Collegium Helveticum

nccr evolving language
NATIONAL CENTRE OF COMPETENCE IN RESEARCH

University of Zurich^UZH
URPP Language and Space

LiRI - Linguistic Research Infrastructure

University of Zurich^UZH
Zurich Center for Linguistics

CFPA
Centre for Forensic Phonetics and Acoustics (CFPA)

UZH alumni

04.-06.07.2022 UNIVERSITY OF ZURICH ROOMS: AFL-F- 121 & OLIVENHALLE AFFOLTERNSTRASSE 56 8005 ZÜRICH

Interdisciplinary

PRODUCTION PERCEPTION AND COMPUTATIONAL APPROACHES

Voice Identity

Conference

WEB: VOICE-ID.ORG MAIL: INTERVOICEID@GMAIL.COM TWITTER: VOICEID 2022

Book of Abstracts

DAY 1

Session 1

Poster session 1

Session 2

Session 3

DAY 2

Session 4

Session 5

Poster session 2

Session 6

Session 7

DAY 3

Session 8

Session 9

DAY 1

Session 1

Stefan R. Schweinberger

Department for General Psychology and Cognitive Neuroscience,
Friedrich Schiller University Jena, Germany
Voice Research Unit, Friedrich Schiller University Jena, Germany

http://www.allgpsy.uni-jena.de/stefan-r-schweinberger/

New Tools for Assessing Individual Differences in Voice Perception

Auditory morphing can be used to control sensory information in voices (e.g., by interpolating between an average and a specific identity or expression, or by caricaturing). I first introduce current concepts and research using parameter-specific morphing (PSM) technology, by which we can selectively manipulate acoustic parameters (e.g., fundamental frequency (F0), or timbre), thus permitting more objective assessments of their relative roles for perceiving specific signals. I then present selected examples for how PSM can be used to assess voice perception with cochlear implants (CIs), which tend to be optimized for speech perception, with less attention to socio-emotional signals. Although CI users' voice gender perception seems exclusively based on F0, they make more efficient use of timbre in the context of age or emotion perception. Importantly, subjective quality of life with a CI is related to nonverbal voice perception skills. Overall, PSM is a promising new approach to objectively assess profiles of abilities to perceive socio-emotional vocal signals, and can inform perceptual training interventions which generated promising initial results. Finally, I briefly discuss the Jena Voice Learning and Memory Test (JVLMT) as a new and freely available standardized tool for assessing voice learning and recognition skills with pseudospeech utterances with speech-like phonetic variability.

Back to top

Sarah V. Stevenage

University of Southampton, UK

https://www.southampton.ac.uk/psychology/about/staff/svs1.page

Voice identity processing under challenging conditions: responding to singers and impersonators.

Effective vocal identity processing requires that we can tell together different instances from a single speaker, and can tell apart similar instances from two different speakers. Here, two experiments tested the limits of these capabilities by introducing extreme natural challenges. Experiment 1 challenged listeners by *maximising variability within a target voice* - listeners were asked to match speaking with singing clips. Performance significantly declined in this challenging condition, relative to the baseline condition when matching two speaking clips. Moreover, a lack of target familiarity magnified the impact of this challenge. However, performance remained above chance even in the hardest experimental condition. Taking a different approach, Experiment 2 challenged listeners by *minimising the variability across different target voices* - listeners were asked to distinguish celebrity targets from impersonators. Across three tasks, performance declined when telling apart a target from an impersonator, relative to the baseline condition when telling apart two quite different speakers. Again, however, performance remained above chance even in the hardest conditions. Taken together, these results indicated the resilience of vocal identity processing even under challenging natural listening conditions, and suggested a level of sensitivity to vocal cues that had not previously been demonstrated.

Back to top

Katarzyna Pisanski

CNRS, French National Centre for Scientific Research
Dynamics of Language Lab (DDL), Lyon, France
Sensory Neuroethology Lab (ENES), Saint-Etienne, France

http://www.ddl.cnrs.fr/Annuaires/Index.asp?Langue=FR&Page=Katarzyna%20PISANSKI
www.ENESlab.com

Individual differences in human voice pitch are highly stable

Voice pitch is arguably the most intensively studied and salient nonverbal parameter of the human voice. As the perceptual correlate of fundamental frequency (fo), determined by vocal fold size and tension, voice pitch is lower in adults than in children and in men than in women. However, fo also varies considerably within these age-sex classes. Hundreds of studies have linked these individual differences in fo to biologically and socially relevant speaker characteristics, from hormone levels and reproductive fitness to perceived dominance and trustworthiness. Given the dynamic nature of fo, both as people age and as they speak, how stable are between-individual differences in this critical vocal parameter? In a series of within-subject and longitudinal studies, we show that individual differences in human fo remain remarkably conserved across the lifespan and across utterances. The pitch of babies' cries predicts their voice pitch as children, and the pitch of pre-pubertal children's voices predicts their voice pitch throughout adulthood. Individual differences in voice pitch also covary among neutral speech, emotional speech, and nonverbal vocalisations such as cries and screams. Taken together, these results suggest that voice pitch, known to play an important role in social and mating success, is largely determined in early human ontogeny and has predictive power as a robust individual and biosocial marker across disparate communication contexts, with relevance to both human listeners and voice recognition technologies.

Back to top

Non-verbal traits of a public speaker’s charismatic personality

Mariia Boichenko

National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”

boychenko515@gmail.com

The phenomenon of a charismatic leader has been the interest of inter-disciplinary researches in humanities, namely sociology (Howell & Higgins, 1990; Atwater *et al.*, 1991; House, 1992), psychology (Vugt & Ronay, 2014), and linguistics (Signorello *et al.*, 2012; Reh, 2016). The aforementioned research and any potential investigation of the charismatic speaker’s personality is not possible without the feasible determination of a person as charismatic; to do so scientists have been trying to establish a comprehensive set of traits helping for a person to be perceived as a charismatic leader.

By analyzing currently functioning in the aforementioned branches of humanities multiple traits of a charismatic personality, we were able to create the matrix of their alternative variants. Due to the numerous approaches to the issue by different scientists over the years, it was decided to sort out such traits and determine those of them which are synonymous with or secondary to, or consequential of the included into the matrix more abstract notions. At the same time, it was recognized that all of the analysed traits may be conveniently divided into three groups according to their nature, i.e. psychological, sociological, and physiological.

It was found out that to present a comprehensive scientific description of the results on a multidisciplinary research into phenomena of a public speaker’s charismatic personality’s non-verbal behaviour it is reasonable to use among other non-verbal characteristics (e.g., eye-contact, voice, gestures etc.) and their complexes, such notions as intelligence, self-confidence, persistence, ability to inspire, sociability, dominance, narcissism, and vision. These characteristics are relevant to the charismatic traits of a speakers’ personality since they have been historically conventionalized.

The analysis provides prospects of a further analysis of the interdisciplinary traits of a speaker which must help with correct choice of materials for interdisciplinary research into a charismatic speaker’s non-verbal behaviour.

References

Atwater, L., Penn, R., & Rucker, L. (1991). Personal qualities of charismatic leaders. *Leadership and Organization Development Journal, 12*(2), 7-10. https://doi:10.1108/01437739110143330

Bono, J. E., & Judge, T. A. (2004). Personality and Transformational and Transactional Leadership: A Meta-Analysis. *Journal of Applied Psychology, 89*(5), 901-910. https://doi:10.1037/0021-9010.89.5.901

House, R. J., & Howell, J. M. (1992). Personality and charismatic leadership. *The Leadership Quarterly, 3*(2), 81-108. https://doi:10.1016/1048-9843(92)90028-e

Howell, J. M., & Higgins, C. A. (1990). Leadership behaviors, influence tactics, and career experiences of champions of technological innovation. *The Leadership Quarterly, 1*(4), 249-264. https://doi:10.1016/1048-9843(90)90004-2

Reh, S., Giessner, S. R., & Quaquebeke, N. V. (2015). Leader Charisma: An Embodiment Perspective. *Academy of Management Proceedings, 2015*(1), pp.42. https://doi:10.5465/ambpp.2015.18205abstract

Signorello, R., D’Errico, F., Poggi, I., Demolin, D., & Mairano, P. (2012). Charisma perception in political speech: a case study. *International Conference on Speech and Corpora* (GSCP 2012), 2012, 343–348.

Vugt, M. V., & Ronay, R. (2013). The evolutionary psychology of leadership. *Organizational Psychology Review*, *4*(1), 74-95. https://.

Back to top

Differences in the vocal space of deceptive and non-deceptive speech

Alessandro De Luca[1], *Sarah Lim*[1], *Volker Dellwo*[1]

[1]Institute of Computational Linguistics, University of Zurich

alessandro.deluca@uzh.ch

Deception is a common behaviour in not only humans, but also other taxa (Griffin & Ristau, 2013; Whiten & Byrne, 1988). Although deception can include a wide variety of different behaviours, here we only focus on lying by deliberately not telling the truth in human speech communication. We start from the assumption that humans have no interest in being identifiable when lying and hypothesise that a loss of acoustic congruency between utterances within the speaker may affect listeners' memory to what was being said and to the person saying it (Campeanu et al., 2013, 2015). We thus predict that voices should be less recognizable in situations in which speakers do not tell the truth. In a discrimination task with speech from the CSC Deceptive Speech database (Columbia University et al., 2013), human listeners decided between two voice samples of unfamiliar speakers whether they were produced by one or by two speakers either when speech was produced under deceptive or non-deceptive conditions. Results revealed a significant loss of discrimination performance under deceptive compared to truthful speech. To understand what the acoustic correlates of this effect were, we compared the area occupied by the two speech conditions in a latent feature space obtained by UMAP dimension reduction (McInnes et al., 2020) of a 39-dimensional MFCC, D, and D^2 feature set, extracted from utterances of the CSC Deceptive Speech database. Based on Latinus et al. (2013) and Dellwo et al. (2019) we hypothesise that the lower discrimination performance under deceptive speech may rise from speakers being less distinctive when lying by moving towards an average voice and thus occupying a smaller area in the vocal space when being deceptive. Preliminary results showed that instead deceptive speech occupied a slightly larger area than truthful speech in the latent feature space. To confirm this, we are currently working on replicating this analysis using different dimensionality reduction methods and feature sets.

References:

Campeanu, S., Craik, F. I. M., & Alain, C. (2013). Voice Congruency Facilitates Word Recognition. *PLOS ONE*, 8(3), e58778. https://doi.org/10.1371/JOURNAL.PONE.0058778

Campeanu, S., Craik, F. I. M., & Alain, C. (2015). Speaker's voice as a memory cue. *International Journal of Psychophysiology*, 95(2), 167–174. https://doi.org/10.1016/J.IJPSYCHO.2014.08.988

Columbia University, SRI International, & University of Colorado Boulder. (2013). CSC Deceptive Speech. https://doi.org/10.35111/q500-9a28

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. https://doi.org/10.14712/24646830.2019.18

Griffin, D. R. (Donald R., & Ristau, C. A. (2013). *Truth and Deception In Animal Communication*. 147–172. https://doi.org/10.4324/9780203761700-12

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-Based Coding of Voice Identity in Human Auditory Cortex. *Current Biology*, 23(12), 1075–1080. https://doi.org/10.1016/J.CUB.2013.04.055

McInnes, L., Healy, J., & Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences*, 11(2), 233–244. https://doi.org/10.1017/S0140525X00049682

Back to top

Are temporal features of voice identity influenced by jaw size?

