

Computational Analysis and Simulation of the Human Voice

Sten Ternström^{*1}, Nathalie Henrich Bernardoni^{*2}, Peter Birkholz^{*3},
Oriol Guasch^{*4}, and Amelia Gully^{†5}

- 1 KTH Royal Institute of Technology – Stockholm, SE. stern@kth.se
- 2 Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, FR. nathalie.henrich@gipsa-lab.fr
- 3 TU Dresden, DE. peter.birkholz@tu-dresden.de
- 4 Ramon Llull University – Barcelona, ES. oriol.guasch@salle.url.edu
- 5 University of York, GB. amelia.gully@york.ac.uk

Abstract

This report documents the program and the outcomes of Dagstuhl Seminar 24242 “Computational Analysis and Simulation of the Human Voice”, which was held from the 9th to the 14th of June, 2024. The seminar addressed key issues for a better understanding of the human voice by focusing on four main areas: voice analysis, visualisation techniques, simulation methods, and data analysis with machine learning. There has been enormous progress in recent years in all these fields. The seminar brought together a number of experts from fields as diverse as computer science, logopedics and phoniatrics, clinicians, acoustics and audio engineering, electronics, musicology, speech and hearing sciences, physics and mathematics. The schedule was quite flexible, including inspirational talks in the main areas, interactive working groups, sharing of conclusions and discussions, presentation of successes and failures to learn from, and a large number of free talks that emerged throughout the days. The variety of topics and participants created a highly enriching environment from which novel proposals for future research and collaboration emerged, as well as the collective writing of a paper on the state of the art and future perspectives in human voice research.

Seminar June 9–14, 2024 – <https://www.dagstuhl.de/24242>

2012 ACM Subject Classification Human-centered computing → Sound-based input / output; Human-centered computing → Auditory feedback; Human-centered computing → Visualization theory, concepts and paradigms; Computing methodologies → Speech recognition; Applied computing → Molecular structural biology; Applied computing → Health informatics; Applied computing → Performing arts; Applied computing → Sound and music computing; Applied computing → Physics

Keywords and phrases voice science, voice analysis, voice simulation, visualization, big data, machine learning, clinical voice treatment

Digital Object Identifier 10.4230/DagRep.14.6.84

* Editor / Organizer

† Editorial Assistant / Collector



Except where otherwise noted, content of this report is licensed under a Creative Commons BY 4.0 International license

Computational Analysis and Simulation of the Human Voice, *Dagstuhl Reports*, Vol. 14, Issue 6, pp. 84–107
Editors: Sten Ternström, Nathalie Henrich Bernardoni, Oriol Guasch, and Peter Birkholz



Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Sten Ternström

Nathalie Henrich Bernardoni

Oriol Guasch

Peter Birkholz

License © Creative Commons BY 4.0 International license
© Sten Ternström, Nathalie Henrich Bernardoni, Oriol Guasch, and Peter Birkholz

The human voice is able to produce a very rich set of different sounds, making it the single most important channel for communication human-to-human, and also potentially for human-computer interaction. Spoken communication can be thought of as a stack of layered transport protocols that includes language, speech, voice, and sound. This Dagstuhl Seminar was concerned with the voice and its function as a transducer from neurally encoded speech patterns to sound. This very complex mechanism remains insufficiently explained both in terms of analysing voice sounds, as for example in medical assessment of vocal function, and of simulating them from first principles, as in talking or singing machines. There were four main themes to the seminar:

Voice Analysis. Measures derived from voice recordings are clinically attractive, being non-invasive and relatively inexpensive. For clinical voice assessment, however, quantitative objective measures of vocal status have been researched for some seven decades, yet perceptual assessment by listening is still the dominating method. Isolating the properties of a voice (the machine) from those of its owner's speech or singing (the process) is far from trivial. Computational approaches are expected to facilitate a functional decomposition that can advance beyond conventional cut-off values of metrics and indices.

Voice Visualization. Trained listeners can deduce some of what is going on in the larynx and the vocal tract, but we cannot easily see it or document it. The multidimensionality of the voice poses interesting challenges to the making of effective visualizations. Most current visualizations are textbook transforms of the acoustic signal, but they are not as clinically or pedagogically relevant as they could be. Can functionally or perceptually informed visualizations improve on this situation?

Voice Simulation. Balancing low- and high-order models. A "complete" physics-based computational model of the voice organ would have to account for bidirectional energy exchange between fluids and moving structures at high temporal and spatial resolutions, in 3D. Computational brute force is still not an option to represent voice production in all its complexity, and a proper balance between high and low order approaches has to be found. We discussed strategies for choosing effective partitionings or hybrids of the simulation tasks that could be suitable for specific sub-problems.

Data science and voice research. With today's machine learning and deep neural network methods, end-to-end systems for both text-to-speech and speech recognition have become remarkably successful, but they remain quite ignorant of the basics of vocal function. Yet machine learning and big data science approaches should be very useful for helping us deal with and account for the variability in voices. Rather than seeking for automated discrimination between normal and pathological voice, clinicians wish for objective assessments of the progress of an intervention, while researchers wish for ways to distil succinct models of voice production from multi-modal big-data observations. We have explored how techniques such as domain-specific feature selection and auto-encoding can make progress toward these goals.

This seminar has resulted in (1) shared knowledge and data about the science of voice from the perspectives of scientists in fields as diverse as computer science, voice pathology and therapy, clinicians, acoustics and audio engineering, electronics, musicology, speech and hearing sciences, physics and mathematics, (2) identifying areas of common interest where significant progress is being made and needs to be made, such as individual voice variability, physical replicas for modelling and validation, synthesis and computational modelling, motor control, and availability of data and resources, (3) sharing and discussing failures to learn lessons and ideas for future developments, and (4) envisioning the future of progress in human voice analysis and simulation in the medium to long term: what is needed to make a big leap forward in this field? These ideas will be captured by the publication of a collaborative article in a leading voice journal.

