice memory task.

It is hypothesised that (1) a voice memory test based on naturalistic stimuli will produce a greater range of performances than the voice recognition test by Schäfer & Foulkes (2022), which excluded memory processes. Moreover, it is hypothesised that (2) individual performances in the present experiment only weakly correlate with the performances in the previous voice recognition test (Schäfer & Foulkes, 2022), as psychological studies suggest a rather weak correlation between voice recognition and voice memory (Mühl et al., 2018).

Methodology

100 British participants were invited to take part in an experiment hosted on Pavlovia. The stimuli were taken from a subset of 32 speakers of task 1 of the DyVis corpus (Nolan et al., 2009). The quality of these stimuli is comparable to stimuli used in a VP. Two 10s-long extracts were taken from each speaker, one for the first exposure to the voice and one for subsequent identification. Two stimulus lists of comparable difficulty were created based on the f0 difference between speakers.

In a memorisation phase, participants were presented with 8 voices and asked to memorise them to the best of their abilities. Each voice was replayed 3 times. In the subsequent identification phase, 16 voices (8 new) were presented, and participants provided an old/new judgement (reaction times were also measured). To answer hypothesis (2), 30 participants who took part in the previous voice recognition test (Schäfer & Foulkes, 2022) were reinvited for the voice memory test. The results, currently under analysis, may help assess the suitability of individual witnesses for the standardised VP procedure.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. Behavior Research Methods, 49(1), 97–110. https://doi.org/10.3758/s13428-015-0689-6

McDougall, K. (2021). Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence. Proceedings of XVII AISV. https://www.phonetics.mmll.cam.ac.uk/ivip/

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. G. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. Behavior Research Methods, 50(6), 2184–2192. https://doi.org/10.3758/s13428-017-0985-4

Nolan, F., McDougall, K., de Jong, G., & Hudson, T. (2009). The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. International Journal of Speech Language and the Law, 16(1), 31–57. https://doi.org/10.1558/ijsll.v16i1.31

Robson, J. (2017). A Fair Hearing? The Use of Voice Identification Parades in Criminal Investigations in England and Wales. Criminal Law Review, 1, 36–50.

Schäfer, S. & Foulkes, P. (2022, April 4-8). *Assessing the individual voice recognition skills of earwitnesses* [Poster presentation]. British Association of Academic Phoneticians, York. https://sites.google.com/york.ac.uk/baap2022york/home?authuser=0

Smith, H. M. J., Baguley, T. S., Robson, J., Dunn, A. K., & Stacey, P. C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. Applied Cognitive Psychology, 33(2), 272–287. https://doi.org/10.1002/acp.3478

Smith, H. M. J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2020). Voice parade procedures: optimising witness performance. Memory, 28(1), 2–17. https://doi.org/10.1080/09658211.2019.1673427

Unmasking identity through acoustic analysis: A case study of Indian English

Ravina Toppo[1], *and Sweta Sinha*[2]
[1,2]*Department of Humanities and Social Sciences, Indian Institute of Technology Patna, India*
[1]1821hs04@iitp.ac.in/ravina.toppo@gmail.com
[2]sweta@iitp.ac.in

Human speech is already known to exhibit characteristics that can reveal a speaker's identity, such as their places of origin and socialization, as well as their linguistic backgrounds, such as non-nativeness (Jessen, 2007; Watt, 2010). India is a multilingual country; therefore, Indian English (IE) has its own regional varieties which are the outcome of complex contact situations. Each of such IE varieties appears to be perceptually different for the listeners. Varieties of IE are influenced by the phonological and phonetic measures of native languages. This study aims to identify the unique characteristics of three Indo Aryan languages (Hindi, Bangla, and Odia) and study their influence on L2 IE vowels. The main objective is to establish vowels as the distinguishing features among the three IE varieties with three different L1. Acoustic phonetic properties of vowels can be employed to reveal dialectal differences by analyzing formant frequencies, duration, and pitch. The first two formant frequencies exhibit how vowels are dispersed in vowel space and how their formation varies as a function of dialect.

Overview of the most important differences: Hindi vs Bangla vs Odia

Hindi has many so called "regional dialects" (Masica, 1993). Grierson (1966), divided Hindi into Eastern and Western Hindi. Data was collected from the speakers of Eastern Hindi. Vowel inventory of Hindi are /ə/, /a/, /i/, /i:/, /u/, /u:/, /e/, /æ/, / o/, / ɔ/ (Tiwari, 1966).

There are seven vowel inventories in Bangla (/i/, /e/, /æ/, /a/, /ɔ/, /o/, /u/) with non-contrastive length supplemented by quality differences in the short and long high vowels (Chatterji, 2002). Thus, a change in vowel length leads to no effect on the meaning of the words in which they occur.

There are six vowel phonemes on Odia (/i/, /e/, /a/, /ɔ/, /o/, /u/) (Mahapatra, 2002). Vowel length in Oriya is not considered phonemic, although there are several instances where the vowels are phonetically long and may also contrast with their short counterparts.

Method

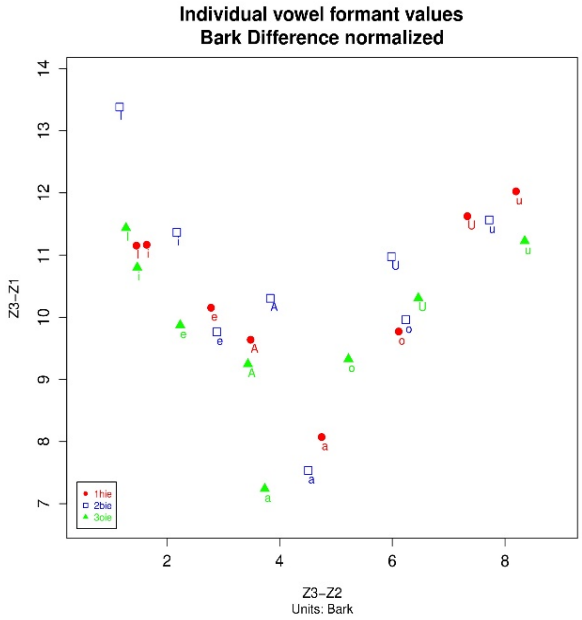
Speakers (N= 60) read a phonetically balanced passage in a natural speaking voice. The recording was done in an anechoic chamber. Gaps and mistakes were allowed between sentences. The continuous speech recording method was used as it may result in more natural production than having the speakers repeating single words. Vowels were then analyzed in PRAAT. It helped us access sound, waveforms, spectrograms, and transcription at the same time. In PRAAT, formant frequencies and fundamental frequencies were measured at the midpoints of the vowels using Linear Predictive Coding. Formant values, fundamental frequency, and duration for IE were measured and compared in speech samples of native speakers of Hindi, Bangla, and Odia.

Results and discussions

Based on the preliminary analysis, the most prominent features were

Shortening of long vowels in Hindi Indian English (HIE), /i/ and /i:/ have similar vowel height and frontness. Bangla Indian English (BIE) tends to produce both long and short high vowels, though, Bangla vowels are not marked for length. This very phenomenon might be possible because Bangla orthography has the provision of short and long vowel symbols. Odia Indian English (OIE) demonstrates shortening of long vowel /i/ vs /i:/, but tends to produce both long and short high vowel /u/ vs /u:/

Figure 1. The vowel space of HIE, BIE and OIE. (*note*: i = /i/, I = /i:/, e = /e/, A = /æ/, a = /a/, o = /ɔ/, u = /u/, and U = /u:/).



References

Chatterji, S. K. (2002). *The origin and development of the Bengali language*. New Delhi: Rupa.

Grierson, G. A. (Eds.). (1966). *Linguistic survey of India: Vol*. Motilal Banarsidass.

Jessen, M. (2007). Speaker Classification in Forensic Phonetics and Acoustics. In C. Mller (Eds.), *Speaker Classification I* (pp. 180-204). Springer Berlin Heidelberg.

Mahapatra, B. P. (2002). *Linguistics survey of India: Orissa* (Vol.1). Language division, Office of the Registrar General.

Masica, C. P. (1993). *The indo-aryan languages*. Cambridge University Press.

Tiwari, B. (1966). *Hindi bhasha*. Kitab Mahal.

Watt, D. (2010). The identification of the individual through speech. C. Llamas & D. Watts (eds) *Language and Identities* (pp. 76-85). Edinburgh: Edinburgh University Press.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Speaker discrimination and classification in breath noises by human listeners

Raphael Werner, Jürgen Trouvain, and Bernd Möbius
Language Science and Technology, Saarland University, Saarbrücken, Germany
{rwerner|trouvain|moebius}@lst.uni-saarland.de

Audible breath noises are frequent companions to speech, occurring roughly every 3 to 4 seconds (Rochet-Capellan & Fuchs, 2013; Kuhlmann & Iwarsson, 2021), and may also be present outside of speech during effortful actions (Trouvain & Truong, 2015). Being a vital function, breathing is arguably less affected by speakers trying to disguise their voice and neural networks have shown promising results on speaker identification based on breath noises (Lu et al, 2020; Zhao, Gao, & Singh, 2017). However, breathing has remained largely untapped for forensic purposes, with few exceptions (eg. Kienast & Glitza, 2003). In this paper we want to investigate the potential that breath noises have for speaker discrimination and classification by human listeners.

We annotated breath noises in dyadic conversations (van Son et al, 2008). For high comparability and since they are most frequent around speech (Lester & Hoit, 2014), we here use 5 audible oral (and probably simultaneously nasal) inhalations each from 6 younger (age range: 20–29; 3m, 3f) and 6 older (age range: 59–65; 3m, 3f) speakers. These noises were then used as stimuli in two tasks: 1) Discrimination task: participants heard 2 breath noises (separated by 500 ms of silence; 14 pairs by participant) and were asked whether they were produced by the same speaker or not. We also recorded participants' confidence on a 5-point Likert scale. 2) Speaker classification task: participants listened to one breath noise at a time (20 noises by participant) and were asked whether the breath noise was produced by a young vs old and male vs female speaker and how confident they were in each of these answers. We recruited and paid 33 speakers (22 f, 10 m, 1 other; age range: 20-71, median: 31), who reported wearing headphones in a quiet environment and having no hearing difficulties, via Prolific (2014) and ran the experiment on Labvanced (Finger et al, 2017).

Preliminary analysis suggests that the discrimination task was answered correctly at 64.3%. In speaker classification, the speaker's age group was correct at a rate of 50.2%, whereas for sex it was 66.7%. The general direction of sex being easier to guess than age here seems to follow the pattern described by Jessen (2007), even though not using speech here and speaker age being a binary decision between two groups. In the further analysis, we will examine what participant or speaker variables contribute to correctness in the discrimination task, as well as look into participant performance by their age and gender.

The findings will have implications for naturalistic synthetic speech and how breath noises there need to be geared to the artificial speaker to be perceived as natural. For forensic purposes, they explore to what extent breath noises may be exploitable for speaker classification and discrimination tasks. It should be borne in mind, however, that all stimuli used here were made under the same recording setup and are thus highly comparable, whereas in real-world forensic applications many factors may complicate comparisons.

References

Finger, H., C. Goeke, D. Diekamp, K. Standvoss, and P. König (2017). LabVanced: A Unified JavaScript Framework for Online Studies. In International Conference on Computational Social Science, 2016–2018.

Jessen, M. (2007). Speaker Classification in Forensic Phonetics and Acoustics. In: Müller, C. (eds) Speaker Classification I. Lecture Notes in Computer Science, vol 4343. Springer, Berlin, Heidelberg.

Kienast, M., & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. ICPhS, 1607–1610.

Kuhlmann, L. L., & Iwarsson, J. (2021). Effects of Speaking Rate on Breathing and Voice Behavior. Journal of Voice.

Lester, R. A., & Hoit, J. D. (2014). Nasal and oral inspiration during natural speech breathing. Journal of Speech, Language, and Hearing Research, 57(3), 734–742.