Daniel Friedrichs, Volker Dellwo

Department of Computational Linguistics, University of Zurich, Switzerland

daniel.friedrichs@uzh.ch, volker.dellwo@uzh.ch

Anatomical properties of the articulators impact the qualitative-spectral characteristics of speech (Fant, 1960). Here, we asked whether jaw anatomy can explain temporal dynamics of the speech amplitude envelope, which have been shown in prior research to vary between talkers (He & Dellwo, 2017). Two groups of adult talkers (N=7 each) with relatively short or long mandibles (mean length group 1: 17cm, s.d.=0.5; group 2: 18.5cm, s.d.=0.4cm) produced consonant-vowel (CV) repetitions for at least five seconds at a comfortable and maximum speech rate. CV combinations consisted of a labial consonant and the back vowel /a/ (/ma/, /ba/, /fa/) to ensure a reasonable degree of mandibular movement. Electromagnetic articulography (EMA) was used to record jaw movements during mouth opening and closure. Articulatory kinematics and the speech amplitude envelope were analyzed within a time window of four seconds starting at production onset. Results showed that talkers with shorter mandibles produced syllables at a significantly higher rate than those with longer mandibles in the fast speech condition, while no differences were found between groups for the comfortable speech rate. Reconstruction of jaw kinematics revealed that talkers with longer mandibles needed considerably more time to raise their mandibles (mouth closure), which resulted in a significantly longer amplitude fall time. These findings suggest that the jaw anatomy influences the syllabic and supra-segmental timing of speech. Further analyses are required to determine whether motor control changes after reaching an articulatory target, but it seems likely that raising the mandible is affected by other factors such as its resonant frequency in fast speech (cf. He & Dellwo, 2017). Knowledge about anatomical properties encoded in the speech signal is relevant in explaining speaker-specific timing characteristics but may also play a role in understanding how segmental timing evolved in the world's languages (cf. Blasi et al., 2019).

References

Blasi, D. E., Moran, S., Moisik, S. R., Widmer, P., Dediu, D., & Bickel, B. (2019). Human sound systems are shaped by post-Neolithic changes in bite configuration. *Science*, 363(6432), eaav3218. https://doi.org/10.1126/science.aav3218

Fant, G. (1960). *Acoustic theory of speech production.* Mouton, The Hague, The Netherlands.

He, L. & Dellwo, V. (2017). Between-speaker variability in temporal organizations of intensity contours. *The Journal of the Acoustical Society of America*, 141(5), EL488–EL494. https://doi.org/10.1121/1.4983398

Back to top

Navigating interactions through vocal control: voluntary social trait expression in voices

Stella Guldner[1], Clare Lally[2], Nadine Lavan[3], Lisa Wittmann[4], Frauke Nees[5], Herta Flor[6], Carolyn McGettigan[2]

[1]Institute of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Mannheim, Germany
[2]Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK
[3]Department of Psychology, Queen Mary University of London, London, UK
[4]Institute of Psychology, University of Regensburg, Regensburg, Germany
[5]Institute of Medical Psychology and Medical Sociology, University Medical Center Schleswig Holstein, Kiel University, Kiel, Germany
[6]Institute of Cognitive and Clinical Neuroscience, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany

stella.guldner@zi-mannheim.de

The voice is a variable and dynamic social behaviour with functional relevance for conveying our identity, intentions and feelings to others. Inferences about personality traits of a speaker are drawn spontaneously (McAleer *et al.*, 2014) and reliably (Mahrholz *et al.*, 2018). These judgments can be influenced through intentional modulations of the voice (Hughes *et al.*, 2010), but it is unclear how specifically voice modulations evoke intended trait judgments, and whether voice modulations have implications in real-world social scenarios. In a series of three experiments, we investigated the mechanisms and efficacy of voluntary voice modulation. We recorded 40 healthy adult speakers during vocal expressions of six social traits (e.g. likeability, confidence) or while speaking in a neutral voice. In Study 1, we show that speakers' voice modulations evoke specific exaggerations of trait precepts in naïve-listeners relative to their neutral voice. The evoked trait ratings clustered on two principal components relating to perceived affiliation (likeability, hostility) and competence (confidence). Moreover, voluntary voice modulations were also clearly discriminable by naïve listeners, with high specificity and sensitivity (Study 2). Lastly, we tested whether voice modulations evoked implicit trait judgements that are relevant to real-world social scenarios. In Study 3, 40 naïve listeners were asked to choose a voice recording most suited for a given social scenario. Listeners were significantly more likely to match a relevant voice modulation to the social scenario over irrelevant voice modulations. That is, listeners matched the voices in line with the speaker's originally intended trait expression (e.g. a confident voice modulation to negotiate a promotion). These findings imply that intended voice modulations can evoke recognizable and specific trait impressions in listeners and that these impressions can be advantageous in navigating social interactions.

References

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS ONE*, *13*(10). https://doi.org/10.1371/journal.pone.0204991

Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The Perception and Parameters of Intentional Voice Manipulation. Journal of Nonverbal Behavior, 38(1), 107–127. https://doi.org/10.1007/s10919-013-0163-z

Back to top

Neural responses in human fusiform gyrus support a model of heteromodal representation of familiar speakers

Jasmine L. Hect[1], *Kyle Rupp*[1], *Ariane Rhone*[2], *Matthew Howard, IIIrd*[2], *Taylor J. Abel*[1,3]

[1] Department of Neurological Surgery, University of Pittsburgh, Pittsburgh PA
[2] Department of Neurosurgery, University of Iowa, Iowa City, IA
[3] Department of Bioengineering, University of Pittsburgh, Pittsburgh PA

jasmine.hect@pitt.edu

Voice and face processing are theorized to occur through convergent neural systems to facilitate speaker recognition. Prior neuroimaging studies suggest processing of a familiar voice engages the bilateral fusiform gyri (FG). However, it remains unknown what role the FG may play in voice processing, and at which point it becomes engaged during voice perception. The purpose of this study was to investigate neural responses to familiar voices and faces in human FG, using direct electrophysiologic recordings. In this exploratory analysis, we tested the hypothesis that neural populations in extrastriate visual cortex respond to familiar voice. Recordings were acquired from n=5 epilepsy surgery patients during a speaker identification task using visual and auditory stimuli of familiar speakers (U.S. presidents Barack Obama, George W. Bush, and Bill Clinton). Patients were presented with pictures of presidents or clips of their voices and asked to identify the portrait/speaker. Non-familiar faces were presented in a separate task as a post-hoc control, and a subset of patients (n=3) also completed a passive listening task with isolated words spoken by an unfamiliar speaker. We found that extrastriate visual cortex, including FG, exhibits responses to voice and establish the temporal dynamics of this response. These responses are relatively delayed (300-500ms) and predominately occur at sites responsive to visual stimuli of familiar faces. We found that responses to familiar voice were relatively lower in magnitude compared to familiar faces. In a subset of patients, we demonstrate these sites show responses to familiar voice, but not to words spoken by an unfamiliar talker. These findings suggest a model of heteromodal representation in extrastriate cortex, a region traditionally considered to be a part of the unimodal visual processing hierarchy. Furthermore, the task/stimulus dependence (naming of familiar voices) and latencies of voice responses in FG suggest a role in higher-order identity processing.

Back to top

Neural representations of naturalistic person identities while watching a feature film

Clare Lally[1], Nadine Lavan[2], Carolyn McGettigan[1]

[1]University College London
[2]Queen Mary University of London

c.lally@ucl.ac.uk

Recognising others is fundamental for many social interactions. Past research has mapped out brain regions implicated during face and voice identification, although this has often been based on tightly controlled experiments that reduce authentic aspects of person identification. We used representational similarity analysis to uncover neural representations of identity during task-free, naturalistic stimulation. We conducted our analyses on an open-access MRI dataset, in which participants watched feature length movies (Aliko et al., 2020). We hypothesised that regions representing person identity should produce similar patterns of activity in response to multiple instances of the same person, and dissimilar patterns based on instances of different people. This was observed in right hemisphere regions associated with face, voice and person perception. This was replicated across two independent groups of participants in response to different sets of identities. The final analysis dissociated contributions of face vs voice information to neural representations of identity, revealing areas of preferential sensitivity for each modality, with more extensive evidence of regions preferring face information. By simultaneously modelling sensitivity to between-person differences as well as within-person generalisation, we were able to take recent theoretical developments in person perception and apply these to cognitive neuroscience. Further, this work is one of the first to investigate the neural representations that underpin identity perception in the absence of an experimental task, enabling us to characterise how the brain represents information about other people in the real world.

References

Aliko, S., Huang, J., Gheorghiu, F., Meliss, S., & Skipper, J. I. (2020). A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data*, *7*(1), 1-21. https://doi.org/10.1038/s41597-020-00680-2

Back to top

Voice identity as predicted from the acoustic properties of fillers

Emily O'Hara[1], *Alessandro Vinciarelli* [1], *Phil McAleer* [1]

[1]University of Glasgow

e.ohara.1@research.gla.ac.uk

The human voice contains a multitude of social information, allowing listeners to rapidly infer speaker identity as well as impressions of characteristics such as personality and affective state (Belin *et al.,* 2011). Recently, perception research has moved towards more naturalistic representations of voices, with greater emphasis on within-speaker variability and socially relevant stimuli (Lavan *et al*., 2019). In this current study we focused on conversational fillers i.e., filled pauses such as 'uhm' (əm) and 'uh' (ʌh). These are highly common but understudied aspects of natural speech with minimal linguistic cues. We aimed to investigate if even these brief non-verbal utterances contained sufficient information for the formation of stable percepts. Little previous research has used fillers, but Kanber *et al.* (2020) found a strong effect of familiarity on recognition accuracy, illustrating that it is currently unclear in fillers which cues are necessary for recognition and perception. As a starting point, we sought to establish this on a basic acoustic level, by performing k-means clustering to determine whether it was possible for the algorithm to predict identity from acoustic properties. Using 10 fillers each from 93 speakers, we ran an acoustic analysis extracting information on 19 features for each filler. Entropy was then calculated on each participant's clustering result, which determined the amount of information contained within said result, based on the probability of its occurrence. Results yielded a mean entropy score of 0.426, SD = 0.091, which was statistically significant ($p < 0.001$) from a random distribution of 1. These results show that algorithmically it is possible to predict identity from fillers using their underlying acoustics. Furthermore, due to successful extraction of acoustic properties considered in the voice perception literature to be essential cues, the results suggest that stable personality and affective percepts should also be obtainable from fillers.

References

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding Voice Perception: Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

Kanber, E., Lavan, N., & McGettigan, C. (2022). Highly accurate and robust identity perception from personally familiar voices. *Journal of Experimental Psychology: General*, *151*(4), 897–911. https://doi.org/10.1037/xge0001112

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90–102. https://doi.org/10.3758/s13423-018-1497-7

Back to top

Incorporating a new auditory identity into the self-concept

Bryony Payne[1], Angus Addlesee[2], & Carolyn McGettigan[3]

[1] Department of Psychology, King's College London
[2] School of Mathematical and Computer Sciences, Heriot Watt University
[3] Department of Speech, Hearing and Phonetic Sciences, University College London

bryony.payne@kcl.ac.uk

Our voice is central to our self-identity; an important medium through which we express ourselves to others and achieve social and communicative goals. Here we ask whether it is possible to give people a *new* voice – a voice that can become processed as self-relevant and enable a sense of agency over its use. All participants (n=88) were given the chance to choose a new synthesised voice identity to own, and half the group (n=44) were further able to *use* their new voice in a real-life social interaction. Specifically, these participants were given flexible control over their new voice via text-to-speech synthesis, enabling them to play a communicative two-player game online. We used a perceptual matching paradigm to measure whether participants perceived their new voice as self-relevant and attributed self-bias to it. Further, we used a temporal binding paradigm to measure participants' sense of agency over their new voice. Finally, we assessed whether self-bias or sense of agency were modulated specifically by using the new voice in a social interaction, versus only owning the voice. The results show that participants attributed self-bias to their new voice identity, relative to the voice of another, and also showed an increased sense of agency over it. This suggests that it is possible to incorporate a new voice into the self-concept. Surprisingly, these effects were not modulated by whether the new voice had been used interactively to represent the self or not. Indeed, self-bias and sense of agency were similar between participants who only owned the voice and those who owned it and had also been able to use it. The results suggest that the fundamental knowledge of what is self-owned (i.e., what is 'mine') may be sufficient to generate self-bias and a sense of agency over a new auditory identity.