2 Table of Contents

Executive Summary

Sten Ternström, Nathalie Henrich Bernardoni, Oriol Guasch, and Peter Birkholz . 85

Inspirational Talks

Inspirational Talk on Analysis: Voice analysis: looking at the subsystems of the vocal apparatus <i>Filipa M. B. Lã</i>	89
Inspirational Talk on Visualisation: Collaboration between Domain Scientists and Visualization Experts <i>Tino Weinkauff</i>	89
Inspirational Talk on Simulation: Articulatory Speech Synthesis: Modelling and Simulation – or – Can we make a voice instrument? <i>Sidney Fels</i>	90
Inspirational Talk on Data Science and ML: Data Science and Machine Learning on Clinical Applicability of Voice Studies <i>Pedro Gómez-Vilda</i>	91

Free papers

Revisiting Laver’s voice qualities <i>Philip Aichinger</i>	92
Voice conversion approaches to synthetic substitution voices <i>Philip Aichinger</i>	92
Sampling rate bias of vocal cycle perturbations <i>Jean Schoentgen</i>	93
Inverse filtering <i>Johan Sundberg</i>	93
Model based speech research <i>Brad Story</i>	94
The Laryngonaut <i>Scott Reid Moisiak</i>	94
Pressure in, SPL out (where is the flow?) <i>Peter Pabon</i>	95
First steps to a pacemaker for phonation? <i>Oriol Guasch</i>	95
Formant tuning of 3D vocal tracts for FEM vowel synthesis <i>Marc Arnella</i>	96
Development of an AI assisted data assimilation framework for voice research and clinics <i>Qian Xue</i>	96

Working Groups

Spontaneous workshops	97
---------------------------------	----

Workshops and talks on dreams and failures

What would you need to make a leap?	98
My favourite failure: Learning from failures and outliers: “Oh my... that was embarrassing...” <i>Eric J Hunter</i>	99
My favourite failure: can subglottal pressure be estimated from intra-oral pressure in speech and singing? <i>Nathalie Henrich Bernardoni</i>	99
My favourite failure: Spontaneous confessions <i>Sten Ternström</i>	100
My dream come true: the story of Pinocchio <i>Nathalie Henrich Bernardoni</i>	100
My dream come true: Design of self-oscillating biomimetic vocal folds <i>Lucie Bailly</i>	101
My dream come true: Discussion on particle-based simulations <i>Peter Birkholz</i>	102

Workshop on the elusive Inversion Problem

Acoustic-articulatory inversion with rtMRI <i>Yves Laprie</i>	102
Two examples using VocalTractLab and neural networks <i>Peter Birkholz</i>	103
Inversion in voice production <i>Zhaoyan Zhang</i>	103

Challenges in voice analysis and simulation

What do clinicians need? <i>Meike Brockmann-Bauser</i>	104
Numerical simulations <i>Michael Döllinger</i>	104

Concluding Plenary

Towards the future <i>Amelia Gully</i>	105
Final discussions	106

Participants	107
-------------------------------	-----

3 Inspirational Talks

For each of the four main topics of the seminar, a participant was invited to prepare an inspirational talk of about 30 minutes.

3.1 Inspirational Talk on Analysis: Voice analysis: looking at the subsystems of the vocal apparatus

Filipa M. B. Lã (UNED – Madrid, ES)

License © Creative Commons BY 4.0 International license
© Filipa M. B. Lã

This presentation shows examples of voice analysis that are currently available in the voice studio to assist the teaching of singers. It also sets the ground for discussing possibilities to simplify current recording setups and voice analysis, especially when the aim is to describe and understand not only acoustical but also physiological and aerodynamic aspects of voice production that are relevant in voice education.

Discussion. Discussion took place across five groups. Group 1 emphasized that averages are not reliable for representing data like vibrato, and that SPL significantly impacts measurements. Group 2 stressed the need for organizing and simplifying tools for studio and clinical use, noting the overwhelming nature of existing tools. Group 3 discussed the potential of simulations to isolate and understand specific signals, suggesting a forum for sharing tools and methods. Group 4 pointed out the challenges in translating measurements into actionable information, the need for benchmark data, and standardization. Group 5 explored the potential of AI and visualization to enhance feedback and teaching methods, suggesting that additional or combined metrics might improve current practices. Overall, the discussions underscored the importance of better data representation, tool simplification, standardization, and the integration of advanced technologies in acoustic analysis of the voice.

3.2 Inspirational Talk on Visualisation: Collaboration between Domain Scientists and Visualization Experts

Tino Weinkauff (KTH Royal Institute of Technology – Stockholm, SE)

License © Creative Commons BY 4.0 International license
© Tino Weinkauff

Visualization and data analysis have become increasingly integral components of research workflows in all academic disciplines. Data is sourced from experiments, sensors, modeling, simulations, data repositories, and other means, making it a ubiquitous presence in the scientific landscape. Scientific progress heavily relies on comprehending this data. Across all domains, scientists face growing challenges in data analysis, such as dealing with massive datasets, high dimensionality, noisy data, and intricate complexity. Addressing these challenges often extends beyond the expertise of the respective scientists, and rather requires specialized knowledge in data analysis and visualization. Hence, it is necessary to develop a joint language between visualization experts and domain scientists with the goal of clearly

expressing the visualization tasks and the properties of the data. This talk investigates the opportunities and potential pitfalls in these regards, and gives practical advice on how to facilitate the interdisciplinary work.

Discussion. The discussion took place in 4 groups. Group 1 discussed how spectrograms were a breakthrough and that modern visualisation techniques offer a similar paradigm shift. They also noted the need for context-specific visualization tools. Group 2 emphasized the importance of domain knowledge for effective visualization, the potential role of AI, and the need for a repository of visualization tools. Group 3 focused on the value of visualization in distinguishing noise from meaningful data, understanding AI findings, and exploring data. Group 4 considered how visualization can reveal unknowns, facilitate discussions, and present data in new narrative forms, stressing the importance of multi-modal understanding. Across the groups, there was a call for developing a domain-specific visualization toolkit and addressing issues like data reduction and meaningful representation.

3.3 Inspirational Talk on Simulation: Articulatory Speech Synthesis: Modelling and Simulation – or – Can we make a voice instrument?

Sidney Fels (University of British Columbia – Vancouver, CA)

License  Creative Commons BY 4.0 International license
© Sidney Fels

In this inspirational talk, I make the case that the goal of creating a human controlled voice instrument embodies many of the major challenges in speech modeling and synthesis research. Starting with a inspiration drawn from the successes and shortcomings of various voice instruments of the past, like Bell's skull-based actuator, the Vodor and my own Glove-TalkII instrument, I discuss the driving aspects of my own research on biomechanical modelling of the human vocal tract, forays into making a real-time speech synthesis engine and human control of complex systems. The first shortcoming in Glove-TalkII was that the control space used formant which led to a difficult mapping between the relatively slow hand gesture movements and the faster dynamics of the formant space. The thinking was, if we could build an articulatory speech synthesizer that uses synthetic muscle activations, rather than kinematics, the mapping between hand gestures to simulated muscles would be easier to learn for the person and neural networks. This led to, first, creating a computational, muscle activated, hybrid rigid-body/FEM biomechanical model of the human vocal tract and the necessary modelling and simulation infrastructure, Artisynth.com, suitable for simulating the complexity of the human vocal tract dynamics in real time. Second, the geometry of the simulated 3D vocal tract models the dynamic 3D geometry of the human vocal tract, thus, like a number of researchers in our field, we are working on generating the voice acoustics from the 3D geometry of the vocal tract. We require real-time performance, hence, 3D FEM approach are not suitable. Plus, we need the full range of vocal sounds, including frication and stops, leading to our 2.5D approach rather than modification of the 1D wave equation. Finally, we have been investigating how the human brain represents and performs speech motor control to discover ways to reduce the information requirements needed for control. Our thinking is that the human motor control system takes advantage of constraints in the physical world to reduce the needed degrees of freedom to make takes have a reduce index of difficulty, thus, reducing the information load for control. Thus, the real-time articulatory speech synthesis engine embodies the physics of the world that can be used by a machine

