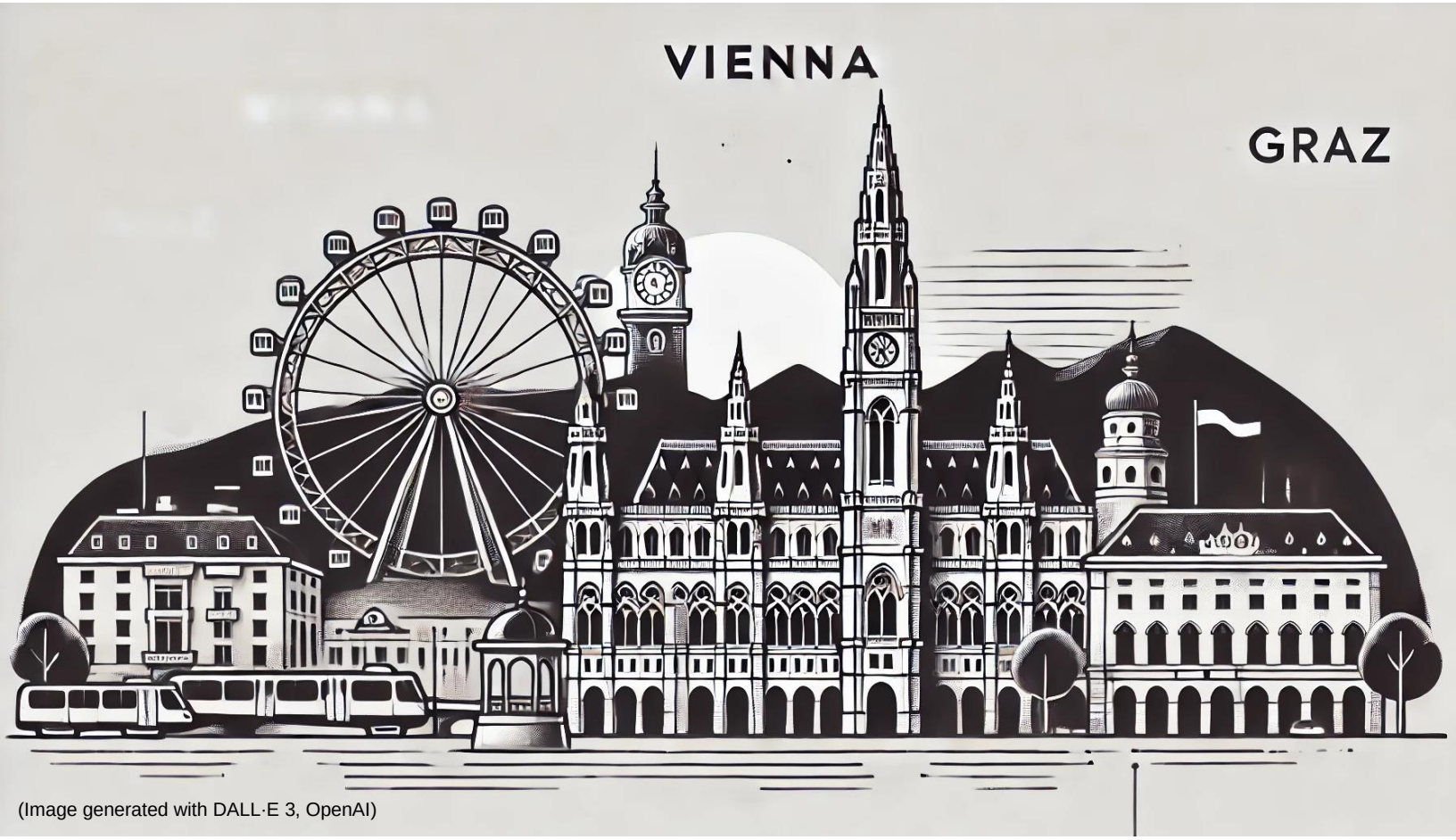
Book of Abstracts

3rd Graz-Vienna Speech Workshop:
Connecting with Health Sciences

Vienna, April 23-25, 2025



(Image generated with DALL·E 3, OpenAI)

Impressum

Book of Abstracts – 3rd Graz-Vienna Speech Workshop: Connecting with Health Sciences
April 23–25, 2025
Medical University of Vienna, Austria

Publisher and Responsible for Content:
Philipp Aichinger
Medical University of Vienna
Währinger Gürtel 18–20
1090 Vienna, Austria

Editorial Team:
Philipp Aichinger (Medical University of Vienna)
Eva Reinisch (Austrian Academy of Sciences)
Barbara Schuppler (Graz University of Technology)

Design and Layout:
Organizing Committee

Contact:
Email: philipp.aichinger@meduniwien.ac.at

Publication Details:
This publication is available through the institutional repository of the Medical University of Vienna Library and will be permanently archived and accessible for reference.
ISBN: 978-3-903477-11-7 (ebook)

Disclaimer:
All abstracts in this volume are published as submitted by the authors. The editors and organizers are not responsible for the accuracy, completeness, or opinions expressed in individual contributions. Inclusion does not imply endorsement.

Content

Editorial 5

Travelling to and navigating through the building: 7

Workshop program 8

Session 1: Child speech

C. Schmid (MedUni Vienna): On the Vowel Development of Bilingual Kindergarten Children in their L2 German 11

A.-L. Feichter (Uni Graz): Expressive Vocabulary Development in Children with Permanent Hearing Loss: Early Nonverbal Predictors 13

M. Galović (TU Graz): Phonetic Analysis of Dysarthric Child Speech 15

S. Alwaisi (BME Hungary): Multi-Style Child TTS: An Expressive Text-to-Speech System for Children 17

Session 2: Conversational speech

Y. Huang (ARI/ÖAW): An Acoustic and Articulatory Study of Voice Quality in Coarticulated Hanoi Vietnamese Tones 19

L. Hladek (ARI/ÖAW): Does Spatial Auditory Attention Change After Having a Face-To-Face Conversation in a Noisy Environment? 21

S. Wepner (TU Graz): (When) Does it Harm to Be Incomplete? Encoding ASR Mistranscriptions of Syntactically Disfluent Structures 24

Y. Meng (BME Budapest): ASR of Non-lexical and Disfluent Events - an Investigation Across Tasks 25

E. Berger (TU Graz): Speech Enhancement of Conversational Speech in Cocktail Party Noise 27

S.M. Pearsell (SDU Sonderborg): Building a Robust HMI Command Inventory for Various Noisy Environments 29

Session 3: Automatic speech recognition

L. Yue (BME Budapest): Comparison of Self-Training Strategies for ASR 31

D. Mengke (BME Budapest): Improving Khalkha Mongolian ASR via Transliteration-Based Transfer Learning 33

M. Gedeon (BME Budapest): Speech Event Extraction: An ASR+NLP Challenge 35

A. Žgank (Univ. Maribor): Towards Creating a Carinthian Slovenian Spoken Language Resource 37

Practical session A:

L. Eckert (TU Graz): How to Use SLICER for Stimuli Extraction from Large Speech Corpora 39

A. Viehhauser (TU Graz): Annotating Creak in Healthy and Pathological Voices 41

J. Linke (TU Graz): SpeechScape: An Easy-to-Use Tool for Clustering Speech Data Based on Self-Supervised Representations 43

S. Lafenthaler (CDK Salzburg): Speech Pauses in Alzheimer Disease: Exploring Challenges in Training Automatic Speech Recognition Systems 45

M. Fleischer (Charité, Berlin): From Medical Data to Models: Own Experiences and Challenges 47

Practical session B:

S. Ternström (KTH Sweden): Evidencing Physiological and Acoustic Outcomes of Clinical Voice Interventions Using Voice Maps 49

J. Yun (TU Dresden): Optopalatography for Phonetic Research 51

E. Reinisch & Team (ARI/ÖAW): Using Ultrasound Tongue Imaging for Phonetic Research 53

M. Gubian (LMU Munich): Analysis of Uni- and Multi-Dimensional Contours with GAMs and Functional PCA: an Application to Ultrasound Tongue Imaging 55

Session 4: Digital health and AI

B. Mayrhofer (TU Graz): Voice Conversion in Pathological Speech: Applications and Challenges 57

P. Cyrta (Uhura Bionics Ltd.): Realtime Personalized Deep Learning Voice Morphing for Electrolaryngeal Speech in Polish Language 59

I. Jánoki (BME Budapest): A Medical Application of ASR and LLM 61

P. Long (MedUni Vienna): The Mere-Measurement Effect in Patient-Reported Outcomes: A Randomized Controlled Trial with Speech Pathology Patients 63

Practical sessions C & D:

P. Pombala (Zana.ai, Karlsruhe): Speech-Based Cardiorespiratory Health Monitoring with VOICE-BIOME: a scalable voice biomarker platform 65

D. Nadrchal (JKU Linz): Deep Learning ASR for a Patient with Permanent Tracheostomy 67

Session 5: Voice science

C. Drioli (Univ. Udine): Physically-Based Machine Learning for Vocal Fold Video Data Interpretation 69

J. Schoentgen (ULB, Brussels): Sampling Rate Bias of Vocal Jitter and Shimmer 71

A. Van Hirtum (Univ. Grenoble): Resonance Frequencies in Non-Rigid Compressed Waveguides 73

X. Pelorson (Univ. Grenoble): Physical Model of Phonation with Reduced and Measurable Parameters. 75

Editorial

Welcome to the 3rd Graz-Vienna Speech Workshop, which is held this year at the Medical University of Vienna from April 23–25, 2025! The Graz-Vienna Speech Workshop was held for the first time in Graz in 2021, aiming at bringing speech scientists and speech technologists from Graz and Vienna together with close international collaboration partners. In contrast to other conferences and workshops, its' focus has since the beginning been on sharing work in progress, offering plenary talks with plenty of discussion time, as well as practical sessions, in which participants learn new tools and methods. Following the spirit of its predecessors, this third edition continues the tradition of fostering interdisciplinary dialogue, collaboration, and innovation in the field of speech sciences and technology -- this time with a dedicated focus on health sciences and applications -- this time with participants from 22 organizations from 10 nations.

The theme "*Connecting with Health Sciences*" reflects the increasingly vital interface between speech research and healthcare. Our program brings together a diverse and dynamic community — from phonetics, linguistics, and engineering to speech-language pathology, psychology, and digital health. This convergence of disciplines aims not only to deepen scientific understanding but also to accelerate real-world applications through technological advances.