Back to top

Can voice recognizability be controlled by speakers?
A study on identity marked speech

Valeriia Perepelytsia[1]*, Leah Bradshaw*[1]*, Volker Dellwo*[1]

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

valeriia.perepelytsia@uzh.ch

Previous studies have shown that voice recognizability can be controlled by speakers. For example, they can deliberately change individuality cues to disguise their voices (Eriksson, 2010; Eriksson & Wretling, 1997) by either manipulating their own individuality cues (e.g., Hove & Dellwo, 2014), or by imitating those of other speakers (e.g., Kitamura, 2008). Therefore, speakers have some intuition regarding which cues are best manipulated to reveal less information about their voice identity. However, the evidence is limited regarding whether speakers are also capable of altering their vocal properties to reveal *more* information about their identity. One study by Dellwo et al. (2019) asked participants to speak to a mock automatic speech recognition system, eliciting more intelligible speech (*clear speech*) and to a mock automatic speaker recognition system, eliciting more recognizable speech (*identity-marked speech*). They found preliminary evidence that speech within these two styles differed and argued that speakers exhibit some control over cues to their voice identity and ability make themselves more recognizable.

This study offers an expansion of the work by Dellwo et al. (2019), in which we will further explore the phenomenon of *identity-marked speech*. We will present an updated experimental technique with improved ecological validity to make the task of talking to a mock speech and speaker recognition system more realistic. We will collect speech data in three speaking styles: read, clear, and identity-marked speech, which will be used to examine whether human and computer recognition performance benefits from identity-marked speech compared to clear and read speech. We will also present acoustic analyses of these speaking styles, as well as results of data clustering using t-distributed Stochastic Neighborhood Embedding method (van der Maaten & Hinton, 2008). These findings will contribute to the discussion of the extent to which speakers are able to manipulate their indexical vocal cues.

References

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.

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

Eriksson, A., Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. *Fifth European Conference on Speech Communication and Technology*.

Hove, I., Dellwo, V. (2014). The effect of voice disguise on f0 and on the formants. *Proceedings of IAFPA*.

Kitamura, T. (2008). Acoustic analysis of imitated voice produced by a professional impersonator. *Proceedings of Interspeech 2008*, 813-816.

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

Back to top

Hierarchical machine learning classifiers better predict source identity from marmoset vocalizations

Nikhil Phaniraj[1,2]*, Kaja Wierucka*[1]*, Yvonne Zürcher*[1]*, Judith M Burkart*[1]

[1]University of Zürich, [2]Indian Institute of Science Education and Research (IISER) Pune

nikhil.phaniraj@uzh.ch

Animals living in dense habitats or with poor vision that have little visual contact between group members often have to signal identity in their vocalizations along with other social, emotional, and contextual information (Boughman & Wilkinson, 1998; Fukushima et al., 2015; Prat et al., 2016; Soltis et al., 2005; Tooze et al., 1990). In such cases, it is possible that animals use broader-category cues that can convey information such as age or sex of the source for effective individual recognition. Determining source identity from vocalizations is important not only for receiver individuals but also for researchers studying these animals. As manual vocal data acquisition and real-time labeling by researchers is a cumbersome task, automizing this process is beneficial (Blumstein et al., 2011; Rickwood & Taylor, 2008). This has led to the development of a plethora of machine learning techniques for source identification, each competing for higher accuracy (Cheng et al., 2010, 2012; Stowell et al., 2019; Vieira et al., 2015). However, whether machine learning algorithms can use broader-category cues present in vocalizations for efficient source identification remains poorly explored. In this study, we used common marmosets, an arboreal primate species relying mainly on vocal cues to communicate, to test this hypothesis (Eliades & Miller, 2017). Because marmosets tend to live in smaller groups with an average of around eight individuals and have a social structure that includes helpers and breeders (Erb & Porter, 2017), it is plausible that most of the individual variability of calls within groups is explained by variation in sex and social structure. We found that a hierarchical classification approach in which the machine first learns to determine the social status, then sex, and finally the source identity outperformed the non-hierarchical classifier. This provided evidence that marmoset calls contain information about sex, social status, and individual identity of the caller. Finally, we assessed the impact of sample size on classifier accuracy and provided sample size guidelines for future studies. Hierarchical classifiers appear to be a promising tool for automatic source identification from animal vocalizations.

References

Blumstein, D. T., Mennill, D. J., Clemins, P., Girod, L., Yao, K., Patricelli, G., Deppe, J. L., Krakauer, A. H., Clark, C., Cortopassi, K. A., Hanser, S. F., McCowan, B., Ali, A. M., & Kirschel, A. N. G. (2011). Acoustic monitoring in terrestrial environments using microphone arrays: Applications, technological considerations and prospectus. *Journal of Applied Ecology*, *48*(3), 758–767. https://doi.org/10.1111/j.1365-2664.2011.01993.x

Boughman, J. W., & Wilkinson, G. S. (1998). Greater spear-nosed bats discriminate group mates by vocalizations. *Animal Behaviour*, *55*(6), 1717–1732.

Cheng, J., Sun, Y., & Ji, L. (2010). A call-independent and automatic acoustic system for the individual recognition of animals: A novel model using four passerines. *Pattern Recognition*, *43*(11), 3846–3852. https://doi.org/10.1016/j.patcog.2010.04.026

Cheng, J., Xie, B., Lin, C., & Ji, L. (2012). A comparative study in birds: Call-type-independent species and individual recognition using four machine-learning methods and two acoustic features. *Bioacoustics*, *21*(2), 157–171.

Eliades, S. J., & Miller, C. T. (2017). Marmoset vocal communication: Behavior and neurobiology. *Developmental Neurobiology*, *77*(3), 286–299.

Erb, W. M., & Porter, L. M. (2017). Mother's little helpers: What we know (and don't know) about cooperative infant care in callitrichines. *Evolutionary Anthropology: Issues, News, and Reviews*, *26*(1), 25–37.

Fukushima, M., Doyle, A. M., Mullarkey, M. P., Mishkin, M., & Averbeck, B. B. (2015). Distributed acoustic cues for caller identity in macaque vocalization. *Royal Society Open Science*, *2*(12), 150432.

Prat, Y., Taub, M., & Yovel, Y. (2016). Everyday bat vocalizations contain information about emitter, addressee, context, and behavior. *Scientific Reports*, *6*(1), 1–10.

Rickwood, P., & Taylor, A. (2008). Methods for automatically analyzing humpback song units. *The Journal of the Acoustical Society of America*, *123*(3), 1763–1772. https://doi.org/10.1121/1.2836748

Soltis, J., Leong, K., & Savage, A. (2005). African elephant vocal communication II: Rumble variation reflects the individual identity and emotional state of callers. *Animal Behaviour*, *70*(3), 589–599.

Stowell, D., Petrusková, T., Šálek, M., & Linhart, P. (2019). Automatic acoustic identification of individuals in multiple species: Improving identification across recording conditions. *Journal of The Royal Society Interface*, *16*(153), 20180940. https://doi.org/10.1098/rsif.2018.0940

Tooze, Z. J., Harrington, F. H., & Fentress, J. C. (1990). Individually distinct vocalizations in timber wolves, Canis lupus. *Animal Behaviour*, *40*(4), 723–730.

Vieira, M., Fonseca, P. J., Amorim, M. C. P., & Teixeira, C. J. (2015). Call recognition and individual identification of fish vocalizations based on automatic speech recognition: An example with the Lusitanian toadfish. *The Journal of the Acoustical Society of America*, *138*(6), 3941–3950.

Back to top

Same data, different results? An evaluation of the robustness of approaches used for establishing individual distinctiveness in mammalian acoustic cues

Kaja Wierucka[1,2], *Judith M. Burkart*[1,2]

[1]Department of Anthropology, University of Zurich
[2]The Swiss National Centre of Competence in Research (*NCCR*) *Evolving Language*

kaja.wierucka@uzh.ch

Complex, automated acoustic data analysis methods are becoming increasingly popular for the study of animal vocal communication. Establishing caller identity is a critical element of many behavioural studies and call classification models are used to demonstrate the ability of cues to convey individual identity information. This is often not only a basis for hypothesis testing through bioassays, but also a key factor in evaluating cognitive and communication abilities in many animal species. However, feature extraction methods as well as classifiers used vary greatly between studies making it impossible to accurately compare results across species, conduct comparative research and draw comprehensive conclusions of evolutionary significance. While previous studies have evaluated the accuracy with which models can predict caller identity, none have investigated the effect that data preparation and feature extraction have on the consistency of these results and whether certain combinations of processing and classification methods produce more reliable results. By incorporating data from multiple species (including both primates and non-primates), we evaluate the robustness and consistency of accuracy scores across different feature extraction methods (frequency measurements, mel-frequency cepstral coefficients, and highly comparative time-series analysis) and classifiers (discriminant function analysis, permutational multivariate analysis of variance, support vector machines, gaussian mixture models, neural networks and random forests). We also explore the effect of various data processing methods (balancing, outlier removal, normalisation and consequent sample size) on the robustness of obtained results. We identify trends that are generalisable across species and provide guidelines for the processing and analysis of mammalian vocalisations in relation to determining the individual distinctiveness of acoustic cues and establishing caller identity.

Back to top

Session 2

Neural responses in human superior temporal cortex support coding of voice representations

Taylor J. Abel[1,2], Jasmine L. Hect[1], Madison Remick[1], Avniel Ghuman[1], Bharath Chandrasekaran[3], Lori L. Holt[4], Kyle Rupp[1]

[1]Department of Neurological Surgery, University of Pittsburgh
[2]Department of Bioengineering, University of Pittsburgh
[2] Department of Communication Science and Disorders, University of Pittsburgh
[3]Department of Psychology, Carnegie Mellon University

abeltj@upmc.edu

The ability to recognize abstract features of voice during auditory perception is an intricate feat of human audition. For the listener, this occurs in near-automatic fashion to seamlessly extract complex cues from a highly variable auditory signal. Voice perception depends on specialized regions of auditory cortex, including superior temporal gyrus (STG) and superior temporal sulcus (STS). However, the nature of voice encoding at the cortical level remains poorly understood. We leverage intracerebral recordings across human auditory cortex during presentation of voice and non-voice acoustic stimuli to examine voice encoding at the cortical level, in eight patient-participants undergoing epilepsy surgery evaluation. We show that voice-selectivity increases along the auditory hierarchy from supratemporal plane (STP) to the STG and STS. Results show accurate decoding of vocalizations from human auditory cortical activity even in the complete absence of linguistic content. These findings show an early, less-selective temporal window of neural activity in the STG and STS followed by a sustained, strongly voice-selective window. Encoding models demonstrate divergence in the encoding of acoustic features along the auditory hierarchy, wherein STG/STS responses are best explained by voice category and acoustics, as opposed to acoustic features of voice stimuli alone. This is in contrast to neural activity recorded from STP, in which responses were accounted for by acoustic features. These findings support a model of voice perception that engages categorical encoding mechanisms within STG and STS to facilitate feature extraction.

Back to top

Sascha Frühholz

Department of Psychology, University of Oslo, Norway

Department of Psychology, University of Zurich, Switzerland

https://www.sv.uio.no/psi/english/people/aca/saschaf/

“May I activate your amygdala please” – Realtime modulation of the limbic brain system by live affective speech

Affect signaling in communication involves cortico-limbic brain systems for affect information decoding, such as expressed in a speaker’s vocal tone. To more realistically address the socio-dyadic and neural context of affective communication, we used a novel real-time neuroimaging setup that adaptively linked live speakers’ affective voice production with limbic brain signals in listeners as a proxy for affect recognition. We show that affective communication is acoustically more distinctive, adaptive, and individualized in dyadic than in non-dyadic settings and more efficiently capitalized on neural affect decoding mechanisms in limbic and associated networks. Only vocal affect produced in adaption to listeners~~'~~’ limbic signals was linked to emotion recognition in listeners. While live vocal aggression directly modulated limbic activity in listeners, live vocal joy modulated limbic activity in connection with neural pleasure nodes in the ventral striatum. This suggests that evolved neural systems for affect recognition are largely optimized for dyadic communicative contexts.