learned adaptive mapping and human motor control system to make the voice instrument easier to play, whether by brain signals or gesture. During the talk, I cover how much progress has been done by us and others and how much is left to do. Finally, I conclude with the assertion that creating a new voice instrument for human vocal expression pushes limits of knowledge for modeling, simulation and control of voice and speech.

Discussion. The discussion took place across five groups. Group 1 considered the challenges in defining constraints, and wondered how much knowledge of soft tissue biomechanics is required for appropriate emergent behaviour. The interaction between different systems with different levels of detail was considered a very positive step. Group 2 explored what makes a musical instrument expressive, emphasizing effort, perceiver presence, control, and the creation and perception of time patterns, linking pleasure in performance to flow and transcendence. Group 3 discussed the balance between passion projects and career sustainability, the need to contribute to a larger goal, and the implications of the model presented for training, healthcare, and physiology. Group 4 highlighted the prediction of surgical outcomes, the role of auditory feedback, and the interplay between speech production and perception, including the importance of visual aspects. Group 5 focused on the reasons for pursuing expressive instruments, clinical applications, patient-specific simulations, and the historical significance of voice models, noting the need for storytelling to secure support and funding.

3.4 Inspirational Talk on Data Science and ML: Data Science and Machine Learning on Clinical Applicability of Voice Studies

Pedro Gómez-Vilda (NeuSpeLab – Las Rozas de Madrid, ES)

License  Creative Commons BY 4.0 International license
© Pedro Gómez-Vilda

Machine Learning (ML), Deep Neural Network (DNN) architectures, and End-to-End Systems (EES) for both text-to-speech and speech recognition have become remarkably successful, but they remain quite agnostic of the basics of neurological foundations of vocal and speech function. Nevertheless, ML, DNNs and EES are very useful for dealing with pathological speech characterization, either organic, functional, or neurogenic. Rather than seeking for automated discrimination between normal and pathological voice, clinicians wish for objective assessments of the progress of an intervention. AI researchers, on the other hand, look for ways to distil predictive scores about voice pathology from multi-modal big-data observations. A comprehensive search on domain-specific shallow architectures having an impact toward these goals is nowadays the Holy Grail of ML.

Discussion. The discussions took place in four groups. All groups reported some variation of the idea that AI should assist, rather than replace, human decision-making, since AI tools are adept at distinguishing patterns but should lack human expertise and intuition. Group 1 questioned the clinical relevance of binary (pathological / non-pathological) classification, and noted the use of AI tools for learning something about features of relevance. Group 2 discussed an analogy with cars, suggesting something akin to a driving licence for using AI – we don't need to know exactly how cars work, but we do need to follow certain rules to ensure everyone's safety. Group 3 pointed out that the humans using the AI are the major concern, and the risks of AI amplifying errors due to human misuse, also suggesting

for certifications for AI users. Group 4 discussed the task-sensitive nature of AI, its role in simplifying clinical decisions, and the varying levels of trust in AI among clinicians and patients. An additional point was raised about the concentration of power and resources required for big data, highlighting the need for ethical considerations.

4 Free papers

Participants were invited to volunteer presentations without constraints, which resulted in ten free talks of 10-15 minutes each. The first three free talks were given on day 2 and the remaining ones on day 4.

4.1 Revisiting Laver's voice qualities

Philip Aichinger (Medizinische Universität Wien, AT)

License  Creative Commons BY 4.0 International license
© Philip Aichinger

Voice quality is a relevant feature in healthy as well as pathological speech. In particular, the use of different voice qualities in connected speech may be impeded in pathological speakers. Laver had defined phonatory settings relating to different voice qualities in terms of longitudinal tension, medial compression, and adductive tension [1]. This enabled modal, whispery, breathy, creaky, and harsh voice, as well as falsetto. A laryngeal high-speed video synthesizer [2] is extended here to visualize both the cartilaginous and the muscular parts of the glottis. A few example videos showing different voice qualities are presented.

References

- 1 J. Laver, *The Phonetic Description of Voice Quality*. Cambridge University Press, 2009.
- 2 P. Aichinger; S. Kumar; S. Lehoux; J. Švec, *Simulated Laryngeal High-Speed Videos for the Study of Normal and Dysphonic Vocal Fold Vibration*, *J. Speech, Lang. Hear. Res.*, 65:7, 2431–2445, 2022.

4.2 Voice conversion approaches to synthetic substitution voices

Philip Aichinger (Medizinische Universität Wien, AT)

License  Creative Commons BY 4.0 International license
© Philip Aichinger

Communication disorders, including speech pathologies, have a 12-month incidence of approximately 10%, varying in severity. In severe, persistent cases, these disorders can significantly impact quality of life, potentially leading to social isolation. Assistive devices such as speech synthesizers can restore communication during and after rehabilitation. However, their use is often limited by difficulties in controlling them and dissatisfaction with the voice sound. Voice conversion using deep learning offers a new approach by emulating a target speaker's identity based on a reference microphone recording. Essentially, the user speaks normally, and the voice converter outputs improved speech. This study aims to assess the quality of speech output by voice converters, focusing on perceived voice quality improvement while monitoring potential adverse effects on intelligibility.

Test material includes speech audio of pathological speakers and a healthy speaker using an electrolarynx. A few examples of speech converted to normal are presented. In conclusion, current advancements in voice conversion technology promise improvements in output quality and reduced latency, making voice conversion technology a potential game-changer in the field of speech-assisting devices, especially in telecom applications.