This year's workshop is hosted by the Speech and Hearing Science Lab (SHSLab), at a time of exciting institutional growth. Research in Artificial Intelligence is expanding across campus. A new Comprehensive Centre for AI in Medicine, of which SHSLab is a founding member, was founded in January 2025. This momentum is further supported by a series of new research buildings currently under construction — including the Campus Mariannengasse (35,000 m², 750 scientists), the Eric Kandel Institute – Center for Precision Medicine (6,700 m²), the Centre for Translational Medicine (14,000 m²), the Anna Spiegel II building (4,766 m²), and the planned Centre for Technology Transfer (7,000 m²). These state-of-the-art facilities will soon provide ample space and new possibilities for collaborative, interdisciplinary research in speech science and health.

Our scientific program features a rich mix of plenary talks and hands-on practical sessions, organized into thematic tracks. Topics span from early speech development to advanced voice technology and include:

- **Child speech:** Covering bilingual development, hearing loss, dysarthria, and expressive speech synthesis.
- **Conversational speech:** Exploring what sets conversational speech apart from other forms, including voice quality, auditory attention, and the challenge of disfluent and spontaneous speech for both humans and machines.

- **Automatic speech recognition (ASR):** Presenting novel methods, low-resource language strategies, and the limitations of ASR systems when applied to pathological or conversational speech — challenges that even large companies like Google are tackling with initiatives such as Project Euphonia.
- **Practical sessions:** Offering hands-on engagement with tools and data — from ultrasound tongue imaging and voice maps to clustering algorithms, vocal tract data acquisition, speech rhythm in Alzheimer's disease, and MRI analysis.
- **Digital Health and AI:** Focusing on speech-based biomarkers, assistive technologies, medical uses of large language models, as well as internet-enabled data collection, this track reflects the expanding role of AI in healthcare-oriented speech research.
- **Voice science:** Highlighting current approaches to voice modeling, acoustics, and vocal fold dynamics, this session bridges physiology and computation to advance both theory and clinical application.

A particular highlight will be Friday's kick-off of a new research initiative dedicated to assistive voice technologies for individuals with speech pathologies. This project, developed in collaboration with TU Graz, focuses on voice conversion AI and will expand into Vienna as soon as the new SHSLab spaces become available. When the time may come that participants return to SHSLabs Vienna, we look forward to welcoming them into these modern, integrated research environments.

We extend our heartfelt thanks to all contributors for their insights and enthusiasm, and to the institutions and support teams that made this workshop possible. The abstracts collected in this volume represent the breadth, depth, and vibrancy of our community — and the exciting directions in which it is heading.

We hope this workshop will be a space of discovery and connection, and that it will spark ideas and collaborations that continue to grow well beyond these three days in Vienna.

With warm regards,

The Organizing Committee
Barbara Schuppler (Graz University of Technology)
Eva Reinisch (Austrian Academy of Sciences)
Philipp Aichinger (Medical University of Vienna)

Travelling to and navigating through the building:

The workshop is conducted in the Teaching center ("Hörsaalzentrum"), which is located in the General Hospital of Vienna ("Allgemeines Krankenhaus (AKH) Wien"). Please pay attention that Google maps may lead you to the wrong building if you navigate to "Teaching Center", better

Go to
"689W+XQ Vienna" or "Hörsaalzentrum der Medizinischen Universität Wien"

Follow the signs and enter the hospital through the **main entrance hall**. From there

Go to Level 8 using any elevator, and follow the signs to "Leitstelle 8L".

Follow the signs to "Hörsaal (HS) 5".

(The rooms for the practical sessions are right next to the Lecture hall ("Hörsaal") 5.)

By public transportation:

A fast way to travel to the building is by

Underground (metro) line U6 stopping at "Michelbeuern AKH".

(Heavy construction work is currently complicating other options to enter the building.)

By car:

If you come by car, it is most convenient to use the main garage entering from the major street "Währinger Gürtel". Use google maps to

Search for "AKH Einfahrt", or "689V+RC Vienna".

→ Follow the sign AKH Garage WIPARK. ←

Please note that parking there is around EUR 28 per day.

Workshop program

Venue: Teaching center, AKH / MedUni Vienna, Währinger Gürtel 18-20, 1090 Vienna

	Wednesday 23.04.2025 (Room: HS 5, Level 8)	
12:00 - 13:00	Registration	
13:00 - 13:10	Welcome by the organizers	
	Session 1: Child speech	**Chair: Veronika Mattes**
13:10 - 13.30	C. Schmid (MedUni Vienna):	On the Vowel Development of Bilingual Kindergarten Children in their L2 German
13:30 - 13:50	A.-L. Feichter (Uni Graz):	Expressive Vocabulary Development in Children with Permanent Hearing Loss: Early Nonverbal Predictors
13:50 - 14:10	M. Galović (TU Graz):	Phonetic Analysis of Dysarthric Child Speech
14:10 - 14:30	B. Schuppler (TU Graz):	Towards Developing an Assistive Communication Tool for Dysarthric Children
14:30 - 14:50	S. Alwaisi (BME Hungary):	Multi-Style Child TTS: An Expressive Text-to-Speech System for Children
14:50 - 15:10	Wrap up & Discussion	
	Coffee Break (Foyer, Level 8)	
	Session 2: Conversational speech	**Chair: Barbara Schuppler**
15:40 - 16:00	Y. Huang (ARI/ÖAW):	An Acoustic and Articulatory Study of Voice Quality in Coarticulated Hanoi Vietnamese Tones
16:00 - 16:20	L. Hladek (ARI/ÖAW):	Does Spatial Auditory Attention Change After Having a Face-To-Face Conversation in a Noisy Environment?
16:20- 16:40	S. Wepner (TU Graz):	(When) Does It Harm to Be Incomplete? Human and Automatic Speech Recognition of Syntactically Disfluent Structures
16:40 - 17:00	Y. Meng (BME Budapest):	ASR of Non-lexical and Disfluent Events -an Investigation Across Tasks
17:00 - 17:20	E. Berger (TU Graz):	Speech Enhancement of Conversational Speech in Cocktail Party Noise
17:20 - 17:40	S.M. Pearsell (SDU Sonderborg):	Building a Robust HMI Command Inventory for Various Noisy Environments
17:40 - 18:00	Wrap up & Discussion	
18:30	**Dinner at the restaurant "Zum Gangl" Campus Altes AKH (Alser Straße 4/Hof 1)**	

	Thursday 24.04.2025 (Room: HS 5, Level 8)	
	Session 3: Automatic speech recognition	**Chair: Peter Mihajlik**
9:00 - 9:20	L. Yue (BME Budapest):	Comparison of Self-Training Strategies for ASR
9:20 - 9:40	D. Mengke (BME Budapest):	Improving Khalkha Mongolian ASR via Transliteration-Based Transfer Learning
9:40 - 10:00	M. Gedeon (BME Budapest):	Speech Event Extraction: An ASR+NLP Challenge
10:00 - 10:20	A. Žgank (Univ. Maribor):	Towards Creating a Carinthian Slovenian Spoken Language Resource
10:20 - 10:40	Wrap up & Discussion	
	Group photo	
	Coffee Break (Foyer, Level 8)	
	Practical sessions A (Room: KR 21)	**Practical Sessions B (Room: KR 22)**
11:10 - 11:55	L. Eckert (TU Graz): How to Use SLICER for Stimuli Extraction from Large Speech Corpora	S. Ternström (KTH Sweden): Evidencing Physiological and Acoustic Outcomes of Clinical Voice Interventions Using Voice Maps *(2 slots)*
12:00 - 12:45	A. Viehhauser (TU Graz): Annotating Creak in Healthy and Pathological Voices	
	Lunch Break (Mensa, Level 5)	
14:00 - 14:45	J. Linke (TU Graz): SpeechScape: An Easy-to-Use Tool for Clustering Speech Data Based on Self-Supervised Representations	J. Yun (TU Dresden): Optopalatography for Phonetic Research (show and tell)
14:50 - 15:35	S. Lafenthaler (CDK Salzburg): Speech Pauses in Alzheimer Disease: Exploring Challenges in Training Automatic Speech Recognition Systems	N. Elsässer (ARI): Using Ultrasound Tongue Imaging for Phonetic Research
15:40 - 16:25	M. Fleischer (Charité, Berlin): From Medical Data to Models: Own Experiences and Challenges	M. Gubian (LMU): Analysis of Uni- and Multi-Dimensional Contours with GAMs and Functional PCA: an Application to Ultrasound Tongue Imaging
16:30 - 17:30	Discussion and exchange in small self-organized groups	
18:00	**Dinner at the restaurant "Zum Gangl" Campus Altes AKH (Alser Straße 4/Hof 1)**	

	Friday 25.04.2025 (Room: HS 5, Level 8)	
	Session 4: Digital health and AI	**Chair: Philipp Aichinger**
9:00 - 9:20	P. Aichinger (MedUni Vienna), M. Hagmüller (TU Graz):	Project Kick-Off: Voice Conversion for Pathological Speech
9:20 - 9:40	B. Mayrhofer (TU Graz):	Voice Conversion in Pathological Speech: Applications and Challenges
9:40 - 10:00	P. Cyrta (Uhura Bionics):	Realtime Personalized Deep Learning Voice Morphing for Electrolaryngeal Speech in Polish Language
10:00 - 10:20	I. Jánoki (BME Budapest):	A Medical Application of ASR and LLM
10:20 - 10:40	P.A. Long (MedUni Vienna):	The Mere-Measurement Effect in Patient-Reported Outcomes: A Randomized Control Trial with Speech Pathology Patients
	Coffee Break (Foyer, Level 8)	
	Practical sessions A (Kursraum 21)	**Practical Sessions B (Room: KR 22)**
11:00 - 11:45	J. Hoxha (Zana.ai): Speech-Based Cardiorespiratory Health Monitoring with VOICE-BIOME: a scalable voice biomarker platform	D. Nadrchal (JKU Linz): Deep Learning ASR for a Patient with Permanent Tracheostomy
	Session 5: Voice science	**Chair: tba**
11:50 - 12:10	C. Drioli (Univ. Udine):	Physically-Based Machine Learning for Vocal Fold Video Data Interpretation
12:10 - 12.30	J. Schoentgen (ULB):	Sampling Rate Bias of Vocal Jitter and Shimmer
12:30 - 12:50	A. Van Hirtum (Univ. Grenoble):	Resonance Frequencies in Non-Rigid Compressed Waveguides
12:50 - 13:10	X. Pelorson (Univ. Grenoble):	Physical Model of Phonation with Reduced and Measurable Parameters
13:10 - 13:30	Wrap up & Discussion	
13:30	Closing	
	Farewell Lunch (Mensa, Level 5, closes 14:30)	

Check for program updates:

On the Vowel Development of Bilingual Kindergarten Children in their L2 German

C. Schmid

Department of Pediatrics and Adolescent Medicine, Medical University of Vienna, Vienna, Austria

carolin.schmid@meduniwien.ac.at

Background and objectives: Common practice in speech diagnostics in Austria is monolingual and therefore carries the risk of misdiagnosing multilingual children, who often present the majority in large cities such as Vienna [1]. Moreover the acoustic-phonetic level is rarely investigated in speech diagnostics, and vowels receive little or no attention [2], even though they might differentiate between typically developing children and clinical populations who are described to display characteristic features of vowel articulation (e.g., childhood apraxia of speech [3], autism spectrum disorder [4], or speech disorders [5]). This study aims to give first reference values for vowel productions of typically developing bilingual children as compared to their age-matched monolingual peers. The data should serve as a basis for the early identification of disorders.

