VoiceID 24

Conference Programme

Marburg
28. - 30. August 2024

Forschungszentrum Deutscher Sprachatlas

Philipps Universität Marburg

Queen Mary University of London

DFG

UCL

Universität Zürich UZH

Ursula-Kuhlmann-Fonds Marburger Universitätsbund e.V.

A welcome from the organisers

Dear Participants of the 2nd International Conference on Voice Identity (voiceID),

It is our pleasure to welcome you to Marburg!

This charming student town with its winding cobble-stoned streets is a place with a rich academic heritage. The almost 500 year-old Philipps-Universität, also one of the oldest protestant universities in Germany, is known for its pioneering research in multiple research fields (many relevant to voice identity research).

The most well-known alumni of Marburg are probably Emil von Behring and the Grimm Brothers. The latter studied here and collected many of their fairytales from around the region. A fairy tale trail through the city will lead you to all sorts of historical sights and quirky works of art. A fancy QR-code provides you with the associated audio recording, as it happens, produced by our phonetics students. Von Behring was the founder of serum therapy. He developed vaccines against diphtheria and tetanus. Interesting fact: as Von Behring founded his company "Behringwerke" to produce antitoxins and vaccines, he also made Marburg worldwide one of the main centres for immunology research. During the pandemic you actually may have received your BioNTech-Covid-vaccine from Marburg.

For the linguists amongst us, however, it may be alumnus Georg Wenker who is remembered most for his contribution. He founded the Research Centre »Deutscher Sprachatlas« in 1876, one of the oldest linguistic research centres in the world. It has been instrumental in documenting and analysing the regional dialects of the German language. Its archive contains data from over 50,000 locations in Germany, making it an important database for dialectology and linguistic variation research. Wenker was honoured with a statue that can be found in the library of our conference location.

For the many blind students in Marburg it may be Alfred Bielschowsky and Carl Strehl they remember most. It was thanks to Bielschowsky, Director of the Marburger Eye clinic, and Carl Strehl, Director of the "Deutsche Blindenstudienanstalt", that special courses and schools were organised for the many young soldiers who returned from World War I with eye injuries. Their main goal was to enable them to join the workforce again and to obtain a degree. In 1921 the first gymnasium in Germany for people with a visual impairment was founded. In no other place in Germany is the number of blind and visually impaired citizens compared to the total number of citizens that high. The clattering sound of guiding canes is therefore common in Marburg, as blind people navigate the town aided by beeping traffic lights, pavements and floors with ridges and bumps that act as tactile signals of hazards or barriers. Detailed miniature bronze models of major sights such as Marburg's castle and marketplace allow blind visitors to feel the structure of each landmark. Due to its long history as a hub for accessibility, Marburg now proudly calls itself a "Blindenstadt".

Marburg is also known for its pioneering research in phonetics, particularly acoustic phonetics. Early on, researchers engaged in working with instruments recording and analysing human speech. It is not surprising, then, that acoustic-phonetic analyses have also extended to forensic research, making Marburg an important centre for forensic phonetics in Germany. Starting in the 70s and 80s at the time that the RAF was active in Germany, the contribution of Marburger phoneticians was extremely relevant during the investigation process. Nowadays, combining the digitised dialectal Wenker maps mentioned earlier, with sophisticated data management tools in addition to recently added audio recordings, forensic phoneticians are able to provide a fairly accurate estimation of a (criminal) speaker's region of origin based on his/her dialectal features. This is particularly useful in the case of a speaker profile, where an audio recording exists but no suspect has yet been identified.

Finally, the host of this year's conference, the Phonetics research group, engages in interdisciplinary research, combining classical methods from phonetics with cutting-edge methods from the cognitive neurosciences. This line of research is also pursued in the newly established research training group (RTG), whose quest is the modelling and neurobiologically plausible description of representations in speech and language.

Wishing you an interesting, inspiring, successful but also enjoyable conference, and a lovely stay in Marburg!

The Voice-ID organising team

About the organisers

International Organising Committee

De Jong-Lendle, Gea	Philipps-University Marburg	Germany
Dellwo, Volker	University of Zurich	Switzerland
Lavan, Nadine	Queen Mary University London	UK
McGettigan, Carolyn	University College London	UK
Rinke, Paula	Philipps-University Marburg	Germany
Roswandowitz, Claudia	University of Zurich	Switzerland
Scharinger, Mathias	Philipps-University Marburg	Germany

Local Organisers

Feike, Ingrid	Research Training Group »Representations«
Jesberg, Markus	Research Group Phonetics
Krieglstein, Celine	Institute of German Linguistics / Research Centre »Deutscher Sprachatlas«
Möller, Sophie	Research Group Phonetics
Oran, Marc	Institute of German Linguistics
Tzallas, Mina	Institute of German Linguistics
Student assistants	Institute of German Linguistics, Research Training Group »Representations«

Session Chairs

Dellwo, Volker	University of Zurich
Gerards, Jonas	Research Training Group »Representations«
Hoffmeister, Toke	Research Training Group »Representations«
Leonidou, Alexia Despina	Research Training Group »Representations«
Möller, Sophie	Research Group Phonetics
Rinke, Paula	Research Group Phonetics
Scharinger, Mathias	Research Group Phonetics
Tzallas, Mina	Institute of German Linguistics

Programme Supporters

Hahn, Matthias	Guided Tour »Wenker-Archives«	Senior Researcher, Research Centre »Deutscher Sprachatlas«
Lameli, Alfred	Words of Greetings	Director of the Research Centre »Deutscher Sprachatlas«
Hoffmeister, Toke	Guided Tour »Wenker-Archives«	Senior Researcher, Research Centre »Deutscher Sprachatlas«
Nauss, Thomas	Words of Greetings	President of the Philipps-University
Oberdorfer, Georg	Guided Tour »Wenker-Archives«	Senior Researcher, Research Centre »Deutscher Sprachatlas«
The Glorious Formants & Friends	Evening Programme	A-Capella-Group, Institute of German Linguistics / Research Centre »Deutscher Sprachatlas«

About the sponsors & funding agencies

Sponsors

Queen Mary University, London, UK
University College, London, UK
LiRI – Linguistic Research Infrastructure, University of Zurich, Switzerland

Funding Agencies

Deutsche Forschungsgemeinschaft (DFG)
Marburger Universitätsbund – Ursula-Kuhlmann-Fonds

Programme Overview

Date: Wednesday, August 28th, 2024

8:45am - 9:15am	**Registration & Coffee**
9:15am - 9:45am	**Opening of Conference**
9:45am - 11:45am Main Conference Hall (001)	**Session 1: Forensic Phonetics** Location: **Main Conference Hall (001)** Session Chair: **Sophie Möller**
11:45am - 12:45pm	**Lunch**
12:45pm - 2:15pm Poster Room (101/102)	**Poster Session 1** Location: **Poster Room (101/102)**
2:15pm - 3:45pm Main Conference Hall (001)	**Session 2: Voice Processing in Animals** Location: **Main Conference Hall (001)** Session Chair: **Jonas Gerards**
3:45pm - 4:15pm	**Coffee Break**
4:15pm - 5:45pm Main Conference Hall (001)	**Session 3: Voice Perception** Location: **Main Conference Hall (001)** Session Chair: **Volker Dellwo**
6:30pm - 10:00pm	**Finger food & Drinks / Evening Programme**

Date: Thursday, August 29th, 2024

9:00am - 9:20am	**Get Together Day 2 & Coffee**
9:20am - 10:50am Main Conference Hall (001)	**Session 4: Artificial Voice Identity** Location: **Main Conference Hall (001)** Session Chair: **Jasemina Tzallas**
11:00am - 12:30pm Poster Room (101/102)	**Poster Session 2** Location: **Poster Room (101/102)**
12:30pm - 1:30pm	**Lunch**
1:30pm - 3:10pm Main Conference Hall (001)	**Session 5: Talker Identification** Location: **Main Conference Hall (001)** Session Chair: **Alexia Despina Leonidou**
3:10pm - 3:40pm	**Coffee Break**
3:40pm - 4:50pm Main Conference Hall (001)	**Session 6: Neurobiology of Voice** Location: **Main Conference Hall (001)** Session Chair: **Mathias Scharinger**
4:50pm - 5:00pm Main Conference Hall (001)	**The Voice Communication Sciences Project** Location: **Main Conference Hall (001)** Session Chair: **Pascal Belin**
5:00pm - 6:00pm	**Guided Tour "DSA - Georg Wenker Archive"**
7:00pm - 10:00pm	**Cuban Dinner in Havana**

Date: Friday, August 30th, 2024

9:00am - 9:30am	**Get Together Day 3 & Coffee**
9:30am - 11:00am Main Conference Hall (001)	**Session 7: Voice and Speech and Music** Location: **Main Conference Hall (001)** Session Chair: **Toke Hoffmeister**
11:00am - 11:30am	**Coffee Break**
11:30am - 12:40pm Main Conference Hall (001)	**Session 8: Challenges in Talker Identification** Location: **Main Conference Hall (001)** Session Chair: **Paula Rinke**
12:40pm - 1:10pm Main Conference Hall (001)	**Final Group Discussion** Location: **Main Conference Hall (001)**
1:10pm - 1:20pm Main Conference Hall (001)	**Closing Comments** Location: **Main Conference Hall (001)** Session Chair: **Nadine Lavan** Session Chair: **Carolyn McGettigan**

Abstracts

Session 1: Forensic Phonetics

Time: Wednesday, 28/Aug/2024: 9:45am - 11:45am · *Location:* Main Conference Hall (001)
Session Chair: Sophie Möller

9:45am - 10:15am
ID: 157
Invited Talk

Voice distinctiveness: An investigation of the role of speakers’ position in a population with respect to f0

Kirsty McDougall
Cambridge University, United Kingdom; kem37@cam.ac.uk

Long-term f0 is understood to be important in listeners’ judgements of speaker similarity. Not known is whether a speaker’s position in a population distribution of long-term f0 values affects similarity judgements, i.e. whether a very low-pitched speaker is perceived as more distinctive than a mid-pitched one, and whether two speakers a given distance apart in f0 are judged as equally distinct if their f0 values are both in the middle of the distribution or both in one tail.

Three experiments were conducted using 12 male SSBE speakers: 4 each with high f0, mid f0 and low f0. In Experiment 1, listeners rated the similarity of all speaker pairs on a 9-point scale; judgements correlated with similarity in f0.

In Experiment 2, listeners rated the similarity of the same speakers whose f0 had been manipulated to high, mid and low regardless of original pitch. Speaker pairs ‘heard low’ or ‘heard high’ were rated more similar than the same speakers ‘heard mid’ suggesting greater sensitivity to pitch in the distribution’s centre.

In Experiment 3, the original stimuli were manipulated such that a pair’s samples either had the same f0 (control pairings) or f0 3st apart (test pairings). Listeners rated a balance of control and test pairings. Test pairings were perceived as less similar than controls, and the 3st shift had a stronger distancing effect for mid speakers, consistent with Experiment 2, confirming more sensitivity to pitch in the centre of the f0 distribution (where more speakers lie) than in the tails.

10:15am - 10:35am
ID: 117
Abstract

Keywords: voice imitation, speech acoustics, automatic speaker recognition, Czech

Acoustic and ASR analysis of voice impressions by two professional imitators

Radek Skarnitzl
Charles University, Czech Republic; radek.skarnitzl@ff.cuni.cz

Professional voice imitators continue to be of interest for forensic phoneticians and for speech science in general, particularly because it provides data about the flexibility of the human voice production mechanism. The present study examines voice impressions by two Czech professional imitators who performed their impressions in the sound-treated recording studio of the Institute of Phonetics in Prague. Each of them imitated one famous speaker (the Czech ex-president Zeman, whose speech is extremely idiosyncratic and whom both artists regularly imitated), one previously unknown speaker (a senator from North Moravia with marked regional characteristics but otherwise relatively few idiosyncratic features), and one speaker each that they had in their “voice repertoire” but were not quite happy with how they sounded when impersonating them. Acoustic properties of the imitators’ original voice, of their impressions, as well as the original voices of the target speakers (obtained from publicly available sources) have been analyzed (f0 baseline and variability, formant values in vowel midpoints as well as long-term formant distributions, and spectral characteristics of salient noise segments). Also, results of an automatic speaker recognition (ASR) analysis will be presented, based on the latest Phonexia Speaker Identification (SID4 XL5) DNN-based system, with each of the voices compared against each other.

10:35am - 11:05am
ID: 158
Invited Talk

Deepfake voices in forensic speaker recognition: An existential threat or manageable challenge?

Anil Alexander
Oxford Wave Research, United Kingdom; anil@oxfordwaveresearch.com

It is now possible to create highly realistic, natural-sounding synthetic speech samples, either using text-to-speech synthesis or voice conversion that are almost indistinguishable from real speech. While the creation of convincing synthetic speech typically required significant technical expertise and knowledge, widely available commercial and open-source toolkits have made this capability easily accessible to non-technical users, as evidenced by media reports of deepfake recordings, successfully used for fraud, political propaganda and reputational damage. There is a real risk that trust in the spoken word could diminish to the point that speech is not considered a credible form of evidence. In cases that involve disputed utterances, in addition to considering whether a voice came from a particular speaker, it will also be necessary to consider the hypothesis that the voice was a fake. This naturally raises questions about the role of the forensic audio expert in the detection of such material, as well as the potential impact on forensic speaker comparison. Deepfake audio recordings have now reached the level of naturalness and sophistication that they pose an existential threat to forensic speech science.

We will provide an overview of the current audio deepfake landscape, explore the creation of deepfakes using commercial and publicly-available resources, and demonstrate some possibilities for the detection of synthetic speech. We will argue that human and computer-based countermeasures to detect and authenticate voice must urgently be developed to address the newer algorithms and threats that appear with increasing frequency.

11:05am - 11:25am
ID: 119
Abstract

Keywords: forensic voice comparison, human expert, automatic speaker recognition

A focus on whole analysis processes for forensic voice comparison

Georgina Brown[1,2], Christin Kirchhübel[2]
[1]Lancaster University, United Kingdom; [2]Soundscape Voice Evidence; g.brown5@lancaster.ac.uk

Forensic voice comparison (FVC) is a form of voice identification performed on evidential recordings. Forensic speech scientists have arrived at two broad approaches to FVC: Auditory Phonetic and Acoustic analysis led by a human expert (AuPhA), and automatic speaker recognition. There are principled arguments for adopting one method or the other. Generally though, in jurisdictions where both are admissible and feasible, the complementarity between the two approaches is recognised and they are combined. Forensically motivated studies also support the complementarity between the two FVC approaches, but such studies do not necessarily reflect FVC in practice. For example, it is common to observe the performance of only a fragment of an AuPhA analysis because it is practical to single out the features that can be most easily measured and statistically analysed. This form of research tends to neglect the extensive decision-making and interpretation that run through a FVC analysis, irrespective of approach. The present work therefore further embraces the complementarity of the two analysis approaches by testing: 1) AuPhA as a whole analysis process in isolation, 2) automatic speaker recognition as a whole analysis process in isolation, and 3) combining the two approaches. By applying each of these to 60 carefully constructed voice comparison trials, this is the largest study to date that tests the AuPhA approach as a whole analytical process in isolation and in conjunction with an automatic speaker recognition system.