4.3 Sampling rate bias of vocal cycle perturbations

Jean Schoentgen (Free University of Brussels, BE)

License  Creative Commons BY 4.0 International license
© Jean Schoentgen

Acoustic features that report vocal cycle length or vocal cycle amplitude perturbations are biased when they involve length or amplitude data that are sampled at the pace of the vocal cycles. The reason of the bias is that the features involve high-pass filtering of the cycle lengths or amplitudes to separate fast (i.e. jitter or shimmer) from slow perturbations followed by averaging the magnitude of the filtered data. Indeed, the cut-off frequency of a digital filter depends on the sampling frequency that is equal to the vocal frequency (f_0) when the speech signal is sampled once every cycle. The cutoff frequency above which the vocal perturbations are reported therefore depends on the vocal frequency. As a consequence, the jitter or shimmer bandwidth is token-dependent loose from any physiological or anatomical causes. For the same reason, the vocal perturbation bandwidth evolves within a same token when the intonation is not flat. This difficulty concerns most of the known features that report vocal jitter or shimmer. Typically, these features are obtained by means of popular analysis software such as MDVP or PRAAT.

We have compared five biased features that are reported by PRAAT and that describe the jitter of cycle lengths sampled at the rate of f_0 to one unbiased feature describing the jitter of cycle lengths sampled at a fixed rate. The purpose of the comparison has been to examine whether the dispersion of the feature values within a corpus of speech tokens differs for fixed and variable rate features as well as whether variable-rate features may be substituted for fixed-rate features. A second topic has been the correlation of the magnitude and frequency of vocal jitter with vocal frequency f_0 .

4.4 Inverse filtering

Johan Sundberg (KTH Royal Institute of Technology – Stockholm, SE)

License  Creative Commons BY 4.0 International license
© Johan Sundberg

This talk presented the finding that after inverse filtering, harmonic partials located at or close to $F1$ are reduced in amplitude, likening the effect to the partial “doing the limbo” under the formant. A discussion then ensued about the cause of this observation. In the discussion it was noted that the inertic part of the impedance rises with frequency towards the peak of a resonance, and then drops rapidly, becoming compliant at the resonance frequency itself. The fact that this occurs over a wider frequency range than just at the resonance frequency is due to the wider bandwidths associated with real vocal tracts. This suggestion is borne out by the observation that the harmonic amplitude increases slightly with the inertance in the frequency region below the resonance.

4.5 Model based speech research

Brad Story (University of Arizona – Tucson, US)

License  Creative Commons BY 4.0 International license
© Brad Story

During the production of speech, a talker-specific, baseline configuration of the airway is modulated almost continuously by the movements of the tongue, jaw, lips, velum, and larynx. From an acoustic perspective, the airway can be considered a non-uniform conduit whose shape at a given instant of time supports a specific pattern of acoustic resonances that transmit information related to both the intended message and the identity of the talker. This presentation begins with recollection of attempts to simulate connected speech about 30 years, and then summarizes the recent development of a model in which individual speech segments that comprise a word, phrase, or sentence are specified as relative deflections of the resonance frequencies of the baseline vocal tract configuration, and then transformed to time-dependent modulations of the airway. The output of the model is artificial speech that can be presented to listeners. Examples will demonstrate the construction of speech with the model, results from a recent perceptual experiment, effects that may occur with constraints imposed on the vocal tract, and engage the audience in an interactive listening experience. [Research supported in part by NIH 5R01DC017998 and Galileo Circle Fellows grant from the University of Arizona.]

4.6 The Laryngonaut

Scott Reid Moisiik (Nanyang TU – Singapore, SG)

License  Creative Commons BY 4.0 International license
© Scott Reid Moisiik

Laryngeal simulation and visualization overwhelmingly emphasize the vocal folds. The structures of the larynx above the level of the vocal folds, known collectively as the epilarynx (which includes the ventricular folds, aryepiglottic folds, and epiglottis), are often removed to provide a better view of the vocal folds below, despite the important contribution of epilaryngeal structures to overall laryngeal behaviour in speech, singing, and life-supporting functions. In this work, I discuss a broadly targeted phonetics-oriented laryngeal simulation platform that provides interactive 3D visualization and sound synthesis. I will discuss how the work draws on a range of numerical methods for fast simulation of the physics of rigid and deformable bodies, representing the laryngeal structures, and a simplified 1D aeroacoustic model coupled to Titze's (1973) dual particle-chain model for vocal fold vibration, all with the express purpose of real-time responsiveness to user interaction. The entire larynx is represented, including the often-neglected structures of the epilarynx, which, along with the vocal folds, undergo natural-looking deformations in response to changes of muscle activity and virtual manipulation of the structures. With this system, it is possible to explore a range of laryngeal states which map onto important phonetic categories (including various phonation types, such as creaky voice, and articulations, such as epiglottal stop) while simultaneously generating an appropriate approximation of the sound output expected for the current laryngeal state. While considerable progress has been made, much still remains, and some future plans will be discussed, such as extending the system to provide a simulation of the various types of epilaryngeal vibration (such as aryepiglottic trilling).

4.7 Pressure in, SPL out (where is the flow?)

Peter Pabon (Utrecht, NL)

License © Creative Commons BY 4.0 International license
© Peter Pabon

Variation seen in acoustic voice quality measurement is often seen as an inevitable random factor happening along the time line. However, bringing in the SPL- and fo-based localisation principle of voice mapping refutes the notion that a voice can be characterized with some single representative quality value with some inevitable and uncontrollable variation. Rather, metrics exhibit much less intra-subject variation when mapped against SPL and fo, while inter-subject differences prove to be larger than expected. The localising principle rests on (1) a calibrated SPL scale (dual mic headset, constant microphone distance checking and calibration), (2) pitch-period synced information processing and metrics sampled in sync, (3) clean, unbiased, preferable salient metrics on log scales (no hidden constraints by hard links to axes) – “where does this voice hit its own constraints?” – limited control space; and (4) by not allowing disinformation, that is, include information (including noise) from the voice only.

References

- 1 Pabon, P., and Ternström, S. (2020). “Feature Maps of the Acoustic Spectrum of the Voice,” *J. Voice*, 34, 161.e1-161.e26. doi:10.1016/j.jvoice.2018.08.014

4.8 First steps to a pacemaker for phonation?

Oriol Guasch (Ramon Llul University – Barcelona, ES)

License © Creative Commons BY 4.0 International license
© Oriol Guasch

We present first ideas about the possibility of building a pacemaker for phonation. Vocal fold mass models show that small changes in muscle restoring forces, such as excessive stiffness typical of Parkinson’s patients, or too much subglottal pressure, can cause chaotic oscillations of the VF, resulting in abnormal glottal volume flow. Regular oscillations could be restored by a pacemaker made of a smart material with adjustable damping to control chaos [1]. To evaluate the effectiveness of the pacemaker on voiced sounds, the glottal volume velocity for normal, chaotic, and controlled vocal fold oscillations can be calculated and convolved with MRI-derived vocal tract impulse responses for the vowels /a/, /i/, and /u/ [2]. Audiovisual files in spectral and temporal analysis show that chaotic oscillations significantly distort vowel sounds, but the control strategy restores them to normal.