Back to top

Decoding attended talker solely from listening-state EEG signals

Mohamed Elminshawi, Julia Kostina, Emanuël A. P Habets, Neeraj Kumar Sharma

International Audio Laboratories Erlangen, Germany
(A joint institution of the Friedrich-Alexander-Universität Erlangen-Nürnberg and Fraunhofer IIS, Germany)

mohamed.elminshawi@audiolabs-erlangen.de

Evolutionarily, the human brain specializes in efficient reception, processing, and interpretation of sensory signals. Focusing on aspects of human listening, the brain extracts a multitude of information from an attended audio signal. The problem of decoding the attended audio source attributes while listening to an audio signal is commonly referred to as auditory attention decoding (AAD). Conventional auditory attention decoding (AAD) approaches require access to both EEG and individual audio source signals in order to decode the attended audio source attributes (O' Sullivan et al., 2015). In this work, we propose an alternate approach to AAD which is based on classifying attended audio source attributes using only the EEG signal. This enlarges the scope of AAD as such an approach does not require access to the individual audio source signals. In the study, we analyze two AAD tasks, namely, attended speaker decoding and attended direction decoding. We hypothesize that attended speaker and location signatures can be decoded directly from short-time EEG signals. We validate this hypothesis using two publicly available listening-state AAD datasets (Fuglsang et al., 2020, and Mundanad et al., 2021). We find that both speaker and spatial attention can be decoded with significantly above chance performance from EEG. This holds even when using short (0.1 - 0.5 sec) decision windows. We also find that in comparison to hearing impaired subjects, normal hearing subjects perform better on the speaker detection task. Furthermore, the duration as well as temporal distribution of the test data is found to significantly impact the classification performance. Our findings contribute towards furthering the understanding of attended speaker decoding from EEG signals, and open new questions on what aspects of the speech influence this decoding.

References

O'Sullivan, James A., Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund C. Lalor. "Attentional selection in a cocktail party environment can be decoded from single-trial EEG." *Cerebral cortex* 25, no. 7 (2015): 1697-1706.

S. A. Fuglsang, J. Märcher-Rørsted, T. Dau, and J. Hjortkjær, 'Effects of Sensorineural Hearing Loss on Cortical Synchronization to Competing Speech during Selective Attention', *J. Neurosci*., vol. 40, no. 12, pp. 2562–2572, Mar. 2020, doi: 10.1523/JNEUROSCI.1936-19.2020.

Narayanan, Abhijith Mundanad, Rob Zink, and Alexander Bertrand. "EEG miniaturization limits for stimulus decoding with EEG sensor networks." *Journal of Neural Engineering* 18, no. 5 (2021): 056042.

Back to top

Judith M. Burkart

Department of Anthropology, University of Zurich

https://www.aim.uzh.ch/de/members/professors/judithburkart.html

VoiceID in marmoset monkeys: Flexibility and trade-offs in vocal accommodation

Marmosets are highly voluble monkeys, renowned for the vocal flexibility. Even though their vocal repertoires are fixed, they engage in some vocal learning in the form of vocal accommodation, i.e. changes in the acoustic structure of their calls. We find that different captive colonies of marmosets have different “dialects”, and translocation experiments where animals are moved from one colony to the other show that these dialects are socially learned. Intriguingly, not all call types accommodate in the same way: long distance contact calls, for which signaling identity is crucial because the animals can not see each other, tend to accommodate less and without compromising individual recognizability of the calls. In contrast, for short distance contact calls, signaling identity is less important because individual recognition is warranted by visual and olfactory cues. These short distance contact calls accommodate more, which leads to a decrease in individual recognizability. These results suggest a trade-off in vocal accommodation, between the need to signal social closeness by becoming more similar to each other and the need to maintain individual recognizability. To further scrutinize these trade-offs, we have developed on the one hand more sensitive ML based approaches to analyze and classify marmoset vocalizations and currently test their generalizability to other primates and mammals. On the other hand, we follow vocal changes in wild marmosets in Brazil during migrations between groups, which allows us to better estimate the ultimate function of vocal accommodation and signaling individuality under natural conditions.

Back to top

Marta Manser

Department of Evolutionary Biology and Environmental Studies, University of Zurich, Switzerland

https://www.ieu.uzh.ch/en/staff/member/manser_marta.html

Selection levels on vocal individuality: strategic use or byproduct

In animals, large variation for vocal individuality between and within call types exist, yet we know little on what level selection is taking place. Identifying the selection pressures causing this variation in individuality will provide insight into the evolutionary relationships between cognitive and behavioral processes and communication systems, particularly in group-living species where repeated interactions are common. Analyzing a species' full, large vocal repertoire on individual signatures, its biological function, and the respective selection pressures is challenging. Here, we emphasize that comparing the acoustic individual distinctiveness between life-history stages and different subjects within a call type will allow the identification of selection pressures and enhance the understanding of variation in individuality and its potential strategic use by senders.

Back to top

Junichi Yamagashi

National Institute of Informatics, Japan
https://doi.org/10.1016/j.csl.2020.101114

Differences between human- and machine-based audio deepfake detection – analysis on the ASVspoof 2019 database

To automatically detect audio deepfake and prevent spoofing attacks, we have built a large corpus, ASVSpoof2019, which pairs natural human speech with speech waveforms generated by several types of synthesis algorithms. The speech synthesis methods are diverse and include text-to-speech synthesis and voice conversion.

In this talk, we will first present the results of large-scale listening tests conducted on this database to discriminate between natural and synthetic human speech. In the test, the subjects were asked to conduct two role-playing tasks. In one task, they were asked to judge whether the utterance was produced by a human or machine, given an imagined scenario where they must detect abnormal telephone calls in the customer service center of a commercial bank. In the other task, the subjects listened to two utterances and were asked to judge whether they sounded like the same person's voice.

Next, the results of several automatic detection algorithms for similar tasks on the same database are presented. Finally, the differences between human- and machine-based audio deepfake detection are discussed.

Back to top

Claudia Roswandowitz

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

https://www.suz.uzh.ch/cl/de/people/team/phonetics/roswandowitz.html

Do humans distinguish deepfake from real vocal identity? Insights from the perceptual and neurocognitive system

Deepfakes artificially re-create and manipulate original human data, with the main purpose of spreading social and political misinformation. High-quality deepfakes are viral ingredients of digital environments, and they can trick human cognition into misperceiving the fake as real. However, experimental research on how the human neurocognitive system processes deepfake information has been greatly neglected so far. In this talk, I will present perceptual and neuroimaging data on the sensitivity of the human brain to detect or be deceived by instances of deepfake voice identities. By using advanced deepfake technologies, we created voice identity clones that are acoustically like the natural human voices. During an identity recognition task, humans were mainly deceived by deepfake voice identities, but showed some remaining resources for deepfake detection. On the brain level, we identified a potential „deepfake sensor" including the subcortical ventral striatum, which assigns social reward to natural but not to deepfake identities, and sensory auditory cortex evaluating the acoustic degree of artificiality in human utterances. With our study, we present neurocognitive findings on the potential but also limitations of emerging deepfakes as artificial social signals for humans. Our findings highlight the relevance of the reward level of social cues for successful and effective human-computer interactions.

Back to top

Persuasive synthetic speech: voice perception and user behaviour

Mateusz Dubiel[1], *Pilar Oplustil Gallegos*[2], *Martin Halvey*[3], *Simon King*[2]

[1]University of Luxembourg, [2]University of Edinburgh, [3]University of Strathclyde

mateusz.dubiel@uni.lu

Previous research indicates that synthetic speech can be as persuasive as human speech (Stern et al., 1999). However, there is a lack of empirical validation on interactive goal-oriented tasks. In our two-stage study (Dubiel et al., 2020), which comprised of online listening test and interactive evaluation, we compared participants' perception of the persuasiveness of synthetic voices created from speech in a debating style (*IBM Debater* dataset (Mirkin et al., 2018)) vs. speech created from audio-books (*Libri TTS* dataset (Zen et al., 2019)). The goal of the first stage was to select the single most persuasive synthetic voice per dataset, thus providing a persuasive voice for the second stage, and a strong baseline voice for comparison. In the second stage (interactive evaluation), participants undertook a series of search tasks, interacting with a Conversational Agent (CA) to achieve the goal of selecting a flight. In each task, the CA attempted to persuade participants to change their original flight selection by providing counterarguments. We evaluated the persuasiveness of the CA via: (1) questionnaires adapted from Stern et al. (1999), covering perception of the message, voice, and personal qualities of the speaker; (2) the number of times a participant followed the recommendation of the CA. We found that participants who interacted with the CA using the voice created from the debating style speech rated it as significantly more truthful and more involved (qualities of persuasive speakers) than the CA using the audio-book-based voice. However, there was no difference in how frequently each group followed the CA's recommendations during interactive goal-oriented tasks. While our findings are preliminary and further experiments in different domains would be required to assess their validity, our proposed experiment is an important step towards more ecologically-valid evaluation of text-to-speech quality for systems designed to support decision making in goal-oriented tasks.

References

Dubiel, M., Halvey, M., Gallegos, P. O., & King, S. (2020). Persuasive synthetic speech: Voice perception and user behaviour. *In Proceedings of the 2nd Conference on Conversational User Interfaces* (pp. 1-9). https://doi.org/10.1145/3405755.3406120

Mirkin, S., Jacovi, M., Lavee, T., Kuo, H-K., Thomas, S., Sager, L., Kotlerman, L., Venezian, E.and Slonim, N. (2018). *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* https://aclanthology.org/L18-1037

Stern, S. E., Mullennix, J. W., Dyson, C. L., & Wilson, S. J. (1999). The persuasiveness of synthetic speech versus human speech. *Human Factors, 41(4),* 588-595. https://doi.org/10.1518%2F001872099779656680

Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., ... & Wu, Y. (2019). LibriTTS: A corpus derived from LibriSpeech for text-to-speech. *arXiv preprint arXiv:1904.02882*.

Back to top

Carolyn Mcgettigan

UCL Speech Hearing and Phonetic Sciences, UK

www.carolynmcgettigan.com
https://www.ucl.ac.uk/pals/people/carolyn-mcgettigan

Perceiving familiar voice identities

Identity perception from voices can be error-prone, and is generally thought to be inferior to face identity processing. While performance can improve with greater exposure to vocal identities, the continued popularity of "mystery voice" contests in radio broadcasts keenly demonstrates the fragility of our mental representations of even very well-known voices. In this talk, I will present some of our behavioural work investigating how different types of familiarity affect the accuracy of voice identity perception, particularly in the presence of perceptual challenges such as natural within-person variation, reduced verbal cues, and artificial modulations of vocal acoustics. Our findings indicate that familiarity is not a binary state but more likely reflects a continual process of developing a perceptual representation via greater experience with a voice. At its best - for example when hearing the voice of a close relative or partner - the identification of a person from the voice can in fact be highly accurate and robust. However, familiarity benefits do not generalise to all tasks – in a speech-in-noise recognition task, we found equivalent performance for personally familiar and unfamiliar targets.

Back to top

Behavioral and neural patterns underlying self-other voice discrimination

Pavo Orepic[1], Giannina Rita Iannotti[2,3], Oliver A. Kannape[1], Denis Brunet[2], Thomas Koenig[4], Nathan Faivre[5], Sixto Alcoba-Banqueri[1], Dorian F.A.Garin[3], Karl Schaller[3], Christoph M. Michel[2], Olaf Blanke[1]

[1]Laboratory of Cognitive Neuroscience, Center for Neuroprosthetics and Brain Mind Institute, Faculty of Life Sciences, École polytechnique fédérale de Lausanne (EPFL), 1202, Switzerland
[2]Functional Brain Mapping Lab, Department of Fundamental Neurosciences, University of Geneva, 1202, Switzerland
[3]Department of Neurosurgery, University Hospitals of Geneva and Faculty of Medicine, University of Geneva, 1205, Switzerland
[4]Translational Research Center,University Hospital of Psychiatry and Psychotherapy, University of Bern, Bern 3000, Switzerland
[5]University Grenoble Alpes, University Savoie Mont Blanc, CNRS, LPNC, Grenoble, France.