11:25am - 11:45am

ID: 125

Abstract

Keywords: Voice research in forensic settings, Acoustic analysis and modelling of synthetic voices, Dataset generation, Criminal conversational datasets

Generating synthetic datasets for the validation and training of automatic speech analysis systems in the context of organized crime

David Grünert[1], Dominic Pfister[1], Alexandre de Spindler[1], Volker Dellwo[2]

[1]ZHAW, Zurich University of Applied Sciences, Switzerland; [2]Department of Computational Linguistics, University of Zurich, Switzerland; grud@zhaw.ch

Much progress has been made with tools for automatic speech analysis in forensic audio processing over the past years. However, it is difficult for researchers to evaluate novel approaches due to missing real world data. Data recorded during criminal investigations is often confidential and therefore unavailable, and existing datasets from other domains do not share the characteristics of crime-related data.

ROXSD [1] is one of the few synthetic datasets that was created with the help of law enforcement agencies to include characteristics of realistic audio-surveillance. It consists of recordings and metadata from three hypothetical cases. The compilation of such datasets is time consuming, and their number is therefore very limited.

In general, crime data contains recordings from telephone conversations or from audio surveillance equipment with varying audio quality, multilingual and emotional speech, as well as background noise with relevant information. Furthermore, higher level analyses such as communication structure detection [2], [3] require the spoken content to match the context of criminal cases.

We developed a system that can generate synthetic datasets from a case description. The system first uses LLMs to generate transcripts annotated with emotions and timing aspects. It then uses text-to-speech models to generate the utterances. Next, these utterances are combined and augmented with background noise. Finally, signal processing is applied to achieve realistic acoustic environments and variations in audio quality. Our prototype implementation demonstrates the feasibility of the approach and shows the reduction of effort to create such datasets allowing researchers to use more or even tailormade datasets.

References

- [1] K. Maly, G. Backfried, F. Calderoni et. al, ROXSD: a Simulated Dataset of Communication in Organized Crime, September 2021, SPSC Symposium.
- [2] D. Grünert, A. de Spindler, and V. Dellwo, Speaker Diarization Systems in the Context of Forensic Audio Analysis. IAFPA 2023
- [3] M. Fabien, S. S. Sarfjoo, P. Motlicek, and S. Madikeri, Improving Speaker Identification using Network Knowledge in Criminal Conversational Data, Jun. 2020. [Online]. Available: http://arxiv.org/abs/2006.02093

Session 2: Voice Processing in Animals

Time: Wednesday, 28/Aug/2024: 2:15pm - 3:45pm · *Location:* Main Conference Hall (001)
Session Chair: Jonas Gerards

2:15pm - 2:45pm

ID: 159

Invited Talk

Voice identity perception in dogs: Neural bases, capacities and sensitivities

Attila Andics

Neuroethology of Communication Lab, Department of Ethology, Eötvös Loránd University (ELTE), Budapest; attila.andics@ttk.elte.hu

A central objective of comparative cognitive neuroscience is to reveal how and when certain processing preferences and neural sensitivities emerged during evolution: what is uniquely human, what is shared across species, and how do brains adapt to environmental changes? When it comes to mechanisms underlying social functions, and specifically voice processing, such questions require comparisons not only to evolutionarily proximal species (primates), but also to socially proximal ones (domestic animals). Companion dogs are an ideal study population to complement primate research on the phylogenetic and experiential contributions to the neural bases of voice identity perception. Over the last decade our lab has been conducting non-invasive EEG and fMRI on awake, cooperating dogs using no restraints or sedation. This is only done in very few labs worldwide and has not yet been done with any other animals. In this talk I will overview our latest behavioural and neuroimaging findings on dogs' sensitivity to voice-like sounds, to conspecific vocalizations, to certain human voices, and to auditory regularities in the speech signal; on dogs' capacities for voice identity discrimination and recognition; on decoding vocally encoded emotions and motivations; and on how vocal emotions may affect dogs' voice identity processing. I will highlight exciting similarities but also remarkable differences to corresponding human capacities.

2:45pm - 3:05pm

ID: 149

Abstract

Keywords: conspecific vocalization, event-related potential, dog, pig

ERP evidence for conspecific voice sensitivity in mammals

Boglárka Morvai[1], Marianna Boros[1], Elodie Ferrando[1], Fruzsina Horváth[1], Lilla Magyari[1,2,3], Attila Andics[1,4]

[1]Neuroethology of Communication Lab, Department of Ethology, Eötvös Loránd University, Budapest; [2]Norwegian Reading Centre for Reading Education and Research, Faculty of Arts and Education, University of Stavanger, Stavanger; [3]Department of Social Studies, Faculty of Social Sciences, University of Stavanger; [4]ELTE NAP Canine Brain Research Group; boglarka.morvai@ttk.elte.hu

Vocalizations produced by conspecifics convey crucial information supporting the survival and reproduction of individuals. While there is substantial evidence for distinct neural processing of conspecific vocalizations in primates, less is known about whether it is a common feature across mammals. In the present study we used a comparative approach to investigate the neural responses to various sound stimuli in three distinct mammalian species: humans (Homo sapiens), dogs (Canis familiaris) and pigs (Sus scrofa domesticus). Given dogs' extensive selection for cooperation with humans, we hypothesized that dogs might show neural specializations for processing human voices, in addition to conspecific voice sensitivity. Using non-invasive, awake electroencephalography (EEG), we recorded event-related potentials (ERPs) as subjects listened to conspecific, heterospecific, and environmental sounds. Our findings revealed specific time windows in each species where ERPs differed for conspecific vocalizations compared to all other sound categories (humans: 430-510 ms, dogs: 318-508 ms, pigs: 126-410 ms after stimulus onset). Additionally, pigs also exhibited general voice sensitivity, i.e. distinct ERPs for processing any vocalizations vs. non-vocal environmental sounds (260-356 ms). Furthermore, trial-by-trial analysis in dogs revealed an adaptation effect in case of relevant vocal stimuli: the ERP magnitude decreased with repeated exposure to human and dog vocalizations within an early time window (160-380 ms). To our knowledge, this study provides the first electrophysiological evidence of conspecific voice sensitivity in non-human mammals. Moreover, our results indicate that the special selection for vocal cooperation with us during domestication may have led to neural specializations for processing human vocalizations in dogs.

3:05pm - 3:25pm

ID: 129

Abstract

Keywords: Temporal Cortex, fMRI-guided Electrophysiology, Non-Human Primates

Encoding of voice identity by neurons in the macaque anterior Temporal Voice Area

Margherita Giamundo[1,2], Regis Trapeau[1], Etienne Thoret[1,2], Yoan Esposito[1], Luc Renaud[1], Thomas Brochier[1], Pascal Belin[1,2]

[1]Institut de Neurosciences de la Timone, France; [2]Institute of Language Communication and the Brain (ILCB), France; pascal.belin@univ-amu.fr

Social interactions in primates are possible through the ability to extract relevant information from voices, for example their identity. The anterior Temporal Voice Area (aTVA) is a region in the anterior temporal lobe of humans, macaques and marmosets specialized in the processing of voices, but the exact voice information represented by individual neurons in the aTVA remains obscure.

Here we asked how aTVA neurons encode voice identity information, separating different identities and grouping together calls from a same identity. We implanted two rhesus macaques with high-density multi-electrode arrays in their fMRI-localized aTVA. Spiking activity was recorded during an auditory stimulation task in which we presented 50 natural stimuli including 5 different coo calls from each of 5 macaques, and 5 different calls from each of 5 humans.

Of 160 auditory-responsive neurons, 77% were modulated by voice identity (one-way ANOVA). Representational Dissimilarity Matrices, capturing pairwise spiking activity differences between the stimuli, showed significant association with an ideal categorical model separating between identities and grouping different calls from each identity, from 50ms post-stimulus onset.

A principal components analysis (PCA) applied to the mean population activity over time revealed that population responses to the same identity followed similar trajectories in the multidimensional state space, with the third PC showing marked differentiation of the different identities in the sustained response and allowing higher accuracy in the discrimination between identities than the other PCs.

These results contribute to elucidating the mechanisms by which abstract representations of identities, allowing speaker recognition, emerge in the primate brain.

3:25pm - 3:45pm
ID: 150
Abstract

Keywords: voice identity discrimination, neural entrainment, comparative, EEG, fast periodic auditory stimulation

Fast periodic auditory stimulation EEG reveals recent-over-ancient tuning effects on voice identity processing in dogs and humans

Dorottya Szilvia Rácz[1], Boglárka Morvai[1], Attila Andics[1,2]
[1]Neuroethology of Communication Lab, Department of Ethology, Eötvös Loránd University, Budapest; [2]ELTE NAP Canine Brain Research Group; dorottyaszilviaracz@gmail.com

Voices of conspecifics, typically the most relevant social partners, are processed preferentially in many species' brains. This processing preference may reflect either ancient or more recent tuning to a socially relevant stimulus class. To disentangle these accounts, the special interspecific relationship of dogs and humans provides an ideal study case: for dogs, that are not only selected to communicate effectively with humans but also live with them, human vocalizations may have recently gained greater social relevance than anciently relevant conspecific vocalizations; and among humans, dog vocalizations may be more relevant to dog owners than non-owners. In a fast periodic auditory stimulation-based EEG study, we presented dogs (N=16) and humans (N=21) with continuous, rhythmically patterning vocalization streams (dog/human): every third sound differed in vocalizer identity. We found that both dog and human brains detected human voice identity changes more efficiently, despite that dog and human voices were acoustically similarly discriminable. Dog owners, however, did not outperform non-owners. Recent relevance may thus outweigh ancient preferences in voice identity processing, but individual experience with a vocalizer species does not necessarily improve identity discrimination abilities.

Session 3: Voice Perception

Time: Wednesday, 28/Aug/2024: 4:15pm - 5:45pm · *Location:* Main Conference Hall (001)
Session Chair: Volker Dellwo

4:15pm - 4:45pm
ID: 139
Invited Talk

Prototype theory and acoustic variability in voice perception

Jody Kreiman
University of California, Los Angeles, United States of America; jkreiman@ucla.edu

In general terms, prototype models of speaker recognition assert that voices are recognized (or learned, or remembered) with reference to an abstract voice sample that is average for the population (or the individual, or the entire universe of voices), or else to a sample that sounds very ordinary/not at all distinctive. This model has existed for many years, and many variants of this basic framework have emerged. In this paper I will attempt a comprehensive review of prototype models, covering definitions of a prototype, what a prototype is intended to explain, the data to be modeled (perceptual? acoustic?), the domain of averaging (single voice? population of voices?), experimental results that support the concept, and other details. This review will be framed in context of recent findings regarding the nature and extent of acoustic variability within and across speakers. It is hoped that this review will clarify the state of prototype models in voice studies and bring some order to what is a tantalizing but somewhat muddled set of concepts.

4:45pm - 5:15pm
ID: 161
Invited Talk

Memory for voices and individual differences

Romi Zäske[1,2]
[1]University Hospital Jena, Germany; [2]Friedrich Schiller University, Jena, Germany; romi.zaeske@med.uni-jena.de

The ability to recognize others by their voice ranges from phonagnosia to super-recognition. To understand the neural mechanisms and individual differences underlying voice memory, we conducted a series of studies using recognition memory paradigms. Results suggest the rapid acquisition of voice representations after minimal voice exposure to brief sentences. In the EEG, learned – compared to novel – voices elicited a suppression in beta-band oscillations from ~300 ms independent of speech content, indicating the detection of learned speaker identities across novel utterances. In fMRI, explicit voice recognition independent of speech content recruited both voice-sensitive cortex areas of the right superior temporal gyrus and extra-temporal areas. In terms of individual differences, we observed higher recognition performance for young vs. older adults, and a general recognition advantage for old adult voices. Finally, to assess individual differences in voice memory, we developed and validated the Jena Voice Learning and Memory Test (JVLMT), a standardized and freely available research tool, based on item-response theory. Its format is broadly analogous to the Cambridge Face Memory Test. The JVLMT is suited to screen for phonagnosia and super-recognition abilities within ~22min and is applicable across languages due to the use of pseudo-speech. As an outlook, we aim at refining the JVLMT to discriminate within the upper and lower ability spectrum and plan to develop further tests to assess various objective and self-reported aspects of voice memory and voice cognition. Prospectively, these tools can inform both researchers into individual differences and clinicians about the effectiveness of interventions in hearing-impaired patients.

5:15pm - 5:45pm

ID: 162

Invited Talk

A data-driven exploration of the mental representational voice identity space

Claudia Roswandowitz

University of Zurich (UZH), Switzerland; claudia.roswandowitz@uzh.ch

Empirical investigations on the mental representation of vocal identities are largely challenged by two factors: i) the large number of voices each person is familiar with, and ii) the multidimensional nature of the physical voice signal. Therefore, a realistic characterization of voice representation should encompass a wide variety of voices and not be limited to a small, predefined set of acoustic features. In this talk, I present findings of a data-driven analysis approach that processes voice similarities among a large number of speakers, as rated by numerous participant. I collected sentences from 337 male speakers, and 359 participants rated the voice similarities in a web-based triplet-odd one out task. The model, trained on the similarity ratings, identified five dimensions that predict voice similarity ratings with an accuracy of 76%. Through acoustic analyses and descriptive voice labelling, I identified meaningful acoustic and perceptual labels for these five dimensions, proposing them as relevant characteristics of the representational voice space.

Session 4: Artificial Voice Identity

Time: Thursday, 29/Aug/2024: 9:20am - 10:50am · *Location:* Main Conference Hall (001)
Session Chair: Jasemina Tzallas

9:20am - 9:50am

ID: 163

Invited Talk

Investigating effects of self-relevance and familiarity on first impressions of voice clones

Carolyn McGettigan, Emma Soopramanien, Victor Rosi

University College London, United Kingdom; c.mcgettigan@ucl.ac.uk

Voice cloning technology has advanced rapidly, such that highly naturalistic outcomes can now be achieved with only seconds of input data and often at very low (or no) cost. We investigated whether listeners' first impressions of such state-of-the-art voice clones depend on whether the clone being heard is a replica of the self, a friend, or a total stranger. We recorded and cloned the voices of 50 English-speaking adult participants, recruited as 25 pairs of friends. Forty-seven of these experimental participants (and 47 unfamiliar controls) gave first-impression ratings of trustworthiness, attractiveness, competence, and dominance for cloned and recorded samples of their own voice and their friend's voice. We consistently observed that while the experimental group found clones to sound less trustworthy, attractive, competent, and dominant than recordings, control listeners showed the opposite profile and rated clones more highly than recordings. Participants also tended to prefer the friend's voice to their own. Greater perceived similarity of clones to the "real" self/friend predicted higher ratings of trustworthiness, attractiveness, and competence, but there was no clear relationship for dominance. Overall, we find that familiar listeners' impressions are sensitive to the accuracy and authenticity of cloning for voice they know well, while unfamiliar listeners tend to prefer the synthetic versions of the those same voice identities. The latter observation may relate to the tendency of generative voice synthesis models to homogenise speaking accents and styles, such that they more closely approximate (preferred) norms.