References

- 1 O Guasch; A. Van Hirtum; A.I Fernández; M. Arnela, *Controlling chaotic oscillations in a symmetric two-mass model of the vocal folds*, *Chaos Solit. Fractals*, 159, 112188, 2022.
- 2 O Guasch; M. Freixes; M. Arnela; A. Van Hirtum, *Controlling chaotic vocal fold oscillations in the numerical production of vowel sounds*, *Chaos Solit. Fractals*, 182, 114740, 2024.

4.9 Formant tuning of 3D vocal tracts for FEM vowel synthesis

Marc Arnela (*Ramon Llul University, ES*)

License  Creative Commons BY 4.0 International license
© Marc Arnela


This talk introduces a methodology for tuning formants in 3D vocal tract geometries obtained via MRI to produce specific voice or singing effects. The process involves converting 3D MRI-based geometries into 1D area functions, iteratively adjusting these functions with an algorithm based in sensitivity functions, and reconstructing the 3D vocal tract with modified cross-sections. This method enables the tuning of vowel formants while maintaining the high energy spectrum of 3D models at low computational cost. Examples include shifting the first formant ($F1$) and generating formant clusters, such as those found in singing.

References

- 1 O Guasch; M. Arnela; A. Pont, *Resonance tuning in vocal tract acoustics from modal perturbation analysis instead of nonlinear radiation pressure*, J. Sound Vib., 493, 115826, 2021.
- 2 M. Arnela; O Guasch, *Formant frequency tuning of three-dimensional MRI-based vocal tracts for the finite element synthesis of vowels*, IEEE/ACM Trans. Audio Speech Lang. Process, 32, 2790–2799, 2024.

4.10 Development of an AI assisted data assimilation framework for voice research and clinics

Qian Xue (*Rochester Institute of Technology, US*)

License  Creative Commons BY 4.0 International license
© Qian Xue

Capturing real-time high-resolution 3D images of tissue dynamics remains a challenge due to several inherent factors including organ accessibility and low temporal/ spatial resolution of imaging and image reconstruction. This study introduces a novel hybrid physics-informed neural network (PINN) differentiable learning algorithm that integrates a recurrent neural network model of 3D continuum soft tissue with a differentiable fluid solver to infer 3D flow-induced tissue dynamics and other physical quantities from sparse 2D images. To enhance the scalability and convergence, the designed algorithm leverages the prior knowledge in solid mechanics by projecting the governing equation onto the numerical eigenmode space, reducing the infinite dimensions of the continuous solution space to a finite dimension of discrete search space. The dimensions of the problem are further reduced by only using truncated eigenmodes, which can effectively represent the whole dynamics with negligible errors. To better capture the temporal dependence of flow-structure interaction (FSI) dynamics and enhance the predictive accuracy, a Long short-term memory (LSTM)-based recurrent encoder-decoder connected with a fully connected neural network (FCNN) is designed to learn the time history of modal coefficients, which, combined with eigenmodes, enable the spatiotemporal predictions of tissue dynamics. The effectiveness and merit of the proposed algorithm is demonstrated in subject-specific models by using synthetic data from a canine vocal fold model and experimental data from four excised pigeon syringes. The results showed that, by only using sparse 2D vibration profiles, the algorithm was able to accurately reconstruct the full 3D tissue dynamics as well as other high-dimensional physical quantities, such as

aerodynamics and acoustics quantities, which are otherwise very difficult/impossible to measure. The algorithm can advance disease diagnosis beyond the current morphological and 2D dynamics criterion. It also allows a significant expansion of the measurable quantities in both experimental/clinical research, which could broadly enhance biomedical research capabilities.

5 Working Groups

5.1 Spontaneous workshops

On the first day of the seminar, a whiteboard poll was conducted for identifying the topics of greatest interest to the participants. Working groups were then formed around these topics:

- Individual variability in the voice
- Physical replicas for modelling and validation
- Synthesis and computational modelling
- Motor control
- Data and resources

In a breakout session, the groups discussed these topics for about 30 minutes and then returned for a plenary discussion.

5.1.1 Individual variability in the voice

The first two groups focused on individual variability, considering sources of variation, and research questions around voice pathologies and how to compare voices. There was also discussion of how voices vary from moment to moment, leading to questions about robustness and the validation of variability measurements. Longitudinal data (comparing an individual to themselves) was preferred over normative values, which have little value at the individual level.

5.1.2 Physical replicas for modelling and validation

The discussion of physical replicas considered their purpose (primarily for validating computational models, but also to understand voice production when existing computational models break down, such as during fricatives or with large deformations). The current issues and needs in this area were identified as being the choice of suitable materials (e.g. active materials), reproducibility, and a question of how much detail is necessary.

5.1.3 Synthesis and computational modelling

The discussion of synthesis focused on how much detail is “enough” in simulation, concluding that there are several very different cases where simulation is used, for which the modelling cost varies greatly. A spectrum exists, from “drastically simplified but real-time”, via “perceptually sufficient but physically lacking”, to “detailed reconstructions of the physics taking many hours to compute”. There are considerable issues around validation, but the data necessary for this is lacking, and depends to a large extent how detailed the simulation method is. Nevertheless it was agreed that some benchmark data could be agreed – if available – to further this goal.

5.1.4 Motor control

Finally, the discussion of motor control noted how – including here! – simulation considerations are often decoupled from the idea of motor control, and suggested a future where the two could be considered together. It remains unclear what control variables the system is actually working with, which is a priority for future research, and may not relate to the variables currently used for simulation. The group also highlighted how perception should be included in models of voice motor control, as the message passes from one brain to another by way of muscle movement.

6 Workshops and talks on dreams and failures

The idea of these workshops and talks was to explore the journey of innovation in voice science through three key questions, each addressed in a sub-section. *What do you need to make a leap?* focuses on the essential elements and conditions required for significant advancements in the field. By understanding these factors, we can better navigate the complexities of research and development. *My favorite failure* reflects on setbacks that provided invaluable lessons and insights for future developments. Failures often pave the way for breakthroughs, making them crucial for sustained innovation. Finally, *My dream come true* envisions the future of progress in human voice analysis and simulation, setting ambitious goals for the years ahead.

The workshops were organised around spontaneous presentations followed by discussions, with the aim of encouraging participants to dream big and push the boundaries of what is currently possible in voice science.


6.1 What would you need to make a leap?

This was a one-hour break-out session with group discussions, summarized *in plenum*.