Pavo.orepic@unige.ch

There is growing evidence showing that the representation of the human "self" recruits special systems across different functions and modalities. Compared to self-face and self-body representations, few studies have investigated neural underpinnings specific to self-voice. Moreover, self-voice stimuli in those studies were consistently presented through air and lacking bone conduction, rendering self-voice stimuli different to the self-voice perceived during natural speech. Accordingly, factors that contribute to self-voice perception and enable its discrimination from other familiar and unfamiliar voices remain largely unknown. In a series of four studies, we combined psychophysics, voice-morphing technology, and high-density EEG in order to identify perceptual and neural patterns underlying self-other voice discrimination (SOVD), both with air- and bone-conducted stimuli. We demonstrate that specifically self-other but not familiar-other voice discrimination improved for stimuli presented using bone as compared to air conduction. Furthermore, our data outline independent contributions of familiarity and acoustic processing to separating own from another's voice: although vocal differences increased general voice discrimination, self-voices were more confused with familiar than unfamiliar voices, regardless of their acoustic similarity. Finally, we identified a self-voice-specific EEG topographic map occurring around 345 ms post-stimulus and activating a network involving insula, cingulate cortex, and medial temporal lobe structures (Iannotti & Orepic, 2021). Occurrence of this map was modulated both with SOVD task performance and bone conduction. Specifically, the better participants performed at SOVD task, the less frequently they activated this network. In addition, the same network was recruited less frequently with bone conduction, which, accordingly, increased the SOVD task performance. Collectively, our findings show that concomitant vibrotactile stimulation improves auditory self-identification, thereby portraying self-voice as a fundamentally multimodal construct. This work could have an important clinical impact. Indeed, it reveals neural correlates of SOVD impairments, believed to account for auditory-verbal hallucinations, a common and highly distressing psychiatric symptom.

References

Iannotti, G. R., Orepic, P., Brunet, D., Koenig, T., Alcoba-Banqueri, S., Garin, D. F. A., Schaller, K., Blanke, O., & Michel, C. M. (2021). EEG Spatiotemporal Patterns Underlying Self-other Voice Discrimination. *Cerebral Cortex*, bhab329. https://doi.org/10.1093/CERCOR/BHAB329

Back to top

Rapid pre-attentive voice recognition of a famous speaker: neural correlates of voice familiarity

Paula Rinke[1], Kjartan Beier[1], Ramona Kaul[1], Tatjana Schmidt[1], Mathias Scharinger[1]

[1]Institute of German Linguistics, Philipps-University Marburg

Email
paula.rinke@uni-marburg.de, beier_k@gmx.net, kaulr@students.uni-marburg.de,
tatjana.schmidt@hotmail.de, mathias.scharinger@staff.uni-marburg.de

Recognizing familiar voices and identifying speakers is a basic, yet remarkable human ability. Previous research has established a voice area in the right temporal cortex that helps extracting relevant acoustic features while listening to speech (Belin et al., 2004).

A previous Mismatch Negativity (MMN) study by Beauchemin et al. (2006) contrasted brain responses to personally familiar and unfamiliar voices and reported an effect of voice familiarity on voice processing seen in larger MMNs by familiar voices.

The present study investigates the neural patterns of voice familiarity processing for publicly, but not personally familiar voices. Therefore, a classic oddball paradigm contrasted two two-syllable German utterances (‘Kinder’ and ‘Tochter’) by the former German chancellor Angela Merkel with the same words being uttered by two unknown female speakers with matched age, regional background and voice quality. Angela Merkel can be considered a famous speaker publicly familiar to most German native speakers. Electroencephalogram was recorded from 32 active electrodes while 21 participants were presented with the different voices in standard or deviant position. Voice processing indices were quantified as MMNs and P3a differences, and cortical sources of both difference wave forms were estimated with variable resolution electromagnetic tomography.

The results showed differences in latency and amplitude for both MMN and P3a: Merkel’s voice elicited a smaller but earlier MMN than the control voices. The P3a, by contrast, was both larger and later in response to Merkel. The topographic distribution of the MMN in response to Merkel suggested right hemispheric activation, overlapping with the voice area, and activation in the left superior temporal gyrus for the control voices.

These results indicate that voice recognition is an automatic and pre-attentive process, occurring within the first 150 ms of the acoustic signal, yielding similar recognition patterns for famous compared to personally familiar voices.

References

Beauchemin, M., De Beaumont, L., Vannasing, P., Turcotte, A., Arcand, C., Belin, P., & Lassonde, M. (2006). Electrophysiological markers of voice familiarity. *European Journal of Neuroscience*, 23(11), 3081–3086. https://doi.org/10.1111/j.1460-9568.2006.04856.x

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Back to top

Practical analyses for real-time magnetic resonance (rtMRI) of voice production

Michel Belyk [1,2], *Carolyn McGettigan* [2]

[1]Edge Hill University, Department of Psychology, Ormskirk, UK
[2]University College London, Department of Speech, Hearing and Phonetic Sciences, London, UK

belykm@edgehill.ac.uk

Real-time magnetic resonance imaging (rtMRI) is a technique that provides high contrast videographic data of the vocal tract. These dynamic MRI scans capture 2D T1-weighted mid-sagittal slices that provide a rich description of complex movements inside the mouth and throat. These anatomical structures provide the physiological basis for behaviours such as speech, singing, expressions of emotion, and swallowing, which are otherwise not accessible for external observation. However, while this technology produces images of the dynamic morphological features of the vocal tract, taking quantitative measurements from these images is notoriously difficult. Traditionally, researchers have taken limited measurements of select morphological features, done so on a limited number of images, and in small sample sizes constrained by the extreme laboriousness of these procedures. We introduce a semi-automated processing pipeline that produces outlines of the vocal tract as a means to quantify the dynamic morphology of the vocal tract. Our approach uses simple tissue classification operating within researcher-specified constraints to facilitate feature extraction while retaining the involvement of a human analyst. We demonstrate that this pipeline generalises well to new datasets covering behaviours such as speech, vocal size exaggeration, laughter, and whistling as well as producing reliable outcomes across analysts. This approach provides considerable advantages in the ability to scale analyses to larger datasets relative to traditional analysis strategies. We make this pipeline available for immediate use by the research community, and further suggest that it may contribute to the continued development of fully automated methods based on deep learning algorithms.

Back to top

Using the visual world paradigm to explore voice identity processing

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

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).

The Visual World Paradigm (VWP – Allopenna *et al.*, 1998) is a popular eye-tracking experimental technique in psycholinguistic and phonetic research domains, used 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 recognition (Schindler & Reinisch, 2015).

Given its capacity to assess online processing, the VWP represents a viable technique for exploring, for instance, 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 useful 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 as a method for exploring voice identity processing.

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

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9–26. https://doi.org/10.1016/S0167-6393(99)00028-X

Schindler, C. & Reinisch, E. (2015). Tracking the temporal relation between speaker recognition and processing of phonetic information. *Proceedings of the 18th International Congress of Phonetic Sciences.* Glasgow, UK.

Back to top

Comparison of machine learning methods for the vocal identification of meerkats (*Suricata suricatta*)

Alessandro De Luca[1,2], *Ariana Strandburg-Peshkin*[3,4], *Britta Walkenhorst*[1], *Marta Manser*[1,2]

[1]Department of Evolutionary Biology and Environmental Studies, University of Zurich

[2]Kalahari Meerkat Project, Kuruman River Reserve

[3]Department for the Ecology of Animal Societies, Max Planck Institute of Animal Behavior

[4]Department of Biology, University of Konstanz

alessandro.deluca@uzh.ch

The topic of individuality in animals has long been at the forefront of animal behaviour research (Clark & Ehlinger, 1987). Individuality can serve many functions and is particularly crucial in cooperation (Beecher, 1982; Charrier et al., 2002). One of the widespread modalities of signalling and recognizing individual signatures is through vocal communication (Owings et al., 1998). Vocal individuality signalling is thought to be the basis of complex communication systems in which repeated interactions occur between numerous individuals (Beecher, 1982; Cheney & Seyfarth, 2018). Meerkats are a cooperatively breeding mongoose species which use a large vocal repertoire to coordinate activities with group members (Clutton-Brock & Manser, 2016). Due to their sociality, cooperative system, and common usage of vocal communication, meerkats are an ideal species to investigate vocal individuality and automatic individual recognition. Here, I compare how several machine learning algorithms perform on the task of identifying individual meerkats from their vocalizations, based on two different types of acoustic features. The best classification performance is obtained by using mel-spectrogram images with a random forest classifier. In addition, context may have an influence on the classification performance. The individuals' age at which the data was gathered instead did not influence the performance significantly in a train-on-one-age-class and test-separately-on-all-age-classes paradigm, indicating that the individual signature is stable throughout lifetime. Automatic individual identification through vocalizations can have many practical applications in conservation, such as accurate passive monitoring and tracking at the individual level (Terry et al., 2005; Wijers et al., 2020). My results support the possibility of using a machine learning approach for passive acoustic monitoring of meerkats, but to achieve such a generalized framework more analyses investigating the influence of biological and technical factors on the classification performance are needed.

References:

Beecher, M. D. (1982). Signature systems and kin recognition. *Integrative and Comparative Biology*, 22(3), 477–490. https://doi.org/10.1093/icb/22.3.477

Charrier, I., Mathevon, N., & Jouventin, P. (2002). How does a fur seal mother recognize the voice of her pup? An experimental study of Arctocephalus tropicalis. *Journal of Experimental Biology*, 205(5), 603–612.

Cheney, D. L., & Seyfarth, R. M. (2018). *How monkeys see the world: Inside the mind of another species*. University of Chicago Press.

Clark, A. B., & Ehlinger, T. J. (1987). Pattern and Adaptation in Individual Behavioral Differences. In P. P. G. Bateson & P. H. Klopfer (Eds.), *Perspectives in Ethology* (pp. 1–47). Springer. https://doi.org/10.1007/978-1-4613-1815-6_1

Clutton-Brock, T. H., & Manser, M. B. (2016). Meerkats: cooperative breeding in the Kalahari. *Cooperative Breeding in Vertebrates*, 294, 317.

Owings, D. H., Morton, E. S., & others. (1998). *Animal vocal communication: a new approach*. Cambridge University Press.

Terry, A. M. R., Peake, T. M., & McGregor, P. K. (2005). The role of vocal individuality in conservation. *Frontiers in Zoology*, *2*, 1–16. https://doi.org/10.1186/1742-9994-2-10

Wijers, M., Trethowan, P., du Preez, B., Chamaillé-Jammes, S., Loveridge, A. J., Macdonald, D. W., & Markham, A. (2020). Vocal discrimination of African lions and its potential for collar-free tracking. *Bioacoustics*, *00*(00), 1–19. https://doi.org/10.1080/09524622.2020.1829050

Back to top

Developing a challenging speaker discrimination test

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.froelich@uzh.ch, volker.dellwo@uzh.ch, meike.ramon@unil.ch

In the field of vision sciences, so-called “Super-Recognizers” (SRs) - individuals with exceptional face identity processing abilities - have been described (Russell et al., 2009) and receive increasing interest from international law enforcement (Ramon, 2021). Whether superior abilities exist in the auditory domain remains an open question. We are currently developing highly challenging tests of voice processing.

Identifying voice SRs has potential impact for forensic phonetics, where practitioners are charged with increasing amounts of data. Traditional auditory and acoustic-phonetic methodsare time-consuming and therefore not suitable to perform large numbers of speaker comparisons for police investigations where time is limited. Automatic voice comparison systems can rapidly complete such tasks. However, they are not reliable if the audio quality ispoor or samples are of short duration. Therefore we seek to identify individuals with exceptional voice processing skills to assist forensic phoneticians in such scenarios.

An voice SR would excel at voice processing tasks such as discrimination and recognition. To find people with extraordinary voice processing skills we start by first creating a challenging discrimination task. We employed a pre-trained ECAPA-TDNN model (Ravanelli*et al.* 2021), and combined its embeddings with F0 deltas of the speaker pairs to find the mostsimilar stimuli per trial. The pilot study was run with participants from different classes of police cadets using stimuli from the TEVOID-Corpus (Dellwo *et al.* 2012).

Our results show that in a discrimination task with randomly combined pairs, participants on average correctly classified 83.3% of the pairs, whereas in the challenging discrimination taskonly 69.5% of the pairs were classified correctly. These results show that we have successfully established a method to create challenging discrimination tasks to be used in the search for auditory SRs. We are currently exploring further test scenarios such as recognition and clustering.