Lu, L., Liu, L., Hussain, M. J., & Liu, Y. (2020). I Sense You by Breath: Speaker Recognition via Breath Biometrics. IEEE Transactions on Dependable and Secure Computing, 17(2), 306–319.

Prolific. (2014). URL https://www.prolific.co. Accessed: 17/05/2022.

Rochet-Capellan, A., & Fuchs, S. (2013). The interplay of linguistic structure and breathing in German spontaneous speech. Interspeech, 2014–2018.

Trouvain, J., & Truong, K. P. (2015). Prosodic characteristics of read speech before and after treadmill running. Interspeech, 3700–3704.

van Son, R., Wesseling, W., Sanders, E., & van den Heuvel, H. (2008). The IFADV Corpus: a Free Dialog Video Corpus. LREC. 501–508.

Zhao, W., Gao, Y., & Singh, R. (2017). Speaker identification from the sound of the human breath. *arXiv preprint arXiv:1712.00171*.

Acoustic variation within Persian-English bilingual speakers

Homa Asadi[1]*, Maral Asiaee*[2]*, and Volker Dellwo*[3]

[1]*Department of Linguistics, University of Isfahan, Isfahan, Iran*
h.asadi@fgn.ui.ac.ir

[2]*Department of Linguistics, Alzahra University, Tehran, Iran*
m.asiaee@alzahra.ac.ir

[3]*Department of Computational Linguistics, University of Zurich, Zurich, Switzerland*
volker.dellwo@uxh.ch

An important part of human social interaction is the ability to hear and identify voices on a daily basis. Our voice not only conveys information about the message being spoken but also provides clues about the identity and emotional attributes of an individual. There are many individuals around the world who are speaking in two or more than two languages. This phenomenon adds an intriguing dimension of variability to the speech, both in perception and production. But do bilinguals change their voice while switching from one language to another? From the speech production perspective, it is suggested that while some aspects of speech signal vary due to linguistic reasons, some indexical features remain intact across different languages (Johnson et al., 2020). The situation can become much more complicated when bilinguals speak in different speaking styles. Nevertheless, little is known about the influence of language and speaking style on within- and between-speaker vocal variability.

Here, we investigated how acoustic parameters of voice quality vary across different languages and speaking styles of Persian-English bilingual speakers and to what extent such features can discriminate between bilingual speakers. For this purpose, a total of 20 native speakers of Persian, 10 males and 10 females, who were fluent speakers of English were recorded on two different sessions (mean age: 27.6, sd: 3.1, range:24-37; device: ZOOM H4n; sr: 44100, 16bit, sound treated recording environment). Two speaking styles, i.e. read and spontaneous speech were recorded. All voiced segments (including vowels and consonants) were extracted using Vocal Toolkit (Corretge, 2022) in Praat (Boersma & Weenink, 2022). The acoustic parameters were selected based on the psychoacoustic model of voice quality proposed by Kreiman et al. (2014): F0, F1, F2, F3, F4, Formant Dispersion (FD), H1*-H2*, H2*-H4*, H4*–H2kHz*, H2kHz*–H5kHz, Cepstral peak prominence (CPP), Energy, and subharmonics-harmonics ratio (SHR) were measured via the VoiceSauce (Shue et al., 2009) using 5 ms intervals.

A linear mixed-effects model showed that the interaction between language and style on the voice quality features of Persian-English bilingual speakers was significant ($p<0.0001$). In order to predict the ranking of the most important parameter, we ran a random forest to classify speakers based on language, style and voice quality features using R package *randomForest*. Based on the results, language is more important in speaker classification compared to style. For male speakers, CPP, Energy, F0 and F1 contributed most to between-speaker variability, while Energy, F0, F1 and F3 were the most important acoustic parameters in showing variation across female speakers. Despite the variability of voice quality features in different languages and speaking styles, our results showed that they still have the potential to classify speakers based on their voice.

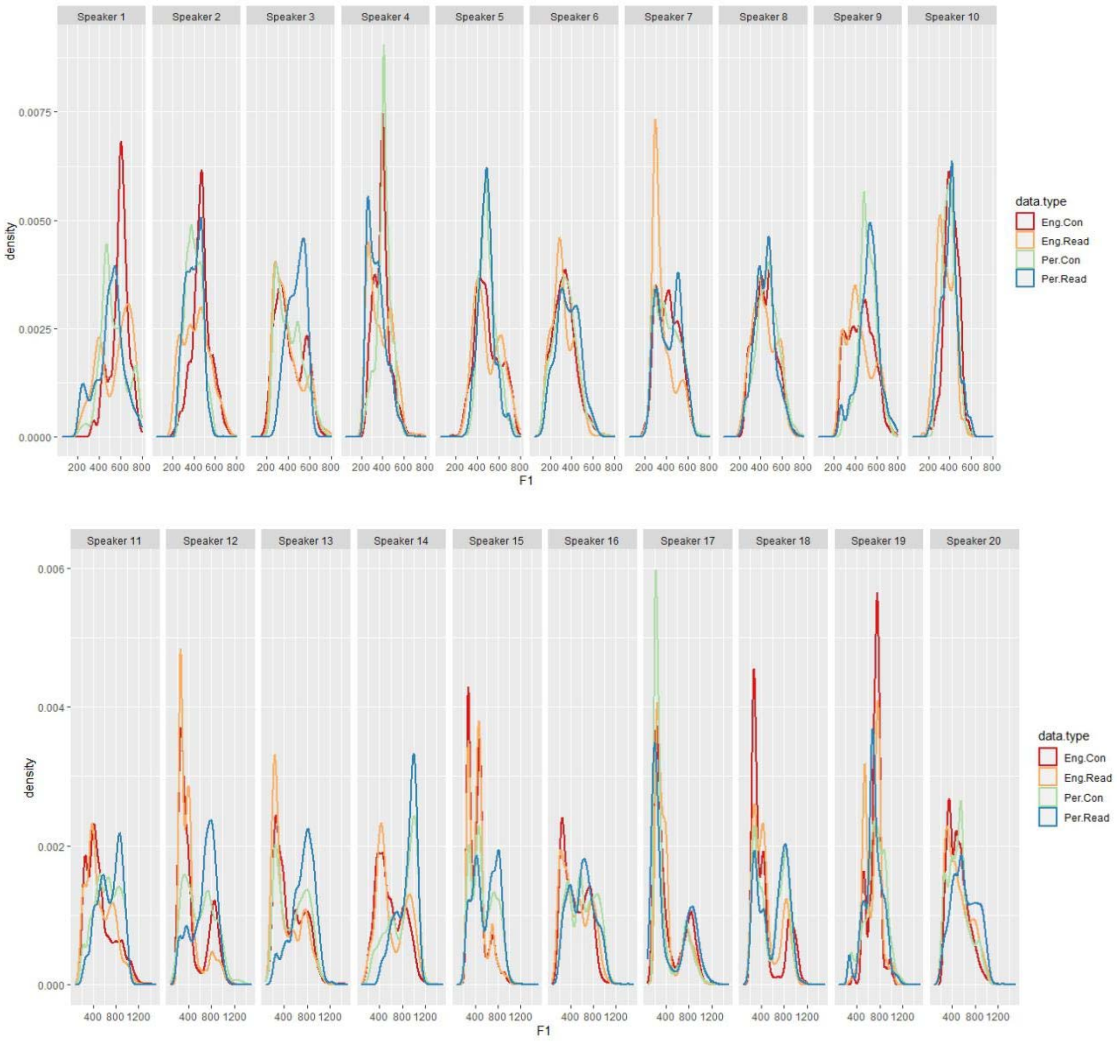


Figure 1: Density plot showing the distribution of F1 across each language and speaking styles within and between male (top) and female (bottom) bilingual Persian-English speakers.

References

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.1.39). http://www.praat.org/.

Corretge, R. (2022). *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com.

Johnson, K. A., Babel, M., & Fuhrman, R. A. (2020). Bilingual acoustic voice variation is similarly structured across languages. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2020-Octob*, 2387–2391.

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1), e009.

Shue, Y.-L., Keating, P., & Vicenik, C. (2009). VOICESAUCE: A program for voice analysis. *The Journal of the Acoustical Society of America*, *126*(4), 2221.

Prague, July 10–13, 2022
Student Paper

Language-dependency of /s/ in L1 Dutch and L2 English

Meike de Boer and Willemijn Heeren
Leiden University Centre for Linguistics, Leiden, The Netherlands
{m.m.de.boer|w.f.l.heeren}@hum.leidenuniv.nl

Forensic casework increasingly often involves speech samples in more than one language [1], showing a need to explore language-independent characteristics within speakers. However, knowledge on such characteristics is limited, and at this point the general consensus is merely to 'exercise particular caution with cross-language comparisons' [2]. As part of a bigger study, the current research investigates cross-linguistic within-speaker consistency of /s/ among speakers with Dutch as a first language (L1) and English as a second language (L2). Research on speaker-dependency in consonants shows that /s/ contains speaker-specific information [5, 6]. The voiceless alveolar fricative /s/ is phonetically similar but not identical in the Dutch and English language. According to the Speech Learning Model [3], this increases the chance that L2 speakers fail to realize that the Dutch and English /s/ have phonetic differences and that they use their Dutch /s/ also when speaking English. Such L1 transfer would be helpful in forensic phonetics, as it would allow for the inclusion of /s/ as a feature in cross-linguistic comparisons.

According to [4], who looked into read speech /s/ by L1 Dutch speakers with a relatively high proficiency of L2 English, speakers use different /s/ realizations in the L1 and L2. This implies that /s/ is not useful as a feature in cross-linguistic speaker comparisons. However, these same speakers have also been recorded producing spontaneous speech, which may be more representative for forensic casework data and may evoke less formal language use. Hence, this study investigates the language-dependency of /s/ in spontaneous speech.

Method. Using a sample from the same speakers as [4], this study investigates the language-dependency of /s/ in 2-minute spontaneous monologues in L1 Dutch and L2 English (N = 52; n = 4,904). Linear mixed-effects models were built for the spectral Centre of Gravity (CoG, 550–8000 Hz), its standard deviation (SD), skewness, kurtosis, and the spectral tilt of /s/,[1] testing the fixed factor Language (levels: Dutch, English), random by-speaker slopes, and random slopes of Language over Speaker.

Results. For all models but SD, the optimal model included the fixed factor Language and random slopes for Language over Speaker. Table 1 shows the intercepts and adaptations when the speakers spoke L2 English. For example, the CoG was on average 783 Hz higher in L2 English than in L1 Dutch (intercept: 5,007 Hz). Although the effect of Language was speaker-dependent, for all speakers, the English CoG was on average higher than the Dutch one (see Fig. 1). For SD, again, random slopes for Language were included in the optimal model. However, there was no overall Language effect.

	Intercept	*SE*	*t*	*Language*	*SE*	*t*
CoG	5,007	99	50.83	783	78	10.09
SD	1,307	29	45.08	-	-	-
Skewness	0.90	0.05	17.87	0.34	0.07	5.02
Kurtosis	4.91	0.48	10.30	1.06	0.40	2.63
Spectral tilt	12.03	0.73	16.47	3.33	0.54	6.21

Table 1. Overview of the fixed parts of the optimal models.

[1] Note that these features were all correlated except for skewness and kurtosis.