Summary of Discussions. The five discussion groups presented various ideas and desires to advance voice science and technology. Group 1 proposed creating a 100M€ voice center, developing non-invasive physiological signal collection systems, and employing a “voice influencer” to fund research. They also envisioned a robot head with soft tissue activation and advanced imaging systems for vocal tract shapes. Group 2 focused on AI-driven data analysis, distance voice teaching with machine learning, and portable devices for comprehensive physiological monitoring. Group 3 aimed for a digital twin of the human body, systems visualizing interconnected data, predictive tools, and general models that can be personalized. Group 4 sought real-time imaging of muscle activation, non-invasive measurement of subglottal pressure, and pacemakers for vocal folds. Group 5 emphasized the need for funding, advanced measurement techniques, comprehensive computational models, and collaborative efforts. They stressed the importance of storytelling in securing funding, the need for standardized data-sharing protocols, and a community-driven approach to set priorities and justify research efforts.

6.2 My favourite failure: Learning from failures and outliers: “Oh my... that was embarrassing...”

Eric J Hunter (University of Iowa, US)

License  Creative Commons BY 4.0 International license
© Eric J Hunter

Isaac Asimov said, “The most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ but rather ‘hmm... that’s funny...’”. In its simplest form, the practice of science can be thought of as learning from our mistakes and noticing those things that don’t quite fit theory. While we work to approach research rationally, avoiding dogmatic opinions, and remaining open to opposing views, we really can never claim absolute truth yet always in pursuit of it. Applying the scientific method involves objective analysis of issues, recognizing that induction plays a role, but the true method for advancing knowledge is often filled outliers, oddities, and even some immense and embarrassing failures. Yet, learning from these is integral to both scientific progress and personal growth. This presentation includes a few of my stories.

6.3 My favourite failure: can subglottal pressure be estimated from intra-oral pressure in speech and singing?

Nathalie Henrich Bernardoni (Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, FR)

License  Creative Commons BY 4.0 International license
© Nathalie Henrich Bernardoni

The direct measurement of subglottal pressure is challenging, because it requires a very invasive approach. It consists in placing a pressure transducer below the glottis by tracheal puncture between the cricoid cartilage and the trachea first ring. Other methods have been proposed, which estimate the subglottal pressure using less invasive approaches. The most common one is to estimate subglottal pressure from intra-oral pressure measured during the closed phase of voiceless consonants (Smitheran and Hixon, 1981; Hertegard et al., 1995). This approach has been validated in the case of normal speaking voice. However, few studies have explored its validity for soft or loud voice, for whisper or pressed voice, and in the case of singing. This study explores the possibilities and limitations of estimating subglottal pressure from intra-oral pressure in speech and singing. Two subjects (a speaker and a trained singer) were recorded while uttering CV segments (plosive consonant followed by a vowel) with different voice qualities (normal, soft, loud, whisper, pressed). The singer sung sentences at several pitches covering his comfortable tessitura in the two main laryngeal mechanisms. Two recording sessions were conducted with a one-year time interval in between. Several methods for estimating subglottal pressure from intra-oral pressure signal are compared. A good agreement between estimates and direct measures is found in many cases. Bad agreement may be found in the case of soft phonation, for productions in laryngeal mechanism M2, and in the case of pressed speech. We thought a lot about the reasons for these discrepancies, and found that they may be due to the absence of a nose clip during measurements. The conclusion of this favourite failure is that it is of much importance to be concerned about the nose in speech and singing production, even in cases where it should not interfere ...

References

- 1 Smitheran JR, Hixon TJ. (1981) A clinical method for estimating laryngeal airway resistance during vowel production. *J Speech Hear Disord.* 46(2):138-46.
- 2 Hertegård S, Gauffin J, Lindestad PA. (1995) A comparison of subglottal and intraoral pressure measurements during phonation. *J Voice* 9(2):149-55.

6.4 My favourite failure: Spontaneous confessions

Sten Ternström (KTH Royal Institute of Technology – Stockholm, SE)

License  Creative Commons BY 4.0 International license
© Sten Ternström

In addition to the above accounts, there were a few spontaneous and amusing participant “confessions” to failures.

Jean Schoentgen related how he had once been trying to synthesise rough voices, with some success – they sounded rough and reasonably natural. Then he had the “good bad idea” to apply the general knowledge to include soft f_0 contours that decrease 10-30% over 1 s. But then the roughness disappeared! He thought it was an error, but... information in the frequency contour is in the spectral side-bands – and so is the roughness. Somehow one information masked the other. If this is correct / relevant – then if you are looking for voice quality in connected speech, don’t look for roughness in there.

Sten Ternström described how for the Singing Synthesis Competition at Interspeech 2007 in Antwerpen he had prepared a source-filter synthesis, working with headphones with a soprano voice in the left ear only. But on playback over loudspeakers at the venue, the soprano was too shrill, resulting in a low rating in the competition. Sten had forgotten to account for an imbalance in his own hearing, which has a dip at 3-5 kHz in the left ear. If he had tried reversing the headphones, or first played it to someone else, the competing entry might have done better!

Peter Pabon recounted how he once had been on the brink of a very high-profile failure, on a live television show in the Netherlands. He had been tasked to break a huge crystal glass, one meter high, by catastrophic resonance, which involves matching a strong loudspeaker tone very precisely to the previously measured eigenfrequency of the glass. The glass was vibrating vigorously, but refused to break. Then Peter noticed that the spectrum analyzer of the audio was showing a slightly different frequency – the tone had been shifted by the studio’s public-address system, as a counter-feedback measure. He tweaked the frequency just that little, and the glass broke, spectacularly. And the shooter with an air-gun, waiting in the wings, did not have to fake it after all.

6.5 My dream come true: the story of Pinocchio

Nathalie Henrich Bernardoni Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, FR)

License  Creative Commons BY 4.0 International license
© Nathalie Henrich Bernardoni

Speech production results from a fine neuro-controlled coordination between breathing, phonatory and articulatory movements. In physical terms, these movements induce fluid-

structure-acoustic interactions within the vocal tract. In the framework of source-filter theory (Fant 1960), the aerodynamic phase of speech is often overlooked, yet being of much importance for voiced and unvoiced speech sound production (Catford 1977). In the case of voiced speech sound, it is taken into account by myoelastic-aerodynamic theory of phonation (Jan G. Švec et al. 2021), in which glottal constriction and vocal folds vibration generate aeroacoustic sources. Our aim is to advance our understanding of speech sound production by developing a biomimetic in vitro test-bed. Over the last thirty years, vocal-fold testbeds have evolved in complexity and biomimicry. However, most testbeds explore the physics of phonation on geometrically-fixed replicas capable of self-sustained oscillations in fluid-structure interaction, but unable to produce intonative variations. Most of the time, the testbeds are not coupled to vocal tract, and whenever they are, the resonant cavities are static 3D-printed tracts. Reproducing in vitro the dynamic movement of speech articulators, such as jaw, tongue, velum and larynx, together with phonation, remains a challenge. We present here the first steps in the design of a biomimetic mechatronic testbed that would integrate all phonatory and articulatory aspects important for voice and speech production.