References

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

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

Ravanelli, M. *et al.* (2021): SpeechBrain: A General-Purpose Speech Toolkit, *arXiv:2106.04624.*

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

Back to top

Intracerebral investigation of the neural representation of voice in human auditory cortex using voice-like acoustic stimuli

Jasmine Hect[1], Kyle Rupp[1], Emily Harford[1], Kanupriya Gupta[1], Hope Reecher[1], Frederic Dick[2], Lori Holt[4], Taylor Abel[1,5]

[1]Department of Neurological Surgery, University of Pittsburgh
[2]Department of Psychological Sciences, Birkbeck College
[3]Department of Experimental Psychology, UCL
[4]Department of Psychology, Carnegie Mellon University
[5]Department of Bioengineering, University of Pittsburgh

Jasmine.hect@pitt.edu

Voice perception engages characteristic regions of auditory cortex. However, it remains unclear to what extent these regions rely on shared or unique mechanisms for processing voice and non-voice sounds (1–5) and for assessing for sound patterns specific to voice (6–9). We used direct intracerebral recordings (from auditory cortex of four patient-participants with epilepsy undergoing chronic monitoring) to test the hypothesis that temporal voice areas (TVAs) rely on shared representations of voice and other natural sounds across superior temporal gyrus (STG) and superior temporal sulcus (STS). Data were recorded while participants listened to 1) voice and non-voice stimuli (Voice Localizer (10)) and 2) synthetic sounds generated from modulated noise (11), called Gaussian Sound Patterns (GSPs). The GSPs stimuli mirror spectrotemporal features of natural sounds, while remaining perceptually distinct, in line with prior fMRI work (1). We used a CNN to classify GSPs into sound categories (12) and selected stimuli most and least likely to be classified as voice (250 total). We extracted broadband high-gamma activity (HGA; 70-150 Hz) and identified sound-responsive channels (two-sample t-test, FDR-corrected, $q < 0.01$). We tested decoding (80% train, 20% test) of stimulus category from HGA prior to cross-task decoding, used to examine similarities in the neuronal representation between voice and GSPs within sound-responsive channels. Decoding of voice from non-voice was significant for all patients (61-76%, $p < 0.001$). Decoding accuracy of GSPs was significantly above chance for two patients (63% and 68% accuracy, $p < 0.001$), and nonsignificant for two (46% and 55% accuracy). We found cross-task decoding did not perform above chance when GSPs (41-48%) or Voice Localizer (41-52%) were used as the training set. These preliminary data suggest TVAs may employ unique representations of voice, even when spectrotemporal properties are controlled between artificial and natural sound stimuli.

References

1. Staib M, Frühholz S. Cortical voice processing is grounded in elementary sound analyses for vocalization relevant sound patterns. *Prog Neurobiol*. 2021 May 1;200:101982.
2. Norman-Haignere S, Kanwisher NG, McDermott JH. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*. 2015 Dec 16;88(6):1281–96.
3. Peretz I, Vuvan D, Lagrois MÉ, Armony JL. Neural overlap in processing music and speech. *Philos Trans R Soc B Biol Sci*. 2015 Mar 19;370(1664):20140090.
4. Zatorre RJ, Belin P, Penhune VB. Structure and function of auditory cortex: music and speech. *Trends Cogn Sci*. 2002 Jan 1;6(1):37–46.
5. Staib M, Frühholz S. Distinct functional levels of human voice processing in the auditory cortex. *Cereb Cortex*. 2022 Mar 26;bhac128.
6. Chevillet M, Riesenhuber M, Rauschecker JP. Functional Correlates of the Anterolateral Processing Hierarchy in Human Auditory Cortex. *J Neurosci*. 2011 Jun 22;31(25):9345–52.

7. Hickok G, Poeppel D. The cortical organization of speech processing. *Nat Rev Neurosci*. 2007 May;8(5):393–402.

8. Staeren N, Renvall H, De Martino F, Goebel R, Formisano E. Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Curr Biol*. 2009 Mar 24;19(6):498–502.

9. Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud AL. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat Commun*. 2014 Sep 2;5(1):4694.

10. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. *Nature*. 2000 Jan;403(6767):309–12.

11. McDermott JH, Wrobleski D, Oxenham AJ. Recovering sound sources from embedded repetition. *Proc Natl Acad Sci*. 2011 Jan 18;108(3):1188–93.

12. Gemmeke JF, Ellis DPW, Freedman D, Jansen A, Lawrence W, Moore RC, et al. Audio Set: An ontology and human-labeled dataset for audio events. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017. p. 776–80.

Back to top

A comparative study of automatic classifiers to recognize speakers based on fricatives

Nina Hosseini-Kivanani[1], *Homa Asadi*[2&3], *Christoph Schommer*[1], *Volker Dellwo*[3]

[1]University of Luxembourg, [2]University of Isfahan, [3]University of Zurich

Nina.hosseinikivanani@uni.lu, h.asadi@fgn.ui.ac.ir, Christoph.schommer@uni.lu,
volker.dellwo@uzh.ch

Speakers' voices are highly individual and for this reason speakers can be identified based on their voice. Nevertheless, voices are often more variable within the same speaker than they are between speakers, which makes it difficult for humans and machines to differentiate between speakers (Hansen, J. H., & Hasan, T., 2015). To date, various machine learning methods have been developed to recognize speakers based on the acoustic characteristics of their speech; however, not all of them have proven equally effective in speaker identification, and depending on the obtained techniques, the system achieves a different result. Here, different machine learning classifiers have been applied to identify the best classification model (i.e., Naïve Bayes (NB), support vector machines (SVM), random forests (RF), & k-nearest neighbors (KNN)) for categorizing 4 speaking styles based on the segment types (voiceless fricatives) considering acoustic features of center of gravity, standard deviation, and skewness. We used a dataset consisting of speech samples from 7 native Persian subjects speaking in 4 different speaking styles: read, spontaneous, clear, and child-directed speech. The results revealed that the best performing model to predict the speakers based on the segment type was RF model with an accuracy of 81,3%, followed by SVM (76.3%), NB (75.4%), and KNN (74%) (Table 1). Our results showed that the RF performed the best for voiceless fricatives /f/, /s/, and / ʃ / which may indicate that these segments are much more speaker-specific than others (Gordon et al., 2002), and the model performance was low for the voiceless fricatives of /h/ and /x/. Performance can be seen in the confusion matrix (Figure 1), which produced high precision and recall values (above 80%) for /f/, /s/ and / ʃ / (Table 2). We found that the model performance improved when the data related to clear speaking style; the information in individual speakers (i.e., voiceless fricatives) are more distinguishable in clear style than other styles (Table 1).

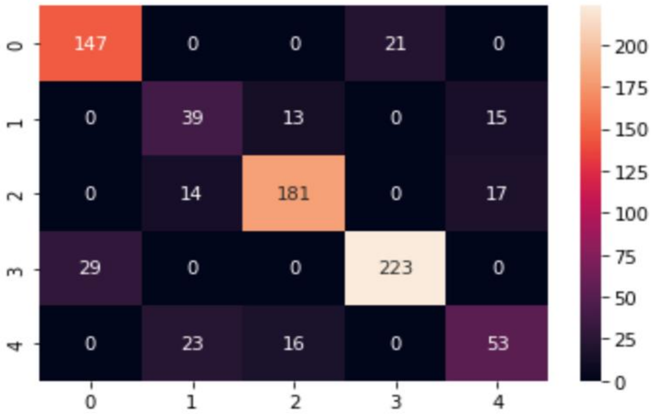
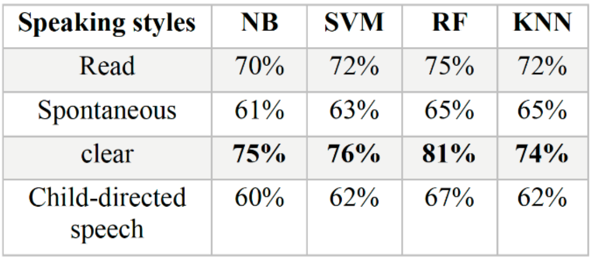


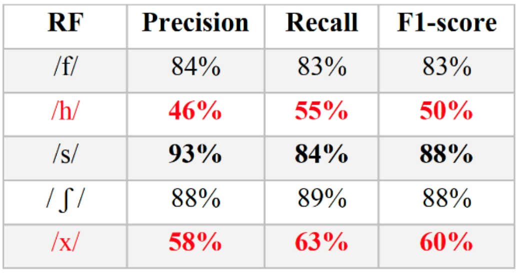
Figure 1. Confusion Matrix of RF (0: f, 1: h, 2: s, 4: x).

Table 1. The output of ML models per speaking styles.



Speaking styles	NB	SVM	RF	KNN
Read	70%	72%	75%	72%
Spontaneous	61%	63%	65%	65%
clear	**75%**	**76%**	**81%**	**74%**
Child-directed speech	60%	62%	67%	62%

Table 2. The output of RF models per segment types: clear speaking style.



RF	Precision	Recall	F1-score
/f/	84%	83%	83%
/h/	**46%**	**55%**	**50%**
/s/	**93%**	**84%**	**88%**
/ ʃ /	88%	89%	88%
/x/	**58%**	**63%**	**60%**

References

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

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. *Journal of the International Phonetic Association*, 32(2), 141-174.

Back to top

Examining how the amount of training exposure affects recognition of voice identities

Elise Kanber[1], Carolyn McGettigan[1]

[1]University College London

Elise.Kanber@ucl.ac.uk

In order to fully learn a voice, a listener must be able to recognise it in all its variations. Previous voice learning studies have found that listeners can accurately recognise new voice identities after a short amount of training, provided the conditions are roughly the same at training and test (Holmes et al., 2021). Yet recognition of lab-trained voices has been found to be fairly unstable, with discrimination and identification easily disrupted by changes in speaking style (Lavan et al., 2016; Lavan et al., 2019). Further, we recently showed a substantial disadvantage for recognising lab-trained voices compared to personally familiar (i.e., romantic partner) voices when they are acoustically modulated (Kanber et al., 2021). However, in our study, lab-trained familiarity was based on a relatively short amount of exposure to that identity during training – it may be that a greater degree of training might have allowed the listeners to form more stable identity representations. Here, I will present work that addressed this issue. We manipulated the amount of training listeners received when learning lab-trained voices and tested their subsequent abilities to recognise acoustically modulated and unmodulated samples of those voices. Participants were trained with either 20 or 80 stimuli per voice, and recognition was tested on excerpts with no modulation, or a range of adjustments to glottal pulse rate (GPR) and apparent vocal tract length (VTL). The results showed that increasing exposure can produce overall improved accuracy for acoustically modulated voices and reduce the reliance on low-level acoustic cues. This suggested that even a small amount of additional exposure may be beneficial for forming voice representations. Yet the findings also provided some evidence that task demands affected the precise patterns observed. We thus concluded that representations of lab-trained voices remain far inferior to those observed for personally familiar ones.

References:

Holmes, E., To, G., & Johnsrude, I. S. (2021). How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to voice training. *Psychological Science*, *32*(6), 903-915. https://doi.org/10.1177/0956797621991137

Kanber, E., Lavan, N., & McGettigan, C. (2021). Highly accurate and robust identity perception from personally familiar voices. Journal of Experimental Psychology: General. Advance online publication. https://doi.org/10.1037/xge0001112

Lavan N., Scott S.K. & McGettigan C. (2016). Impaired generalization of speaker identity in familiar and unfamiliar voices. Journal of Experimental Psychology: General, 145(12), 1604-1614. https://doi.org/10.1037/xge0000223

Lavan N., Burston, L.F.K., Merriman, S.E., Ladwa P., 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

Back to top

Consistency and bias: characterizing individual variability in the production of American English /æ/ and /ɑ/

Carolina Lins Machado[1], *Lei He*[1,2]

[1]Department of Computational Linguistics, Zurich University, Zurich, Switzerland
[2]Department of Phoniatrics and Speech Pathology, Clinic for Otorhinolaryngology, Head and Neck Surgery, University Hospital Zurich (USZ), Zurich, Switzerland

cmachado@ifi.uzh.ch, lei.he@uzh.ch

Differences in individual behavior can be evaluated by describing the outcome of movements in relation to a reference target or to another performer in the same environmental condition (Schmidt et al. 2018). Error scores are a set of measures which can assess bias and consistency of movement outcomes. Constant Error (CE) relates to performance tendency indicating the amount of deviation in movement relative to a target; Variable Error (VE) measures the inconsistency in movement outcome, indicating how precise individuals are in their performance (Henry, 1974). Combined, these measures can characterize variation in speakers' behaviors. In this study, we computed CE and VE scores of tongue blade and dorsum kinematic variables (TBy, TBx, TDy, TDx) and the first two formants (F1, F2) measured at five equidistant points to investigate differences in the production of the vowels /æ/ and /ɑ/ by 20 native speakers of American English selected from the EMA-MAE corpus (Ji et al., 2014). Preliminary results of CE and VE scores of acoustic and kinematic variables revealed that variation in error scores seems to be larger in the acoustic than in the articulatory dimension. Furthermore, for F1, TBy, TDy, and TDx there was a significant effect of vowel in CE scores, which were less dispersed in /æ/, indicating that speakers tended to remain closer to the mean values of this vowel and were, therefore, more target oriented in their production of /æ/. Differences between speakers are less straight-forward, however error scores of the kinematic and acoustic variables indicated that some speakers tended to be consistently short of the target (low VE and negative CE scores), while others were inconsistently overshooting the target goal (high VE and positive CE scores). Ultimately, error scores are valuable tools to characterize speakers' tendencies and consistencies in speech production.