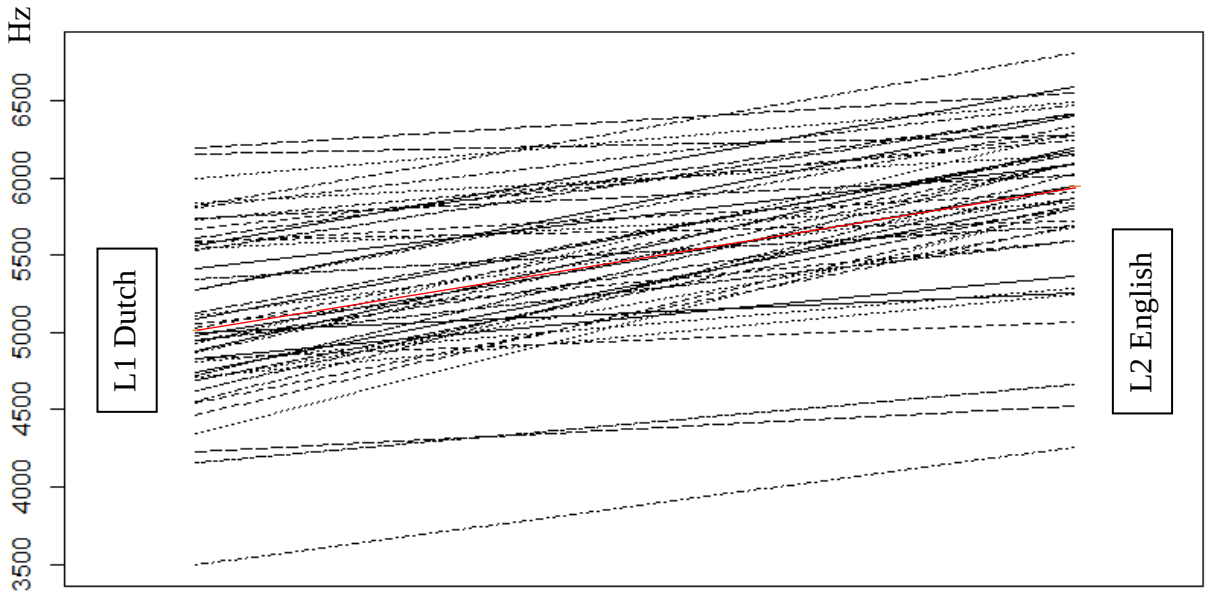


Fig. 1. By-speaker Means of the Centre of Gravity of /s/ in L1 Dutch (left) and L2 English (right); lines connect Means belonging to the same speaker. The grand mean is provided in red.

A follow-up analysis includes left and right phonetic context as fixed factors (levels: rounded/labial, unrounded/non-labial, cf. [5]) and will be presented at the conference. Preliminary results show that the language effect remains; the differences between L1 Dutch and L2 English cannot be attributed to context effects.

Discussion. We found that the Centre of Gravity and Spectral Tilt of /s/ are language-dependent within speakers. In addition, the language effect may vary with speaker, which makes the effect unpredictable. Hence, based on these findings, /s/ does not seem to be a suitable feature to be used in cross-linguistic forensic speaker comparisons, at least for highly proficient L2 speakers. A follow-up study using a likelihood ratio approach will investigate this matter in more detail.

References

[1] Van der Vloed, D.L., Bouten, J.S., & Van Leeuwen, D.A. (2014). NFI-FRITS: A forensic speaker recognition database and some first experiments. *Proceedings of Odyssey Speaker and Language Recognition Workshop 2014*, Joensuu, Finland, June 16-19, 2014, 6-13.
[2] IAFPA (2020). Code of Practice. https://www.iafpa.net/the-association/code-of-practice/
[3] Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. *Speech perception and linguistic experience: Issues in cross-language research, 92*, 233-277.
[4] Smorenburg, L., & Heeren, W. (2020). The distribution of speaker information in Dutch fricatives /s/ and /x/ from telephone dialogues. *Journal of the Acoustical Society of America, 147*, 949-960.
[5] Kavanagh, C. M. (2012). *New consonantal acoustic parameters for forensic speaker comparison*. Doctoral dissertation, University of York.
[6] Quene, H., Orr, R., & van Leeuwen, D. (2017). Phonetic similarity of/s/in native and second language: Individual differences in learning curves. *Journal of the Acoustical Society of America, 142*(6), EL519-EL524.

Prague, July 10–13, 2022
Work-in-Progress

Evidential value of long-term laryngeal voice quality acoustics

Ricky K.W. Chan[1] *and Bruce Xiao Wang*[2]
[1]*Speech, Language and Cognition Laboratory, School of English, University of Hong Kong, HK*
rickykwc@hku.hk
[2]*Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, HK*
bruce.wang@alumni.york.ac.uk

Introduction. One of the main goals in forensic voice comparison (FVC) research is to identity speech features that are useful for distinguishing voices under forensically relevant conditions. Voice quality (VQ) was reported to be one of the most popular and useful features for FVC (e.g. Gold & French, 2011), but empirical studies that test this claim are surprisingly limited, especially for the acoustics aspects of VQ. This contribution focuses on the acoustics of laryngeal voice quality (aka phonation types), and tests how the use of non-contemporaneous recordings affect their evidential value under the likelihood-ratio framework.

Methods. 75 male speakers aged 18-45 were selected from a forensically-oriented database of 552 Australian English speakers (Morrison et al., 2015). These speakers were from Sydney/New South Wales and were recorded on more than one occasion, performing three speaking tasks each time. For each speaker, four recordings—the casual telephone conversation (CNV) and pseudo police interview (INT) tasks recorded in two separate sessions (by at least a one-week interval) (i.e. CNV1, CNV2, INT1, INT2)—were selected. Around 33 seconds of vocalic material per recording was analyzed. The VQ parameters reported in Hughes et al. (2019) (see Tables 1 and 2) were extracted using VoiceSauce (Shue et al., 2011) and served as input for score generation and LR computation. The fvclrr package (Lo, 2018) was used to implement the multivariate kernel-density (MVKD; Aitken & Lucy, 2004) formula for same-speaker and different-speaker comparisons. Calibrations were conducted using logistic regression. The 75 speakers were randomly assigned to training, test, or reference set (25 speakers in each set). The procedure above was replicated 100 times with different speakers in the training, test, and reference sets, as it has been demonstrated that the reliability of system performance hinges on the speaker samples involved (Wang et al., 2019). This contribution reports two comparisons: CNV1 vs. INT1 (contemporaneous recordings) and CNV1 vs. INT2 (non-contemporaneous recordings).

Results and discussion. Overall, all the input parameters yielded a small standard deviation in C_{llr} (less than 0.1) and EER (mostly less than 5%) values across the 100 replications, suggesting that system performance using these parameters as input were stable. Individual VQ parameters performed rather poorly, with mean C_{llr} values close to 1 and mean EER value mostly greater than 40%. This suggests that individual VQ parameters carry little speaker-discriminatory information. System performances improved considerably when combining all the spectral tilt parameters or all the additive noise parameters, but the results are still less promising than those reported in Hughes et al. (2019). Surprisingly, using all spectral tilt and additive noise parameters as input led to worse performance, suggesting that these two types of measures provide overlapping or conflicting information for distinguishing speakers. More comprehensive analysis, theoretical and forensic implications, and suggestions for future research will be presented in the conference.

Acknowledgements. This research has been generously supported by the Hong Kong Research Grant Council Early Career Scheme (Project no.: 21606918). We are grateful to Prof. Philip Rose for his invaluable feedback at the early stage of this research project.

CNV1 vs. INT1

	C_{llr}				EER			
VQ parameter	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
H1 - H2	0.97	1.09	1.00	0.02	36.00	64.42	48.26	4.55
H2 - H4	0.89	1.41	1.00	0.06	32.75	56.42	44.11	4.83
H1 - A1	0.99	1.03	1.00	0.00	40.08	60.00	49.30	3.45
H1 - A2	0.99	1.04	1.00	0.01	39.50	60.33	49.15	3.72

H1 - A3	0.92	1.11	0.98	0.03	32.83	52.00	42.51	3.73
Spectral tilt	0.91	1.04	0.96	0.02	34.67	51.42	41.11	3.61
CPP	0.93	1.12	0.98	0.03	32.00	52.00	42.59	3.87
HNR05	0.91	1.15	0.96	0.03	36.00	53.00	44.84	4.07
HNR15	0.88	1.13	0.97	0.05	28.25	48.75	41.32	4.27
HNR25	0.89	1.21	0.98	0.05	29.00	52.00	41.40	4.13
HNR35	0.92	1.35	0.98	0.07	32.00	48.42	40.89	3.29
Additive noise	0.84	1.05	0.92	0.04	28.00	47.83	36.77	4.25
Spectral tilt + additive noise	0.85	1.02	0.93	0.04	25.58	44.67	35.31	4.00

CNV1 vs. INT2

	C_{llr}				EER			
VQ parameter	**Min**	**Max**	**Mean**	**SD**	**Min**	**Max**	**Mean**	**SD**
H1 - H2	0.93	1.33	1.01	0.06	32.00	64.08	43.73	5.37
H2 - H4	0.97	1.07	1.00	0.02	37.00	59.08	48.49	3.66
H1 - A1	0.99	1.11	1.00	0.01	40.67	56.92	50.10	3.36
H1 - A2	0.95	1.08	0.99	0.02	40.00	56.00	49.27	3.36
H1 - A3	0.95	1.08	0.99	0.02	40.00	56.00	47.88	3.26
Spectral tilt	0.91	1.05	0.97	0.03	32.00	48.00	39.76	3.41
CPP	0.93	1.12	0.98	0.03	32.00	52.00	42.59	3.87
HNR05	0.91	1.15	0.96	0.03	36.00	53.00	44.84	4.07
HNR15	0.88	1.13	0.97	0.05	28.25	48.75	41.32	4.27
HNR25	0.89	1.21	0.98	0.05	29.00	52.00	41.40	4.13
HNR35	0.92	1.35	0.98	0.07	32.00	48.42	40.89	3.29
Additive noise	0.76	1.01	0.88	0.05	23.17	40.00	30.75	3.71
Spectral tilt + additive noise	0.87	1.02	0.93	0.03	27.33	44.08	35.84	3.61

Tables 1 and 2: statistics of Cllr and EER values across 100 replications with VQ parameters as input in CNV1 vs. INT1, and CNV1 vs. INT2 respectively. Spectral tilt: combination of H1-H2, H2-H4, H1-A1, H1-A2, H1-A3; Additive noise: combination of CPP and HNR05-35.

References

Aitken, C. G. 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. https://doi.org/10.1046/j.0035-9254.2003.05271.x

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

Hughes, V., Cardoso, A., Foulkes, P., French, J. P., Harrison, P. and Gully, A. (2019) Forensic voice comparison using long-term acoustic measures of voice quality. *Proceedings of the 19th International Congress of Phonetic Sciences (ICPhS)*. Melbourne, Australia. pp. 1455-1459.

Lo, J. (2018). FVClrr: likelihood ratio calculation and testing in forensic voice comparison (2.0.1) [Computer software]. https://github.com/justinjhlo/fvclrr.

Morrison G.S., Zhang C., Enzinger E., Ochoa F., Bleach D., Johnson M., Folkes B.K., De Souza S., Cummins N., Chow D., Szczekulska A. (2021). *Forensic database of voice recordings of 500+ Australian English speakers (AusEng 500+)*. [Available: http://databases.forensic-voice-comparison.net/]

Shue, Y.-L., P. Keating , C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Exploring covariation as a marker of speaker specificity

Lois Fairclough
Department of Linguistics and English Language, Lancaster University, UK
l.fairclough@lancaster.ac.uk

A fundamental hypothesis of forensic speech science is that speakers show idiosyncratic realisations of speech sounds. While many studies have documented the scope and nature of speaker-specific variability across a range of individual features, it is likely that speaker individuality may more concretely reside in the ways in which features co-occur. In sociophonetic research, this is often referred to as 'style' (Podesva, 2008), with a given feature differing in its social meanings depending on the other features that comprise that style. Co-variation also shows systematicity in phonological systems; for example, a speaker's production of stop consonants tend to be highly related to one another, even when across a corpus of speech these features exhibit considerable variability in VOT (Chodroff & Wilson, 2018). This suggests that analysing co-variation of phonetic features may reveal deep structure in both phonology and speaker-specificity. Accordingly, this study extends previous work in order to assess whether structured co-variation of phonetic features is a useful tool for speaker identification.