Materials and methods: A picture naming test (PLAKSS-II, [6]) will be performed with 60 typically developing children (half bilingual) between 3;0 and 5;11 years of age. Austrian vowel phonemes are segmented in Praat [7] and acoustically analyzed in terms of vowel space size, vowel formant frequencies, precision, and stability. The effect of bilingualism and age group (3;0-4;5 and 4;6-5;11) is statistically investigated with linear models.

Results and discussion: The analysis of the data is still ongoing. Up to date analyses could be finished for 28 children of the older age group. The data of the younger children is already recorded but still needs to be analyzed. In the group of the older children, there are significant differences between monolingual and bilingual children concerning their German vowel productions. Monolingual children show a larger vowel space than bilingual children. Vowel categories are produced more precisely and less variably by monolingual children than by bilingual children. The results also show that some vowel phonemes are not significantly differentiated by bilingual children, concerning either F1 and/or F2.

Conclusions: This study contributes to basic research on the vowel development of monolingual and bilingual preschool children in their surrounding language German. It thus provides the first acoustic reference values for typically developing children and forms a basis for follow-up studies on the pronunciation development of monolingual and bilingual children in Vienna. Further analysis will include the group of younger children and, in a next step, investigate clinical implications by studying whether and in what way children with clinical conditions differ from typically developing children in their vowel production. The preliminary results are in line with a large body of research on cross-language mappings in bilinguals [8], indicating that similar vowel categories of all of the children's languages respectively influence one another. Future perception experiments should investigate whether the differences in vowel production between monolingual and bilingual children indeed result in a lesser intelligibility of bilingual children. A limitation of the present study is, however, that the bilingual children of the study have different

first languages, so that there might be language-dependently different influences on their German vowel production. Additional data should enable an analysis of the first language as a statistical variable.

References:

[1] Statistik Austria. (2022). Bevölkerungsprognosen für Österreich und die Bundesländer. [Online], Available: https://www.statistik.at/statistiken/bevoelkerung-und-soziales/bevoelkerung/demographische-prognosen/bevoelkerungsprognosen-fuer-oesterreich-und-die-bundeslaender
[2] Kent, R. D., & Rountrey, C. (2020). What Acoustic Studies Tell Us About Vowels in Developing and Disordered Speech. American Journal of Speech-Language Pathology, 29(3), 1749–1778. https://doi.org/10.1044/2020_AJSLP-19-00178
[3] Lenoci, G., Celata, C., Ricci, I., Chilosi, A., & Barone, V. (2021). Vowel variability and contrast in Childhood Apraxia of Speech: Acoustics and articulation. Clinical Linguistics & Phonetics, 35(11), 1011–1035. https://doi.org/10.1080/02699206.2020.1853811
[4] Wynn, C. J., Josephson, E. R., & Borrie, S. A. (2022). An Examination of Articulatory Precision in Autistic Children and Adults. Journal of Speech, Language, and Hearing Research, 65(4), 1416–1425. https://doi.org/10.1044/2021_JSLHR-21-00490
[5] Roepke, E., & Brosseau-Lapré, F. (2021). Vowel errors produced by preschool-age children on a single-word test of articulation. Clinical Linguistics & Phonetics, 35(12), 1161–1183.
[6] Fox-Boyer, A. (2014). PLAKSS II. London: PEARSON.
[7] Boersma, P., & Weenink, D. (1992). Praat: Doing phonetics by computer [Computer program]. Available: von https://www.praat.org
[8] Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Hrsg.), Second Language Speech Learning: Theoretical and Empirical Progress (S. 3–83). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108886901.002

Expressive Vocabulary Development in Children with Permanent Hearing Loss: Early Nonverbal Predictors

A.-L. Feichter

Department of Linguistics, Karl-Franzens-University, Graz, Austria
Institut für Sinnes- und Sprachneurologie, Linz, Austria

a.feichter@edu.uni-graz.at

Background and objectives: Children with hearing loss face significant challenges in their language development due to restricted access to auditory language input. Several studies [1, 2, 3] found that, as a result, they experience disadvantages in expressive vocabulary development compared to normal hearing peers. This master's thesis investigates the expressive vocabulary development of Austrian children with permanent hearing loss. The primary aim is to identify early nonverbal risk factors that play a crucial role in the development of expressive vocabulary. Specifically, this research examines nonverbal communication and symbolic behavior at the age of 9 months to determine whether they serve as early predictors of expressive vocabulary development by 27 months. The findings could help clinical linguists and speech therapists in identifying children at risk for delayed vocabulary development at an early stage.

Materials and methods: In this prospective study, collected data between 2020 and 2024 will be analyzed. All data are retrieved from the longitudinal epidemiological database of the AChild study (Austrian Children with Hearing Impairment – Longitudinal Databank) [4]. The participants in this study are monolingual German-speaking children living in Upper or Lower Austria. Children with additional disabilities, an IQ below 71, or unilateral hearing loss will be excluded. Data on the examined predictors will be gathered at the age of 9 months, while outcome data will be drawn from the data set at 27 months. For assessing nonverbal communication, the tests Communication and Symbolic Behavior Scales (CSBS) [5] and Austrian Communicative Development Inventories (ACDI) [6] were selected for this study. Additionally, hearing threshold, auditory performance (Little Ears [7]), hearing aid/CI use, and nonverbal cognitive development (Bayley [8]) will be included as potential predictors. To measure the expressive vocabulary at of 27 months, data will be obtained using the production subtest of the SETK-2 (Sprachentwicklungstest für zweijährige Kinder) [9] for direct assessment and the FRAKIS (Fragebogen zur frühkindlichen Sprachentwicklung) [10] as a proxy measure (parent report).

Results and discussion: The data have been collected, and the next steps involve analyzing and interpreting the results. Currently, we are in the process of preparing the statistical analysis and conducting initial data cleaning and validation. It is expected that the findings will reveal key nonverbal predictors and risk factors. This master's thesis aims to contribute to previous research in early language development [1, 2, 3], which emphasizes the importance of early intervention.

Conclusions: Examining children's development longitudinally is important to identify early risk factors and predictors of language impairment that could be used in tailored intervention planning. However, it is crucial to note the limitations of this study. There are no German normed scores for the ACDI, necessitating the use of the American English normed scores. Additionally, the FRAKIS only provides percentile ranges for statistical analysis. Future research should aim to include a more representative sample, including

multilingual, unilateral hearing impaired children and children with an IQ below 71 or additional disabilities.

References:

[1] P. Carew, D. A. Shepherd, L. Smith, T. Howell, M. Lin, E. L. Bavin, S. Reilly, M. Wake and V. Sung. (2023, July). "Spoken Expressive Vocabulary in 2-Year-Old Children with Hearing Loss: A Community Study." *Children* [Online]. vol. 10, issue 7. Available: https://doi.org/10.3390/children10071223

[2] D. Holzinger, M. Dall, S. Kiblböck, E. Dirks, P. Carew, L. Smith, L. Downie, D. A. Shepherd and V. Sung. (2022, July). "Predictors of Early Language Outcomes in Children with Connexin 26 Hearing Loss across Three Countries." *Children* [Online]. vol. 9, issue 7. Available: https://doi.org/10.3390/children9070990

[3] C. Yoshinaga-Itano, A. L. Sedey, M. Wiggin and W. Chung. (2017, July). "Early Hearing Detection and Vocabulary of Children With Hearing Loss". "Early Hearing Detection and Vocabulary of Children With Hearing Loss," *Pediatrics*, vol. 140, no. 2. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5595069/

[4] M. Dall, S. Kiblböck, D. Müllegger, J. Fellinger, J. Hofer, R. Kapplmüller, S. Breitwieser, K. Schossleitner, C. Weber, R. Zöhrer and D. Holzinger, "Understanding the Impact of Child, Intervention, and Family Factors on Developmental Trajectories of Children with Hearing Loss at Preschool Age: Design of the AChild Study". *Journal of Clinical Medicine* [Online]. vol. 11, issue 6. Available: https://www.mdpi.com/2077-0383/11/6/1508

[5] B. M. Prizant. *CSBS Developmental Profile Forms*; Paul, H., Ed. Baltimore: Brookes Publishing Company, 2002.

[6] P. Marschick, R. Vollmann and C. Einspieler. *ACDI (Austrian Communicative Development Inventory) Aufgaben und Gesten*. Graz: Karl-Franzens-Universität Graz, 2004

[7] MED-EL. *Little Ears Hör-Fragebogen. Elternfragebogen zur Erfassung auditiven Verhaltens.* Innsbruck: MED-EL, 2003.

[8] N. Bayley. *Bayley Scales of Infant and Toddler Development—Deutsche Fassung*, 3rd ed. Frankfurt am Main: Pearson Deutschland GmbH, 2015

[9] H. Grimm. *SETK-2 Sprachentwicklungstest für Zweijährige Kinder*, 2nd ed. Göttingen: Hogrefe, 2016

[10] G. Szagun, B. Stumper and S. A. Schramm, *FRAKIS Fragebogen zur Frühkindlichen Sprachentwicklung*, 2nd ed. Frankfurt am Main: Pearson Deutschland GmbH, 2009.

Phonetic Analysis of Dysarthric Child Speech

M. Galović

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

mgalovic@student.tugraz.at or martina.galovic36@gmail.com

Background and objectives: Dysarthria is a motoric speech disorder that occurs when damages of the nervous system result in difficulties controlling the muscles necessary for speech production. Dysarthria can be developmental or acquired, depending on when the brain damage occurred [1]. This disorder affects articulation, prosody and the overall intelligibility of speech, but is for many patients characterized by consistent speech errors. Despite the consistent errors occurring within a single participant, there are significant differences in speech production within the group of individuals with dysarthria [2]. Two of the most prevalent characteristics are a reduced speaking rate consequently from prolonged phoneme duration, and that speakers with dysarthria often insert a neutral schwa vowel. This phenomenon is the central focus of this preliminary study on the phonetic characteristics of dysarthric speech in children.