9:50am - 10:10am

ID: 111

Abstract

Keywords: Voice Cloning, Voice Identity, Conversational agents

Echoes of self: Exploring the acceptability of voice identity cloning.

Victor Rosi, Emma Soopramanien, Carolyn McGettigan

University College London, United Kingdom; v.rosi@ucl.ac.uk

We investigated how people evaluate the acceptability of using voice identity cloning across diverse contexts. In Experiment 1, 120 participants rated the acceptability (0-100) of different use cases via reading about scenarios. Scenario texts were designed to manipulate 2 factors per use case: 1) Accessibility i.e. who can use the cloned voice (original speaker only; speaker plus friends/family; anyone) and 2) Availability of the original speaker's physical voice (typical voice; stammer; no voice due to illness; no voice because deceased). Participants tended to endorse cloning most strongly when the original speaker had a stammer or illness, and when the cloned voice was used only by the original speaker. Ratings were also lower when participants considered themselves as the cloned speaker (p<.001). Experiment 2 explored more personalised use cases by creating bespoke cloned voice samples of each participant and a friend (N=20 participants so far). For each cloned voice identity (self, friend) and scenario, participants rated acceptability, how much they liked the cloned voice, and its similarity to the original voice. Participants consistently preferred the original speaker having exclusive use of the cloned voice, and this effect was more pronounced when rating their own voice. Intriguingly, higher acceptability was predicted by greater liking of the cloned voice samples (r=.32), but not by their perceived similarity to the original voice. Our study offers key first insights on the ethics of speech technologies and the influences of personal relevance.

10:10am - 10:30am

ID: 110

Abstract

Keywords: naturalness, human-likeness, synthetic voices, voice distortion

Naturalness of voices – from human to artificial agents

Christine Nussbaum[1,2], Stefan R. Schweinberger[1,2,3]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany; [2]Voice Research Unit, Friedrich Schiller University, Jena, Germany; [3]Swiss Center for Affective Sciences, University of Geneva, Switzerland; christine.nussbaum@uni-jena.de

Perceived naturalness of a voice is a prominent feature which affects our interaction with both human and artificial agents. In humans, naturalness can be affected by voice distortion or transformation. In artificial agents, its tremendous importance has been acknowledged through multiple efforts to create and constantly improve synthetic voices to resemble human vocal expression. Yet, we are far away from a systematic understanding of voice naturalness. To address this issue, we first identify four key problems in current naturalness research: (a) conceptual underspecification, (b) inconsistent operationalization, (c) lack of exchange between research on humans and artificial agents and (d) insufficient anchoring in voice perception theory. Next, we formulate concrete suggestions to overcome these problems. Of crucial importance, we propose a concise conceptual framework for the definition of naturalness. In parallel, we illustrate how insights into voice naturalness will profit both from pooling evidence from a wide interdisciplinary background and from rooting naturalness in current voice perception models. We conclude with an outlook on core gaps in our current understanding of voice naturalness and discuss different approaches to promote a fuller understanding of the various dimensions of voice naturalness.

10:30am - 10:50am

ID: 109

Abstract

Keywords: self-voice, self-other voice discrimination, bone conduction, auditory-verbal hallucinations, hallucination engineering

Voices and Robots: From self-voice misperceptions to robotically-induced hallucinations

Pavo Orepic

University of Geneva, Switzerland; pavo.orepic@gmail.com

Why does our own voice sound unnatural to us and what does it have to do with hallucinations? Inspired by the prominent theory suggesting that auditory-verbal hallucinations (AVH) – colloquially "hearing voices" – arise as misattributions of self-generated speech, my PhD thesis revolved around self-voice perception and the experimental induction of self-other voice misattribution in healthy participants, mimicking the AVH phenomenology. In this talk, I will present key findings from my thesis. First, I will outline some behavioral, neural, and clinical factors underlying self-other voice discrimination. Second, I will illustrate how we induced AVH-like sensations in healthy, non-hallucinating individuals using a robotic procedure. Together, I will argue that self-voice is much more than just an auditory stimulus – it is a multisensory construct whose (mis)perception could potentially serve as a clinical biomarker for deficits in self-consciousness.

Session 5: Talker Identification

Time: Thursday, 29/Aug/2024: 1:30pm - 3:10pm · *Location:* Main Conference Hall (001)
Session Chair: Alexia Despina Leonidou

1:30pm - 2:00pm

ID: 164

Invited Talk

Speaker familiarity and spoken language processing

Susannah V. Levi

NYU, United States of America; svlevi@nyu.edu

To understand speech, listeners must parse the highly variable acoustic space into the appropriate, language-specific phonological categories and generalize these categories to novel stimuli. Given this high degree of variability in the speech signal, an important question is how listeners can become more accurate perceivers. In this presentation, I will discuss previous studies that demonstrate how speech perception can be improved through familiarity with a particular speaker or type of speaker. In addition, I will discuss the extent to which this familiarity can improve higher level processing, as well as some limitations that seem to arise in studies of speaker familiarity.

2:00pm - 2:20pm

ID: 123

Abstract

Keywords: voice similarity, acoustics, perception, speaker verification

Comparing voice similarity through acoustics, human perception and deep neural network (DNN) speaker verification systems

Suyuan Liu, Molly Babel, Jian Zhu

University of British Columbia, Canada; suyuan.liu@ubc.ca

Voice is composed of rich signals that deliver biological, physiological, psychological, social, and linguistic meaning. The similarity between voices can be a challenging concept to coherently wrangle. The goal of this study is to better understand voice similarity by comparing similarity metrics across acoustic analysis, perceptual judgments by human listeners, and automatic speaker verification systems. We focus on spontaneous speech from the English portion of the Speech in Cantonese and English (SpiCE) corpus (Johnson et al. 2020). In our comparison of vocal similarity, we compare (i) similarity scores generated from 24 acoustic dimensions (Kreiman and Sidtis 2011); (ii) speaker verification scores generated by seven pretrained speaker verification models using Wespeaker (Wang et al 2023); (iii) perceptual similarity from human listeners in an AX discrimination task, and (iv) perceptual (dis)similarity from an independent group of human listeners in a rating task. The output of our Bayesian regression models suggests that when controlling for the specific talkers being compared, the speaker verification models correlate with the psychoacoustic similarity scores, but not with either listener-based measure. When the pairs of voices being compared are not controlled, there is a relationship between listeners and speaker verification models. We take this to suggest that assessments of similarity manifest differently when the focus is on the gross- versus fine-phonetic levels. We discuss these results in the context of quantifying similarity for measuring phonetic imitation and understanding it as a linguistic process.

References

- Johnson, K. A., Babel, M., Fong, I., & Yiu, N. (2020). SpiCE: A new open-access corpus of conversational bilingual speech in Cantonese and English. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 4089-4095).
- Kreiman, J., & Sidtis, D. (2011). Foundations of voice studies: An interdisciplinary approach to voice production and perception. John Wiley & Sons.
- Wang, H., Liang, C., Wang, S., Chen, Z., Zhang, B., Xiang, X., Deng, Y., and Qian, Y. (2023). Wespeaker: A research and production oriented speaker embedding learning toolkit. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1-5). IEEE.

2:20pm - 2:50pm

ID: 165

Invited Talk

Do high variability speaking styles per se enhance voice recognition?

Volker Dellwo

Universität Zürich, Switzerland; volker.dellwo@uzh.ch

Humans have the innate ability to recognise individuals by their voice. Recent evidence suggests that increased acoustic variability in vocal dimensions such as fundamental frequency (f0) may enhances voice recognition processes. According to this, we predicted that speaking styles triggering high f0 variability such as infant-directed and clear speech are better suited for voice recognition. Our results, however, do not support this hypothesis: while learning speaker identities from infant-directed speech has a positive impact on voice recognition performance, clear speech impedes it. To understand this, we are now analysing f0 variability in infant-directed, clear, and read speech produced by the same speakers across three different languages (English, Chinese and Persian). Preliminary findings indicate both clear and infant-directed speech contain higher f0 variability compared to read speech, however, infant-directed speech exhibits greater f0 variability within steady-state vocalic segments, while clear speech shows increased f0 variability between vowels. Thus, high variability speaking styles differ in the way f0 contours are distributed withing the utterances. We argue that these systematic variability differences between speaking styles have evolved to support more or less the recognition of individuals by their voice: infant-directed speech may facilitate stronger individual recognition for stronger mother-child bonding, while clear speech may focus on reducing voice individuality features in favour of enhancing linguistic information.

2:50pm - 3:10pm

ID: 102

Abstract

Keywords: speaker identity coding, self-supervised models, automatic speaker recognition, speaker perception, signal perturbation

Investigating speaker identity representations in speech-based neural network models

Gasser Elbanna[1,2], **Fabio Catania**[2], **Satrajit S. Ghosh**[1,2]

[1]Harvard University, United States of America; [2]MIT, United States of America; gelbanna@mit.edu

Speaker identity processing involves the recognition and/or discrimination of voices [1, 2, 3]. Humans perform this task on a daily basis, yet the presence of between-speaker and within-speaker variability profoundly influences our perception [4, 5]. Previous attempts to identify acoustic correlates of identity perception have yielded inconclusive results [6]. Thanks to the advances in deep learning, particularly self-supervised models (SSMs), several data-driven features have shown superior performance over acoustic ones. While a significant body of research has assessed the predictive power of SSMs on various audio tasks [7] as well as speech tasks [8], little work has been done to examine the encoding spaces of these models. This study explores the suitability of data-driven models, specifically SSMs, as potential models for human perception. We analyze a suite of SSMs, including generative, predictive, and contrastive models, alongside expert supervised models for speaker recognition [9], and handcrafted acoustic features with a focus on exploiting commonalities and differences of how models represent voice identities. Also, we examine the equivariances and limitations of all models through experiments that perturb specific aspects of speech to demonstrate their impact on preserving identity. By evaluating speaker identification accuracy across acoustic, phonemic, prosodic, and linguistic variants, we draw similarities between model performance and human identity perception. These empirical findings provide both enhanced interpretability to these data-driven representational spaces and also support using this family of models as candidates to study speaker identity perception in humans.

References

- [1] Diana Roupas Van Lancker and Gerald J Canter. “Impairment of voice and face recognition in patients with hemispheric damage”. In: Brain and cognition 1 (1982).
- [2] Diana Van Lancker and Jody Kreiman. “Voice discrimination and recognition are separate abilities”. In: Neuropsychologia 25 (1987).
- [3] Diana Roupas Van Lancker, Jody Kreiman, and Jeffrey Cummings. “Voice perception deficits: Neuroanatomical correlates of phonagnosia”. In: Journal of clinical and experimental neuropsychology 11 (1989).
- [4] Nadine Lavan, Luke FK Burston, and L ́ucia Garrido. “How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices”. In: British Journal of Psychology 110 (2019).
- [5] Sarah V Stevenage, Ashley E Symons, Abi Fletcher, and Chantelle Coen. “Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task”. In: Quarterly Journal of Experimental Psychology 73 (2020).
- [6] Stefan R Schweinberger and Romi Z ̈aske. “Perceiving speaker identity from the voice”. In: (2018).
- [7] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Bj ̈orn W Schuller, Christian J Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al. “HEAR: Holistic Evaluation of Audio Representations”. In: NeurIPS 2021 Competitions and Demonstrations Track. PMLR. 2022.
- [8] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. “Superb: Speech processing universal performance benchmark”. In: arXiv preprint arXiv:2105.01051 (2021).
- [9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. “Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification”. In: arXiv preprint arXiv:2005.07143 (2020).

Session 6: Neurobiology of Voice

Time: Thursday, 29/Aug/2024: 3:40pm - 4:50pm · *Location:* Main Conference Hall (001)
Session Chair: Mathias Scharinger

3:40pm - 4:10pm
ID: 156
Invited Talk

The time course of first impression formation in the brain

Nadine Lavan

Queen Mary University, London, United Kingdom; n.lavan@qmul.ac.uk

When listeners hear a voice, they rapidly form a complex first impression of who the person behind that voice might be. We characterize how these multivariate first impressions from voices emerge over time across different levels of abstraction using electroencephalography and representational similarity analysis. We find that for eight perceived physical (gender, age, and health), trait (attractiveness, dominance, and trustworthiness), and social characteristics (educatedness and professionalism), representations emerge early (~80 ms after stimulus onset), with voice acoustics contributing to those representations between ~100 ms and 400 ms. While impressions of person characteristics are highly correlated, we can find evidence for highly abstracted, independent representations of individual person characteristics. These abstracted representationse merge gradually over time. That is, representations of physical characteristics (age, gender) arise early (from ~120 ms), while representations of some trait and social characteristics emerge later (~360 ms onward). The findings align with recent theoretical models and shed light on the computations underpinning person perception from voices.

4:10pm - 4:30pm
ID: 103
Abstract

Keywords: voice area, deep neural network, voice reconstruction, representational similarity analysis, fmri

Reconstructing voice from noninvasive auditory cortex recordings

Charly Lamothe[1,2], Etienne Thoret[1,2,3,4], Régis Trapeau[1], Bruno Giordano[1], Julien Sein[1,5], Sylvain Takerkart[1], Stéphane Ayache[2], Thierry Artières[2,6], Pascal Belin[1]

[1]La Timone Neuroscience Institute UMR 7289, CNRS, Aix-Marseille University, Marseille, France; [2]Laboratoire d'Informatique et Systèmes UMR 7020, CNRS, Aix-Marseille University, Marseille, France; [3]Perception, Representation, Image, Sound, Music UMR 7061, CNRS, Marseille, France; [4]Institute of Language Communication & the Brain, Marseille; [5]Centre IRM-INT@CERIMED, Marseille, France; [6]École Centrale de Marseille, Marseille, France; charly.lamothe@pasteur.fr

The cerebral processing of voice information is known to engage, in human as well as non-human primates, "temporal voice areas" (TVAs) that respond preferentially to conspecific vocalizations. However, how voice information is represented by neuronal populations in these areas, particularly speaker identity information, remains poorly understood. Here, we used a deep neural network (DNN) to generate a high-level, small-dimension representational space for voice identity—the 'voice latent space' (VLS)—and examined its linear relation with cerebral activity via encoding, representational similarity, and decoding analyses. We find that the VLS maps onto fMRI measures of cerebral activity in response to tens of thousands of voice stimuli from hundreds of different speaker identities and better accounts for the representational geometry for speaker identity in the TVAs than in A1. Moreover, the VLS allowed TVA-based reconstructions of voice stimuli that preserved essential aspects of speaker identity as assessed by both machine classifiers and human listeners. These results indicate that the DNN-derived VLS provides high-level representations of voice identity information in the TVAs.

4:30pm - 4:50pm

ID: 142

Abstract

Keywords: speaker processing, voice processing, EEG, ERP

Early neural interactions of speech and speaker information

Paula Rinke[1,3], Nadine Lavan[2], Mathias Scharinger[1,3]

[1]Philipps-University Marburg, Germany; [2]School of Biological and Behavioural Sciences, Queen Mary University of London, United Kingdom; [3]Center for Mind, Brain & Behaviour »CMBB«, Universities of Gießen & Marburg, Germany; rinkepa@staff.uni-marburg.de

Recent neurophysiological research reports that linguistic and extralinguistic information are perceived rapidly within the first hundreds of milliseconds of a speech signal. Speech perception, such as speech sound categorization, was found to be influenced by speaker characteristics such as age or gender indicating an interaction of speech and speaker processing.