The key aim of my research is to assess whether speakers exhibit structured covariation in spontaneous speech, and whether listeners are sensitive to this. In this study, I will sketch out some foundations for this work, by analysing variability across vowels and their co-variation. Data will be taken from WYRED corpus (Gold et al., 2018), which contains forensically relevant data recordings from 120 West Yorkshire English speakers. A selection of five vowels; FLEECE, schwa, GOAT, FACE, and GOOSE are analysed from spontaneous speech in simulated police interview recordings. FLEECE and schwa are used as anchors due to their relative stability within the vowel system (Watt and Fabricus, 2002), while GOAT, FACE and GOOSE represent highly variable vowels that have been subject to variation and change. For instance, GOOSE fronting is a well-known phenomenon of British English, but the trajectory of change varies by dialect (Lawson et al., 2019). Importantly, these three vowels are highly variable, while also having regional and social associations in Yorkshire (Haddican et al., 2013).

The vowels will be analysed acoustically, extracting time-varying F1-F4 over the vowel's duration, which are then parametrised using GAMMs. As a first step, the vowels will be assessed in terms of their variability within and between speakers. Following this, I then examine the correlation between vowel variation across speakers, in order to test the hypothesis that socially meaningful vowel variation exhibits structured covariation.

In doing so, this work will examine how features vary within and between speakers, while also investigating how co-variation patterns manifest across speakers. Consequently, this research will aim to situate speaker individuality in terms of structured constellations of features and, in doing so, aims to unify models of speaker individuality across forensic phonetics and sociophonetics (Fairclough, forthcoming). The implications of this work include application to speaker identification tasks, such as how the presence of different combinations of features may influence analyst perceptions. This production-based study therefore provides foundation to assess perception of covarying features, which will be undertaken as part of my PhD research.

References

Chodroff, E., & Wilson, C. (2018). Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice. *Linguistics Vanguard*, *4*(s2).

Fairclough, L. (2022). Making waves: suggestions for methodological and conceptual collaboration between third wave sociophonetics, and forensic phonetic research and casework. *Modern Languages Open.* (forthcoming)

Gold, E., Ross, S., & Earnshaw, K. (2020). *WYRED - West Yorkshire Regional English Database 2016-2019.* [Data Collection]. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854354

Haddican, B., Foulkes, P., Hughes, V., & Richards, H. (2013). Interaction of social and linguistic constraints on two vowel changes in northern England. *Language Variation and Change*, *25*(3), 371-403.

Lawson, E., Stuart-Smith, J., & Rodger, L. (2019). A comparison of acoustic and articulatory parameters for the GOOSE vowel across British Isles Englishes. *The Journal of the Acoustical Society of America*, *146*(6), 4363-4381.

Podesva, R. J. (2008). Three sources of stylistic meaning. In *Texas Linguistic Forum*. *51*, 1-10.

Watt, D., & Fabricius, A. (2002). Evaluation of a technique for improving the mapping of multiple speakers' vowel spaces in the F1~ F2 plane. *Leeds working papers in linguistics and phonetics*, *9*(9), 159-173.

The Quest to Find Auditory 'Super-Recognizers'

Results from a pilot study

Andrea Fröhlich[1,2,3], *Volker Dellwo*[1], *Meike Ramon*[3]
[1]*Department of Computational Linguistics, University of Zürich, Switzerland*
[2]*Zurich Forensic Science Institute, Switzerland*
[3]*Applied Face Cognition Lab, University of Lausanne, Switzerland*

andrea.froehlich@uzh.ch
volker.dellwo@uzh.ch
meike.ramon@unil.ch

In 2009 Russel *et al.* introduced the term "Super-Recognizer" (SR), describing individuals with excellent unfamiliar face matching (discrimination), face memory (recognition) and identification abilities. Since then, general interest in this area has been rapidly increasing (Ramon, 2021), with law enforcement organizations interested in using SR's extraordinary capabilities for investigative case work. Inspired by this work on unique visual abilities, we are exploring whether SRs also exist in the auditory domain, specifically in the field of voice processing. Within the scope of this project, we developed an initial test to potentially identify auditory SRs.

Relevant Work

Until today, three studies have been published with the goal of finding people with potential super-recognition skills in voice processing. One of them is a discrimination test (Mühl *et al.* 2018) and two are recognition tests (Aglieri *et al.* 2017; Humble *et al.* 2021). In 2021, Jenkins *et al.* further investigated whether visual SRs also show exceptional abilities in voice processing using the Bangor Voice Matching Test, the Glasgow Voice Memory Test and a bespoke Famous Voice Recognition Test. However, these test designs cannot be directly compared to a casework scenario in forensics phonetics and their SRs were identified using suboptimal procedures.

Aim and problem statement

One of the great current challenges in forensic speaker recognition is the processing of large amounts of data with poor audio quality. Automatic voice comparison systems can be employed for tasks like voice clustering of speakers or to perform one-to-many speaker comparisons. However, if the audio recordings are of very poor quality, the performance of these systems decreases drastically. In such scenarios, auditory SRs could be of great value to pre-process the data based on an initial discrimination or recognition for later evaluation by forensic experts. To meet the specific requirements of forensic phonetic case work, a new test was developed to identify auditory SRs.

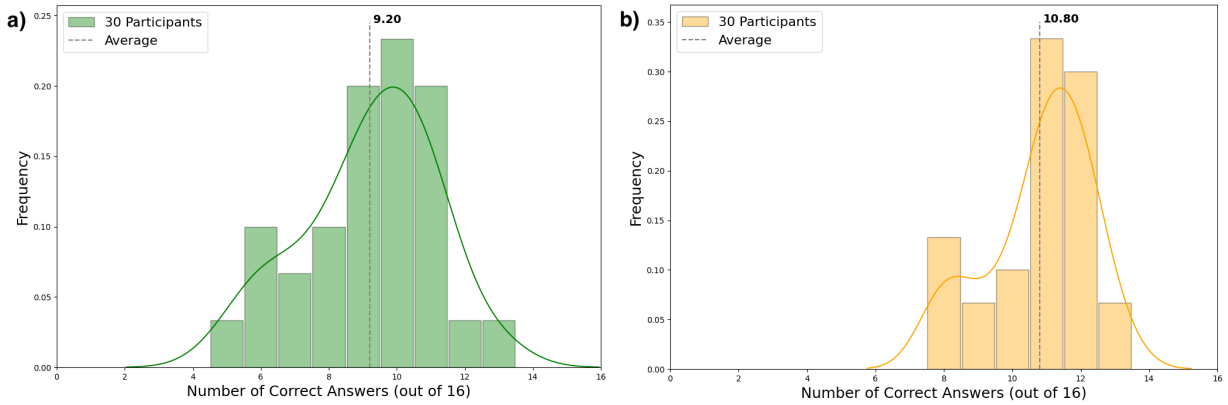
Method

A test with a discrimination and two recognition tasks was developed and optimized. The initial pilot study we report here was conducted with participants from different classes of police cadets

using stimuli from the TEVOID-Corpus (Dellwo *et al.* 2012). Our test comprises a discrimination test (inducing passive voice learning), followed by a surreptious recognition test (of the passively learned and novel voices), as well as an active voice learning phase followed by a recognition test, which was completed using a within-subjects design.

Results

Figure 1. Recognition results from the passive or implicit (a) and active or explicit (b) learning task.



Conclusion

In this preliminary study, we observed a difference in neurotypical participants' behavior for voice recognition. Specifically, explicit voice learning for later recognition was associated with higher performance as compared to that observed after implicit encoding. These initial findings are critical for informing the development of tools to identify auditory SRs. It is conceivable that auditory SRs' learning of voice identity is independent of instruction, inline with findings in visual SRs (Nador *et al.* 2021). Further investigations involving critical improvements to test this hypothesis are underway.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior research methods*, *49*(1), 97-110.

Dellwo, V., Leemann, A., & Kolly, M. J. (2012, September). Speaker idiosyncratic rhythmic features in the speech signal. Interspeech Conference Proceedings.

Humble, D., Schweinberger, S. R., Mayer, A., Dobel, C., & Zäske, R. (2021). The Jena Voice Learning and Memory Test (JVLMT): A standardized tool for assessing the ability to learn and recognize voices.

Jenkins, R. E., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2021). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *Applied Cognitive Psychology*, *35*(3), 590-605.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior research methods*, *50*(6), 2184-2192.

Nador, J. D., Alsheimer, T. A., Gay, A., & Ramon, M. (2021). Image or Identity? Only Super-recognizers' (Memor)Ability is Consistently Viewpoint-Invariant. Swiss Psychology Open: The Official Journal of the Swiss Psychological Society, 1(1), 2. DOI: http://doi.org/10.5334/spo.28

Ramon, M. (2021). Super-Recognizers–a novel diagnostic framework, 70 cases, and guidelines for future work. *Neuropsychologia*, *158*, 107809.

Russell, R., Duchaine, B., & Nakayama, K. (2009). Super-recognizers: People with extraordinary face recognition ability. *Psychonomic bulletin & review*, *16*(2), 252-257.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Can DeepFake voices steal high-profile identities?

Bence Mark Halpern[123] *and Finnian Kelly*[4]
[1]*University of Amsterdam, Amsterdam, The Netherlands*
[2]*Netherlands Cancer Institute, Amsterdam, The Netherlands*
[3]*Delft University of Technology, Delft, The Netherlands*
b.m.halpern@uva.nl
[4]*Oxford Wave Research, Oxford, UK*
finnian@oxfordwaveresearch.com

Computer-generated synthetic voices are increasingly growing indistinguishable from human voices. While these high-quality synthetic voices open new horizons for the entertainment industry, they can be also used with malicious intent. Examples of the latter include obtaining unauthorised access to bank accounts using fake-voice biometrics (Wang et al., 2020), or rapidly spreading disinformation via deepfake videos of political leaders (Wakefield, 2022; Lorenzu-Trueba, 2018; Thomas, 2020). As the amount of data required to build convincing synthetic voices decreases, it is becoming increasingly important to develop automatic tools that can reliably detect malicious usage of this technology.

Celebrity voices are a great example of a scenario where large amounts of data is available to build a synthetic voice. In this paper, we consider the evaluation of a Deep Neural Network (Dilated ResNet) based spoofing detector (Halpern et al., 2020) with a celebrity deepfake speech corpus. The corpus, collected for the present study from various online sources, consists of one deepfake recording and one genuine recording for each of 30 celebrities. We note that this data is uncontrolled, with varying levels of noise, compression, and other artefacts.

The evaluation of the corpus resulted in a spoof detection Equal Error Rate (EER) of 16.7%. Speaker-wise, all except two of the 30 speakers, namely Bill Clinton and Winston Churchill, correctly produced higher detection scores for their genuine recording than for their deepfake recording. We hypothesise that older genuine recordings, and that of Winston Churchill in particular, may contain artefacts resulting from post-hoc speech enhancement, which influence the detector.

We further consider the use of the evaluation scores from the celebrity deepfake corpus to calculate a genuine/spoof likelihood ratio (LR) for a questioned sample from a new speaker. Using the probability densities of genuine and spoof evaluation scores to represent genuine and spoof hypotheses respectively, we calculate a genuine/spoof LR for a Zelenskyy deepfake (Wakefield, 2022), as shown in Fig. 1. We additionally calculate the LR for a genuine recording of Zelenskyy (one with a similar SNR to the deepfake). Converting the detector score to an LR in this way, by considering the competing genuine and spoof hypotheses given relevant data, produces a result that can be directly interpreted. In the present example, the LR for the deepfake recording is less than one (0.18) and the LR for the genuine recording is greater than one (6.3). These LRs therefore provide correct support in both deepfake and genuine cases.

Ongoing work is investigating the influence of environmental noise, recording devices, compression, as well as the speech duration, on the performance of deepfake detection.