Materials and methods: The dataset consists of speech recordings of one primary school-aged boy with dysarthria speaking Austrian German, with a total duration of 70 minutes. The recordings were made either during sessions between the participant and the speech therapist or at home with the parents. The materials include word and picture lists, read stories, quasi-spontaneous speech (sentence creation based on images, storytelling), and spontaneous communication guided by a speech therapist. Some of the audio data is accompanied by the therapist's notes, which include her observations and marked errors, providing valuable insights. In this preliminary study, the focus was on words from the wordlists. Two wordlists of nouns (one with 85 examples and the other with 99 examples) and one verb list (containing 20 examples) were analyzed using the Praat software [3]. The majority of the nouns appeared in both observed lists. The child was asked by the speech therapist to name each word (noun or verb) based on a presented image. Words that proved problematic, specifically those where the insertion of the neutral schwa vowel occurred, were also examined in sentence form whenever possible, provided that a sentence containing the word from wordlist existed in the corpus. In the observed part of the corpus, seven such examples were found.

Results and discussion: In the observed examples (204 tokens), 33 words were found where the participant inserted the schwa vowel: 16 monosyllabic, 13 disyllabic, and 4 trisyllabic. The schwa was mostly inserted between two consonants, creating CəC sequences instead of CC. In fewer cases, one of the surrounding sounds was another vowel. In monosyllabic words, the schwa typically appeared at the beginning, between the initial sounds and before the stressed vowel. Only two exceptions, the words *Mensch (‘human’)* and *fünf (‘five’)*, had schwa inserted between the final two sounds, after /n/. The consonants surrounding the schwa repeated in the examples, with errors occurring when the second consonant in the CC sequence was a sonorant (e.g., /r/, /l/, or /n/) requiring precise motor control, while the first consonant was always an obstruent (e.g., /k/, /g/, /s/, /t/). In disyllabic words, the schwa was mainly inserted at the start of the word, between two consonants and before the stressed vowel. As in monosyllabic words, it also appeared between a vowel and consonant, with the schwa in monosyllabic words occurring in CV sequences with vowel /i/, and in disyllabic words in VC sequences with vowel /a/ or diphthong /ai/. Of the four trisyllabic words, two had schwa added at the beginning between fricative /ʃ/ and nasal consonants (/m/ and /n/), while the other two had schwa inserted in the middle, between a nasal consonant (/n/) and an obstruent. Words that were

problematic in isolation were also mispronounced in sentences, highlighting the consistency of the errors in the speaker with dysarthria.

Conclusions: To explore how the observed irregularities in speech might be applicable in automatic speech recognition systems, their acoustic features need to be studied. Considering that schwa is a vowel, the most important acoustic indicators are its duration and the values of the first and second formants. Variations were observed in the values examined. The duration and formant values of schwa vary depending on the phonetic context of each word, as shown in the table. Several examples were observed where the schwa sound was added between the first two consonants in a word, with the first consonant being an obstruent and the second a lateral approximant /l/. Based on these preliminary data, it appears that the manner of articulation has a greater impact on the acoustic characteristics of schwa (F1 and F2 values), while within the consonant group, voicing affects its duration. These results also suggest that schwa, although a central vowel, does not always have a uniform realization and is not consistently a central vowel. This is evident from the measured values of the first and second formants of schwa in the German language, with the first formant averaging 744 Hz and the second 2248 Hz [4]. Future research will expand the study to other vowels and consonants to provide a more comprehensive quantitative acoustic analysis. The goal is to estimate the likelihood of specific deviation phenomena occurring in certain phonetic contexts. These analyses will not only enhance our understanding of the phonetic characteristics of dysarthric child speech but also contribute to improving automatic speech recognition (ASR) for dysarthric child speech.

Table 1: *Values of the duration and F1 and F2 of schwa in the example of two monosyllabic and two disyllabic words, where schwa occurs between the initial two sounds. The first sound is an obstruent, with some shared and some differing characteristics, while the second sound is a lateral approximant /l/.*

	Clown	Glas	Flasche	Schlüssel
Target	k l ˈ aʊ n	ɡ l ˈ aː s	f l ˈ a ʃ . ə	ʃ l ˈ ʏ s . ə l
Realisation	**k** ə l ˈ aʊ n	**ɡ** ə l ˈ aː s	**f** ə l ˈ a ʃ . ə	**ʃ** ə l ˈ ʏ s . ə l
Manner of articulation	plosive	plosive	fricative	fricative
Place of articulation	velar	velar	labiodental	postalveolar
Voiceness	voiceless	voiced	voiceless	voiceless
Duration of schwa (ms)	112	95	170	192
F1 of schwa (Hz)	899	780	869	868
F2 of schwa (Hz)	1804	1860	1510	1474

References:

[1] Ziegler W, Schölderle T, Aichert I, Staiger A. Motor speech disorders. The Oxford handbook of neurolinguistics. 2019 Mar 14:448-71.
[2] Haas E. Developmental courses of childhood dysarthria (Doctoral dissertation, lmu).
[3] Boersma P. Praat: doing phonetics by computer [Computer program]. http://www.praat.org/. 2011.
[4] Winkler-Kehoe M. An acoustic study of schwa syllables in monolingual and bilingual German-speaking children. Linguistische Berichte. 2019 Oct 1(260).

Multi-Style Child TTS: An Expressive Text-to-Speech System for Children

S. Alwaisi, M. S. Al-Radhi, and G. Németh

Department of Telecommunication and Artificial intelligence, Hungary

shaima.alwaisi@edu.bme.hu, {malradhi,nemeth}@tmit.bme.hu

Background and objectives: This work highlights the significant gap in the development of expressive Text-to-Speech (TTS) systems specifically designed for children, despite the rapid advancements in adult TTS technologies. Children represent a crucial user group for voice technology, yet the scarcity of suitable training datasets poses a major challenge in creating effective child TTS systems. This study aims to address this gap by introducing the Multi-Style Child TTS, a novel expressive TTS model that leverages a carefully curated dataset of child speech to capture their unique speaking styles. The primary objective is to develop a system that not only generates natural and expressive speech but also outperforms existing baseline models in terms of style expressiveness and naturalness, thereby laying a foundation for future research and applications in child speech synthesis.

Materials and methods: The Multi-Style Child TTS system is developed based on the StyleTTS architecture [1]. The model is trained using the ChildTinyTalks (CTT) dataset [2], which comprises approximately two hours of expressive child speech collected from 25 children aged 6 to 11 years. This dataset encompasses four distinct speaking styles: neutral, happy, excited, and sad. The training procedure was conducted over 200 epochs utilizing two NVIDIA A100 PCIe GPUs, with a batch size of 64 and optimization performed using the AdamW algorithm. To effectively model the diverse speaking styles of children, the system incorporates adaptive normalization as a style conditioning mechanism.

Results and discussion: Multi-Style Child TTS system significantly enhances style expressiveness and naturalness compared to baseline models Multi-speaker ChildTTS [3] and CoquiTTS [4]. In MUSHRA test[5], 68% of listeners found the synthesized speech comparable to ground truth, indicating high perceived quality. Objective evaluations further confirm its effectiveness, with lower Mel-cepstral distortion scores across styles—neutral, happy, excited, and sad—compared to Multi-Speaker Child TTS and CoquiTTS. These results demonstrate the model's ability to capture diverse child speech styles, addressing a key gap in expressive TTS systems and paving the way for future advancements.

Conclusions: This study presents the Multi-Style Child TTS, an advanced expressive text-to-speech model designed specifically for children, addressing a critical gap in existing TTS technologies. Built upon the StyleTTS architecture, the model is trained on the ChildTinyTalks (CTT) dataset, which contains approximately two hours of speech from 25 children and encompasses four distinct speaking styles: neutral, happy, excited, and sad. Evaluation results demonstrate that the Multi-Style Child TTS outperforms baseline models in both style expressiveness and naturalness, highlighting its effectiveness in child speech synthesis. These findings not only underscore the model's capacity to enhance the quality of synthesized speech but also establish a strong foundation for future advancements in the field. Future research will focus on further refining the quality of synthesized output to ensure continued improvements in meeting the needs of young users.

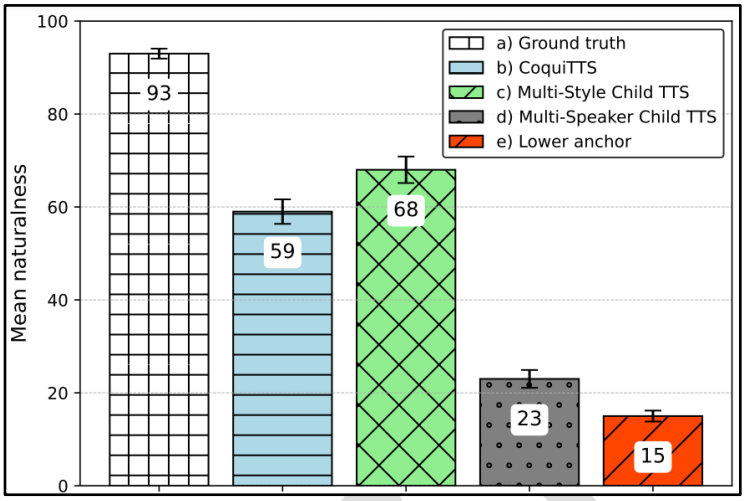


Figure 1: *The MUSHRA scores for the Mean naturalness are presented for (a) Ground truth (b) CoquiTTS (c) Multi-Style ChildTTS (d) Multi-speaker ChildTTS and (e) Lower anchor, with the average results shown. A higher value indicates better overall quality.*

References:

[1] Y. A. Li, C. Han, and N. Mesgarani, "Styletts: A style-based generative model for natural and diverse text-to-speech synthesis," *IEEE J Sel Top Signal Process*, 2025.
[2] S. Alwaisi, M. S. Al-Radhi, and G. Németh, "ChildTinyTalks (CTT): A Benchmark Dataset and Baseline for Expressive Child Speech Synthesis," in *International Conference on Speech and Computer*, Springer, 2024, pp. 230–240.
[3] R. Jain, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis," *IEEE Access*, vol. 10, pp. 47628–47642, 2022, doi: 10.1109/ACCESS.2022.3170836.
[4] G. Eren and The Coqui TTS Team, ", 'Coqui TTS.' ," 2021.
[5] I. Recommendation, "1534-1,'Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA),'" *International Telecommunications Union, Geneva, Switzerland*, vol. 2, 2001.