In the present EEG study using event-related potentials (ERPs), we investigate the neural interaction of speech (different vowel categories) and speaker information (speakers' age and gender) in the early N1 time window (80-150 ms after stimulus onset). 35 German-speaking participants (mean age = 25.22 years) were presented with 96 recordings of the German vowels [a], [i] and [u] by 32 male and female (n=16) speakers of either older (n=16) or younger age. Participants completed a 1-bask task while brain activity was recorded from 32 active electrodes.

Linear Mixed Models based on fronto-central electrodes revealed that there was an interaction between vowel category and the speaker's gender for the N1 amplitude and latency. Male voices elicited stronger and later N1 responses than female voices. For vowel category, [a] elicited the smallest and earliest N1 peak while [i] evoked the strongest and [u] the latest peak.

Speaker's age did not affect N1 properties. However, there were significant interactions of speaker's age with gender and vowel category on N1 amplitude: gender effects were more pronounced for younger speakers while vowel effects were more pronounced for older speakers.

The results indicate an early interaction of speech and speaker information in the N1 time window.

Session 7: Voice and Speech and Music

Time: Friday, 30/Aug/2024: 9:30am - 11:00am · *Location:* Main Conference Hall (001)
Session Chair: Toke Hoffmeister

9:30am - 10:00am

ID: 166

Invited Talk

How does voice familiarity affect speech intelligibility?

Emma Holmes

UCL, United Kingdom; emma.holmes@ucl.ac.uk

People often face the challenge of understanding speech when other sounds are present ("speech-in-noise perception")—which involves a variety of cognitive processes, such as attention and prior knowledge. We have consistently found that familiarity with a person's voice improves the ability to understand speech-in-noise, using both naturally familiar (e.g., friends and partners) and lab-trained voices. In this talk, I will describe experiments in which we aimed to gain insights into the processes underlying the familiar-voice intelligibility benefit, contrasting explanations based on predictions of voice acoustics with those involving cognitive demands. I will discuss the implications of our findings for theories of speech perception, and the implications for populations who typically find speech perception particularly challenging (e.g., older adults).

10:00am - 10:30am

ID: 167

Invited Talk

Neural and computational mechanisms underlying talker-phonetic interactions

Sahil Luthra

Carnegie Mellon University, United States of America; sahilluthra@cmu.edu

Individual talkers differ idiosyncratically in how they produce their speech sounds, even within a single dialect. Theoretical accounts agree that listeners must accommodate this talker-specific phonetic variation as they map from the acoustic signal to mental representations of speech sounds; however, the exact cognitive mechanisms that support talker-invariant speech perception remains underspecified. Neurobiology poses an important constraint on cognition, in the same way that hardware can constrain software performance; as such, the study of the brain can inform mechanistic accounts of speech perception. In this talk, I will review recent neural evidence that phonetic processing and talker processing are supported by separate but overlapping neural systems — specifically, a left-lateralized system for speech perception and a right-lateralized system for vocal identity processing. I will show fMRI data indicating that interactions between talker and phonetic processing are achieved in bilateral portions of the superior temporal gyrus, allowing listeners to identify a talker's "phonetic signature" even when one hemisphere is inhibited using TMS. Finally, I will present simulation results that highlight how computational constraints may create pressure for a neurobiological system to adopt a dual-stream architecture in the first place. Overall, this research helps clarify the extent to which talker processing and phonetic processing interact, informing mechanistic accounts of speech perception.

10:30am - 11:00am

ID: 113

Abstract

Keywords: individual differences, voice perception, singing, aesthetic preferences, acoustics

Individual differences in singing voice perception

Camila Bruder[1], Pauline Larrouy-Maestri[1,2]

[1]Max Planck Institute for Empirical Aesthetics, Germany; [2]Max Planck-NYU Center for Language, Music, and Emotion (CLaME), USA & Germany; camila.bruder@ae.mpg.de

Singing is highly significant to human experience, but research about the basis of singing voice preferences is scarce. To investigate the relationship between attributes of singing performances and their appreciation, we asked 42 lay listeners to rate 96 a cappella singing performances in terms of how much they liked them, as well as on 10 different perceptual scales (Bruder et al., 2024). Results showed that participants' liking ratings could (partially) be explained based on perceptual ratings of the stimuli, but not based on computationally-extracted descriptions of the acoustic signal. Crucially, we found high individual differences not only in participants' preferences, but also in how participants perceived the voices, as indicated by the remarkably low interrater agreement in all tested scales (Intraclass Correlations [type 2,1] of .14 for liking, and between .11 and .27 for the perceptual scales). Inspection of the relationship between acoustic measurements (e.g., pitch interval deviation, vibrato extent, cepstral peak prominence) and corresponding perceptual ratings (e.g., perceived pitch accuracy, amount of vibrato, and breathiness, respectively) for each participant showed correlations ranging from nonexistent to moderate, suggesting that some participants were more "objective" or "acoustically sensitive" to certain voice attributes. Interestingly, for ratings of vibrato and breathiness, these "acoustic sensitivities" were positively associated with participants' general music sophistication. Ongoing work delves further into individual differences in the aesthetic evaluation of both singing and speaking voices with an integrative approach that aims to bridge findings on speaking voice attractiveness with the present research on the singing voice.

References

- Bruder, C., Poeppel, D., & Larrouy-Maestri, P. (in press). Perceptual (but not acoustic) features predict singing voice preferences. Scientific Reports (in press).

Session 8: Challenges in Talker Identification

Time: Friday, 30/Aug/2024: 11:30am - 12:40pm · *Location:* Main Conference Hall (001)
Session Chair: Paula Rinke

11:30am - 12:00pm
ID: 168
Invited Talk

Voice cue sensitivity and speech-on-speech perception in children with hearing aids

Laura Rachman
University Medical Center Groningen, Netherlands, The; l.rachman@umcg.nl

Voice cues, such as fundamental frequency (F0) and vocal-tract length (VTL), allow listeners to distinguish speakers, which can facilitate speech perception in challenging listening conditions. Hearing loss may hinder access to relevant acoustic cues, and in children, perceptual cognitive mechanisms may still be developing, all affecting speech-on-speech perception. We investigated F0 and VTL sensitivity and speech perception in single-talker maskers with differences in F0 and VTL in school-age children (5-18) with hearing aids and with normal hearing. Both groups of children show trends for development of voice cue sensitivity and speech-on-speech perception as a function of age. The developmental patterns for F0 and VTL sensitivity seem to differ. Hearing-aided children seem to catch up with age-typical development around teenage years for F0 sensitivity, but not for VTL sensitivity. For speech-on-speech perception, both groups show a benefit of F0 and VTL differences between target and masker speech, but the VTL difference benefit is relatively smaller for the hearing-aided participants. Our results show a large variability in hearing-aided children, with some children performing at the level of age-matched normal-hearing children, while others performing lower. Given the age effect, baseline developmental data is essential for the evaluation of speech-on-speech perception in individual hearing-aided children with respect to their age, and for the adjustment of rehabilitation to fit their needs. The finding of many hearing-aided children performing at age-expected levels indicates that hearing aids can provide good compensation for hearing loss for voice cue sensitivity and for speech-on-speech perception.

12:00pm - 12:20pm
ID: 148
Abstract

Keywords: aging, hearing impairment, talker identification, speech perception, talker change detection

Talker identification under adversity: Effects of noise, aging, and hearing loss

Tyler Perrachione
Boston University, United States of America; tkp@bu.edu

Talker identification research typically investigates this ability under ideal conditions, presenting voices in quiet to young listeners with normal hearing. However, ecological talker identification takes place in noisy environments and by listeners of all ages and hearing abilities. How do these factors—masking noise, aging, and hearing loss—affect talker identification? Here, I discuss four studies using three experimental paradigms examining talker identification under adverse listening conditions. First, using a talker adaptation task to measure how speech perception in noise improves by 'tuning in' to a talker's voice, we found that the threshold signal-to-noise ratio for word identification was better when listening a single continuous talker compared to changing talkers. Second, in two explicit talker identification tasks, we found that (1) different kinds of masking noise (speech-shaped noise, multi-talker babble, and a single competing talker) impose unique challenges for talker identification, (2) identification of some voices is more susceptible to masking noise, and (3) talker identification is affected by both aging and hearing loss, with older hearing-impaired listeners identifying voices least accurately. Interestingly, older normal-hearing listeners also struggle with talker identification even in quiet, suggesting roles for both central and peripheral factors in talker identification. Finally, using a talker change detection task in quiet and noise, we found that aging and hearing loss differentially affect "telling together" and "telling apart" aspects of voice perception. Together, these findings suggest that talker identification plays an important communicative role by supporting perceptual organization of auditory features for speech communication in adverse listening environments.

12:20pm - 12:40pm

ID: 108

Abstract

Keywords: Cochlear Implant; Diagnostics; Socio-Emotional Communication; Quality of Life

Identifying individual profiles of voice perception abilities in cochlear implant users and their relationship to quality of life (QoL)

Stefan R. Schweinberger[1,2,3,4], **Verena G. Skuk**[1,2], **Romi Zäske**[1,2,5], **Celina I. von Eiff**[1,2]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany; [2]Voice Research Unit, Friedrich Schiller University, Jena, Germany; [3]Swiss Center for Affective Sciences, University of Geneva, Switzerland; [4]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany; [5]Department of Experimental Otorhinolaryngology, Jena University Hospital, Germany; stefan.schweinberger@uni-jena.de

Cochlear implants (CIs) give hearing to many deaf people with sensorineural hearing loss, but signals conveyed to the auditory nerve via a CI differ substantially from biologically transmitted signals. While extensive research exists on speech perception with a CI, research on the ability to perceive nonverbal social signals (emotional prosody, speaker gender, age, or identity) is largely missing. Where quality of life (QoL) is at the focus, this is more than unfortunate, because recent research from our group and other laboratories suggests that nonverbal abilities for communicating socio-emotional cues can be at least as important for QoL with a CI as speech perception. This holds not only for children (Schorr et al., 2009) but also for younger and older adults with a CI (Luo et al., 2018; Schweinberger & von Eiff, 2022; Skuk et al., 2020, von Eiff et al. (2022a,b). In this presentation, we integrate these recent findings, before developing a new perspective for what could be labelled CI precision diagnostics. Specifically, we discuss evidence to support our proposal that future CI diagnostics, by default, should consider a large range of communicative abilities and their relation to QoL. We particularly consider skills in four relevant domains of voice perception (emotion, identity, gender, age), but also speech perception and musical skills. We discuss perspectives for using CI precision diagnostics for perceptual trainings, and how the present approach could be relevant for future CI design and development.

References

- Luo, X., Kern, A., & Pulling, K. R. (2018). Vocal emotion recognition performance predicts the quality of life in adult cochlear implant users. Journal of the Acoustical Society of America, 144(5), EL429-EL435. doi:10.1121/1.5079575
- Schorr, E. A., Roth, F. P., & Fox, N. A. (2009). Quality of Life for Children With Cochlear Implants: Perceived Benefits and Problems and the Perception of Single Words and Emotional Sounds. Journal of Speech, Language, and Hearing Research, 52(1), 141-152.
- Schweinberger, S. R., & von Eiff, C. I. (2022). Enhancing socio-emotional communication and quality of life in young cochlear implant recipients: Perspectives from parameter-specific morphing and caricaturing. Frontiers in Neuroscience, 16, 956917. doi:10.3389/fnins.2022.956917
- Skuk, V.G., Kirchen, L., Oberhoffner, T., Guntinas-Lichius, O., Dobel, C., & Schweinberger, S.R. (2020). Parameter-specific Morphing Reveals Contributions of Timbre and F0 Cues to the Perception of Voice Gender and Age in Cochlear Implant Users. Journal of Speech, Language, and Hearing Research, 63(9), 3155-3175.
- von Eiff, C. I., Frühholz, S., Korth, D., Guntinas-Lichius, O., & Schweinberger, S. R. (2022). Crossmodal benefits to vocal emotion perception in cochlear implant users. iScience, 25(12), 105711. doi:10.1016/j.isci.2022.105711
- von Eiff, C.I., Skuk, V.G., Zäske, R., Nussbaum, C., Frühholz, S., Feuer, U., Guntinas-Lichius, O., & Schweinberger, S.R. (2022). Parameter-specific morphing reveals contributions of timbre to the perception of vocal emotions in cochlear implant users. Ear and Hearing, 43(4), 1178-1188. https://doi.org/10.1097/AUD.0000000000001181

Poster Session 1

Time: Wednesday, 28/Aug/2024: 12:45pm - 2:15pm · *Location:* Poster Room (101/102)

ID: 120

Abstract

Keywords: voice quality, speech perception, discrimination, similarity

Matching perceptual similarity between voices in Polish

Geoffrey Schwartz, Ewelina Wojtkowiak, Kamil Kaźmierski, Maral Asiaee
Adam Mickiewicz University, Poznań, Poland; geoff@amu.edu.pl

Underlying the study of voice quality are theoretical questions concerning the relationship between phonetics and phonology. In particular, can the voice play any role in the phonological structure of a language without phonemic phonation contrasts (Schwartz et al. 2023), or is voice quality non-linguistic? A related question concerns a given language's common phonation properties (Asiaee et al. 2023) that distinguish it from other languages. In our research, we examine these questions in Polish, and in the speech of Polish learners of English.

One way of investigating these questions is to test listeners' judgements about similarity between voices, and investigating how those data are linked with acoustic measures of voice quality. We report here on a *similarity matching* experiment, similar in design to AXB discrimination. In each trial, listeners heard three voices producing the same Polish phrase, and were asked to respond whether the voice in the second item (X) was more similar to the first item (A) or the third item (B). For a variety of voice quality measures, the acoustic differences between X and A, and X and B, was obtained, and the likelihood that acoustic proximity matched listeners' similarity judgments was calculated for each parameter. Acoustic proximity in HNR and H2K*-H5K showed the highest proportion of matches with listener judgements. Lower-frequency spectral tilt measures (H1*-H2*; H2*-H4*) showed the lowest proportion of matches. Hierarchical clustering analysis revealed three groups of listeners, with different relative weights of the acoustic measures.

ID: 115

Abstract

Keywords: gender, nonbinary speakers, gender-role self-concept, sociophonetics, acoustic phonetics

Diversify identity: Acoustic correlates of multiple gender facets in nonbinary and cisgendered speakers

Leah Nieber[1], Sven Kachel[2,3], Mathias Scharinger[1]
[1]Philipps-Universität Marburg; [2]Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau; [3]University of Helsinki;
leah.nieber@gmail.com

Phonetic gender research is a growing field within sociolinguistics. While early studies were primarily concerned with acoustic differences between women and men, the current study considers gender as a multifaceted and multidimensional construct. Correspondingly, different aspects of the gendered self are considered, including social and biological dimensions. One component of social gender that is rarely addressed is the gender-role self-concept (the extent to which someone describes themself as typically masculine and/or feminine). The present research aims to empirically investigate acoustic correlates of different gender facets using a gender-diverse sample of speakers.