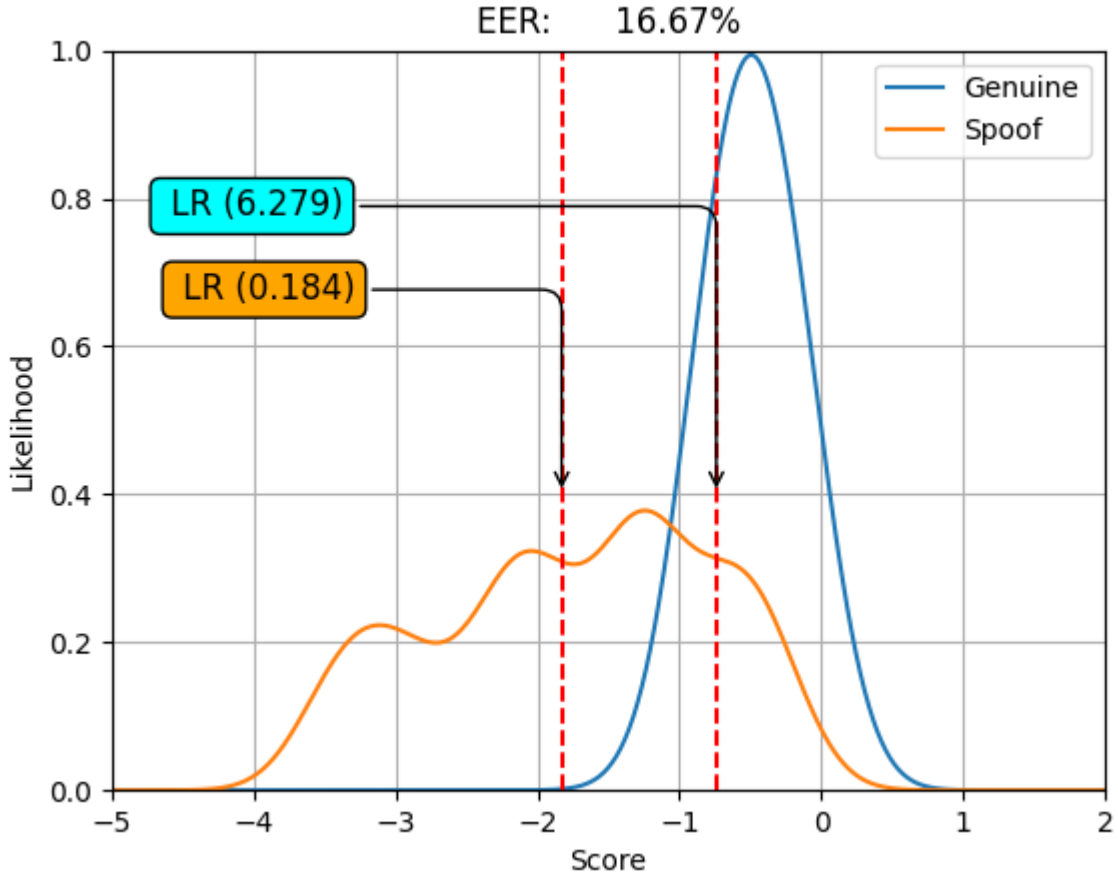


Figure 1 Probability density curves for the genuine and spoof evaluation scores obtained from the celebrity deepfake speech corpus (EER of 16.67%). The orange box indicates the genuine/spoof LR for the Zelenskyy deepfake (0.184) and the blue box indicates the genuine/spoof LR for the Zelenskyy genuine recording (6.279).

References

Halpern, B. M., Kelly, F., van Son, R., & Alexander, A. (2020). Residual networks for resisting noise: analysis of an embeddings-based spoofing countermeasure. *Speaker Odyssey 2020.*

Lorenzo-Trueba, J., Fang, F., Wang, X., Echizen, I., Yamagishi, J., & Kinnunen, T. (2018). Can we steal your vocal identity from the Internet?: Initial investigation of cloning Obama's voice using GAN, WaveNet and low-quality found data. *arXiv preprint arXiv:1803.00860.*

Thomas, D.: Deepfakes: A threat to democracy or just a bit of fun? (2020), https://www.bbc.com/news/business-51204954 (Date of access: 2022. 05. 20)

Wakefield, J.: Deepfake presidents used in Russia-Ukraine war (2022), https://www.bbc.com/news/technology-60780142 (Date of access: 2022. 05. 20)

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., ... & Ling, Z. H. (2020). ASVspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, *64*, 101114.

Prague, July 10–13, 2022
Student Paper

Analysing the performance of automated transcription tools for covert audio recordings

Lauren Harrington[1]*, Robbie Love*[2] *and David Wright*[3]

[1]*Department of Language and Linguistic Science, University of York, UK.*
lauren.harrington@york.ac.uk
[2]*Department of English, Languages and Applied Linguistics, Aston University, UK.*
r.love@aston.ac.uk
[3]*Department of English, Linguistics and Philosophy, Nottingham Trent University, UK.*
david.wright@ntu.ac.uk

The orthographic transcription of audio recordings can provide important evidence in a forensic case (Fraser, 2021), but producing transcripts is an extremely time-consuming task and is often a prerequisite to further analyses. Huge improvements in automatic speech recognition have been observed throughout the past two decades, particularly with the recent development of deep learning (Xiong et al., 2016). The use of an automatic transcription system could significantly decrease the amount of time and effort taken to produce a transcript and this could make such systems an attractive prospect to those in law enforcement. However, there are many factors that are known to negatively affect the accuracy of automatic transcription systems, such as spontaneous speech and increased speech rate (Benzeghiba et al., 2007), overlapping speech (Shriberg et al., 2001; Raj et al., 2021), and background noise (Lippman, 1997; Littlefield & Hashemi-Sakhtsari, 2002). Most of these factors can be directly applied to forensic recordings, which often involve multiple speakers and are of bad quality. Loakes & Fraser (2021) tested two automatic transcription systems on a forensic-like poor-quality recording, and they found that performance was far worse than for a good quality recording, including issues such as consistently identifying non-speech sounds (e.g. drums, laughter) as speech and not transcribing large sections of the recording at all.

This paper reports the design and results of a controlled transcription experiment in which twelve automated transcription tools produced transcripts for the same audio recording. The recording itself is of a conversation between five adults in a busy restaurant taken on a smart phone, and shares many of the typical features of covert forensic recordings, including the presence of multiple speakers, background noise and use of non-specialist recording equipment. It has been found to pose a challenging transcription task for trained human transcribers (Love & Wright, 2021). This paper focuses on the transcripts produced by the twelve systems for 18 specific utterances which are clear enough in their content to be confident of ground truth *and* which most systems attempted to transcribe. All utterances were produced by a single speaker, given the failure of all systems to represent overlapping speech of multiple speakers.

The analysis relied on the timestamps provided by the systems to align the transcripts for comparison on a word-by-word basis (e.g. Figure 1). Based on these comparisons, we examined the output across the systems, identifying instances where there were widespread gaps in the transcripts and common mistranscriptions, as well as those elements of the talk that were consistently transcribed accurately by most or all of the systems. In doing so, we attempt to identify patterns in the transcription tendencies of the systems, account for the variation observed and begin to determine their (relative)

1	Reference	I	can't	see	in	this	light	or	maybe	my	eyes	just	don't	see
2	Microsoft		Oh	see		this	place. Slice	well,	maybe	my	eyes	just	don't	see.
3	Descript		let's	see.				Well,	maybe	my	eyes	just	don't	see
4	Google Cloud								maybe	my	item.			
5	HappyScribe							Or	maybe	my	eyes	just	30	
6	Konch	I	would	say	in	this	place,	well,	maybe	my	eyes	just	don't.	
7	NVivo					This		well,	maybe	my.				
8	Otter							will	maybe	my	eyes	just	don't	see
9	Sonix													
10	Temi			see,				well,	maybe	my	eyes	just	don't	see
11	Transcribear							Well,	maybe	my	eyes	just	don't	see.
12	Transcribe by Wreally													
13	Trint	I		see		this	place.	Well,	maybe	my	eyes	just	don't	see.

Figure 1. An example of an aligned comparison of the transcripts produced by all twelve automated systems for the utterance "I can't see in this light or maybe my eyes just don't see".

strengths and weaknesses. Finally, we discuss the implications of these results for the potential application of automated transcription systems in forensic contexts and, in particular, the role of the human expert in managing and interpreting the output of such systems and the challenges this poses.

References

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C. and Rose, R., Tyagi, V. & Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech communication*, *49*(10-11), 763-786.

Fraser, H. (2021). Forensic transcription: the case for transcription as a dedicated area of linguistic science. in M. Coulthard, A. Johnson, and R. Sousa-Silva (eds.), The Routledge Handbook of Forensic Linguistics. Editors (2nd edn). London: Routledge, pp. 416–431.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech communication*, *22*(1), 1-15.

Littlefield, J., & Hashemi-Sakhtsari, A. (2002). *The effects of background noise on the performance of an automatic speech recogniser*. DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION SALISBURY (AUSTRALIA) INFO SCIENCES LAB.

Loakes, D. & Fraser, H. (2021). Assessing the role of automatic methods for the transcription of indistinct covert recordings. In *29th Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Marburg, Germany [online].

Love, R., & Wright, D. (2021). Specifying challenges in transcribing covert recordings: implications for forensic transcription. *Frontiers in Communication,* 6:797448 (Research Topic: 'Capturing talk: The institutional practices surrounding the transcription of spoken language'). DOI: 10.3389/fcomm.2021.797448

Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., He, M., Watanabe, S., Du, J., Yoshioka, T., Luo, Y. & Kanda, N. (2021, January). Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 897-904.

Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Seventh European Conference on Speech Communication and Technology,* Aalborg, Denmark.

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M. L., Stolcke, A., Yu, D. & Zweig, G. (2017). Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(12), 2410-2423.

Prague, July 10–13, 2022

Work-in-Progress

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

Vincent Hughes[1], *Paul Foulkes*[1], *Philip Harrison*[1], *David van der Vloed*[2] *and Finnian Kelly*[3]

[1]*Department of Language and Linguistic Science, University of York, UK.*
{vincent.hughes|paul.foulkes|philip.harrison}@york.ac.uk

[2]*Netherlands Forensic Institute, Netherlands.*
d.van.der.vloed@nfi.nl

[3]*Oxford Wave Research, Oxford, UK.*
finnian@oxfordwaveresearch.com

In this 'work in progress' paper, we introduce a new ESRC-funded project called *Person-specific automatic speaker recognition: understanding the behaviour of individuals for applications of ASR* (ES/W001241/1). The project will run from 2022 to 2025 and involves collaboration between the University of York, the Netherlands Forensic Institute and Oxford Wave Research.

The project will examine what makes particular voices easy or difficult for automatic speaker recognition (ASR) systems to identify. In doing so, we will assess the performance of systems with individual speakers and develop methods to handle *problematic* types of speakers. The project has four central research questions:

i. What systematic properties of speakers make them more or less susceptible to ASR errors, in terms of voice (e.g. pitch, voice quality) and demographic factors (e.g. accent, ethnicity, age, sex)? And how do the magnitudes of these effects compare to known technical effects?
ii. How consistent are results for individual speakers within and across ASR systems?
iii. How do results produced by techniques that combine ASR and linguistic methods on a person-specific basis compare with the current one-size-fits-all approach?
iv. How generalisable are methods and results across datasets and languages?

The project is organised around three workpackages. In workpackage (1), we will collect controlled recordings of phoneticians adapting aspects of their vocal output using a variety of different channels and recording devices over a number of sessions. This will allow us to compare the relative effects of speaker variation, technical variation, and random occasion-to-occasion variability on ASR output. In workpackage (2), we will conduct much larger scale testing of a range of different ASR systems (varying elements of the ASR processing and calibration) using databases of English (Home Office COST Collection) and Dutch (NFI FRIDA). Finally, in workpackage (3), we will develop novel methods to systematically integrate ASR and expert linguistic analyses. This will involve flagging comparisons containing *problematic* speakers for the ASR system, subjecting them to more detailed linguistic analysis, and then validating the entire approach.