An Acoustic and Articulatory Study of Voice Quality in Coarticulated Hanoi Vietnamese Tones

Y. Huang

Acoustics Research Institute, Austrian Academy of Science, Vienna, Austria

Yaqian.Huang@oeaw.ac.at

Background and objectives: Hanoi Vietnamese has a six-tone system where tone and phonation are fused, which means each lexical tone is realized with a specification in both pitch and voice quality. It is typologically interesting as four of the six tones, *nặng* B2, *hỏi* C1, *ngã* C2, and *huyền* A2, carry non-modal voice quality including creaky or breathy voice. Previous studies gathered various acoustic and articulatory evidence, supporting the existence of various tone-associated phonation types [1-5]. However, those descriptions appear inconsistent, as multiple combinations of tone and voice quality have been suggested, including the coexistence of different voice qualities within the same tone, especially when the tone has variants of pitch contour in isolation [1,6] or in connected speech [2]. This project presents an in-depth examination of the types of voice qualities and their interactions with pitch in this complex tonal system. Specifically, breathy voice and subtypes of creaky voice such as vocal fry (low f0, glottal constriction), period doubling (alternating cycles, rough quality), and tense voice (non-low f0, constriction) are expected to occur with tones with non-modal quality. We aim first to systemize the phonatory and acoustic properties of the types of voice quality in non-modal Hanoi tones; second, we investigate the variation in pitch and voice quality and their interactions depending on tonal sequences.

Materials and methods: A read speech corpus consisting of full contextual variation of the six Hanoi Vietnamese tones was designed. The stimuli consist of meaningful trisyllabic Vietnamese words such that each tone is flanked by one of the six tones for a full range of 6*6*6=216 combinations. The stimuli were recorded as a standalone word and also embedded in a carrier sentence, namely, “I want to hear STIMULUS one more time”. The experiment includes two repetitions of the 216 words and sentences (432 in total). We recorded 28 native speakers (15 F, 13 M; mean age=22.6) who were born and grew up in or near Hanoi, Vietnam and have not had substantial exposure to a third language besides English. Simultaneous recordings of audio and electroglottography (EGG) were obtained in a sound-treated room in Hanoi.

Results and discussion: The data collection was completed at the end of October 2024. Currently, audio and EGG recordings are being processed. The audio analysis of six Hanoi tones in isolation is completed and used as the baseline for comparison. Figure 1 shows their f0 contours, and Figure 2 illustrates two representative acoustic measures of voice quality: spectral tilt (H1-H2) and harmonics-to-noise ratio <500 Hz (HNR05), measured using PraatSauce [7]. The distinct pitch contours are consistent with previous studies (e.g., [3]) such that *ngang* A1 is high-level, *huyền* A2 is mid-falling, *sắc* B1 is low-rising, *nặng* B2 is mid-low and has the shortest duration, *hỏi* C1 is low-falling, and *ngã* C2 is mid-rising. However, the beginning of C2 seems to start lower (below 150 Hz) while the documented C2 started around 175 Hz in Kirby (2011). Figure 1 shows that A1, A2, and B1 are more modal and less variable than the other tones given their intermediate ranges of H1-H2 and positive HNR. The acoustic measures of C1 and C2 are variable. B2 has a low H1-H2, indicating a larger extent of glottal constriction. EGG data of the six tones are currently being processed. For the middle target tones in trisyllabic tonal sequences and in carrier phrase,

the Montreal forced aligner was used to segment the recordings and manual correction is currently underway. Next steps are to analyze pitch and voice quality correlates of the target tones.

Conclusions: Investigating tonal and phonational variations in Vietnamese tones contributes to our understanding of the interactions between pitch and voice quality in tone systems. It sheds new insights into typology of tone and phonation and expands the work of voice quality classification.

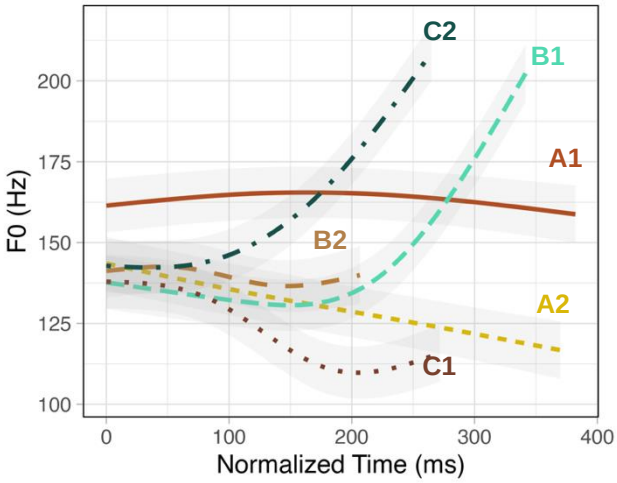


Figure 1: *f0 tracks of six Hanoi Vietnamese tones. (A1=ngang, A2=huyền, B1=sắc, B2=nặng, C1=hỏi, C2=ngã) The values are estimated by natural cubic spline regressions, with lines showing estimated means, and shaded areas showing one standard error above and below the estimated means. Time is normalized across speakers and tokens and plotted based on the mean duration of each tone.*

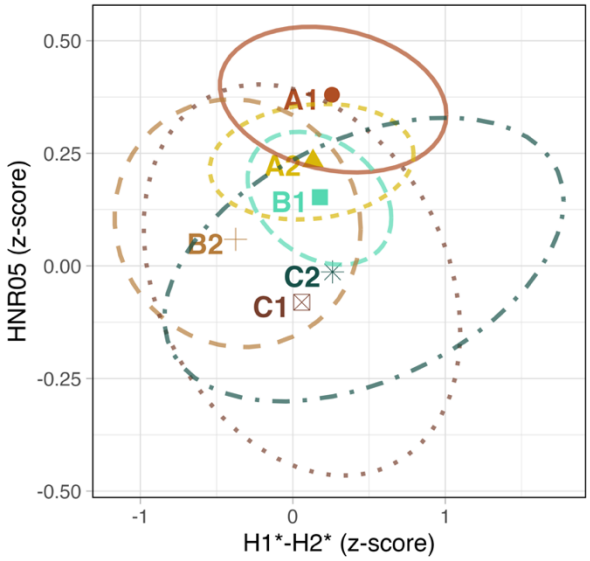


Figure 2: *H1-H2 (formant corrected) and HNR <500Hz of six Hanoi Vietnamese tones. The colored ellipses represent 95% confidence intervals around the mean of each tone following a multivariate t-distribution. The respective means are shown by different shapes and labels. The bottom-left corner represents a creakier voice (more constricted and noisier), and the upper-right corner represents a more modal voice (less constricted and less noisy).*

References:

[1] Edmondson, J., & Lợi, N. V. (1997). Tones and voice quality in modern northern Vietnamese: instrumental case studies.". *Mon-Khmer Studies, 28*(35), 1-18.
[2] Brunelle, M., Nguyễn, D. D., & Nguyễn, K. H. (2010). A laryngographic and laryngoscopic study of Northern Vietnamese tones. *Phonetica, 67*(3), 147-169.
[3] Kirby, J. P. (2011). Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association, 41*(3), 381-392.
[4] Michaud, A. (2004). Final consonants and glottalization: new perspectives from Hanoi Vietnamese. *Phonetica, 61*(2-3), 119-146.
[5] Pham, A. H. (2004). *Vietnamese tone: A new analysis*. Routledge.
[6] Blodgett, A., Fox, M. K., Rytting, C. A., & Twist, A. (2010). Non-Contrastive Voice Quality Characteristics of Northern Vietnamese Tones. In Speech Prosody 2010-Fifth International Conference.
[7] Kirby, J. P. (2018). praatsauce: Praat-based tools for spectral analysis. v.0.2.4, 2018.

Does Spatial Auditory Attention Change After Having a Face-To-Face Conversation in a Noisy Environment?

Ľ. Hládek, P. Majdak & R. Baumgartner

Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

lubos.hladek@oeaw.ac.at

Background and objectives: Engaging in a conversation in a noisy, reverberant room may lead to temporary cognitive fatigue, negatively affecting cognitive functions such as attention. Under the assumption of limited cognitive resources and a certain recovery period needed to restore those [1], the resources allocated to deal with the complex acoustic setting reduce the resources available for paying attention to the conversation partner, especially after extended periods of conversation time. The present experiment tests this hypothesis by performing an auditory spatial selective attention task after having a face-to-face conversation in a reverberant environment with high level of background noise and many competing talkers.

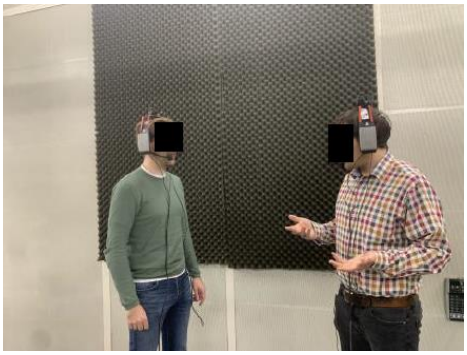


Figure 1: *Two people talk in a virtual acoustic environment spatialized using super open headphones (AKG K1000) with motion tracking and head-set microphones. The mixed virtual reality was created using real-time acoustic simulation (rtSOFE [2]) of an underground station [3] with interfering noise and talkers.*

Materials and methods: Participants with normal hearing and high proficiency in German language had a conversation with another fellow participant on a free small-talk topic. The conversation took place either in quiet or in a virtual acoustic environment delivered over headphones with motion tracking, which accounted for the movements of both conversation partners. In the third condition, participants passively listened to the acoustic environment. The environment was based on a real-time acoustic simulation (rtSOFE [2]) of an underground station [3] with additional background noise and interfering talkers who produced conversational speech. Before and after the conversation, or the passive listening, the participants were tested on their auditory spatial selective attention control [4]. In the test, participants heard two streams of syllables (taken from the set: 'ba', 'da', 'ga') presented simultaneously from their left- and right-hand side (±30° azimuth). The to-be-attended target stream was indicated by an auditory spatial cue 800 ms before the onset. The test had four blocks of 40 trials each.

Results: As data collection is still underway, results are preliminary at this stage. These suggest that the ability to pay attention in the spatial auditory streaming task was temporarily reduced when participants had a face-to-face conversation in the noisy environment, but not when the conversation took place in the quiet environment or when they passively listened to the noisy scene.

Discussion: We have developed a paradigm to effectively test the effect of acoustic conditions on cognitive functions in a realistic and interactive communication scenario. The preliminary results are consistent with

the limited capacity model of attention [4]. Yet, the streaming task tested here focuses solely on spatial attention, leaving other aspects of attention unexplored.

Acknowledgements: Lubos Hladek is a recipient of the Seal of Excellence of the Austrian Academy of Sciences.