In a lab study, N = 52 nonbinary, cis female, and cis male speakers were recorded. Relying on semi-spontaneous speech, speakers were asked to describe a set of pictures used to elicit specific target words. Established acoustic gender markers (fundamental frequency characteristics, spectral moments of sibilants, vowel space features) were examined using quantitative analysis. Additionally, different gender facets were assessed employing a socio-psychological questionnaire (e.g., gender identity, gender-role self-concept, gendered socialization).

Results show speakers to differ in their acoustic features according to their gender identities. Additionally, for nonbinary speakers, differences have been found between those who have been assigned female and those who have been assigned male at birth. Further, within each gender identity group, gender-role self-concept correlated with some acoustic features (e.g., shimmer). Taken together, our study shows the acoustic and gender-related heterogeneity of nonbinary speakers and presents how the acoustic markers are influenced by the various dimensions of the gendered self.

ID: 104

Abstract

Keywords: audio-visual asynchrony; intelligibility; trait perception

Factors affecting speech intelligibility and trait perception in virtual meeting contexts

Ziyun Zhang, Carolyn McGettigan

University College London, United Kingdom; ziyun.zhang.19@ucl.ac.uk, c.mcgettigan@ucl.ac.uk

We frequently use virtual meeting platforms as a replacement for in-person interaction. However, the nature of virtual meetings, especially under bad connections, brings challenges to interactive communication. While previous studies have reported the interactive effects of auditory and visual signal degradation on speech intelligibility (McGettigan et al., 2012) and the relative weighting of audio and visual cues for speech perception (Hazan et al., 2010), there is a surprising gap in our understanding of the effects of audiovisual synchrony on perception during spoken interactions. To address this gap, our study conducted an audio-visual sentence perception experiment in a laboratory setting to simulate common challenges encountered in online meetings, including drops in audio intensity and variable temporal lags between the talker's voice and face. The experiment was conducted with sentences extracted from CANDOR corpus with naturalistic Zoom recordings (Reece et al., 2022). The results showed that signal asynchrony and mismatches in audio-visual playback speed hinder speech intelligibility. Generally, the more severe the manipulation, the worse performance (e.g., synchronous > asynchronous; shorter > longer lags; lower speed > higher speed), though there were some exceptions. Factors affecting trait perceptions varied across different traits (trustworthiness, likability, aggressiveness, and dominance). Ongoing work developed from these findings is exploring the inter-brain neural correlations using fNIRS hyperscanning.

References (optional)

- Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. Speech Communication, 52(11), 996–1009.
- Mcgettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. Neuropsychologia, 50, 762–776.
- Reece, A., Cooney, G., Bull, P., Chung, C., Dawson, B., Fitzpatrick, C., Glazer, T., Knox, D., Liebscher, A., & Marin, S. (2023). The CANDOR corpus: Insights from a large multimodal dataset of naturalistic conversation. Science advances, 9(13), eadf3197.

ID: 107

Abstract

Keywords: self-similarity, kinship, parameter-specific morphing, trustworthiness, familiarity

Kinship perception in human voices: ratings of trust, dominance and familiarity in morphed voices at various levels of self-similarity

Ayaka Tsuchiya[1,2,3], Maria-Kezia Zangemeister[1], Stefan R. Schweinberger[1,2,3]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University, Jena, Germany; [2]Max Planck Research School for the Science of Human History, Jena, Germany; [3]Voice Research Unit, Friedrich Schiller University, Jena, Germany; ayaka.tsuchiya@uni-jena.de

Kinship is one of the crucial elements of human interaction, along with other social information such as identity, emotion, age, gender and personality traits. Kinship recognition helps us understand others better, establish self-identity in the community, and develop cooperative attitudes. In the present study, we investigated how people perceive self-similarity in voices as a potential cue of kinship, by using parameter-specific voice morphing. Voice recordings of 28 participants were paired with reference voices of four sex-matched speakers for the stimulus preparation. The voice pairs were morphed on the basis of either full morphs, F0 morphs, or timbre morphs at the level of 20, 40, 60, or 80 of the participant's own voice. During the experiment, participants were asked to rate these voices in terms of familiarity, trustworthiness and dominance. The results revealed that stimuli with a higher proportion of participants' own voice led to increased ratings of familiarity and trustworthiness, indicating that the participants processed self-similarity in voices as showed positive attitude towards self-similar speakers. Furthermore, we found that both F0 and timbre information contributed substantially to these familiarity and trustworthiness ratings. By comparison, findings were less clear for dominance ratings, with a trend for F0 - but not timbre - to affect dominance ratings. Overall, these findings provide empirical evidence for the perception of self-similarity in human voices and its impact on trait evaluations.

ID: 121
Abstract
Keywords: vocal persona, voice perception, acoustic manipulation, expressive speech, text-to-speech

Manipulating acoustic correlates for vocal persona transition: From neutral to friendly

Chenyi Lin
University of Groningen, Netherlands, The; c.lin.22@student.rug.nl

The concept of vocal persona, representing the identity or character perceived through an individual's voice, exhibits dynamic variability as it adapts to different social contexts. This adaptability, crucial for the naturalness and effectiveness of Text-to-Speech (TTS) systems, remains insufficiently explored, particularly in the realm of attitudinal nuances like transitioning from a neutral to a friendly tone. Addressing the limitations in available speech datasets, which predominantly lack diverse attitudinal tones, our study employs specific acoustic manipulations (namely alterations in pitch, duration, and energy) to facilitate the perceptual transition of vocal personas from neutral to friendly in Mandarin Chinese TTS, employing the FastSpeech2 framework. We propose an examination of these acoustic features' individual and combined effects on enhancing the friendliness of synthesized speech. Through a controlled experimental setup, our research quantifies these perceptual shifts using mean opinion scores (MOS).

Based on the findings of Chen et al. (2004) and Li et al. (2004), we anticipate that increasing the mean pitch of a neutral voice alone will significantly influence friendliness perception. Moreover, integrating it with shorter phone duration and slightly raised energy is expected to further optimize the friendliness MOS. This exploration will deepen our understanding of voice persona modulation, offering valuable insights for advancing TTS technology. By bringing the acoustic underpinnings of vocal persona transitions to light, our findings aim to contribute to more expressive and engaging TTS applications, with broader implications for voice branding, assistive technology, and human-computer interaction.

References

- Chen, F., Li, A., Wang, H., Wang, T., & Fang, Q. (2004, May). Acoustic analysis of friendly speech. In 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (Vol. 1, pp. I-569). IEEE.
- Li, A., Chen, F., Wang, H., & Wang, T. (2004). Perception on synthesized friendly standard Chinese speech. In International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

ID: 153
Abstract
Keywords: Temporal voice areas ; synthetic voice; norm-based coding

Temporal Voices Areas response to synthetic voices in natural conversation support norm-based coding of vocal identity information

Thierry Chaminade, Camilla Di Pasquasio, Arthur Pineaud, Pascal Belin
Aix-Marseillle Université - CNRS, France; thierry.chaminade@univ-amu.fr

With recent technical advances in robotics and artificial intelligence, humanoids have proven to be valuable tools to investigate mechanisms of human social interactions (Wykowska et al., 2016). While many behavioural features are relevant for social interactions, voices, together with faces, are among the most impactful (Belin, 2018) but also the less studied (Fischer & Niebuhr, 2023).

While their responsiveness of temporal vocal areas (TVAs) to vocal stimulations is now clearly established, little is known about the factors that impact their response (Belin et al., 2000). It suggests that response in Temporal Voice Areas increases as individual voices diverges from an average human voice (Latinus et al., 2013). Here, we sought to investigate the response in Temporal Voice Areas to synthetic voices. As these voices are significantly different from human voices, norm-based coding suggests an increased response in TVAs. We performed an analysis using an existing corpus of 25 participants recorded with functional MRI while engaging in conversations with human and robot interlocutors, the later endowed with a text-to-speech synthesis. We focused on periods when participants listened to the interlocutor. Results (Figure, left) indicate that TVAs identified by an increased response to human voices compared to the implicit baseline show an additional increased response to synthetic compared to human voices. Reciprocal comparisons between the artificial and natural interlocutors (Figure, right) revealed that the temporoparietal junctions associated with mental states attributions is more active when listening to the former. Altogether, our results suggest Temporal Voice Areas increased activity signals synthetic voices as unnatural.

References

- Belin, P. (2018). The Vocal Brain: Core and Extended Cerebral Networks for Voice Processing. In S. Fruhholz & P. Belin (Eds.), The Oxford Handbook of Voice Perception (pp. 37-59): Oxford University Press. In The Oxford Handbook of Voice Perception, S. Frühholz & P. Belin (Eds.), (Oxford University Press, pp. 37–59).
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. Nature, 403(6767), 309-12.
- Fischer, K., & Niebuhr, O. (2023). Which Voice for which Robot? Designing Robot Voices that Indicate Robot Size. ACM Transactions on Human-Robot Interaction, 12(4), 1–24. https://doi.org/10.1145/3632124
- 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
- Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. Philosophical Transactions of the Royal Society B: Biological Sciences, 371(1693), 20150375. https://doi.org/10.1098/rstb.2015.0375

ID: 124
Abstract
Keywords: forensic phonetics, big data, voice perception, automatic speaker recognition, voice discrimination

Accurate and fast: Investigating efficiency in voice discrimination

Andrea Fröhlich[1,2], **Meike Ramon**[4], **Peter French**[3], **Volker Dellwo**[2,3]

[1]Zurich Forensic Science Institute; [2]University of Zurich, Department of Computational Linguistics; [3]University of Zurich, Centre for Forensic Phonetics and Acoustics (CFPA); [4]University of Lausanne, Applied Face Cognition Lab; andrea.froehlich@uzh.ch

Dealing with large volumes of low-quality data remains a challenge in forensic speaker comparison, especially considering the time constraints present in ongoing investigations. While automatic speaker recognition systems can handle data quickly, they often fall short with degraded audio quality, necessitating human post-processing. We therefore propose to leverage the perceptual skills of voice super-recognizers to assist in this process (Fröhlich et al. 2023; Ruch, Fröhlich, and Lim 2023).

Recent findings investigating face processing skills in police officers found a correlation between participants' reaction time and accuracy. Furthermore, they found that highly skilled face processors were more efficient (accurate *and* fast) than average or low-performing individuals (Nador et al. 2022).

While developing tests for the identification of voice super-recognizers, we investigated the relationship between reaction time (RT) and accuracy with data from a speaker discrimination experiment with 38 participants (Fröhlich et al. 2023), to see if there is a speed-accuracy trade-off (Bruyer and Brysbaert 2011). If a correlation between performance and RT can be found, these two measurements could be combined into the inverse efficiency score (Townsend and Ashby 1983) to be used as an efficiency-measure for voice super-recognisers.

We found no correlation between the mean accuracy and RT overall for the bespoke data. However, at the individual participant level, we found r-values of -0.4 or more for some participants. These first results indicate that the performance / RT relationship is highly individual. Further investigations are needed and are currently ongoing.

References

- Bruyer, Raymond, and Marc Brysbaert. 2011. 'Combining Speed and Accuracy in Cognitive Psychology: Is the Inverse Efficiency Score (IES) a Better Dependent Variable than the Mean Reaction Time (RT) and the Percentage of Errors (PE)?' PSYCHOLOGICA BELGICA 51(1):5–13.
- Fröhlich, Andrea, Volker Dellwo, Peter French, and Meike Ramon. 2023. 'ASR-BASED DEVELOPMENT OF CHALLENGING SPEAKER DISCRIMINATION TESTS'.
- Nador, Jeffrey D., Michael Vomland, Markus M. Thielgen, and Meike Ramon. 2022. 'Face Recognition in Police Officers: Who Fits the Bill?' Forensic Science International: Reports 5:100267. doi: 10.1016/j.fsir.2022.100267.
- Ruch, Hanna, Andrea Fröhlich, and Sarah Lim. 2023. 'Clustering a Large Number of Unknown Voices'. P. 23 in 31st IAFPA Conference.
- Townsend, James T., and F. Gregory Ashby. 1983. Stochastic Modeling of Elementary Psychological Processes. CUP Archive.

ID: 122
Abstract
Keywords: sociophonetics, acoustic analysis

Diachronic change in a politician's accent: a case study of Jonathan Reynolds MP

Thomas Devlin

University of Huddersfield, United Kingdom; t.devlin@hud.ac.uk

I present an acoustic analysis of the speech of one individual - Jonathan Reynolds MP - across a span of just over a decade from 2011 to 2023, showing how different accent features change or remain stable across this period. The individual is a UK politician, born in north-eastern England, living and representing a constituency in north-western England, and spending a large amount of time in the UK parliament in London. These three geographic influences on pronunciation are used to explore the speech features examined. Similar panel studies (e.g. Harrington et al 2000) have shown a large degree of variation diachronically, suggesting that features can and do change over time even within the same speaker. The analysis comprises calculating the first two formant frequencies or vocal tract responances in a range of vowels (FACE, GOAT and FOOT/STRUT in Wells 1982 lexical sets) in Praat acoustic analysis software (Boersma & Weenink 2024) as well as an auditory analysis of the consonant /t/ word-medially. The results show that the speaker uses a greater amount of fronted FOOT, and more frequent STRUT vowels, suggesting a style switch towards features more usually associated with southern English, as well as reduced instances of glottal /t/ in favour of aspirated /t/, suggesting the development of a more formal speech style. However, usage of FACE and GOAT shifts towards more northern local pronunciation, with greater monophthong usage over diphthongs, over the same period. The results are discussed in terms of salience and regional and professional identity.

References

- Boersma, P. & Weenink, D. 2024. Praat: doing phonetics by computer [Computer program]. Version 5.3.35. Retrieved: 14 April 2024: http://www.praat.org/
- Harrington, J., Palethorpe, S. & Watson, C. 2000. Does the Queen speak the Queen's English?. Nature 408, 927–928, https://doi-org.libaccess.hud.ac.uk/10.1038/35050160
- Wells, J. 1982. The Accents of English. Vol. 1. Cambridge: Cambridge University Press.

ID: 116

Abstract

Keywords: voice, prosody, AI generated speech, text-to-speech, TTS

Effect of syntax and semantics on the perception of "humanness" in synthetic and natural speech.