Prague, July 10–13, 2022
Work-in-Progress

Auditory and machine-based identification of closely related languages: A comparison of methods for LADO procedure

Jacek Kudera[1], *Bernd Möbius*[1]
[1]*Language Science and Technology, Saarland University, Saarbrücken, Germany*
{kudera|moebius}@lst.uni-saarland.de

Aims

This work aims to compare the methods used in auditory and machine-based LADO involving closely related languages (Wilson & Foulkes, 2014; Pellegrino & André-Obrecht, 2000). The goals of the study were to verify the ability of lay-listeners to recognize the linguistic origin of speakers, based on spoken samples with limited segmental and suprasegmental information, and to correlate the signal features with the subjects' performance. Additionally, the work aimed to present ideal competence of lay-listeners in LADO cases regarding Slavic languages (Fraser, 2011; Hoskin, 2018; Language and National Origin Group, 2004; Patrick et al. 2012).

Methods

In the first experiment, the native speakers of Bulgarian, Czech, Polish, and Russian were given a task to identify the L1 of a sex-balanced group of 40 recorded native speakers of the above-mentioned languages (10 speakers per language). A second study involved xVOCALISE package for speech comparison based on formant (F1-F4) dynamics and signal representation using x-vectors. The auditory task to identify a language of origin was given to 228 native speakers of four Slavic languages with no linguistic and phonetic training. The stimuli consisted of CVCV and CVCVCV logatomes, controlled for lexical stress placement. The participants were asked to select one of the four languages which they believed to be the native language of the speaker in the recording and mark the certainty on a confidence scale. A confusion matrix of tested languages was computed based on Perceptual Similarity Index (Thomas, 2011) for disyllabic and trisyllabic sequences separately. Furthermore, the linguistic profile of subjects was analyzed to picture ideal competence of lay-listeners for LADO procedure. To find the most informative signal features, the vowel overlap computed as 3D Pillai-Bartlet trace including duration of vocalic segments (Pillai, 1954) was correlated with lay-listeners' performance. The machine-based procedure involved comparison of speech samples using PLDA (Prince & Adler, 2007) and cosine distance between the vectors representing the recordings selected for LADO.

Results

The results suggest that limited spectral and temporal features of speech signal can provide a cue in the identification of linguistic origin. However, the gathered data suggest that lexical stress distribution is not a discriminable factor for speakers of Slavic languages. The applied machine-

based approach can complement the perceptual LADO. The analyses showed that DNN only outperformed lay listeners whose L1 was Bulgarian. The other groups of native speakers ranked the languages equally good as trained network.

Conclusions

This study provides a clear argument for the involvement of native speakers in LADO/LOID procedures. It appears that highly limited signals can cause an attention shift towards typically less relevant features in spoken language perception such as vowel quality in the spectral and temporal domains. These findings should be considered in LADO as well as in forensic applications.

References

Cambier-Langeveld, T. (2016). Language analysis in the asylum procedure: a specification of the task in practice. International Journal of Speech, Language & the Law, 23(1), 25-41. https://doi.org/10.1558/ijsll.v23i1.17539

Fraser, H. (2011). The Role of Linguistics and Native Speakers in Language Analysis for the Determination of Origin: A Response to Tina Cambier-Langeveld. International Journal of Speech, Language and the Law, 18(1), 121-130. https://doi.org/10.1558/ijsll.v18i1.121

Hoskin, J. (2018). Native speaker non-linguists in LADO: an insider perspective. In I. Nick (Ed.), Forensic Linguistics: Asylum-seekers, Refugees and Immigrants (pp. 23-40). Malaga: Vernon Press.

Language and National Origin Group (2004). Guidelines for the use of language analysis in relation to questions of national origin in refugee cases. International Journal of Speech, Language and the Law, 11(2), 261-266. https://doi.org/10.1558/ijsll.v11i2.261

Patrick, P., Schmid, M., & Zwaan, K. (2012). Language Analysis for the Determination of Origin: Current Perspectives and New Directions. Springer.

Pellegrino, F., & André-Obrecht, R. (2000). Automatic language identification: an alternative approach to phonetic modelling. Signal Processing, 80(7), 1231-1244.

Pillai, K. C. (1954). On some distribution problems in multivariate analysis. North Carolina State University. Dept. of Statistics. https://repository.lib.ncsu.edu/bitstream/handle/1840.4/2164/ISMS_1954_88.pdf?sequence=1

Prince, Simon; Elder, James, 2007. Probabilistic Linear Discriminant Analysis for Inferences about identity. In *IEEE 11th International Conference on Computer Vision (ICCV)*, 2007: 1–8.

Thomas, E. R. (2011). Sociophonetics: An introduction. Palgrave MacMillan.

Wilson, K., & Foulkes, P. (2014). Borders, variation, and identity: Language analysis for the determination of origin (LADO). In D. Watt & C. Llamas (Eds.), Language, Borders and Identity (pp. 218-229). Edinburgh: Edinburgh University Press.

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Acoustic characteristics of filler particles in German

Beeke Muhlack[1]*, Jürgen Trouvain*[1]*, and Michael Jessen*[2]
[1]*Language Science and Technology, Saarland University, Germany*
{muhlack|trouvain}@lst.uni-saarland.de
[2]*Bundeskriminalamt, Germany*
Michael.Jessen@bka.bund.de

It is generally assumed that filler particles (FPs), such as *äh* and *ähm* in German, are mainly unconsciously produced and thus may prove useful in forensic casework (Jessen, 2008; Künzel, 1987). Disfluencies like FPs, in combination with sound-prolongations, repetitions, and self-interruptions, show a speaker-specific pattern with evidence for German (Braun & Rosin, 2015) and English (McDougall & Duckworth, 2018). However, the consistency of this pattern may be instable across dissimilar speaking tasks, e.g. voice messages compared to interviews (Harrington et al., 2021).

For this study, we aim to present the characteristics of FPs for 100 German male speakers in two conditions: in a Lombard condition and in a non-Lombard ('normal') condition. The data was collected in 2001 as part of the Pool2010 Corpus (Jessen et al., 2005) which uses a picture-description task with forbidden "taboo" words to elicit spontaneous speech. The mean recording time for each speaker is ca. 4 minutes in each condition, amounting to a total duration of ca. 13 h for the sub-corpus investigated here. A typical feature of Lombard speech is an increase of the mean fundamental frequency of speakers as the vocal effort is increased (Jessen et al., 2005). But it is yet unclear to what extent the Lombard condition influences the distribution and phonetic characteristics of FPs. Features under investigation are the frequency (items/min) of different types of FPs (*uh*, *uhm*, *hm*, glottal FPs and tongue clicks), the occurrence of silences before and after the FP, the duration of their segments, fundamental frequency, vowel quality of the vocalic portion of FPs, as well as creaky voice/glottal pulses during the FP.

Preliminary results show that the frequency of typical FPs (*uh*, *uhm*, *hm*) decreases from normal to Lombard speech while the frequency of tongue clicks and glottal FPs (produced with creak/creaky voice only) increases in the Lombard condition (see Table 1). Furthermore, the most frequent FP used by these speakers is the vocalic type (*uh*) which occurs more than twice as often as the vocalic-nasal type (*uhm*). Figure 1 shows that *uhm* is generally longer than *uh*. Moreover, the longest FPs occur between silences, i.e. in a pause. Those FPs that are articulated within an inter-pausal unit (IPU) are shortest. FPs in IPU-final position are longer than in IPU-initial position which is in line with the effect of pre-pausal lengthening. The pattern of a "duration hierarchy" (see Figure 1) holds true for both filler particles *uh* and *uhm*.

Observations on the individual-speaker level reveal that each feature shows high between-speaker variation, so that the rate for the FP *uh* ranges from 0-19 items/min, for *uhm* from 0-15 items/min while *hm* generally shows a lower frequency with a range of 0-7 items/min. The extreme values may be particularly interesting for forensic casework, e.g. three speakers with a higher glottal FP-rate were observed (> 3 standard deviations higher than mean). The poster will show the general trend of the participants but also focus on the individual performance of the speakers and the variation within the dataset.

	Normal (%)	*Lombard (%)*	**Sum**
uh	921 (36.7)	857 (31.2)	1778
uhm	395 (15.7)	327 (11.9)	722
hm	182 (7.3)	86 (3.1)	268
glottal FP	237 (9.4)	381 (13.9)	618
clicks	774 (30.9)	1098 (39.9)	1872
Sum	2509 (100)	2749 (100)	5258

Table 1. The frequency distribution of the phenomena under investigation in the normal and the Lombard condition. The values in parentheses are the percentages for each condition.

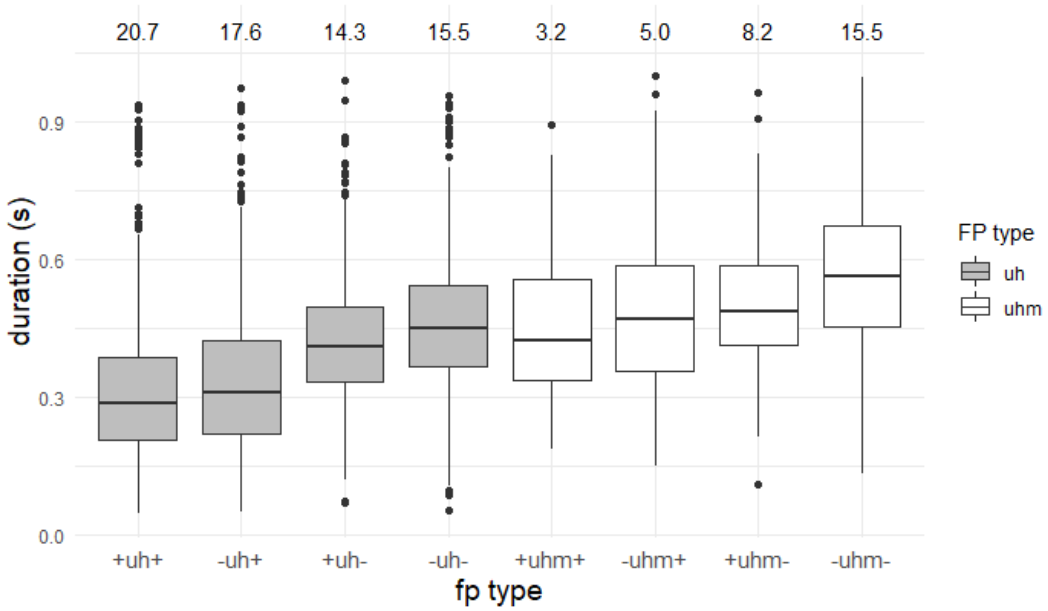


Figure 1. The duration of the filler particles *uh* and *uhm* in their context: speech (+) or silence (-). Thus, a -FP+ occurs in IPU-initial position while +FP- is an FP in IPU-final position, +FP+ occurs within an utterance, -FP- occurs in isolation. The values above each plot represent the percentages per category in the dataset containing only *uh* and *uhm*.

References

Braun, A., & Rosin, A. (2015). On the speaker-specificity of hesitation markers. Proceedings of the 18th International Congress of Phonetic Sciences. Glasgow, UK: The University of Glasgow.

Harrington, L., Rhodes, R., & Hughes, V. (2021). Style variability in disfluency analysis for forensic speaker comparison. International Journal of Speech Language and the Law, 28(1), 31–58. https://doi.org/10.1558/ijsll.20214

Jessen, M. (2008). Forensic Phonetics. Language and Linguistics Compass, 4(2), 671–711. https://doi.org/10.1017/S0022226700012755

Jessen, M., Köster, O., & Gfroerer, S. (2005). Influence of vocal effort on average and variability of fundamental frequency. International Journal of Speech Language and the Law, 12(2), 174–213. https://doi.org/10.1558/sll.2005.12.2.174

Künzel, H. J. (1987). Sprechererkennung. Grundzüge forensischer Sprachverarbeitung. Kriminalistik Verlag.