References

Henry, F. M. (1974). Variable and Constant Performance Errors Within a Group of Individuals. *Journal of Motor Behavior*, *6*(3), 149–154. https://doi.org/10.1080/00222895.1974.10734991

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. https://doi.org/10.1109/ICASSP.2014.6855102

Schmidt, R. A., Lee, T. D., Winstein, C., Wulf, G., & Zelaznik, H. N. (2018). *Motor Control and Learning: A Behavioral Emphasis*. Human Kinetics.

Back to top

The impact of voice recognition skills on earwitness testimony

Sascha Schäfer, Paul Foulkes

University of York

sascha.schaefer|paul.foulkes@york.ac.uk

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

The present study complements these findings with an analysis of inter-listener differences in voice recognition, which cannot be controlled by an investigator ("estimator variables"). Psychological tests for assessing voice recognition have already shown participant performances ranging from developmental phonagnosia to 'super recognition' (Aglieri et al., 2017; Mühl et al., 2018). It is, however, unclear whether these results translate to earwitnesses, as the stimuli were created from isolated sounds/syllables rather than naturalistic speech. The present study addresses this problem:

100 British participants (50 male, mean age = 36, SD = 13.8) took part in an AX discrimination task hosted on Pavlovia. For the stimuli, two 10s-long recordings were taken from 48 speakers of the DyVis corpus (Nolan et al., 2009). Three stimulus lists of comparable difficulty were created based on the f0-difference between speakers. Participants were assigned to one of the stimulus lists and provided a same/different rating for 32 voice pairs (16 same), while reaction times were measured. They also reported their confidence (6pt-scale). Participants differed markedly in recognition accuracy (range 50-93.8%, mean =75%, SD = 9.1%), including two 'super-recognisers' (>= 2 SDs above mean) and four participants at the opposite end of the spectrum (<= 2 SDs below mean). The index *d prime* revealed high differences in listener discriminability (range 0-2.94, mean = 1.38, SD = 0.57). The results indicate that earwitnesses might not be equally suited for a standardised VP.

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

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

Back to top

Spectral Moments as a source of speaker discriminant information

Nikita Suthar[1], *Peter French*[2]

University of York, UK

nikita.suthar|peter.french@york.ac.uk

Forensic speaker discriminant studies aim to determine features and combinations of features that can differentiate speakers from one another. Formant analysis has previously figured in such studies (Cao & Dellwo, 2019; Jessen, 2008; Nolan F., 1983; Rose, 2002). The current work takes formant analysis further by adding spectral moments to formant centre frequency values, - which have been the mainstay of much previous research to increase the discriminant power of formants (Nittrouer, 1995, Forrest et al.,1988). The study has centred round value of four primary spectral moments, each concerning the distribution of within-formant energy: centre of gravity, standard deviation of the energy variance across the spectrum, skewness, and kurtosis.

Forty-five female Marwari monolinguals from the Bikaner district (India) were recruited. The recordings were collected from spontaneous and non-spontaneous speech and focused on 8 different vowels. Three types of data were collected; (i) a list of 80 words (10 tokens per vowel) that the participants were asked to read aloud; (ii) a picture description task; (iii)free conversation where two participants were paired up and asked to have a natural conversation on a topic of their choice or from a provided list. Marwari language was used as a testbed and in theory the analysis could be carried out with any other language.

Spectral moments were extracted with the help of a Praat script. An ANOVA conducted in R showed significant vowel differences. Table 1 summarises the discriminant analysis results and the classification rates of the individual features for the wordlist and story data. All investigated spectral measures increased the classification rates by a minimum factor of 2.5, suggesting that spectral moments carry valuable speaker-specific information. While the results show a clear pattern, the exact relationship between spectral moments and human vocal tract needs further exploration.

Table 1. Discriminant Analysis of Spectral Moments and Centre Formant Frequencies for Wordlist and Story Data

Acoustic Measures	**Wordlist**		**Story**	
	Classification Rate	***Times above chance***	***Classification Rate***	***Times above chance***
F1+F2+F3+F4	15%	6.5 times	11%	4.5 times
COG: F1-F4	15%	6.5 times	12%	5 times
SD 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

References:

Cao, H., & Dellwo, V. (2019). The role of the first five formants in three vowels of mandarin for forensic voice analysis. Melbourne: *International Congress of Phonetic Sciences*, 617-621.Retrieved June 2021, from University of Zurich: https://www.zora.uzh.ch/id/eprint/177494/

Forrest, K., Weismer, G., Milenkovic, P., & Dougall, R. N. (1988). Statistical analysis of word-initial voiceless obstruents: Preliminary data. *The Journal of the Acoustical Society of America*, 84(1), 115–123.

Jessen, M. (2008). Forensic phonetics. Language and Linguistics Compass, *2*(4), 671–711.

Nittrouer, S. (1995). Children learn separate aspects of speech production at different rates: Evidence from spectral moments. *Journal of the Acoustical Society of America*, *97*(1), 520–530. https://doi.org/10.1121/1.412278

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

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

Back to top

Identifying speaker specific properties in Russian fricatives

Natalja Ulrich, Marc Allassonnière-Tang, François Pellegrino

Laboratoire Dynamique Du Langage (DDL) UMR 5596, CNRS/Université Lyon 2, Lyon, France
Laboratoire Eco-Anthropologie (EA) UMR 7206, CNRS/MNHN/Université Paris Cité, Paris, France

natalja.ulrich@univ-lyon2.fr

Phonetic research on speaker-specific information assumes that idiosyncratic characteristics are reflected in the physical properties of speech sounds (He & Dellwo, 2014) and can be exploited by listeners, and through the extraction of acoustic cues (Dellwo et al., 2007). Investigating fricative sounds, studies found considerable inter speaker variation, suggesting their further exploration (Gordon et al., 2002; C. Kavanagh, 2011; C. M. Kavanagh, 2012; Narayanan et al., 1995; Schindler & Draxler, 2013; Silbert & de Jong, 2008).

On the example of the Russian fricatives /f/, /s/, /S/, /v/, /z/, /Z/, /sj/, /Sj/, produced by 58 native speakers of Russian, the current study explores the question of how much speaker-specific information is carried by noise sounds. First, we identify gender-specific acoustic properties. Second, we zoom into the individual level. To do so, acoustic speech features, such as temporal, spectral, and harmonicity measurements and 13 Mel Frequency Cepstral Coefficients (MFCCs) are extracted. The statistical mean, median, and standard deviation are also computed.

Applying various statistical and machine learning methods (such as decision-tree based algorithms), the following overall patterns are detected: i) significant gender and individual differences are to a certain degree sound and cue specific, with alveolar fricatives carrying the most idiosyncratic differences; ii) speakers differ significantly in their productions and variability, with female speaker producing higher spectral energy as well as more variation across the production of the same token; iii) females produce less distance between, for instance fricative categories /f/ and /s/, but a greater distance between /s/ and /S/; iv) applying machine learning models, speakers gender can be identified by acoustic cues with high accuracy for acoustic speech feature and MFCCs.

Besides its academic contribution, the investigation enhances our understanding of the distribution of linguistic and speaker idiosyncratic information, which contributes to the development of Automatic Speech Recognition (ASR) systems and the assessment of the impact of individual variation in linguistic research.

References

Dellwo, V., Huckvale, M., & Ashby, M. (2007). How Is Individuality Expressed in Voice? An Introduction to Speech Production and Description for Speaker Classification. Speaker Classification I, 4343, 1–20. https://doi.org/10.1007/978-3-540-74200-5_1

Gordon, M., Barthmaier, P., & Sands, K. (2002). A cross-linguistic acoustic study of voiceless fricatives. Journal of the International Phonetic Association, 32(2), 141–174. https://doi.org/10.1017/S0025100302001020

He, L., & Dellwo, V. (2014). Speaker idiosyncratic variability of intensity across syllables. Interspeech 2014, 233–237. https://doi.org/10.21437/Interspeech.2014-59

Kavanagh, C. (2011). Intra- and inter-speaker variability in acoustic properties of English /s/. The Journal of the Acoustical Society of America, 130, 2519. https://doi.org/10.1121/1.3655046

Kavanagh, C. M. (2012). New consonantal acoustic parameters for forensic speaker comparison. Unpublished doctoral dissertation: University of York.

Narayanan, S. S., Alwan, A. A., & Haker, K. (1995). An articulatory study of fricative consonants using magnetic resonance imaging. The Journal of the Acoustical Society of America, 98(3), 1325–1347. https://doi.org/10.1121/1.413469

Schindler, C., & Draxler, C. (2013). Using spectral moments as a speaker specific feature in nasals and fricatives. Interspeech 2013, 2793–2796. https://doi.org/10.21437/Interspeech.2013-639

Silbert, N., & de Jong, K. (2008). Focus, prosodic context, and phonological feature specification: Patterns of variation in fricative production. The Journal of the Acoustical Society of America, 123(5), 2769–2779. https://doi.org/10.1121/1.2890736

Back to top

Speech timing cues reveal deceptive speech: an acoustic analysis of communication in social deduction board games

Ziyun Zhang[1], *Carolyn McGettigan*[1], *Michel Belyk*[2]

[1]Department of Speech, Hearing and Phonetic Sciences, University College London, London, United Kingdom
[2]Department of Psychology, Edge Hill University, Ormskirk, United Kingdom

Ziyun.zhang.19@ucl.ac.uk

The faculty of language allows humans to state falsehoods in their choice of words. However, while what is said might easily uphold a lie, *how* it is said may reveal deception. Hence, some features of the voice that are difficult for liars to control may keep speech mostly, if not always, honest. Previous research has identified that speech timing (Anolli & Ciceri, 1997) and voice pitch cues (DePaulo *et al*., 2003) can predict the truthfulness of speech, but this evidence has come primarily from laboratory experiments, which sacrifice ecological validity for experimental control. We obtained ecologically valid recordings of deceptive speech while observing natural utterances from players of a popular social deduction board game, in which players are assigned roles that either induce honest or dishonest interactions. When speakers chose to lie, they were prone to longer and more frequent pauses in their speech. This finding is in line with theoretical predictions that lying is more cognitively demanding (Zuckerman *et al*., 1981). However, lying was not reliably associated with vocal pitch. This contradicts predictions that increased physiological arousal from lying might increase muscular tension in the larynx (Patel *et al*., 2011), but is consistent with human specializations that grant *Homo sapiens sapiens* an unusual degree of control over the voice relative to other primates (Belyk & Brown, 2017). The present study demonstrates the utility of social deduction board games as a means of making naturalistic observations of human behavior from semi- structured social interactions.

References

Anolli, L., and Ciceri, R. (1997). The voice of deception: vocal strategies of naive and able liars. *Journal of Nonverbal Behavior, 21*(4), 259–284.

Belyk, M., and Brown, S. (2017). The origins of the vocal brain in humans. Neuroscience and Biobehavioral Reviews, *77*, 177–193.

DePaulo, B.M., Malone, B.E., Lindsay, J.J., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129*(1), 74–118.

Patel, S., Scherer, K.R., Björkner, E., and Sundberg, J. (2011). Mapping emotions into acoustic space: The role of voice production. *Biological psychology*, *87*(1), 93-98.