Janniek M Wester[1], **Pauline Larrouy-Maestri**[1,2]

[1]Department of Music, Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany; [2]Max Planck NYU, Center for Language, Music, and Emotion (CLaME), New York, United States; janniek.wester@ae.mpg.de

The quality of computer generated speech keeps getting better but people are still able to identify whether speech is produced by a human or not (Bruder & Larrouy-Maestri, 2023). The lack of humanness in synthetic speech might be linked to the naturalness or quality of its prosody. In order to understand this issue we focus on the role of syntactic and semantic aspects of language (which are known to be intrinsically linked to prosody (Cutler et al., 1997)) on the perception of humanness in speech. We created German sentences that were then manipulated to obtain three more conditions conveying syntactic but no semantic information (jabberwocky sentences), no syntactic but semantic information (wordlists), and no syntactic or semantic information (jabberwocky wordlists). This material was generated by text-to-speech tools (Microsoft, Elevenlabs, Murf, and Listnr) and recorded by eight human speakers. Native German listeners will evaluate hundreds of these speech samples in terms of "how human it sounds to them". Using a linear mixed model approach we will explore the effects of voice category (natural versus synthetic) and of syntactic and semantic information on humanness perception. We hypothesize that sentences, which is by definition the condition that is most familiar to listeners, will be perceived as more human than jabberwocky and wordlists. Whether syntactic and semantic information interacts with voice category (in addition to the expected main effect, i.e., synthetic speech perceived as less human than natural speech) will inform us about the potential differences in processing synthetic and natural speech.

References

- Bruder, C., & Larrouy-Maestri, P. (2023, September 06-09). Attractiveness and social appeal of synthetic voices. Paper presented at the 23rd Conference of the European Society for Cognitive Psychology, Porto, Portugal.
- Cutler, A., Dahan, D., & Van Donselaar, W. (1997). Prosody in the Comprehension of Spoken Language: A Literature Review. Language and Speech, 40(2), 141–201. https://doi.org/10.1177/002383099704000203

ID: 112

Abstract

Keywords: voice hearing, hallucination-proneness, speech comprehension, auditory verbal hallucinations

Hearing voices: Altered voice perception and speech comprehension in hallucination-prone individuals

Hannah Ziesenies[1], **Jens Kreitewolf**[2]

[1]Leipzig University, Leipzig, Germany; [2]McGill University, Montreal, Canada; hannah.ziesenies@gmail.com

Auditory verbal hallucinations (AVH) occur in the context of various psychiatric and medical conditions and can cause high levels of distress in patients.

However, the processes leading to the emergence of AVHs remain unclear. Despite a focus on altered speech perception in the hallucination literature, AVHs also convey a wealth of information about the human voice, such as speaker identity; however, the topic of voice perception has so far received surprisingly little attention. Here, we present a study that sought to (a) quantify the degree to which voice perception is altered in individuals along the continuum of hallucination-proneness and (b) investigate a potential link between altered voice perception and deficient speech comprehension in those individuals. To this end, we determined thresholds for hearing voices in ambiguous auditory stimuli on continua ranging from environmental (e.g., water stream) to vocal sounds (e.g., multi-talker babble). These sounds were presented at four different spatial locations (left, right, front and back), allowing us to investigate whether this effect is direction-specific.

We hypothesize that increasing hallucination proneness is positively correlated with voice hearing thresholds. We further hypothesize that voice hearing thresholds are negatively correlated with the ability to comprehend speech at the cocktail party. Contrary to our hypothesis, preliminary data show that voice hearing thresholds increase with hallucination-proneness. These preliminary data also indicate location-specific differences in voice-hearing thresholds.

ID: 130
Abstract
Keywords: Impression Formation, Attractiveness, Multimodal Person Perception, Life Span, Age Estimation

Impression formation across the life-span: Differential contributions of face and voice

Helene Kreysa[1], **Romi Zäske**[1,2], **Stefan R. Schweinberger**[1,3,4]
[1]Friedrich Schiller University Jena, Germany; [2]Jena University Hospital, Germany; [3]Swiss Center for Affective Sciences, University of Geneva, Switzerland; [4]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany; helene.kreysa@uni-jena.de

When forming first impressions, information from face and voice is rapidly integrated into a unified representation of an individual (Young et al., 2020). Although many aspects of a person can be inferred from either modality, one modality may be preferred when both are available (Mileva et al., 2018). To disentangle differential contributions of faces and voices to first impression formation based on an individual's age, we combined photos of middle-aged faces (40-50 years old; Ebner et al., 2010) with voices that were younger or older than these faces (~20 vs. ~70 years old; Zäske et al., 2019). Participants were asked to rate each "audiovisual person" for trustworthiness, attractiveness, and dominance, and to estimate their age. In two experiments, we assessed how participants' own age affected these ratings: Experiment 1 tested 27 students (M = 22 years old); Experiment 2 tested 29 senior citizens (M = 75 years). The students rated faces combined with younger voices as more attractive than the same faces combined with older voices; for senior raters, voice age did not affect attractiveness ratings. For dominance ratings, the pattern reversed: Students experienced faces combined with older voices as more dominant; senior raters found faces with younger voices more dominant. Neither voice age nor rater age affected trustworthiness ratings. Age estimates reflected the age of the face relatively accurately. Despite age differences of ±20 years between faces and voices, face-voice pairings were rarely experienced as mismatching, suggesting that face and voice age can differ considerably without appearing implausible.

References

- Ebner, N.C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods, 42, 351–362. https://doi.org/10.3758/BRM.42.1.351
- Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2018). Audiovisual integration in social evaluation. Journal of Experimental Psychology: Human Perception and Performance, 44(1), 128-138. https://doi.org/10.1037/xhp0000439
- Young, A., Fruehholz, S., & Schweinberger, S. R. (2020). Face and voice perception: Understanding commonalities and differences. Trends in Cognitive Sciences, 24(5), 398-410. https://doi.org/10.1016/j.tics.2020.02.001
- Zäske, R., Skuk, V.G., Golle, J., & Schweinberger, S. R. (2020). The Jena Speaker Set (JESS)—A database of voice stimuli from unfamiliar young and old adult speakers. Behavior Research Methods, 52, 990–1007 (2020). https://doi.org/10.3758/s13428-019-01296-0

ID: 118
Abstract
Keywords: AI, Intellectual Property, copyright

Intellectual property matters relevant to AI-based voice identification

Christin Kirchhübel[1], **Georgina Brown**[1,2]
[1]Soundscape Voice Evidence; [2]Lancaster University; ck@soundscapevoice.com

As Artificial Intelligence (AI) continues to evolve, the law must progress as well. There are questions around who, or what, is responsible for AI decision-making. Similarly, there are questions around who, or what, owns the outputs of generative AI. Legal issues also attach to the training of AI. AI, of course, is behind voice identification technology that features in applications across financial, security and commercial sectors, among others. Focusing on Intellectual Property (IP), this paper looks to some of the current legal cases and discussions around AI before extending these to AI-based voice identification. Specifically, this paper engages with copyright law and what this means for the data used to train voice identification technology powered by data-hungry neural networks. We raise these issues at a time where Getty Images has recently issued copyright infringement proceedings against Stability AI for scraping millions of Getty's images and then using those images to train and develop AI (for example, Getty Images v Stability AI [2023] EWHC 3090). There is also a suggestion that it is relatively commonplace to train AI with copyrighted materials. In a recent House of Lords Inquiry into Large Language Models and Generative AI, Open AI conceded that it was "impossible to train today's leading AI models without using copyrighted materials". As the law has yet to clarify on these matters, this work aims to keep one eye on the relevant IP issues that could make their way to AI-based voice identification.

ID: 106

Abstract

Keywords: heritage speech, ethnolinguistic variation, Vietnamese German, perception

Investigating speaker voice and ethnolinguistic identification of Vietnamese-German heritage speakers

Thanh Lan Truong, Andrea Weber

University of Tübingen, Germany; thanh-lan.truong@uni-tuebingen.de

Every day listeners perceive diverse acoustic properties from various speakers and have learned to distinguish voices by age, gender, and nativeness. But what about the heritage of speakers? Studies found that European Americans can be reliably distinguished from African Americans based solely on their voices, and this was also found for Asian American voices. Since research on Asian heritage speech has only been conducted in a US context, we conducted two experiments using a forced-choice identification task in German. The German sentence *Flöhe können um das Hundertfache ihrer eigenen Körperlänge in die Höhe springen* ('flea can jump a hundredfold time their body length') was recorded by eight non-Vietnamese Germans, who grew up monolingually and monoculturally with German, and eight Vietnamese Germans (2nd generation), who varied in their linguistic competence of the heritage language but were all German native speakers. In Experiment 1, 25 Asian-German listeners and 25 non-Asian German listeners indicated for each unmodified recording if they thought the speaker was Vietnamese German or non-Vietnamese German. The results showed that Vietnamese German speakers can be identified reliably. Furthermore, participants were better at identifying speakers of their own ethnic background, substantiating an advantage of a heritage match. In Experiment 2, the sentences were time-reversed, making them unintelligible but still comparable to forward speech in terms of auditory characteristics. Thus far, we can report data from 16 non-Asian Germans and found no ability to identify the heritage background, suggesting that understanding speech is imperative for the identification of the speaker's heritage.

ID: 126

Abstract

Keywords: Voice Discrimination, Clear Speech, Speech Intelligibility, Voice Recognition

Sacrificing identity for intelligibility: Clear speech renders speakers less recognisable

Leah Bradshaw, Volker Dellwo

University of Zurich, Switzerland; leah.bradshaw@uzh.ch

Speakers are shown to adapt their speech in response to communicative contexts or listener needs, i.e., when speaking to hearing-impaired individuals (Picheny, Durlach & Braida, 1986), children (Bradlow, Kraus & Hayes, 2003), non-native speakers (Uther, Knoll & Burnham, 2007) or in the presence of background noise (Payton, Uchanski & Braida, 1994). In these contexts, and in most circumstances where we observe speech adaptations, the goal is speech intelligibility. Speakers produce what can be broadly defined as "clear speech" in interactions where they aim to be easier understood. Albiet this speaking style is frequently shown to be beneficial for speech intelligibility, little is known about what is also sacrified in this style and specifically, what speaking intelligibly means for voice identity.

We conducted a voice discrimination task, to assess how a clear speech style interferes with recognisability. We used stimuli from the IDEAR corpus (Perepelytsia, Bradshaw & Dellwo, 2023), which contains speech from three different tasks: an interaction with a *speech recognition* system (Clear Speech), an interaction with a *speaker recognition* system (ID-marked speech) and a read speech task (Read Speech). It is assumed that if individuality is sacrificed for intelligibility, listeners will be less accurately discriminated when producing clear speech. Findings showed that participants performed significantly worse than average in the *clear speech* style and responded significantly slower in this style compared to the other speaking styles. Overall, our findings show that speakers are less discriminable when producing *clear speech*, suggesting some individuality is sacrificed in this speaking style.

References

- Bradlow Ann R., Kraus Nina, & Hayes Erin. (2003). Speaking Clearly for Children With Learning Disabilities. Journal of Speech, Language, and Hearing Research, 46(1), 80–97.
- McDougall, K., Nolan, F., & Hudson, T. (2016). Telephone transmission and earwitnesses: Performance on voice parades controlled for voice similarity. Phonetica, 72(4), 257-272.
- Perepelytsia, V., Bradshaw, L., & Dellwo, V. (2023) IDEAR: A speech database of identity-marked, clear, and read speech. In Proceedings of ICPhS (pp. 3217-3220).
- Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. Journal of Speech, Language, and Hearing Research, 29(4), 434-446.
- Uther, M., Knoll, M. A., & Burnham, D. (2007). Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. Speech communication, 49(1), 2-7.

ID: 127
Abstract
Keywords: Blindness, Speaker Identity, Event-related Potentials (ERPs), Voice perception

The effects of long-term visual deprivation on speaker identity perception

Eduardo Xavier, João Sarzedas, Tatiana Conde, Ana P. Pinheiro
University of Lisbon, Portugal; egpxavier@gmail.com

Blind individuals exhibit remarkable sensory processing adaptations, offering valuable insights into experience-driven neuroplasticity. Despite previous research showing superior auditory abilities in the blind, little is known regarding how they decode nonverbal voice cues such as speaker identity. To address this gap, we combined behavioral and event-related potentials (ERPs) to examine how early blind discriminate self-generated, familiar, and unfamiliar voices compared to their controls.

Fourteen early blind individuals and twelve sighted controls participated in this study. Participants completed a three-alternative forced-choice (3AFC) voice recognition task with words spoken in their own voice (self-voice [SV]), a familiar voice (FV), and an unfamiliar one (UV). ERP analysis revealed differences between blind and sighted groups in both early (N1) and late stages of voice processing (N400). Blind individuals exhibited a reduced N1 amplitude in response to the SV compared to sighted controls. For blind individuals only, the SV also elicited reduced N400 amplitudes compared to FV and UV.

These findings suggest that early blindness is associated with differences in speaker identity perception. They also contribute to a broader understanding of brain plasticity and person perception through voices.

ID: 114
Abstract
Keywords: singing phonetics, overtone listening, resonance control

Singing phonetics vs. speech phonetics - A new approach to high-precision resonance control

Wolfgang Saus
Freelancer, Germany; voiceid2024@oberton.org

Singers, unlike speakers, have to adapt their vocal resonances to partials of the singing voice. Vowels can therefore deviate greatly from speaking vowels, especially in high singing ranges. It is therefore sensible for singers to understand how their harmonics interact with the vowel resonances.

This talk introduces a sophisticated and yet simple to use singing phonetics chart that illustrates how vowel resonances interact with vocal harmonics. It displays the distribution of harmonics in any acoustic-phonetic vowel triangle for any pitch and pitch change.

Singing vowels are pitch sensitive. The chart assists in distinguishing psychoacoustic and subjective sensations, such as vowel perception and vocal feel, from measurable acoustic parameters, such as frequencies of harmonics and resonances. And it provides vocal teachers with objective criteria for handling resonances.

A new didactic approach for controlling vocal tract resonances to within a semitone, based on trained overtone listining, will be introduced. Originally developed for opera singers and other professional singers, this tool could also support speech phonetics in redefining vocal timbres and reproducing them with high precision and objectivity. The lecture includes a practical demonstration of the effect of an auditory transformation to overtone listening, so that the audience experiences a widely unknown perception of vowel timbres.

Poster Session 2

Time: Thursday, 29/Aug/2024: 11:00am - 12:30pm · *Location:* Poster Room (101/102)

ID: 131

Abstract

Keywords: voice discrimination, prosody, voice conversion

Prosody plays a role in speaker discrimination: insights from voice conversion

Elisa Pellegrino[1], Debora Beuret[1], Thayabaran Kathiresan[2], Claudia Roswandowitz[1], Sascha Fruholz[3], Volker Dellwo[1]

[1]University of Zurich, Switzerland; [2]University of Melbourne; [3]University of Oslo; elisa.pellegrino@uzh.ch

Decades of research have emphasized the importance of spectral segmental information in voice recognition (Schweinberger et al. 2014). In most common automatic voice recognition systems, prosodic information plays, at best, a subordinate role (Adami et al. 2003). The purpose of this study is thus to examine the role of prosody as opposed to timbre in human voice discrimination. Stimuli for the discrimination task comprised 5-word and 2-word German utterances. These were created using a voice conversion technique wherein the timbre (short-term feature representation, such as instantaneous F0 and spectrum, Wu and Lee, 2014) of one source speaker was converted into that of four different target speakers (target). The newly converted utterances had the timbre of the target speakers and the prosodic characteristics (rate, rhythm, intonation) of the source speaker (Kobayashi and Toda, 2018). In the experiment, the voices in a trial were combined under three conditions:

1. Maintaining consistent timbre across speakers while varying prosody.

2. Varying prosody while keeping timbre consistent.

3. Variations in both timbre and prosody between speakers.

Forty Swiss German listeners, aged between 18 and 35, participated in the experiment. The overall results underscored the significance of both prosody and timbre for voice discrimination and confirmed the relevance of longer utterances on voice recognition (cf. Cook and Wilding, 1997). The results will be further discussed regarding listeners' preferences regarding prosody versus timbre for voice discrimination and the impact of different voice combinations within a trial.