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

Prague, July 10–13, 2022
Work-in-Progress, Student paper

Speaker/ Author Profiling in Maltese

Amanda Muscat[1]
[1]*Institute of Linguistics and Language Technology, University of Malta, Msida, Malta*
amanda.muscat.1@um.edu.mt

The ways people use language can reveal a great deal about their personalities and social background (Andrew Schwartz et al., 2013). Research suggests that relatively stable traits, which can be biological (e.g. a person's sex), cultural (e.g. a person's social class) and/or related to personality types (e.g. the Big Five, more specifically traits such as extroversion), are subtly reflected in a person's linguistic choices (Schuller et al., 2013). The exact nature of this relationship, however, remains somewhat elusive and research into this has not been carried out cross-linguistically to any significant extent. This study aims to investigate inter- and intra-speaker variation in written and spoken Maltese in order to attempt to give an account of how linguistic features correlate with specific definable traits and to develop an objective methodology which enables linguistic profiling for Maltese. For the purpose of this study, data was collected from 10 participants, who were carefully selected to balance in terms of gender and age while being dominant users of Maltese, in 5 different communicative situations. The different communicative situations have been designed to capture variation in both modalities of language, spoken and written. Presumably, these tasks tap into sources of variation such as formal versus informal speech / text, different audience, etc., that different linguistic theories might hold to be central, as can been seen in table 1.1 below.

Tasks	**Modality**	**Code Theory** (Bernstein, 1960, 2003)	**Attention to Speech** (Labov, 1972)	**Audience Design** (Bell, 1984)	**Communication Accommodation Theory** (Giles, 1973)	**Speaker Design Approaches**	**Resource-Constraint Model** (Grant & MacLeod, 2018)
TAT	Speech and writing	Unrestricted topics	informal	-		project identity to the communicative audience	
Personal Experience	Speech and writing	Restricted topics	Formal and informal	imaginary known and unknown audience		project identity to the communicative audience	
Chats	Writing	Restricted topics	informal	unknown audience		project identity to the communicative audience	
Writings	Writing	Unrestricted topics	Formal and informal	known and unknown audience		project identity to the communicative audience	
Map-Task	Speech	Restricted topics	Formal and informal	unknown audience		-	

Table 1. A list of tasks which participants had to complete according to the theories of variation reviewed.

This poster reports on preliminary analysis for a number of linguistic features mainly filled pauses, silent pauses, fundamental frequency and speaking time, with a view to looking for correlations with gender, age and the Big Five personality traits, in conversation, utilising the 44-Big Five personality measure. The analysis confirms previous findings such as that filled pauses are positively correlated with females (Bamman, Eisenstein, & Schnoebelen, 2014) while also revealing correlations between the big five personality traits and filled pauses.

References

Andrew Schwartz, H., Eichstaedt, J. C., Dziurzynski, L., Kern, M. L., Seligman, M. E. P., Ungar, L. H., … Stillwell, D. (2013). Toward personality insights from language exploration in social media. AAAI Spring Symposium - Technical Report, SS-13-01(January), 72–79. Retrieved from https://www.researchgate.net/publication/283270498_Toward_Personality_Insights_from_Language_Exploration_in_Social_Media

Bamman, D., Eisenstein, J., & Schnoebelen, T. (2014). Gender identity and lexical variation in social media. Journal of Sociolinguistics, 18(2), 135–160. https://doi.org/10.1111/josl.12080

Bell, A. (1984). Language style as audience design. Language in Society, 13(2), 145–204.

Bernstein, B. (1960). Language and Social Class. The British Journal of Sociology, 11(3), 271–276. Retrieved from https://www-jstor-org.ejournals.um.edu.mt/stable/586750?sid=primo&origin=crossref&seq=1#metadata_info_tab_contents

Bernstein, B. (2003). Class, Codes and Control. Volume 1: Theoretical Studies Towards a Sociology of Language (Routledge, ed.). London.

Giles, H. (1973). Accent Mobility : A Model and Some Data. Anthropological Linguistics, 15(2), 87–105. Retrieved from https://www.jstor.org/stable/30029508

Grant, T., & MacLeod, N. (2018). Resources and constraints in linguistic identity performance – a theory of authorship. Language and Law, 5(1), 80–96. Retrieved from https://www.researchgate.net/publication/327573867_Resources_and_constraints_in_linguistic_identity_performance_a_theory_of_authorship

Labov, W. (1972). Sociolinguistic Pattern. Philadelphia: University of Pennsylvania Press.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., & Narayanan, S. (2013). Paralinguistics in speech and language - State-of-the-art and the challenge. Computer Speech and Language, 27(1), 4–39. https://doi.org/10.1016/j.csl.2012.02.005

Prague, July 10–13, 2022
Student Paper

The effect of free voice-disguise methods on ASR performance

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

In the forensic speaker comparison (FSC) field, the voice disguise phenomenon has been known and examined for decades. Depending on what type of crime is being committed, the offender is more or less likely to be trying to conceal their voice identity; nevertheless, the numbers (how many do try) vary a great deal (see e.g., Künzel, 2000; or Braun, 2006 for more information). Moreover, perpetrators have come up with a range of ways of altering their voice; from changing the fundamental frequency, imitating regional/foreign accent, placing an object before/into the mouth, to modifying their voices electronically. Luckily for the justice system, criminals opt for rather less sophisticated methods, which might be caused by the fact that it is difficult to combine verbal planning and complex voice disguise means at the same time (Masthoff, 1996).

This study was carried out on the forensic database of Common Czech (Skarnitzl & Vaňková, 2017), which comprises 100 male speakers of a supraregional variety of Czech (aged 19–50, mean 25.6 years); and follows the research of Skarnitzl et al. (2019). The speakers, recorded in quiet environments with a high-quality portable device, were asked to deliver three distinct speech styles – spontaneous speech (everyday topics, approx. 2-minute samples were cut out of 25–45-minute recordings), read speech (a phonetically rich text, approx. 60 seconds in length), disguised speech (a phonetically rich text similar to the previous one, using some identical words to facilitate comparison, approx. 60 seconds in length). As for the last style, having been instructed to report to their kingpin, the speakers were given time to select their own technique to conceal their voice identity. The voice modifications employed differed from almost no audible changes at all to complex disguise strategies (see Růžičková & Skarnitzl, 2017).

As automatic speaker recognition (ASR) has been gaining prominence in FSC in the last years (Gold &French, 2019), the goal of this paper was to compare the three datasets described above using VOCALISE's i-vectors and x-vectors (Oxford Wave Research, 2019a; Oxford Wave Research, 2019b). Since VOCALISE provides its users with a possibility to tune the performance by condition adaptation and reference normalisation, we were interested to know whether, and to what extent, such tweaking of the system's settings would also help with disguised voices. First, we divided each dataset into two random subsets (N=50) and compared them as such; second, we examined the subsets after adapting and normalising the system with always the other 50 recordings of spontaneous and disguised samples (the disguised ones were always used as the so-called comparison files).

In the end, no matter which tuning procedure had been applied, when the disguised set was subjected to comparison, the performance of the systems fell short of the typically reported results, with EER ranging from 14.9% to 30%. It is worth noting that the i-vector system outperformed the newer x-vectors in several individual comparisons. As for the contribution of various tuning procedures, our results do not suggest a greater benefit of tuning with a subset which featured disguised voices.

References

Braun, A. (2006). Stimmverstellung und Stimmenimitation in der forensischen Sprechererkennung. In T. Kopfermann (ed.) *Das Phänomen Stimme: Imitation und Identität: 5. Internationale Stuttgarter Stimmtage 2004. Hellmut K. Geissner zum 80. Geburtstag* (pp. 177–182). St. Ingbert: Röhrig Universitätsverlag.

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

Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics*, 7, 149–179.

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

Oxford Wave Research (2019a). iVOCALISE 2019A.

Oxford Wave Research (2019b). BioMetrics 2019A.

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

Skarnitzl, R., Asiaee, M. & Nourbakhsh, M. (2019). Tuning the performance of automatic speaker recognition in different conditions: Effects of language and simulated voice disguise. *International Journal of Speech, Language and the Law*, 26, 209–229.

Skarnitzl, R. & Vaňková, J. (2017). Fundamental frequency statistics for male speakers of Common Czech. *Acta Universitatis Carolinae – Philologica 3, Phonetica Pragensia XIV*, 7–17.

Prague, July 10–13, 2022
Student Paper

Notorious and new voice: How does a professional imitator fare?

Vojtěch Skořepa, Radek Skarnitzl
Institute of Phonetics, Charles University, Prague, Czech Republic
vojta.skora@gmail.com
radek.skarnitzl@ff.cuni.cz

The success of an imitator depends on his or her ability to mimic the salient features of a particular speaker in order to strengthen the perceptual impression (Zetterholm, 2006). Imitation belongs to the field of theatre arts and can be studied at some drama schools (Singh, 2016). It was only with the advancement of technology and techniques that imitation started to be used for nefarious purposes (impersonation, forging voice passwords, etc.). Perrot et al. (2009) say that using voice identity imitation can discredit someone else in order to divert attention away from oneself. Research on impersonators is also very important in order to identify the possibilities and limits of speech production, and to map the plasticity of the human voice, among other things.

In this study, we used a publicly known Czech professional impersonator to cooperate with. This impersonator (male, 42 years old) is originally from Prague, but has been living in South Bohemia for the last few years. The aim was to find out how successful he is in imitating the vocal identity of a speakers who is well-known to him and of one who is not. In cooperation with the imitator, we selected one speaker that this imitator already has in his repertoire, the current President of the Czech Republic, Miloš Zeman (male, 77 years old), who is originally from Central Bohemia and has alternately lived in Prague and the Highlands. At the same time, we provided the imitator with recordings (audio, video) of an unfamiliar voice (and also a less well-known public figure), that of the former Senator of the Czech Republic Jiří Carbol (male, 63 years old), who is originally from the Moravian-Silesian region, where he has lived most of his life. The imitator was given six weeks to learn the new voice.

The recordings were acquired in the sound-treated studio of the Institute of Phonetics in Prague. The imitator was first asked to read the Czech translation of the Rainbow Passage in the target voices, to ensure a neutral content; it is well known that the content may serve as a strong facilitator of a successful imitation. Second, the experimenters assumed the role of reporters and asked the "imitated person" questions prepared in advance. The interviews took between five and ten minutes. Approximately one minute of the interview was used for subsequent analyses in this study. Recordings of similar duration of the two imitated persons (Zeman and Carbol) were used for comparison.

We conducted listening and acoustic analyses to compare the imitator's renditions of the target voices with the originals. For the listening analysis, we used a recently developed protocol (Skarnitzl, 2022, in print) which combines the descriptive dimensions of SVPA (San Segundo and Mompean, 2017) with features listed in other protocols proposed for listening comparisons of voices in a forensic context. Acoustic analyses were carried out in Praat and included, at this stage, f_0 descriptors, vowel formants and spectral properties of selected obstruent sounds. The poster will present the results of all these analyses.

References

Perrot, P., Morel M., Razik, J. & Collet G. (2009). Vocal forgery in forensic Sciences. *Forensics in Telecommunications, Information and Multimedia*, 179-185.

San Segundo, E. & Mompean, J. A. (2017). A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity. *Journal of voice*, 31 (5), 1-17.

Singh, R., Gencaga, D. & Raj, B. (2016). Formant manipulations in voice disguise by mimicry. In *Proceedings of IWBF '16*, Limassol, Cyprus.