References:

[1] M. K. Pichora-Fuller *et al.*, "Hearing impairment and cognitive energy: The framework for understanding effortful listening (FUEL)," *Ear and Hearing*, vol. 37, no. 1, pp. 5S-27S, Jul. 2016, doi: 10.1097/AUD.0000000000000312.
[2] B. U. Seeber and T. Wang, real-time Simulated Open Field Environment (rtSOFE) software package. (2021). Zenodo.
[3] S. van de Par *et al.*, "Auditory-visual scenes for hearing research," *Acta Acustica*, vol. 6, p. 55, Nov. 2022, doi: 10.1051/aacus/2022032.
[4] Y. Deng, I. Choi, B. Shinn-Cunningham, and R. Baumgartner, "Impoverished auditory cues limit engagement of brain networks controlling spatial selective attention," *NeuroImage*, vol. 202, p. 116151, Nov. 2019, doi: 10.1016/j.neuroimage.2019.116151.

(When) Does it Harm to Be Incomplete? Encoding ASR Mistranscriptions of Syntactically Disfluent Structures

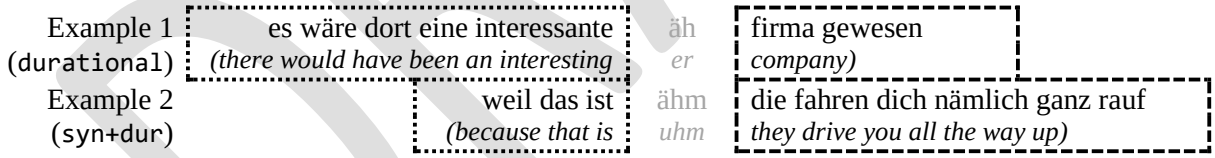
S. Wepner & B. Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

wepner@tugraz.at

Background and objectives: Speech from highly spontaneous conversations is characterised by the frequent occurrence of disfluencies. Despite occurring every few turns in conversation [1], disfluent structures seem underrepresented in state-of-the-art automatic speech recognition (ASR) systems [2], suggesting that "in-real-life WERs [word error rates] are much higher than reported" [3, p. 3290]. The WER is the established measure to evaluate transcription performance of ASR systems. It is simple, but also rough and does not provide (much) insight into where transcription went wrong. We present a visualisation technique supporting the interplay of qualitative and quantitative analysis by encoding word-level features.

Materials and methods: We compare transcriptions by multiple fine-tuned [4] and zero-shot ASR systems, six based on Whisper [5] and three on wav2vec [6]. We extracted utterances from spontaneous conversations in Austrian German [7]. All utterances have in common that they show a disfluency in the middle of a syntactic structure, that is, before reaching a transition relevance place (TRP) [8]. We divided the utterances into different *disfluency types*. Utterances of disfluency type `durational` contain a durational disfluency (i.e., a filled pause) but no syntactic disfluency, meaning that the same syntactic structure that the speaker had started before the pause was continued after the pause (Example 1). Utterances of disfluency type `syn+dur` contain both a durational and a syntactic disfluency, i.e., the speaker interrupted a syntactic structure with a (filled) pause and started a new syntactic structure after the pause (Example 2).



Each utterance was fed into all ASR systems in different *conditions*: Either only the part before the disfluency (`pre-disfluency`, in a dotted frame above), or the part after the disfluency (`post-disfluency`, dashed above), or the whole utterance was processed (condition `whole`). For each utterance and condition, we determined the WERs (for all systems) from which we derived for each word, whether it was transcribed correctly or not. We counted how many ASR systems mistranscribed this word and normalised this number by the number of total transcriptions, yielding the percentage of mistranscription. Then, we mapped every word in an utterance to an integer number, centred around the disfluency. Finally, we subtracted the percentage of mistranscription in `pre-/post-disfluency` from the percentage in condition `whole`.

Results and discussion: Figure 1 shows a visualisation of this difference of the percentage of mistranscription. Words from utterances of type `durational` were generally transcribed better (indicated by more green-bordered circles) than those of `syn+dur`. In `durational`, ASR mostly profited from longer utterances, indicated by blue shaded circles. In `syn+dur`, the picture was mixed: While in some utterances, the context supported correct transcription, in others, it led to more mistranscriptions, indicated by red

shaded circles. Black circles hint to words whose transcription seemed to be generally difficult, regardless of the condition.

Conclusions: The presented visualisation technique allows us to recognise relationships between the words within a stimulus as well as beyond it, that is, by comparing with other stimuli. It allows to spot details on the word or utterance level that are not easy to find, for instance, as outliers that are defined by exceeding the threshold of a certain variable. Of course, other features can also be encoded in this way.

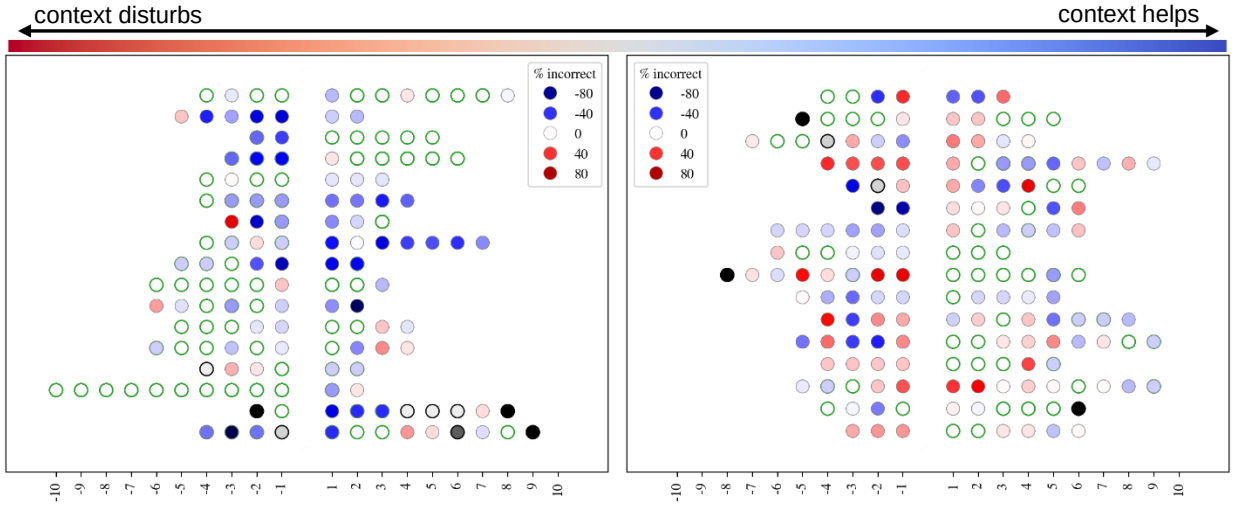


Figure 1: *Difference of the percentage of mistranscriptions for each individual word in condition* ***`whole`*** *vs.* ***`pre-/post-disfluency`****. Each row represents one utterance of the disfluency types* ***`durational`*** *(left) and* ***`syn+dur`*** *(right), each circle represents one word. An empty, green-bordered circle means that this word was never mistranscribed by any of the ASR systems; a black-bordered circle means that this word was equally often mistranscribed in the different conditions (difference zero).*

Acknowledgements: This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P32700].

References:

[1] Ginzburg J, Fernández RM, and David S. "Disfluencies as Intra-Utterance Dialogue Moves." *Semantics & Pragmatics*, 7(9):1–64, 2014.
[2] Riviere M, Copet J, and Synnaeve G. "ASR4REAL: An Extended Benchmark for Speech Models." *arXiv:2110.08583*, 2021.
[3] Szymański P, Żelasko P, Morzy M, Szymczak A, Żyła-Hoppe M, Banaszczak J, Augustyniak L, Mizgajski J, and Carmiel Y. "WER We Are and WER We Think We Are." In *Proc. EMNLP*, 2020.
[4] Linke J, Geiger BC, Kubin G, Schuppler B. "What's so Complex About Conversational Speech? A Comparison of HMM-Based and Transformer-Based ASR Architectures." *Computer Speech & Language*, 2025.
[5] Radford A, Kim J. W, Xu T, Brockman G, McLeavey C, and Sutskever I. Robust. "Speech Recognition via Large-Scale Weak Supervision." In *Proc. ICML*, pages 28492–28518. PMLR, 2023.
[6] Facebook Research. "Fairseq Model (XLSR)." *GitHub Repository*. Available: https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec
[7] Schuppler B, Hagmüller M, and Zahrer A. "A Corpus of Read and Conversational Austrian German." *Speech Communication*, 94:62–74, 2017.
[8] Selting M. "On the Interplay of Syntax and Prosody in the Constitution of Turn-Constructional Units and Turns in Conversation." *Journal of Pragmatics*, 6(3):371–388, 1996.

ASR of Non-lexical and Disfluent Events - an Investigation Across Tasks

Y. Meng[1], P. Mihajlik[1,2], D. Mengke[1] & K. Mády[2]

1) Budapest University of Technology and Economics, Budapest, Hungary
2) HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary

yan.meng@edu.bme.hu

Background and objectives: Spontaneous speech contains a significant number of disfluencies and non-lexical sounds (e.g., backchannels, filled pauses), which are often difficult to transcribe. Disfluency labeling for automatic speech recognition (ASR) aims at editing these phenomena in the transcription to improve overall recognition accuracy. Such labeling techniques typically delete non-lexical/disfluent labels from the prediction, where classical ASR techniques either ignore or treat them as lexical items. In this paper we revisit disfluency labeling techniques addressing the issue of end-to-end ASR of spontaneous speech containing numerous non-lexical and disfluent acoustic events, such as backchannels, hesitation or broken words. We show that in three remarkably different Hungarian conversational speech datasets clear overall error rate improvements can be made. Unlike most of the other disfluency labeling approaches we do not remove any of the disfluency/non-lexical symbols from the ASR output, so that they can be applied for various diagnostic purposes.

Materials and methods: In speech recognition tasks, accurate labeling is crucial for model training and performance optimization, especially when dealing with spontaneous speech. This paper uses Disfluency Labeling, Hesitation Labeling and Backchannel Labeling, three methods to label the disfluency and non-lexical parts of spontaneous speech. Based on the calculation of the overall error rate, this paper introduces a more detailed evaluation system. This system further calculates the error rate at the symbol and text level by calculating the number of insertion, deletion, and substitution errors between the predicted text and the real text, and summarizing them. This method can accurately quantify the error rate of a single unit in the transcribed text, which helps to deeply analyze the impact of different units on the overall error rate.

Results and discussion: In the evaluation phase, we used the Detailed Evaluation System method to conduct a comprehensive analysis of the transcripts of the six experimental and baseline solutions. For three datasets, the error rate of regular word is relatively low in HBDL+ experiments. Especially on the ForVoice evaluation set, the experimental results are the best, with a normal word error rate reduced by 8.76% compared to the baseline. However, for broken words, the error rate remains at a high level. By replacing broken words with the @ symbol, a significant improvement was achieved on the BEA-Base V2 dataset, but the performance on the other two datasets was poor. For hesitation, the transcription of the three datasets has achieved a low error rate. However, when distinguishing oral hesitation from nasal hesitation, the error rate of nasal hesitation is high, which is due to its low frequency of occurrence in the dataset. As for backchannel, it appears less frequently in the BEA-Base V2 and DE datasets, and the transcription effect is not ideal. On the ForVoice dataset, the evaluation set of the HBL experiment achieved the lowest error rate, which is only 11.56%.