6.6 My dream come true: Design of self-oscillating biomimetic vocal folds


Lucie Bailly (Université Grenoble Alpes – Saint Martin d’Hères, FR)

License © Creative Commons BY 4.0 International license
© Lucie Bailly

This work aims to contribute to an in-depth understanding of the link between the histo-mechanical properties of the vocal folds and their remarkable vibratory performances, which still remains elusive. In vitro simulators of the phonatory system have been developed for decades. Yet, most of them are made of materials with structural and mechanical properties still far from those of the native tissues. Thus, this work proposes to design new in vitro self-oscillating biomimetic vocal folds with tailored structural and mechanical properties, and study the impact of material properties on the fluid/structure/acoustics interactions driving phonation. To this end, three complementary families of materials were studied: gelatin-based hydrogels cross-linked with glutaraldehyde, biocompatible polyethylene glycol-based hydrogels and silicone elastomers used in voice research. Formulations were optimised to best fit the mechanics of vocal tissues under physiological multi-axial loadings in tension, compression and shear. Fibre-reinforced composites were then produced to mimic the microscale collagen structure of tissues, and their macroscale non-linear and anisotropic behaviour. Finally, the ability of the tailored materials to vibrate under realistic morphological and aerodynamic conditions was tested using an articulated larynx test-bed. Promising vibratory patterns were obtained for the optimised hydrogel-based replicas, able to self-oscillate with subglottal pressures close to physiological reality. The data have evidenced the link between the material properties and the aeroacoustic performances of several synthetic oscillators, paving the way for future investigation of the impact of the fibrous architecture of vocal tissue on voice quality.

6.7 My dream come true: Discussion on particle-based simulations

Peter Birkholz (TU Dresden, DE)

License  Creative Commons BY 4.0 International license
© Peter Birkholz


A short discussion followed on the possibility of particle-based simulation of the aeroacoustics of the vocal system. Some outline calculations suggested that the computational complexity was feasible at a suggested density of 1000 particles/mm³. Discussion indicated the principle was the same as the lattice Boltzmann method and highlighted the necessity for more complex simulations with more particles for accurate modelling. It was suggested that high-performance computing resources would be suitable for this kind of task.

7 Workshop on the elusive Inversion Problem

An age-old problem with studying the voice organ is that is difficult to access experimentally. Much effort continues to be directed to the inversion problem, that is, to inferring what the voice organ is doing with access only to external signals. This is closely related to acquisition technologies for biophysical data, which are in rapid development. There were two invited talks (Laprie, Birkholz) and one contributed paper (Zhang).

7.1 Acoustic-articulatory inversion with rtMRI

Yves Laprie (LORIA – Nancy, FR)

License  Creative Commons BY 4.0 International license
© Yves Laprie

The use of real-time MRI (rt-MRI) data is dramatically changing the way acoustic-to-articulatory inversion can be addressed. In general, inversion is based on electromagneticographic (EMA) data, but they provide a very limited information in the form of just a few points. rt-MRI data have the advantage of covering the entire vocal tract from glottis to lips, and of enable larger databases for each speaker. On the disadvantage side, rt-MRI data need to be pre-processed to track articulator contours, and the speech signal needs to be denoised. We have developed effective solutions to both these problems. It is also necessary to be able to move from the speech signal acquired in a low-noise environment in the sitting or standing position to the MRI signal produced in a strong noise in the supine position. The inversion itself can be carried out using deep learning techniques, inspired by those already used for EMA-based work.

7.2 Two examples using VocalTractLab and neural networks

Peter Birkholz (TU Dresden, DE)

License © Creative Commons BY 4.0 International license
© Peter Birkholz

The inverse problem tries to estimate parameters of the speech production process from the speech audio signal. In this presentation I show how the articulatory speech synthesizer VocalTractLab (www.vocaltractlab.de) was used to generate synthetic data to train artificial neural networks (ANN) to solve two kinds of inversion problems: the estimation of the glottal flow and the estimation of the articulatory movements from the speech audio signal. In both cases, the ANNs trained on purely synthetically generated data showed a promising performance when used on human speech signals. Directions for future research in this direction are discussed.

7.3 Inversion in voice production

Zhaoyan Zhang (UCLA, US)

License © Creative Commons BY 4.0 International license
© Zhaoyan Zhang

In this talk, I will present a simulation-based machine learning model toward solving the inverse problem in voice production, i.e., to estimate vocal fold geometry, stiffness, position, and subglottal pressure from the produced voice acoustics and aerodynamics, toward clinical and voice technology applications. Unlike previous voice inversion research that often uses lumped-element models of phonation, this study explores the feasibility of voice inversion using data generated from a three-dimensional voice production model. Neural networks are trained to estimate vocal fold properties and subglottal pressure from voice features extracted from the simulation data. Results show reasonably good estimation accuracy, particularly for vocal fold properties with a consistent global effect on voice production, and reasonable agreement with excised human larynx experiment [1]. Human subject studies further showed that the neural network was able to monitor the subglottal pressure with reasonable accuracy and predict the alternating vocal fold adduction and abduction pattern during consonant-vowel-consonant transitions [2]. All subjects simultaneously increased the subglottal pressure and vocal fold approximation when producing a louder voice, although the degrees of laryngeal and respiratory adjustments were speaker-specific. These results demonstrate the potential of this neural network toward monitoring and identifying potentially unhealthy vocal behaviors outside the clinic.

References

- 1 Z. Zhang, *Estimation of vocal fold physiology from voice acoustics using machine learning*, J. Acoust. Soc. Am., 147, EL264–EL270, 2021.
- 2 Z. Zhang, *Estimating subglottal pressure and vocal fold adduction from the produced voice in a single subject study*, J. Acoust. Soc. Am., 151, 1337–1340, 2022.

8 Challenges in voice analysis and simulation

As a researcher, it is easy to fall into “tunnel vision” and lose the perspective of the larger questions. Therefore, two participants were invited to take a step back and present their view of some over-arching issues in voice analysis and voice simulation.