References

- Adami, G., Mihaescu, R. Reynolds, D. A. and Godfrey, J. J. (2003). Modeling prosodic dynamics for speaker recognition, Proceedings ICASSP '03, IV-788.
- Schweinberger S R, Kawahara H, Simpson AP, Skuk V G, Zäske R. (2014). Speaker perception. Wiley Interdiscip Rev Cogn Sci. Jan;5(1):15-25.
- Cook, S., and Wilding, J. (1997). Earwitness testimony: Never mind the variety, hear the length," Appl. Cogn. Psychol. 11(2), 95–111.
- Kobayashi, K. and T. Toda (2018). Sprocket: Open-Source Voice Conversion Software. Proc. Odyssey, 203-210, June 2018. Wu and Lee, 2014, Voice conversion versus speaker verification: an overview, APSIPA Transactions on Signal and Information Processing, Volume 3.

ID: 152

Abstract

Keywords: cochlear implants, emotion perception, vocal caricatures, online training, quality of life

Improving vocal emotion perception for cochlear implant users: An online training approach using vocal caricatures

Celina I. von Eiff[1,2,3], Lukas Erchinger[1], Jenny M. Ruttloff[1], Stefan R. Schweinberger[1,2,3,4]

[1]Department for General Psychology and Cognitive Neuroscience, Friedrich Schiller University Jena, Germany; [2]Voice Research Unit, Friedrich Schiller University Jena, Germany; [3]DFG SPP 2392 Visual Communication (ViCom), Frankfurt am Main, Germany; [4]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany; celina.isabelle.von.eiff@uni-jena.de

Speech comprehension counts as a benchmark outcome of cochlear implants (CIs) – often with little attention to the communicative importance of non-verbal socio-emotional signals. However, successful social interaction depends strongly on the quality of communication. While the ability to recognize vocal emotions is impaired in CI users, vocal emotion perception skills are linked to higher quality of life ratings. We developed a novel training program using caricatures of vocal emotions as training stimuli. Conceptually, our program builds on findings that caricatures of vocal emotions are perceived as more distinctive and intense, and that visual caricatures can enhance face perception in individuals with sensory loss. Fifteen CI users participated in a 30-day online training program, with each session lasting approximately 7 minutes, and 15 normal-hearing individuals acted as controls. Before training, immediately after training, and at a follow-up ~8 weeks after training completion, participants took part in an emotion recognition task in the laboratory. In CI users, consistent training benefits appeared for trained (disgust, fear, happiness, and sadness) but not untrained (anger, surprise) emotions, and these benefits were at least partially maintained at follow-up. Furthermore, training benefits generalized, but in attenuated magnitude, across the same emotions when tested with untrained speakers. These findings suggest that CI users can enhance their vocal emotion recognition abilities through online training utilizing vocal caricatures, which offers distinct advantages over in-person training methods in terms of flexibility, feasibility, and cost-effectiveness.

ID: 134

Abstract

Keywords: voice perception, face perception, personality traits, big five, voice averages

Personality perceptions from vocal and facial trait averages

Verena G. Skuk[1,2], Isabel Jacob[1,2], Rebecca Wientzek[1,2], Robert Ward[3], Stefan R. Schweinberger[1,2,4,5]

[1]Department for General Psychology and Cognitive Neuroscience, Institute of Psychology, Friedrich Schiller University of Jena, Am Steiger 3, 07743 Jena, Germany; [2]Voice Research Unit, Friedrich Schiller University of Jena, Am Steiger 3, Haus 1, 07743 Jena, Germany; [3]Deptartment of Psychology; Bangor University; Bangor LL57 2DG; United Kingdom; [4]Swiss Center for Affective Sciences, University of Geneva, Switzerland; [5]German Center for Mental Health (DZPG), Site Jena-Magdeburg-Halle, Germany; verena.skuk@uni-jena.de

We investigated the perception of Big Five personality traits from trait-average voices when traits were based either on speakers´ self-ratings (Exp. 1, E1) or on other perceivers’ ratings of perceived personality of the original voice samples (E2). Trait-average voices were created from a voice database of 93 speakers (40 male, 53 female) using TANDEM-STRAIGHT n-way morphing. For speaker sex, trait and for two sentences, we created five-voice averages from speakers scoring either high or low on the target trait. We then measured perceivers´ ability to discriminate high and low trait-averages per trait. We also assessed facial trait perception (E3) using the paradigm and the full facial composite images by Kramer and Ward (2010). In trait-average voices based on self-ratings (E1), extraversion (for female speakers) and neuroticism (for male speakers) were the only traits that could be discriminated above chance levels. For trait-average voices which were based on other perceivers´ personality ratings of individual voices (E2), all Big Five traits were discriminated with high accuracy, demonstrating stereotyping in the sense of consistent (though not necessarily valid) personality impressions from voices. By comparison with E1, we found substantially better perception of self-rated traits from faces (E3), for all traits except for openness, replicating Kramer and Ward (2010). Individual differences in trait perception were substantial, and there were small but significant correlations between facial and vocal trait perception skills in both E1 and E2. Overall, the present methodological approach offers a promising window into personality perception from voices.

ID: 132

Abstract

Keywords: deceptive speech, voice recognition, vocal identity, gaussian mixture models

Speakers are less discriminable when producing deceptive speech, and deceptive speech deteriorates speaker models’ performance

Alessandro De Luca, Volker Dellwo

Universtity of Zurich, Switzerland; alessandro.deluca@uzh.ch

Deception, defined as “a deliberate attempt to mislead others” [1], is prevalent in both human and non-human communication [2]. Under the reciprocal altruism theory [3], deception may yield individual gains but can undermine future interactions if the deceiver is recognized. It is plausible that humans seek to be less identifiable when lying. We thus predict that a loss of acoustic congruency between utterances within the speaker has a negative impact on voice recognition performance [4]. In a same-different speaker discrimination task using the CSC Deceptive Speech corpus [5], listeners exhibit significantly lower performance (d-prime) in deceptive comparisons compared to truthful ones [6]. Gaussian Mixture Models trained using exclusively deceptive speech samples show a significant decrease in performance in speaker recognition tasks, compared to exclusively truthful and a mix of both speech registers. In addition, we find a significant decrease in bias for same speaker responses in deceptive speech tests, regardless of the training condition. Based on this result and previous studies [7, 8], we hypothesize that speakers may be “masking” their identity when producing deceptive utterances by sounding more similar to the “average voice”.

References

- [1] B. M. DePaulo, B. E. Malone, J. J. Lindsay, L. Muhlenbruck, K. Charlton, and H. Cooper, “Cues to deception,” Psychological Bulletin, vol. 129, pp. 74–118, 2003.
- [2] Griffin, D. R. (Donald R., & Ristau, C. A. (2013). Truth and Deception In Animal Communication. 147–172. https://doi.org/10.4324/9780203761700-12
- [3] Trivers, R. L. (1971). The evolution of reciprocal altruism. The Quarterly review of biology, 46(1), 35-57.
  [4] Campeanu, S., Craik, F. I., & Alain, C. (2013). Voice congruency facilitates word recognition. PLoS One, 8(3), e58778. https://doi.org/10.1371/JOURNAL.PONE.0058778
- [5] Columbia University, SRI International, & University of Colorado Boulder. (2013). CSC Deceptive Speech. https://doi.org/10.35111/q500-9a28
- [6] Lim, S. (2019). The effects of deceptive speech on speaker discrimination ability. Master of Arts Thesis, University of Zurich, Faulty of Arts and Social Sciences.
- [7] 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
- [8] 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

ID: 143

Abstract

Keywords: speaker discrimination, singing, speaking mode, listener skills, phonation mismatch

A forensic study on the identification of singers

Sophie Möller, Gea de Jong-Lendle

Philipps University Marburg, Germany; Moeller9@students.uni-marburg.de

Introduction

In 2022 a pilot study was conducted to compare the discrimination ability of listeners judging singing-speaking with singing-singing and speaking-speaking speaker pairs. This study was motivated by a forensic case concerning the disputed identity of a singer. In the follow-up study, the effect of speaker/listener characteristics was investigated.

Methodology

10 female and 10 male persons (18-35y) were recorded producing a song and spontaneous speech. 40 native-German listeners (26 female, 14 male; 17-54y.) were subsequently tested in an online discrimination experiment involving 3 conditions (speaking-speaking, speaking-singing, singing-singing) using 10sec. stimuli. Each listener was assigned to either male or female voices. Listener groups matched in terms of mean age, sex and expertise.

Results

Consistent with the pilot study, the correct discrimination rate of 74% for the speaking-singing stimuli was significantly lower ($p<0.001$) compared to singing-singing (91%) and speaking-speaking stimuli (91%), but well above chance level.

Sing-experts performed best (89%) compared to speech-experts and non-experts (85% vs. 84%). Statistical tests, however, were non-significant. Sing- and speech-experts performed equally well for sing- and speech stimuli (76%).

No significant effect was found for listener-sex and accuracy. Discrimination performance, however, was significantly better for male stimuli compared to female stimuli ($p<0.001$). This is presumably due to the closer matched F0-values of the female speakers and the male speakers' singing skills varying more strongly.

Finally, women were significantly better at judging male voices compared to female voices ($p<0.001$). Their performance for male voices was also significantly better compared to male listeners judging female voices ($p<0.001$).

References

- Clermont, F. (2002). Systemic comparison of spoken and sung vowels in formant-frequency space. In C. Bow (Ed.), Proceedings of 9th Australian International Conference on Speech Science & Technology (p.124-129). Australian Speech Science and Technology Association.
- Kreiman, J., & Sidtis, D. (2011). Foundations of voice studies: an interdisciplinary approach to voice production and perception. Wiley-Blackwell.
- Moeller, S. and De Jong-Lendle, G. (2022). ‘When singing becomes illegal’, Poster presented at the International Association for Forensic Phonetics and Acoustics Annual Conference, Prague, 10-13 July 2022.
- Peynircioğlu, Zehra F., Brian E. Rabinovitz & Juliana Repice. (2017). Matching Speaking to Singing Voices and the Influence of Content. Journal of Voice 31(2), https://doi.org/10.1016/j.jvoice.2016.06.004.

ID: 128

Abstract

Keywords: adversarial attack, voice conversion, d-vector cosine similarity, SoVits, VoiceBlock

Exploring adversarial perturbations to defend against deepfake voice cloning

Elaine M Liu, Jih-Wei Yeh, Jen-Hao Lu, Yi-Wen Liu

National Tsing Hua University, Taiwan; eliu@gapp.nthu.edu.tw

Recent advancements in voice conversion systems raise concerns about user privacy and deepfakes, where a target voice is counterfeited. Adversarial attacks on voice conversion aim to defend target voice audio by introducing perturbations inconspicuous to humans but disruptive to machines. We explore the effectiveness of two kinds of adversarial attacks (VoiceBlock [1] and VoiceGuard [2]) on SoVits [3], a recent state-of-the art voice conversion system. VoiceBlock introduces perturbations directly in audio by applying a time-varying finite impulse response filter, while VoiceGuard adds perturbation in the frequency domain, which causes spectrogram-inversion artifacts. Objective evaluations on 20 singing clips from the M4Singer database [4] are performed via d-vector speaker-embedding cosine similarities [5]. We find that the presence of perturbations in the adversarial input significantly drops cosine similarity with the clean target voice by an average of 0.206 for VoiceBlock and 0.158 for VoiceGuard, and thus has potential to de-identify the target from a speaker recognition system. However, adversarial attacks are less effective at disrupting voice conversion: the cosine similarity between the target voice we are defending and the voice conversion output using perturbed target audio, is decreased from that of the voice conversion output using clean target audio by 0.017 for VoiceBlock and merely 0.001 for VoiceGuard. Our results indicate that a discriminative speaker recognition classifier could be more vulnerable to adversarial attacks than a generative voice conversion model. Moreover, VoiceBlock with FIR in the time domain shows more promise than the frequency-domain-based VoiceGuard method for adversarial attacks on voice conversion systems.

References

- [1] Patrick O'Reilly, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo, "VoiceBlock: Privacy through Real-Time Adversarial Attacks with Audio-to-Audio Models," in Neural Information Processing Systems, 2022.
- [2] Chien-yu Huang, Yist Y. Lin, Hung-yi Lee, Lin-shan Lee, "Defending Your Voice: Adversarial Attack on Voice Conversion', IEEE Spoken Language Technology Workshop, 2021.
- [3] https://github.com/svc-develop-team/so-vits-svc
- [4] https://m4singer.github.io/
- [5] https://github.com/resemble-ai/Resemblyzer/tree/master
- [6] Li Wan, Quan Wang, Alan Papir, Ignacio Lopez Moreno, "Generalized end-to-end loss for speaker verification." 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.

ID: 136

Abstract

Keywords: neural speech tracking, voice familiarity, voice identity

Neural correlates of voice familiarity and identity: an EEG study

Valeriia Perepelytsia[1], Nathalie Giroud[1], Martin Meyer[2], Volker Dellwo[1]

[1]Department of Computational Linguistics, University of Zurich, Switzerland; [2]Department of Comparative Language Science, University of Zurich, Switzerland; valeriia.perepelytsia@uzh.ch

The brain tracks incoming speech by aligning the phase of its endogenous oscillatory activity to quasi-rhythmic features of speech such as, for example, slow amplitude modulations. This process is called *neural speech tracking* and can be influenced by various bottom-up and top-down factors such as speech acoustics, speech intelligibility, selective attention, and others. Speech from familiar speakers is more intelligible but it is currently unclear whether voice familiarity can also affect speech tracking. Our study explores the association between voice familiarity and neural speech tracking. Given a higher intelligibility with increasing familiarity, we expected to find enhanced speech tracking for familiar compared to unfamiliar voices. We familiarized 39 young adult listeners with four female voices during high-variability familiarization containing sentence stimuli from different speaking styles and audio qualities. We then recorded EEG while participants listened to sentences from familiar and unfamiliar speakers in quiet and in multitalker babble noise. Speech tracking was quantified with cross-correlation between the neural response and amplitude envelopes of the sentences. Results revealed that familiarity with voices does not affect neural speech tracking. However, we found a significant effect of speaker on speech tracking suggesting that neural speech tracking is modulated by voice identity. We also found that voice recognition ability was related to speech tracking: the better the speakers were recognized, the higher the speech tracking response was to the respective speaker. Our results further support the notion that voice identity is reflected in early acoustic and later linguistic processing stages in the brain.