Skarnitzl, R. (2022, in print). O fonetické identifikaci mluvčího ve forenzním kontextu. *Naše řeč.*

Zetterholm, E. (2006). Same speaker – different voices. A study of one impersonator and some of his different imitations. *In Proceedings of AICSS&T '06,* Auckland, New Zealand, 70–75.

Reducing the degree of uncertainty within automatic speaker recognition systems using a Bayesian calibration model

Bruce Xiao Wang, Vincent Hughes
Department of Language and Linguistic Science, University of York
hollamigo@gmail.com/vincent.hughes@york.ac.uk

In data-driven forensic voice comparison (FVC), empirical testing of a system is an essential step to demonstrate validity (i.e. how well the system performs the task) and reliability (i.e. whether the system would yield the same result if the analysis were repeated). The present study focuses on system reliability, aiming to reduce the degree of uncertainty at the *score* space with small sample size and skewed scores (conditions which are typical of the real world). Wang & Hughes (2021) simulated scores to test different calibration methods showing that the Bayesian model (Brümmer & Swart, 2014) outperformed logistic regression in terms of variability in system validity values (i.e. produced less variable results). However, they simulated scores based on a linguistic system using multivariate kernel density (MVKD), which is likely to have worse overall performance than automatic speaker recognition (ASR) systems utilising Mel-frequency cepstral coefficients (MFCCs) or cepstral measures. We simulated scores generated from i-vector and Gaussian Mixture Model – Universal Background Model (GMM-UBM) ASR systems using real speech data to demonstrate the variability in system reliability as a function of score skewness and sample size. Scores were simulated based on parameters of score distribution from Enzinger et al. (2016) and Morrison & Poh (2018).

Scores were simulated using both skewed and non-skewed parameters (i.e., skewness changed to 0) to investigate the effect of score skewness. To account for sample size, training and test speakers were increased from 20 to 100, with 10-speaker increasements. Logistic regression and a Bayesian model were used for calibration and replicated 100 times per sample size. Performance was evaluated using the mean (overall discrimination) and range (overall variability) of the C_{llr}s across the 100 replications.

Figure 1 shows the C_{llr} mean (dots) and range (lines). Using logistic regression, C_{llr} ranges are 1.3 (i-vector) and 0.69 (GMM-UBM) when scores are skewed (panel (a)) and sample size is small (20 speakers), while the C_{llr} ranges are 0.49 (i-vector) and 0.69 (GMM-UBM) when scores follow normal distributions (panel (b)). Score skewness seems to have a less marked effect on system reliability for the GMM-UBM system when sample size is small, principally because GMM-UBM produced less skewed scores. Panels (c) and (d) show that Bayesian calibration improves system reliability considerably when scores are skewed, e.g., the C_{llr} range is ca. 0.3 (Figure 1 (c)) compared with 1.3 (Figure 1 (a)) when 20 speakers are used. For the GMM-UBM system, Bayesian calibration does not seem to improve system reliability as much it does for the i-vector system, and score skewness seems to have less effect on system reliability when 40 or more speakers are used.

The mean C_{llr} stays stable across score skewness and sample size within systems. However, there appears to be a trade-off, such that overall discrimination (i.e. mean C_{llr}) may be slightly poorer where reliability is slightly better. Thus, it is important for experts to consider what the most important metric of system performance to be and what constitutes 'low enough' mean C_{llr} in making decision about which system to use in a forensic case (see Morrison et al., 2021).

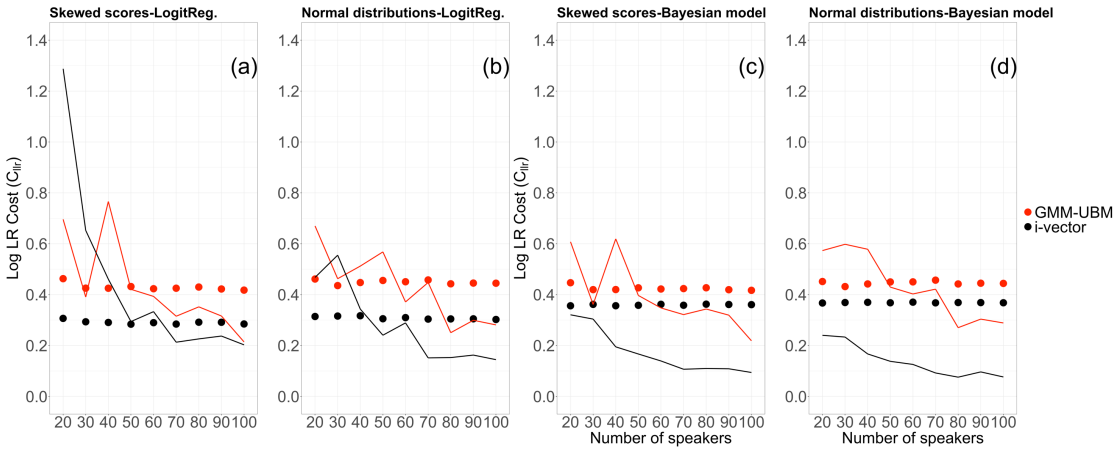


Figure 1. C_{llr} mean and range as a function of score skewness, sample size and calibration methods.

References

Brümmer, N., & Swart, A. (2014). Bayesian Calibration for Forensic Evidence Reporting. *Interspeech*, 388–392.

Enzinger, E., Morrison, G. S., & Ochoa, F. (2016). A demonstration of the application of the new paradigm for the evaluation of forensic evidence under conditions reflecting those of a real forensic-voice-comparison case. *Science & Justice*, *56*(1), 42–57. https://doi.org/10.1016/j.scijus.2015.06.005

Morrison, G., Enzinger, E., Hughes, V., Jessen, M., Meuwly, D., Neumann, C., Planting, S., Thompson, W. C., van der Vloed, D., J F Ypma, R., & Zhang, C. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, *61*(3), 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

Morrison, G., & Poh, N. (2018). Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/Bayes factors. *Science & Justice*, *58*(3), 200–218. https://doi.org/10.1016/j.scijus.2017.12.005

Wang, B. X., & Hughes, V. (2021). System Performance as a Function of Calibration Methods, Sample Size and Sampling Variability in Likelihood Ratio-Based Forensic Voice Comparison. *Interspeech 2021*, 381–385. https://doi.org/10.21437/Interspeech.2021-267

Prague, July 10–13, 2022
Student Paper

Analysis of Forced Aligner Performance on Non-native (L2) English Speech

Samantha Williams[1]
[1]*Department of Language and Linguistic Science, University of York, York, UK*
sejw500@york.ac.uk

Introduction

Forced aligners provide a semi-automatic method of aligning an acoustic signal with phone-level segmentation and have applications in phonetics, forensic speech science, and speech technology research. Provided with an orthographic transcription, a forced aligner can give an estimate of where certain words or segments occur, greatly reducing the amount of manual work required and therefore facilitating much larger scale data analysis. Previous research has compared the performance of different forced aligners (e.g. Gonzalez et al., 2020), investigated the effects of different factors on their accuracy (e.g. Fromont and Watson, 2016), and tested variety mismatch with non-standard varieties of English (e.g. Mackenzie and Turton, 2020). However, the performance of forced aligners on L2 English speech is relatively understudied.

The present study investigates how non-native (L2) English speech is treated by a forced aligner trained on General American English. The intent is to be able to facilitate the comparison of varieties as part of the development of an L1 recognition system.

In this presentation, the following research questions will be addressed with a focus on RQ1:

RQ1. Does the forced aligner perform better on some L2 varieties or types of segments than others?
RQ2. Is variation in forced aligner performance driven more by L2 variety or individual speakers?
RQ3. How does the aligner perform when there are differences between the spoken utterance and the orthographic transcription (e.g., hesitations, repeated phrases, inserted or deleted phones due to variety differences)? Can this inform error checks?

Methods

Data was selected from the Speech Accent Archive, which contains a recording of a read passage and self-reported meta-data for each speaker (Weinburger, 2015). The chosen varieties are all non-native L2 English varieties and are referred to by their L1 as indicated in the corpus: Arabic, Dutch, French, German, Italian, Korean, Mandarin, Portuguese, and Russian. Five speakers from each of the nine varieties were selected (21 M, 24 F). Within a language group, speakers were chosen from the same city where possible. The Montreal Forced Aligner (McAuliffe et al., 2017) was used to align the text to the recordings.

Two displacement measurements were calculated following existing studies: Onset Boundary Displacement (OBD) and Overlap Rate (OvR) (Paulo and Oliveira, 2004). OBD measures the absolute displacement of the segment onset boundary (in milliseconds). OvR is a time-independent measure that indicates the amount of overlap between the automatic and human-aligned segments. In total, 9931

boundary observations were made. For the first two questions, errors unrelated to the performance of the forced aligner were removed, retaining 9852 observations in the analysis.

Results

Results show marked variation in the performance across both language groups (Figure 1) and segment types in both OBD and OvR, with highest accuracy for German and French and lowest accuracy for Russian. For OBD, the aligner's performance on all varieties was comparable to that of General American English and human-human agreement ratings (McAuliffe et al. 2017; Cosi et al. 1991) (Table 1). This indicates the General American English model is sufficient for use with non-native L2 English varieties and can be a useful tool to aid in forensic analysis and the development of automatic systems for L2 English speech with minimal modifications.

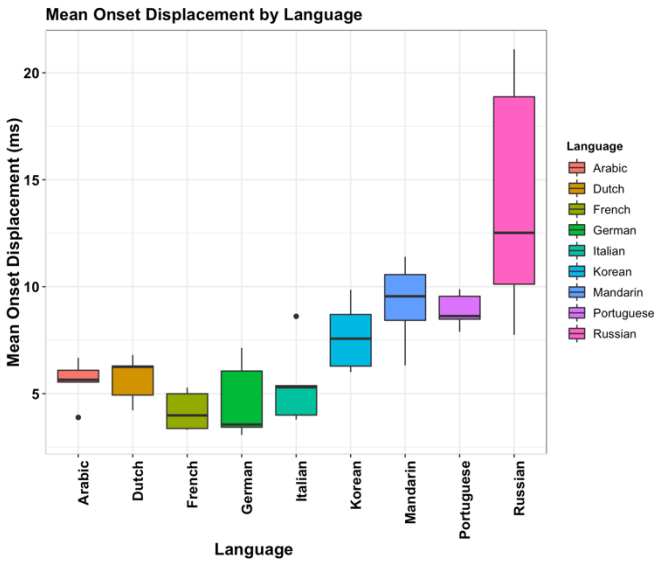


Figure 1. Distribution of OBD speaker means for each language.

Language	*% Onset Displacement <20ms*	*Avg. OvR (%)*
German	93.8	92.1
French	93.4	92.4
Italian	92.7	90.9
Dutch	91.6	89.5
Arabic	89.8	91.9
Korean	87.2	87.6
Portuguese	87.1	83.4
Mandarin	85.4	85.9
Russian	80.0	81.4

Table 1. Percentage of segments where the OBD was less than 20ms and average OvR

References

Cosi, P., Falavigna, D. and Omologo, M. (1991). A preliminary statistical evaluation of manual and automatic segmentation discrepancies. *In Proceedings of Eurospeech 1991,* 693–6. 24–26 September. Genova, Italy.

Fromont, R., & Watson, K. (2016). Factors influencing automatic segmental alignment of sociophonetic corpora. *Corpora*, *11*(3), 401-431.

Gonzalez, S., Grama, J., & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, *6*(1).

MacKenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, *6*(s1).

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M. & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable text-speech alignment using Kaldi. *In Proceedings of the 18th Conference of the International Speech Communication Association*, 498–502.

Paulo, S., & Oliveira, L. C. (2004). Automatic phonetic alignment and its confidence measures. In *International Conference on Natural Language Processing (in Spain)*, 36-44. Springer, Berlin, Heidelberg.

Weinberger, S. (2015). *Speech Accent Archive*. George Mason University. Retrieved from http://accent.gmu.edu