8.1 What do clinicians need?

Meike Brockmann-Bauser (Universitätsspital Zürich, CH)

License  Creative Commons BY 4.0 International license
© Meike Brockmann-Bauser

Instrumental acoustic measurements of quantitative and qualitative human voice characteristics have enormous potential to objectively describe vocal pathology and, thereby, to assist clinical treatment decisions. Despite an increasing application and availability of these techniques, recent research has highlighted a lack of understanding of physiologic and speech related influencing factors. This contribution critically reviews the state of the art in the clinical application of currently recommended instrumental acoustic voice measures and points out future directions. Recent research in vocally healthy and voice disordered adults has shown, that the most widely recommended voice quality measures jitter (%), shimmer (%), harmonics-to-noise ratio (HNR) and Cepstral Peak Prominence (CPP) are affected by natural variations in speaking voice intensity and pitch, vowel and voice task type. Moreover, age, voice training and gender related differences have not been comprehensively described for qualitative, but also quantitative measurements including Voice Range Profiles (SPL and F0 plots). In summary, main limitations include a lack of (a) normative data for known physiologic covariables, such as age, training, speaking voice sound pressure level (SPL) and fundamental frequency (f_0) (b) standardization and reporting of analysis procedures (including voice tasks) and techniques (c) understanding of the relation between audible dysphonia, vocal dysfunction, and instrumental acoustic voice features. Future directions include the exploration of Voice Range Profiles complemented with a third dimension of voice quality or electroglottographic measures as clinical tools for pre-post comparisons of voice functionality, related to specific tasks and pathologies. This calls for further research to transfer currently available techniques into clinical applications.

Details are given in [1].

References

- 1 Brockmann-Bauser, M. and M.F. de Paula Soares, Do We Get What We Need from Clinical Acoustic Voice Measurements? *Applied Sciences*, 2023. 13(2): p. 941.

8.2 Numerical simulations

Michael Döllinger (Universitätsklinikum Erlangen, DE)

License  Creative Commons BY 4.0 International license
© Michael Döllinger

This talk introduced the different types of modelling approach used for numerical simulation of phonatory processes. Models were presented on a spectrum ranging from highly simplified (but fast to run) to highly detailed (but computationally extremely expensive). Lumped

mass models offer the lowest-cost approach, and can simulate basic phonatory principles. Their low cost makes them suitable for optimisation or inversion studies, but they constitute a rough approximation of real physiology, it is difficult to incorporate pathologies, and models are difficult to compare to one another due to a lack of standardisation. Increasing in complexity, computational fluid dynamics (CFD) approaches simulate the 3D flow field in high resolution, but only in relation to proscribed vocal fold movements. Although this is only the second approach on the list, computational cost is already high enough that HPC resources are typically required here, along with detailed expert knowledge about simulation parameters and conditions. Computational aeroacoustics (CAA) methods are based on CFD approaches and also include acoustics, allowing use of the models to investigate sound generation processes, again with a higher computational cost. Increasing in detail further, fluid-structure interaction (FSI) models permit the simulation of flow-induced vocal fold oscillations, provided an appropriate biomechanical tissue model is available for the vocal folds. This method is able to capture the vocal fold-airflow interactions, again at the expense of increased computational cost over CFD/CAA models. Finally, the most detailed models currently in use are fluid-structure-acoustic interaction (FSAI) models, which couple FSI models with acoustics, providing a comprehensive account of the whole phonation process, but with requirements for deep expert knowledge of the systems and parameters involved, and a computational cost higher than all of the above approaches. In all cases, validation is a critical issue. Validation should be performed in order to ensure the models are correct, but there is a critical lack of data to validate against. In summary, there are a range of modelling approaches available, each with pros and cons, and the choice of model should be based upon the research question under study.

9 Concluding Plenary

9.1 Towards the future

Amelia Gully (University of York, GB)

License © Creative Commons BY 4.0 International license
© Amelia Gully

This talk is both a summary of many of the key points raised throughout the week, and a perspective from an early-career researcher looking towards the future of the field. I make three main arguments: 1) we are an interdisciplinary community, and we need to understand our differences so that we can communicate in a shared language; 2) we need to share high-quality resources to avoid re-inventing the wheel; 3) a concerted, international effort to promote voice science is needed in the face of public indifference and AI voices created without any understanding of the underlying systems. I also share some of my experiences developing interdisciplinary communities at a national level. This Dagstuhl Seminar has gone a long way towards building the necessary trust between researchers, and if we can build upon this to promote excellent international voice research, we can make a real difference in the world.

9.2 Final discussions

Participants expressed various perspectives and challenges in their fields. Several highlighted the difficulty of funding and recruiting for research, emphasizing the importance of sustained financial support and collaborative efforts. Issues such as data sharing, standardization, and the need for interdisciplinary collaboration were recurrent themes. There was enthusiasm for initiatives like World Voice Day to raise public awareness, and the suggestion of potential avenues for broader outreach, through media like documentaries or social media platforms. Overall, all participants agreed on the value of personal interactions and ongoing dialogue among researchers, aiming to foster continued collaboration and innovation beyond the event itself. A commitment was made to produce a joint position article on the state of the art and future perspectives in research on the human voice.

Participants

- Philipp Aichinger
Medizinische Universität
Wien, AT
- Marc Arnela
Ramon Llull University –
Barcelona, ES
- Lucie Bailly
Univ. Grenoble Alpes, CNRS,
Grenoble INP, 3SR,
Grenoble, FR
- Peter Birkholz
TU Dresden, DE
- Meike Brockmann-Bauser
Universitätsspital Zürich, CH
- Helena Daffern
University of York, GB
- Michael Döllinger
Universitäts-Klinikum
Erlangen, DE
- Mennatallah El-Assady
ETH Zürich, CH
- Sidney Fels
University of British Columbia,
Vancouver, CA
- Mario Fleischer
Charité – Berlin, DE
- Andrés Gómez-Rodellar
NeuSpeLab, Las Rozas de
Madrid, ES
- Pedro Gomez-Vilda
NeuSpeLab, Las Rozas de
Madrid, ES
- Oriol Guasch
Ramon Llull University –
Barcelona, ES
- Amelia Gully
University of York, GB
- Nathalie Henrich Bernardoni
Univ. Grenoble Alpes, CNRS,
Grenoble INP, GIPSA-lab,
Grenoble, FR
- Eric Hunter
University of Iowa, US
- Filipa M.B. Lã
UNED Madrid, ES
- Yves Laprie
LORIA, Nancy, FR
- Sarah Lehoux
UCLA, US
- Matthias Miller
ETH Zürich, CH
- Scott Reid Moisk
Nanyang TU, Singapore, SG
- Peter Pabon
Utrecht, NL
- Jean Schoentgen
Free University of Brussels, BE
- Brad Story
University of Arizona –
Tucson, US
- Johan Sundberg
KTH Royal Institute of
Technology – Stockholm, SE
- Sten Ternström
KTH Royal Institute of
Technolog – Stockholm, SE
- Tino Weinkauff
KTH Royal Institute of
Technology – Stockholm, SE
- Qian Xue
Rochester Institute of
Technology, US
- Zhaoyan Zhang
UCLA, US