ID: 145

Abstract

Keywords: age estimation, disguises, falsetto

Age estimation: Falsetto versus modal phonation

Vanessa Hübert, Gea de Jong-Lendle

Institut für Germanistische Sprachwissenschaft, Philipps-Universität Marburg; vanessa.huebert1@outlook.de

Introduction

This research was motivated by a case involving a speaker most probably imitating an elderly lady by disguising his voice using falsetto. The question arose, whether age estimation generally is more difficult from falsetto voice compared to modal voice and if there is a significant difference for older or younger persons rating the age.

Methodology

A total of 15 German male speakers (age range 25-75 years) were recorded in a sound-proof booth reading a short text based on the content of the disputed voicemails in modal and falsetto phonation thereby loosely imitating a natural sounding voicemail.

In an online SoSci-study 52 German listeners (age: 18-57, 39 female, 10 male, 3 divers) were asked to listen to 15 male speakers and estimate their age in two cohorts (first falsetto, second modal to avoid recognition). All listeners were German and below 60 years to rule out hearing loss. Subsequently, absolute differences between the speaker's real age (CA) and his perceived age (PA) were calculated for both phonation types for each listener.

Then the ratings of the 15 youngest and oldest listeners were grouped together and compared for each phonation type. Furthermore, the results for the 6 youngest and 4 oldest speakers for the 15 youngest and oldest listeners were compared.

Results

First, a T-Test showed that the summated age estimation deviations for falsetto and modal voice differed significantly ($p < .001$). Second, no difference in accuracy was found between young and old listeners. Third, older speakers were generally rated less accurately.

References

- Künzel, H. J. (2000). Effects of voice disguise on speaking fundamental frequency. International Journal of Speech, Language and the law, 7(2), 150-179.
- Masthoff, H. (1996). A report on a voice disguise experiment. International Journal of Speech, Language and the Law, 3(1), 160-167.
- Moyse, E. (2014). Age estimation from faces and voices: A review. Psychologica Belgica, 54(3), 255–265. https://doi.org/10.5334/pb.aq

ID: 135

Abstract

Keywords: self-voice, ambiguity, auditory verbal hallucinations, vocal congruence, behavioral

Altered self-other voice discrimination as a marker of hallucination proneness

Margarida Marques[1], Maria Amorim[1], Sonja A. Kotz[2], Ana P. Pinheiro[1]

[1]Faculty of Psychology, University of Lisbon, Lisboa, Portugal; [2]Faculty of Psychology and Neuroscience, Maastricht University, Maastricht, The Netherlands; mmargarida@edu.ulisboa.pt

Compared to familiar and unknown voices, the self-voice holds a special status in the auditory system. Understanding how one's voice is distinguished from others is essential as altered self-other distinction is at the core of phenomena such as auditory verbal hallucinations (AVH). However, the existing studies have not accounted for how self-other voice discrimination is modulated by perceptual ambiguity and self-ratings of voice quality, including vocal congruence (perceived similarity between air-conducted and bone-conducted self-voice).

In this study, we used prerecorded spoken words to develop morphing continua from self-voice (SV) to other-voice (OV). After hearing each word, 45 young adult participants answered if the voice they heard was "more mine" or "more other". Participants also completed the Launay-Slade Hallucination Scale probing hallucination proneness.

Mixed-effects regressions showed that, overall, vocal congruence promoted "more mine" responses as the % of SV increased ($p = .008$). Moreover, participants with higher AVH scores identified less frequently a voice as "more mine" as the percentage of SV increased ($p = .005$). They also showed a lower threshold for experiencing uncertainty in voice-identity attribution ($p < .001$).

The findings reveal that individuals with higher AVH proneness exhibit both heightened sensitivity to perceptual ambiguity and a distorted perception of the SV. They confirm that alterations in the capacity to discriminate self and others through their voices can enhance one's vulnerability to perceive voices when there are none.

ID: 133

Abstract

Keywords: Blindness, Voice, Authenticity, Event-related potentials

Blindness affects vocal emotional perception of spontaneous and volitional laughs and cries: Behavioral and ERP insights.

João Sarzedas[1], Magda S. Roberto[1], César F. Lima[2,3], Sophie K. Scott[3], Ana P. Pinheiro[1], Tatiana Conde[1]

[1]CICPSI, Faculty of Psychology, University of Lisbon, Lisboa, Portugal; [2]CIS-IUL, ISCTE - University Institute of Lisbon, Lisboa, Portugal; [3]Institute of Cognitive Neuroscience, University College London, London, UK; jsarzedas@edu.ulisboa.pt

The ability to distinguish spontaneous from volitional emotional expressions is an important social skill. Unlike sighted individuals, to detect the authenticity of an emotional expression, blind listeners cannot rely on facial and body language signs, relying instead on vocal cues alone. In this study, we combined behavioural and ERP measures to investigate authenticity perception in laughter and crying in individuals with early- or late-blindness onset. Early-blind, late-blind, and sighted control participants (n = 17 per group, N = 51) completed authenticity and emotion discrimination tasks while EEG data were recorded. The stimuli consisted of laughs and cries that were either spontaneous or volitional. The ERP analysis focused on the N1, P2, and late positive potential (LPP). Behaviourally, early-blind participants showed intact authenticity perception, but late-blind participants performed worse than controls. There were no group differences in the emotion discrimination task. In brain responses, all groups were sensitive to laughter authenticity at the P2 stage, and to crying authenticity at the early LPP stage. Nevertheless, only early-blind participants were sensitive to crying authenticity at the N1 and middle LPP stages, and to laughter authenticity at the early LPP stage. Furthermore, early-blind and sighted participants were more sensitive than late-blind ones to crying authenticity at the P2 and late LPP stages. Overall, these findings suggest that early blindness relates to facilitated brain processing of authenticity in voices, both at early sensory and late cognitive-evaluative stages. Late-onset blindness, in contrast, relates to decreased sensitivity to authenticity at behavioural and brain levels.

ID: 140

Abstract

Keywords: Fundamental frequency, Spanish, power, distance, speakers' sex

Do Spanish speakers change their fundamental frequency according to the gender and status of their interlocutors?

Paula Barriendo-Cebrián[1], Nuria Polo[2], Filipa M.B. Lã[3]

[1]Department of Spanish Language and General Linguistics, UNED, Spain; [2]Department of Spanish Language and General Linguistics, UNED, Spain; [3]Department of Didactics, School Organization and Special Didactics, UNED, Spain; pbarriendo@flog.uned.es

Previous investigations have looked at conversational analysis in terms of relative talkativeness, relative production of intrusions into speech, interruptions and overlaps, and turn-talking. However, vocal features in conversation, such as, variations in mean fundamental frequency (fo) and related parameters, are still to be explored, especially when regarding Spanish speakers. This study aims at investigating whether the fo of Spanish native speakers varies according to sex, power, and familiarity of their conversation partner.

Two professional actor/actress, in their twenties, were instructed to role-play six daily-life conversation situations presented as the interlocutor stimuli in the subsequent listening experiment: (1) "unfamiliar-powerful"; (2) "unfamiliar-powerless"; (3) "unfamiliar-equal power" person; (4) "familiar-powerful"; (5) "familiar-powerless"; (6) "familiar-equal power". By adopting these roles, the influence of power and distance motivates specific linguistic choices in the answers of the listeners, randomly recruited in Madrid. They were recorded in a sound treated booth using a headset omnidirectional microphone. Participants were led to use the words "lavandería" (laundry) and "domingo" (Sunday) in their responses to allow comparisons between speakers. A script in Praat was used to extract mean fo and related parameters.

The preliminary results suggest that listeners fo and related measures change more according to the conversation situation than to the sex of the interlocutor. The "unfamiliar-powerful" situation presented the biggest drop of fo regardless the sex of the interlocutor. This study contributes to realizing the importance of incorporating situational elements when understanding the relationships between voice, language, and social function.

ID: 137

Abstract

Keywords: Voice quality, Bilingual, Polish, English

One speaker, two languages, two distinct voices?!

Maral Asiaee, Kamil Kaźmierski, Ewelina Wojtkowiak, Geoffrey Schwartz

Adam Mickiewicz University, Poznań, Poland; marasi@amu.edu.pl

The Effect of language on voice quality has been investigated in a number of studies. While some studies suggest that bilingual speakers exhibit distinct voices in the languages they speak (Lee & Sidtis, 2017; Zhu et al., 2022), others have failed to identify significant variations in voice characteristics across languages (Altenberg & Ferrand, 2006; Johnson & Babel, 2023). The inconclusiveness in the results prompted us to further research the relationship between language and voice quality in Polish-English bilinguals.

Speech samples from 10 Polish-English bilinguals were recorded. Acoustic parameters were extracted from the voiced portions of the signal using VoiceSauce (Shue et al., 2009) and an MFCC extraction script. PCA was conducted to capture the similarities and differences across the two languages of each individual.

Results revealed that the majority of bilinguals examined in this study demonstrate comparable spectral tilt measures and lower-dimensional patterns in voice variation, despite differences in segmental and suprasegmental features between Polish and English. However, some speakers (4 out of 10) appeared to have different spectral noise measures across the languages; three speakers exhibited higher values of spectral noise measures in English than in Polish, suggesting a “brighter” voice in English than in Polish, while one speaker showed the opposite pattern. Variation in spectral tilt and formant measures was only observed in one speaker, which can be attributed to the possible effect of language on these measures. The overall results suggest that many individuals possess a consistent "voice" across languages.

ID: 154

Abstract

Keywords: Voice Disorders; Prevention; Stress; Voice Health; App Usage

Regensburger Stimmtraining: A comprehensive approach to prevent voice disorders in teachers

Jonas Hauck, Christian Gegner, Marina Giglberger, Sven Hilbert

University Regensburg, Germany; jonas.hauck@ur.de

The prevalence of developing dysphonia is significantly higher in teachers compared to non-teachers (e.g., Roy et al. 2004). Studies indicate that dysphonia in the teaching profession is associated with missed classes, high financial losses (Richter & Echternach 2010), and high levels of stress experienced by teaching staff (Gassull et al. 2010). Teachers in general are exposed to considerable job-related stress, accompanied by a correspondingly increased risk of stress-related health disorders (Bauer et al. 2007; Zimmermann et al. 2012). Consequently, particular importance should be given to the prevention of voice disorders in the teaching profession.

At the university of Regensburg, a training app for voice hygiene and prevention of voice disorders for public school staff in Bavaria was developed and evaluated in in an experimental and control group design with comprehensive diagnostic options (voice and stress related questionnaires, voice diagnostics). Our evaluation is based on a comprehensive diagnostic approach, incorporating both voice and stress-related questionnaires ($N = 397$) and voice diagnostics ($n = 197$). We aim to present our model that integrates questionnaire data on voice health and stress, as well as usage data from the app. Our aim is to shed light on the complex interplay between voice, stress, and voice training, and their collective impact on voice disorders.

ID: 144

Abstract

Keywords: Bilingual speaker recognition, Machine learning algorithms, Persian-English, MFCC, Voice quality

Speaker recognition in bilinguals: A comparison of source and filter features using machine learning algorithms

Homa Asadi[1], Maral Asiaee[2], Arefeh Kazemi[1]

[1]University of Isfahan, Isfahan, Iran; [2]Adam Mickiewicz University, Poznań, Poland; h.asadi@fgn.ui.ac.ir, marasi@amu.edu.pl

Despite the extensive body of literature on the usefulness of filter parameters in speaker recognition, little attention has been paid to investigating the discriminatory capabilities of voice quality features in forensic phonetics (Hughes et al., 2023). This stands in stark contrast to the widespread use of auditory-perceptual judgments of voice quality by experts in forensic casework (Gold & French, 2011, 2019).

This study investigates the speaker-specific information encoded in source and filter features within read and spontaneous speech samples from 40 Persian-English bilinguals (balanced for gender). Voice quality features were extracted using VoiceSauce (Shue et al., 2009), while Mel-frequency cepstral coefficients (MFCCs) were obtained via a Praat script. Three machine learning algorithms were employed for speaker recognition: K-nearest neighbours (KNN), random forest (RF), and support vector machine (SVM).

Results revealed that, for both genders, MFCCs alone yielded superior performance compared to voice quality features. However, for male speakers, fusing both feature sets significantly enhanced recognition accuracy, suggesting the potential of voice quality features as a complementary tool. Among the algorithms, KNN achieved the highest performance, while RF and SVM performed similarly, with their results closely tracking each other. Across all models, the "read" speech generally exhibited slightly better performance than the "spontaneous" speech for both genders. Additionally, for male speakers, performance across all models demonstrated comparable accuracy between English and Persian. For female speakers, Persian yielded slightly higher accuracy, precision, and recall compared to English.

ID: 146

Abstract

Keywords: vocal tract shape, individual variation, face recognition

Understanding vocal tract shape variation: Lessons from face recognition

Amelia Gully, Nick Pears

University of York, United Kingdom; amelia.gully@york.ac.uk

The field of human face analysis, for applications such as face recognition, is a mature field with a rich history of technological development. These range from early hand-crafted modelling approaches, to the current big data based deep learning approaches, and more recently, hybrid approaches that parameterise explicitly-defined, semantically meaningful models using deep networks. The study of individual vocal tract shape - a critical component in understanding the sources of individual variation in the speech signal - is a newer field, with detailed study of individual shape variation in 3D only recently becoming possible due to the increased availability of medical imaging data for a range of individuals. Although there are important differences between the two domains, particularly that vocal tract shape is usually not observed directly during speech, there are also sufficient similarities that the study of face analysis is instructive when developing models that describe vocal tract shape variation. In particular, concepts such as disentanglement (i.e. separating facial identity from facial expression), and the modelling of individual movement patterns, offer a framework for the development of a vocal tract model that describes both between- and within-subject shape variation. This study considers the lessons that can be learned from the development of face analysis technology for understanding vocal tract shape variation, adding a critical degree of explainability to vocal identification methods based on the acoustic signal alone.

ID: 138

Abstract

Keywords: fundamental frequency, formant frequencies, speech, voice context, context variability

Variability in voice context shapes spoken vowel perception

Carina Ufer, Helen Blank

University of Medical Center Hamburg-Eppendorf, Germany; c.ufer@uke.de

The perception of current sensory input does not only depend on its current features but is also shaped by its context. In our everyday life, we frequently encounter conversations with different speakers with different voice features. Dependent on these voice features the same intended vowel can acoustically vary substantially - a phenomenon described as the invariance problem. Yet, the same physical features explaining between-speaker voice features, i.e. fundamental frequency and formant frequencies, also influence vowel perception. Therefore, we investigated the impact of variability in voice context on the perception of vowels spoken by different speakers. Participants listened to sequences of ambiguous vowels in three experimental setups with varying amount of speaker variability and indicated their perceptual decision via a keypress. Voice features of the previous speaker had a contrastive effect on vowel perception. Moreover, the current decision about the perceived vowel was attracted towards the previous response. Importantly, the contrastive effect of the previous voice features could only be detected once the decisional effect was accounted for. Depending on the voice context, the contrastive effect of previous voice features monotonically increased with decreasing speaker variability, i.e. when the voice context was stable, acoustic differences were detected more accurately. The effect of the previous decision was strongest when speakers were presented in short sequences and was stronger within a sequence than between sequences of different speakers, possibly indicating a chunking of the voice context. Thus, the variability of the voice context differentially affects sensory and decisional processing during speech perception.