Zuckerman, M., Koestner, R., & Driver, R. (1981). Beliefs about cues associated with deception. *Journal of Nonverbal Behavior*, *6*(2), 105–114.

Back to top

Tyler Perrachione

Department of Speech, Language, and Hearing Sciences, Boston University

https://www.bu.edu/sargent/profile/tyler-k-perrachione-ph-d/

The source and the signal: An integrated framework for talker identification and speech processing

There are bidirectional dependencies between talker identification (knowing who is speaking) and speech processing (recognizing what is being said). While classically studied separately, decades of research in psycholinguistics and cognitive psychology now convincingly show that human listeners process these two types of information simultaneously and integrally. Such integral processing of voice and speech is often mutually advantageous to both understanding what was said and recognizing who said it. However, accommodating the cognitive demands of a system that evolved to decode these two signals simultaneously is also sometimes detrimental to fast and accurate talker identification or speech perception. Investigating speech processing and talker identification through an integrated framework provides more parsimonious answers to key questions in both of these areas: Why is listening to several talkers, even one at a time, more effortful than listening to a single talker? How do listeners learn to identify voices when most vocal interactions prioritize speech comprehension? And why are listeners so much better at identifying talkers in their native language than in an unfamiliar foreign language? Ultimately, theoretical advances in both speech perception and talker identification should inform one another: Speech must be recognized in the context of phonetic variability across talkers. And talker identification is enhanced when listeners' linguistic knowledge lets them include talkers' phonetic idiosyncrasies in their representation of talkers' identities.

Back to top

Nadine Lavan

School of Biological and Behavioural Sciences, Queen Mary University of London, UK

https://www.qmul.ac.uk/sbbs/staff/nadine-lavan.html

The time course of person perception from voices

Listeners readily form impressions about a person based on the voice: Is the person old or young? Trustworthy or not? While some studies suggest that these impressions can be formed rapidly (e.g., by 400ms of exposure for traits), it is unclear just how quickly these impressions are formed across a number of person-related impressions. In a gating experiment, we collected ratings of 3 physical characteristics (age, sex, health), 3 trait characteristics (trustworthiness, dominance, attractiveness) and 3 social characteristics (level of education, poshness, professionalism) from recordings of the sustained vowel /a/ of 100 unfamiliar voices. Voice recordings were presented to listeners for 25ms, 50ms, 100ms, 200ms, 400ms, and 800ms. We observe that even from 25ms of exposure, interrater agreement for impressions of physical characteristics is high, with these impressions already being similar to the impressions formed after 800ms. This suggests that impressions of physical characteristics can be established rapidly. In contrast, agreement for trait and social characteristics is low to moderate for short exposure durations and gradually increases over time. These findings thus show that different person-related impressions arise at different points in time, suggesting that, in principle, person impressions that arise earlier may influence impressions that arise later on in time.

Back to top

Pascal Belin

Institut de Neurosciences de la Timone, UMR7289, Centre National de la Recherche Scientifique &Aix Marseille Université, France

Département de Psychologie, Université de Montréal, Canada

https://neuralbasesofcommunication.eu/

How do you say “Hello”? Acoustic-based modulation of voice personality impressions

Research in face perception shows that robust personality impressions—stable in time and consistent across observers (although not necessarily accurate)—emerge within less than a second of exposure to novel faces, and that these impressions are well summarized by a 2-D Trustworthiness-Dominance ‘Social Face Space’. Here I present studies showing that a similar phenomenon exists in the voice domain. Our results indicate that exposure to a single ‘Hello’ is sufficient to elicit robust personality impressions in listeners, and that these impressions are accurately summarized, for both male and female voices, by the same Trustworthiness-Dominance ‘Social Voice Space’ as for faces. Acoustical manipulations based on voice morphing effectively modulate these impressions, while reverse-correlation techniques successfully predict the optimal acoustical pattern for each impression, opening the door to a principled-based ‘vocal make-up’ –modulations of perceived voice personality in real or synthetic voices.

Back to top

Session 7

Song structure, voice identity and digital audio

Tom Parkinson[1]

[1]Royal Holloway, University of London

Tom.Parkinson@rhul.ac.uk

This is a necessarily speculative attempt to link insights from recent experiments into voice identity processing with songwriting in a post-digital context. In the early years of mass adoption of digital music production, the novelty of an absolute volume threshold resulted in the *loudness wars* where music was engineered to be as loud as possible. In the last decade, however, the volume limit of digital audio has been approached as a compositional rather than a technical problem and has manifested new kinds of songwriting. Aside from the volume threshold, there are two other factors at play here: the audible frequency spectrum and vocal clarity. The need to make music that fills the audible spectrum, is loud, and has comprehensible vocals has resulted in linear, rather than simultaneous frequency balancing where, in the most simplistic examples, vocals are present when there is no competition for frequency space with music and loud music happens when there is no singing. In many cases, however, this is a dynamic process and unadorned vocals are present only in the opening section of the song. Across the last decade, the initial sixty seconds of a significant number of commercial pop songs have been characterised by the linear accumulation of acoustic complexity in the frequency space of the voice. It's possible, therefore, that what is at stake is not the clarity of the lyrics but the identity of the voice: once that identity is established, the voice can co-exist with other elements of the music whilst retaining its perceptual coherence. This idea intersects with research on time-based voice identification and proposes further questions about what it means to become familiar with a voice in stylised contexts.

Back to top

DAY 3

Session 8

Volker Dellwo

Department of Computational Linguistics, University of Zurich, Switzerland

https://www.cl.uzh.ch/de/people/team/phonetics/vdellw.html

Vocal identity dynamics: Can speakers control their vocal recognizability?

Identity recognition through voice has so far typically been studied in terms of the performance of a human listener or a machine in identifying a person by their voice. In our line of research, we investigate (a) variable characteristics of human voices and their impact on how well voices can be recognised and (b) whether humans control their vocal identity features to be more or less well recognisable in communication situations. We start from the assumption of a mental acoustic voice space in which voices vary around an average voice (norm-based coding) and hypothesise that speakers can adjust their voices to be closer to the average - and thus less distinguishable from others - or further away from the average to find a more unique place that makes them more distinguishable. We show evidence for such adjustments from within-speaker variability of speaking style in which some speaking styles that require strong social bonding lead to better recognition results (e.g. infant-directed speech) compared to styles in which the speaker typically has no interest in being identified and thus forms more average voices (e.g. deception). We found that styles that are targeted at intelligibility (clear speech) were found to be less speaker-specific and resulted in lower voice recognition performance. We conclude that speakers have control over their recognisability by applying different speaking styles and we will show how more refined control of identity markers may play a role in dialogue processing.

Back to top

Elisa Pellegrino

Department of Computational Linguistics, University of Zurich, Switzerland

https://www.cl.uzh.ch/de/people/team/phonetics/epelle.html

Individualization versus cooperation: The effect of group size on voice individuality

Compared to other species, humans have an unparalleled ability to cooperate with unrelated individuals. Cooperation, acoustically signalled by convergent accommodation, is facilitated when group members are more similar. Nevertheless, convergence constraints may arise when interlocutors need to mark their vocal individuality. Inspired by findings in animal communication showing higher vocal individuality in larger groups, the focus of this presentation will on the effect of group size on vocal individualization in human interactions. We will illustrate the novel data collection method designed to investigate the trade-off between acoustic convergence and voice individualization in cooperative situations wherein voice recognition is at stake. We will describe the computational approaches used to quantify between- and within-speaker acoustic similarity across group sizes (i-Vector PLDA; Principal Component Analysis) with the results based on automatic features (MFCC) and more traditional acoustic features relevant for identity processing (e.g., F0, harmonicity, formant dispersion, duration). We will also show the effect of individualization vs. cooperative accommodation on automatic voice discriminability in terms of Equal Error Rate. Results pointing to vocal convergence in larger groups will be discussed compared to the opposite trend observed in animal species. Alternative interpretations will be offered that are based on the role of feedback, the effect of first exposure, and familiarization between speakers' voices.

Back to top

Homa Asadi

Department of Linguistics, University of Isfahan, Iran.

https://collegium.ethz.ch/en/fellows/ph-d-homa-asadi-university-of-isfahan/

Acoustic variation within and between bilingual speakers

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. Nevertheless, voices are often more variable within the same speaker rather than between different speakers. One of the sources of within-individual vocal variability occurs when speakers communicate in different languages and switch from one language to another. This 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). Nevertheless, little is known about the influence of language on within- and between-speaker vocal variability. In our talk, we will discuss how the acoustic parameters of voice quality vary or remain stable between different speakers' languages. We assume that phonological differences and different sound patterns underlying Persian and English are likely to influence the acoustic parameters of voice quality, and thus it is plausible that acoustic voice structure varies accordingly between the languages of Persian-English bilinguals. Following a psychoacoustic model proposed by Kreiman (2014) and using a series of principal component analyses, we will discuss how acoustic voice quality spaces are structured across the languages of Persian-English bilingual speakers.

Back to top

Peter French

University of York, UK

https://www.york.ac.uk/language/people/academic-research/peter-french/

Voicetype, Phenotype, Genotype

In the criminal justice system, various categories of individual identification evidence are frequently treated as being independent of one another. For example, if evidence against a defendant includes both facial identification and voice identification, the assumption is that the evidence overall is particularly strong as two independent modalities point in same direction. This assumption of independence is called into question by links between voices and cranial/facial anatomy emerging from research studies, which are reviewed in the presentation. Similarly, DNA identification evidence is often presented as an independent biometric. Emerging knowledge of the links between DNA and face/head features is already sufficient to question whether facial identification evidence can legitimately be regarded as independent of DNA evidence. If the three modalities genuinely were to be independent of one another, the procedure for estimating their combined strength would be one of simple multiplication: Voice X Face X DNA. However, while we now know that this procedure would result in an overestimation, we have no basis for 'factoring in' interdependencies. In order to establish such a basis, and thereby an improvement to the quality of justice, further, collaborative research is needed by those working across the modalities. In respect of criminal investigations, the emerging patterns of correlation hold great potential, in that they could allow one to make informed predictions about, say, an offender's facial appearance from a recording of his voice, or vice versa; or – at some future point – to make inferences about both from a DNA sample. The advantages of this are both obvious and inestimable. Those working in forensic speech science have amassed a body of research findings – again reviewed in the presentation - concerning with aspects of voice are most likely to be biologically determined rather than the products of linguistic socialisation. These provide a starting point for further work. We now call for input from other disciplines to take a 'big vision' research initiative forward.

Back to top

Vincent Hughes

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

https://www.york.ac.uk/language/people/academic-research/vincent-hughes/

Forensic voice comparison at the intersection of linguistics and automatic speaker recognition

Within the field of forensic speech science, there has been growing interest in integrating traditional linguistic methods with automatic speaker recognition (ASR) systems. This work has two aims. The first is to better understand what linguistic information is captured by increasingly 'black boxy' ASR systems. The second is to empirically combine the results of linguistic analysis with ASR output, to reduce overall error rates. Some studies have shown promising results. For example, Gonzalez-Rodriguez et al. (2014) and Hughes et al. (2017) found that ASR misclassifications can be resolved by trained phoneticians primarily using laryngeal voice quality analysis. However, key questions remain: As systems continue to produce marked improvements in overall performance with each new paradigm (usually every 3-5 years), will we reach a stage where forensic voice comparison is conducted entirely using ASR? If so, what role will linguistic methods have in forensic casework in the future?I will argue that to answer these questions we must recognise that forensics is a unique application of ASR. As such, we have different concerns and priorities from developers of ASR systems for other commercial applications. Specifically, this means that: (i) Features for analysis should be determined on a case-by-case basis; (ii) The context in which forensic recordings are made is unique, making replication difficult; (iii) Our focus should be on reducing uncertainty rather than maximising potential discriminability. This involves identifying, reporting, and attempting to mitigate for sources of variability in system performance (e.g. sample size); (iv) The state-of-the-art ASR system isn't necessarily the best choice for every forensic case. In this talk, I will review the current state of knowledge at the intersection of linguistics and ASR, and make proposals for ways forward in the quest to find the best ways of analysing voices in the specific context of forensic comparison.

Back to top