Conclusions: In this paper we reconsidered various disfluency (and non-lexical) labeling techniques for end-to-end ASR of three remarkably different conversational speech datasets. Although the ratio of hesitations, broken words and backchannels were significantly different, clear improvements could be made in all types of conversational speech data. Unlike most of the other disfluency labeling approaches we do not remove any of the disfluency/non-lexical symbols from the ASR output, so that they can be used for various diagnostic purposes. The results suggest that the recognition accuracy of the special spontaneous audio events depends highly on their number of occurrences in the training (and test) sets, but otherwise they do not pose an impossible challenge even in low-resource scenarios. As for future work, we plan to explore generating synthetic speech data with disfluencies and so make their recognition more robust.

Model	Type	BEA-Base V2		ForVoice		DE	
		dev	eval	dev	eval	dev	eval
BL	Total	13.22	14.24	16.69	17.12	29.18	29.27
	Regular words	12.15	13.44	16.53	17.01	27.85	27.80
	Broken words	71.65	80.23	95.89	91.92	117.68	141.46
HL	Total	13.14	13.70	16.40	16.89	26.38	28.63
	Regular words	11.24	12.26	16.16	16.63	24.73	26.93
	Broken words	75.10	80.46	95.89	136.36	121.95	150.87
	# (hesitations)	30.59	30.64	73.00	45.70	29.98	29.97
HBL	Total	13.09	13.79	16.18	16.46	**21.75**	**21.52**
	Regular words	11.09	12.00	15.94	16.32	**20.47**	**20.18**
	Broken words	73.95	77.24	**93.15**	**89.90**	**96.34**	**107.32**
	# (hesitations)	30.88	32.24	41.41	40.86	24.03	25.94
	% (backchannels)	113.79	105.74	**14.57**	**11.56**	60.00	44.44
HBDL	Total	12.82	13.76	16.54	15.97	23.86	24.42
	Regular words	11.33	12.40	16.34	15.76	22.50	22.96
	@ (broken words)	52.44	61.02	95.89	90.91	120.12	137.63
	# (hesitations)	31.32	29.68	36.72	48.39	20.14	20.89
	% (backchannels)	113.79	100.00	11.74	16.57	40.00	55.56
HL+	Total	14.25	14.88	16.21	16.15	23.71	24.85
	Regular words	12.07	13.07	15.97	15.96	21.95	23.33
	Broken words	75.29	86.90	87.67	91.92	116.46	126.13
	# (oral hesitations)	25.91	24.32	43.31	39.89	19.42	19.39
	& (nasal hesitations)	73.77	130.27	100.00	75.00	75.00	60.71
HBL+	Total	13.14	13.48	16.57	16.37	23.60	25.82
	Regular words	**11.07**	**11.64**	16.36	16.18	21.84	23.86
	Broken words	70.50	82.07	86.67	91.92	106.71	128.92
	# (oral hesitations)	26.21	21.50	**37.80**	**44.38**	20.00	22.05
	& (nasal hesitations)	**56.11**	**92.97**	100.00	75.00	89.13	62.50
	% (backchannels)	110.34	100.82	14.98	15.41	20.00	51.85
HBDL+	Total	**12.90**	**13.33**	**16.56**	**15.83**	22.52	23.94
	Regular words	11.37	11.81	**16.29**	**15.62**	20.86	22.43
	@ (broken words)	**51.74**	**61.02**	95.89	92.93	111.59	124.04
	# (oral hesitations)	**23.67**	**20.85**	40.94	46.07	**19.13**	**19.58**
	& (nasal hesitations)	58.09	103.78	**100.00**	**62.50**	**65.22**	**52.98**
	% (backchannels)	**100.00**	**103.28**	18.62	16.18	**20.00**	**44.44**

Table 1: *Error Rate (%) Results of Different Labeling Methods Across Three Datasets*

References:

[1] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoga, “End-to-end spontaneous speech recognition using disfluency labeling,” in Proc. Interspeech 2022, 2022, pp. 4108–4112.
[2] P. Mihajlik, Y. Meng, M. S. Kadar, J. Linke, B. Schuppler, and K. M´ady, “On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition,” in Proceedings of Inter-speech 2024, 2024, pp. 1270–1274.
[3] P. Mihajlik, A. Balog, T. E. Graczi, A. Kohari, B. Tarj´an, and K. Mady, “BEA-base: A benchmark for ASR of spontaneous Hungarian,” in Proceedings of the Thirteenth Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, Jun. 2022, pp. 1970–1977.[Online]. Available: https://aclanthology.org/2022.lrec-1.211
[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, “wav2vec2.0: A framework for self-supervised learning of speech repre-sentations,” Advances in neural information processing systems, vol. 33, pp. 12 449–12 460, 2020.
[5] K. Horii, M. Fukuda, K. Ohta, R. Nishimura, A. Ogawa, and N. Kitaoga, “End-to-end spontaneous speech recognition using disfluency labeling,” in Proc. Interspeech 2022, 2022, pp. 4108–4112.

Speech Enhancement of Conversational Speech in Cocktail Party Noise

E. Berger[1], B. Schuppler[1], M. Hagmüller[1], F. Pernkopf[1]

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

emil.berger@tugraz.at

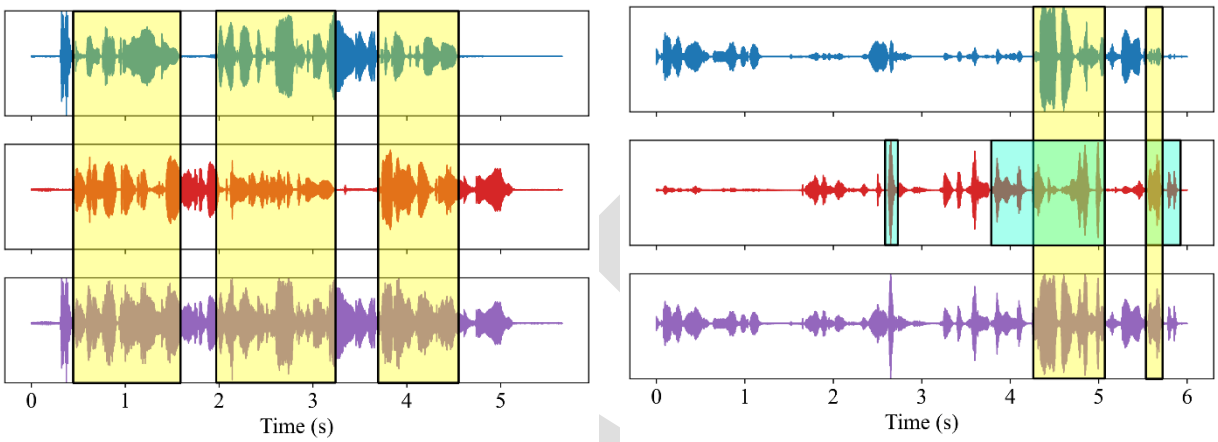


Figure 1: *Waveforms of an example of LibriMix [1] (left) and GRASS StudyFair (right): speaker 1 (upper, blue), speaker 2 (middle, red), mixture (lower, purple). The yellow boxes show segments of overlapping speech, the teal boxes show laughter and laughed speech of speaker 2 [2].*

Background and objectives: The advent of transformers has revolutionised the machine learning landscape, inspiring innovative approaches to a wide range of challenges, including the cocktail party problem. As these models continue to grow in size, their training increasingly relies on massive datasets—often generated by artificially blending read speech with recorded ambient noise. This approach, however, falls short in capturing the intricacies of real-life conversational dynamics. In this study, we introduce a novel speech corpus and evaluate the performance of three distinct architectures for single-channel speech separation of two sources.

Materials and methods: Our experimental framework is built upon the LibriMix [1] dataset, which serves as the pre-training foundation for the models under investigation. To better reflect authentic dialogue, we recorded the GRASS StudyFair corpus (not yet published), featuring 4 hours and 30 minutes of task-oriented, two-person conversations set within a simulated fair environment. The architectures examined include the state-of-the-art transformer-based SepFormer [3], the more lightweight, self-attentive Sandglasset [4], and, for comparative purposes, the non-transformer-based Conv-TasNet [5]. We assessed the models at 8 and 16 kHz sampling rates, examining mixed and same-gender mixtures separately, and validated the results through a comprehensive four-fold cross-validation.

Results and discussion: We measured the scale-invariant signal-to-distortion ratio improvement (SI-SDRi) and observed a peak performance of 13.21 dB for the 16 kHz variant of the SepFormer [3] on female/male mixtures. Overall, the models yielded SI-SDRi values between 4 dB and 8.5 dB, with female/female mixtures consistently outperforming male/male ones. While Conv-TasNet [5] showed slightly lower overall

performance compared to SepFormer [3], it excelled on same-gender mixtures, and Sandglasset [4] recorded the lowest results in every scenario.

Conclusions: This study underscores the persistent challenges of source separation. While current architectures perform impressively under controlled "laboratory conditions" with artificially mixed sources, they fall short when applied to real-life conversations—where hesitations, laughter, and extended pauses (cf. Figure 1) can cause the model to lose its target. Notably, the full potential of the SepFormer [3] remains untapped, possibly due to the limited data available for fine-tuning. Looking ahead, we plan to harness the extensive GRASS corpus [6], which offers longer conversational samples, to further fine-tune and evaluate our models. Moreover, the difficulty in generating clean target signals—a common hurdle in this field—motivates us to benchmark unsupervised approaches, such as those introduced by Wisdom et al. [7].

Acknowledgements: This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P32700].

References:

[1] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge and E. Vincent, "LibriMix: An open-source dataset for generalizable speech separation", arXiv preprint, arXiv:2005.11262, 2020.
[2] E. Berger, B. Schuppler, F. Pernkopf, & M. Hagmueller: "Single Channel Source Separation in the Wild–Conversational Speech in Realistic Environments", Speech Communication; 15th ITG Conference (pp. 96-100). VDE, 2023.
[3] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, & J. Zhong, "Attention is all you need in speech separation", ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25), 2021.
[4] M. W. Lam, J. Wang, D. Su, & D. Yu, "Sandglasset: A light multi-granularity self-attentive network for time-domain speech separation", ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5759-5763), 2021.
[5] Y. Luo, & N. Mesgarani, Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation. IEEE/ACM transactions on audio, speech, and language processing, 27(8), 1256-1266, 2019.
[6] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, & H. Pessentheiner: "GRASS: the Graz corpus of Read And Spontaneous Speech", In Lrec (pp. 1465-1470), 2014.
[7] S. Wisdom, E. Tzinis, H. Erdogan, R. Weiss, K. Wilson, & J. Hershey: "Unsupervised sound separation using mixture invariant training", Advances in neural information processing systems, 33, 3846-3857, 2020.

Building a Robust HMI Command Inventory for Various Noisy Environments

S. M. Pearsell & O. Niebuhr

Institute of Mechanical and Electrical Engineering,
University of Southern Denmark, Sønderborg, Denmark

pearsell@sdu.dk

Background and objectives: Human-Machine Interaction (HMI) in industrial environments could benefit from modern artificial intelligence (AI) advancements, particularly from the integration of voice-based commands. These commands would enhance worker mobility and efficiency. However, with excessive acoustic noise, systems like automatic speech recognition (ASR), have significant issues with performance accuracy. Thus, the necessity to develop an acoustic robust command inventory arises. As part of the EU-funded Arrowhead project, which aims to develop digitization and automation tools for European engineering applications, we hope to address this gap. Through the integration of linguistics and speech signal processing, we have designed a command list for use in industrial environments, allowing for commands to remain intelligible to speech recognition systems, like ASR, in highly noisy environments. The following is the introductory phase of this testing and includes examining ASR performance in industry noise as well the meaning matching (linguistic symbolism) of our commands to real words.

Materials and methods: Our initial studies, which are the focus here, tested ASR recognition in three noisy environment conditions and linguistic symbolism. Our initial command inventory was developed based on acoustic and linguistic aspects such as cross-linguistic accessibility, acoustic distinctiveness, and ease of pronounceability. To ensure the command inventory met all of these requirements, the commands created were *nonwords*. This means the words could exist but do not. The use of nonword commands also avoids any problems which may arise with the semantics and acoustic restrictions associated with real words.

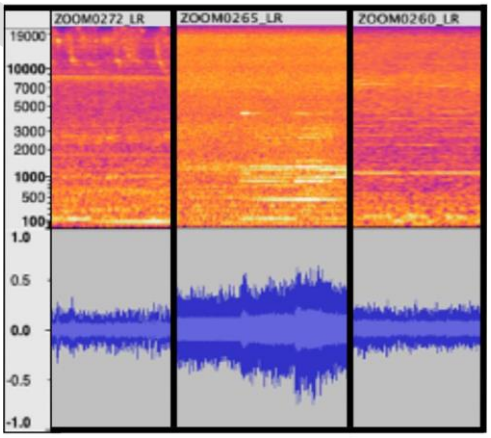


Figure 1: *Spectrograms of the three industrial noises stimuli. From left to right, the industrial noises consist of a grinding machine (ZOOM0272), cutting and polishing machine (ZOOM0265), milling and grinding machine (ZOOM0260).*

To evaluate the robustness of ASR systems in industrial noise environments, we tested Apple Dictation and Google Translate under three noise conditions: white noise, speech shaped noise, and industry noise (using three acoustically, and spectrally distinct industry noises; see Figure 1). We opted for white and speech shaped noise to serve as a baseline for the experiment; white noise containing equal acoustic energy across all frequencies, and speech shaped noise aligning with spectral characteristics of speech. Since ASR systems use real words, we created a small list of real words which were alike acoustically and phonetically to our initial nonword stimuli. Each ASR system was presented with noise at increasing levels of loudness levels alongside word stimuli with each word tested three times to ensure accuracy (scored from 0 – 3).

To understand the learnability of our designed nonwords, we considered the linguistic symbolism of the nonword commands. Essentially, this means understanding how/if listeners associated the nonwords to real words – in this case command words frequently utilized in industrial settings specifically on manual interactions. These commands would entail words like *stop, start, increase speed,* etc. Participants were from ten different language backgrounds (ten per language) resulting in 100 participants total. Audio stimuli of our nonword list were presented for participants to play and participants were asked assign each of the (real word) commands (e.g., *stop, start*) to the nonword they think most closely matched with the command. Each command was only matched once resulting in 17 nonwords matched to 17 commands.

Results and discussion: For our ASR testing, we used a general linear mixed model to examine the effects of noise type, noise level, speaker gender, and ASR system type on the word recognition accuracy. This model included two within-subject factors (noise level and ASR system) and two between-subject factors (speaker gender and noise type). We found significant main effects for noise level, and ASR system. The accuracy of word recognition systematically decreased as noise levels increased. Apple Dictation consistently outperformed Google Translate across most noise conditions (except for speech shaped noise). There was also a significant interaction between ASR system and noise type. Most importantly, Apple Dictation performed better in both white noise and more generally for the industrial noise conditions.

For the linguistic symbolism experiment, the data still needs to be analyzed. From initial inspection, we see several trends. Firstly, certain nonwords have stronger associations with specific commands. An example of this is the nonword *rifad* (/ɹifad/) which was frequently associated with the command word *right*. We also found that some commands (e.g., *hold* or *increase speed*) had scores spread across multiple nonwords meaning no single nonword was perceived as a natural match with the command word. Lastly, we see some commands had higher responses (e.g. *stop, right, rotate, start*) for specific nonwords while others had lower responses (e.g., *cancel, override, reset*) meaning that curtaining in association was lower and more varied.

Conclusions: For the ASR experiment, our results suggest systems like Apple Dictation provide a robustness against spectral noisy industrial and broadband noise suggesting stronger spectral contrast and parsing abilities. For environments like industry, where significant machine noise is present, Apple Dictation provides preferable results. Future iterations of this study will include several addition ASR systems to test a wider array of systems available for use, hopefully providing further insight into the functionality of ASR systems for speech recognition in noisy environments like industrial settings. For our linguistic symbolism experiment, although statistical analysis needs to be performed, we can see there does appear to be some relationships between our proposed nonwords and real word commands. Some nonwords exhibit stronger associations with specific real word providing insight into which nonwords may offer more intuitive usage and memorization for users (i.e., workers) assisting in the ease of learnability.

Through combining the results of these two experiments, as well as other intended future research, including the testing of several other ASR systems, and real-world simulation of noise environments and speech recognition, we can develop a clearer picture of which nonwords to include/exclude from our proposed command inventory. Inclusion of acoustically robust words optimizes speech recognition in technology but also eases learnability and accessibility across various language backgrounds permitting an intuitive nonword inventory. Essentially, the consideration of both optimal word recognition for machines and ease of use for users creates a command list robust enough for machine recognition but easily accessible for humans.

Comparison of Self-Training Strategies for ASR

Y. Luo & P. Mihajlik

Department of Telecommunications and Artificial Intelligence,
Budapest University of Technology and Economics, Hungary

luo.yue@edu.bme.hu

Background and objectives: Automatic Speech Recognition (ASR) has made significant advancements, particularly with the rise of end-to-end ASR models. However, a major challenge still remains: end-to-end ASR models require large amounts of labeled data for training [1]. Obtaining labeled data is expensive and time-consuming, and the difficulty is particularly obvious for low-resource languages like Hungarian. To address this limitation, semi-supervised learning techniques become promising approaches. As a simple and effective method in semi-supervised learning, self-training enables ASR models to leverage vast amounts of unlabeled data by generating pseudo-labels using a pre-trained model and then retraining on these labels [2]. In this study, we assess self-training for Hungarian and Mandarin ASR and compare different self-training strategies for ASR, evaluating their impact on model performance across different test sets. Our research investigates: (1) The effectiveness of self-training with different pseudo-labeling strategies. (2) The influence of different pre-trained ASR models on the quality of generated pseudo-labels. (3) The performance of strategies on datasets of different natures.

Materials and methods: Our experiments used Hungarian and Mandarin ASR datasets. For Hungarian, BEA-Base supervised the seed model training, while BEA-Wavs served as the unlabeled dataset for self-training. For Mandarin, AISHELL-1 provided labeled data, and AISHELL-2 supplied labeled data for supervised learning and unlabeled data for self-training. Test sets and other dataset details are in Table 1.

Table 1: *The statistics of datasets used in experiment.*

Language	Dataset	Abbr.	Train	Dev	Test	Speaker	Use Subsets
Hungarian	BEA-Base	BEA-b	71.2h	4.02h	4.91h	140	train,dev,test
	BEA-Wavs	BEA-w	373.2h	-	-	500	train
	Common Voice-17.0-hu	CV-hu	53.35h	16.68h	17.73h	1614	test
Mandarin	AISHELL-1	AI-1	150h	18h	10h	400	train,dev
	AISHELL-2	AI-2	1000h	-	-	1991	train
	AISHELL-2018A-EVAL	AI-eval	-	2.03h	3.54h	15	test
	Common Voice-17.0-zh	CV-zh	42.34h	15.92h	17.45h	3333	test

Two main self-training strategies were implemented: (1) Pseudo-labels and labeled data (PLL): The model was first trained on the labeled dataset, then used to generate pseudo-labels for the unlabeled speech data. The combined labeled and pseudo-labeled dataset was used to retrain the model. (2) Pseudo-labels only (PL): After generating pseudo-labels from the seed model, the ASR model was trained singly on the pseudo-labeled dataset, without incorporating manually labeled data. To assess the influence of seed models, we compared supervised learning (SL) models trained with limited label data against those using pre-trained models, such as faster Whisper large-V2 [3]. Experiments were conducted using the WeNet [4] ASR toolkit (v2.2.0), leveraging Conformer encoder (12 blocks) and Transformer decoder (6 blocks) with CTC/Attention hybrid model. Evaluations were performed using Word Error Rate (WER) for Hungarian and Character Error Rate (CER) for Mandarin as the primary metric.

Results and discussion: The results detailed in Tables 2 compared two self-training strategies on Hungarian and Mandarin ASR systems. The Hungarian baseline, trained on the spontaneous BEA-Base dataset with fewer than 44M parameters, achieved a WER of 20.13%, outperforming a larger 133M parameter model from benchmark [5]. Notably, the best Hungarian model used PLL, reducing the WER to 15.93%, performance comparable to state-of-the-art methods leveraging massive hours of multilingual data in study [5]. However, models trained solely with pseudo-labels performed poorly regardless of whether the baseline or Whisper was used as the seed. In contrast, the Mandarin baseline on AISHELL-1 achieved a CER of 13.00%. Both self-training strategies markedly improved performance, with the PL and PLL approaches achieving CERs of 7.03% and 7.11%. Remarkably, the model trained only on pseudo-labels nearly matched the performance of models built on 1000h AISHELL-2 labeled datasets.

Table 2: *Error rates of models with different machine learning methods and training dataset.*

Language	Method	Training dataset	seed model	Bea-eval	CV-hu
Hungarian (WER %)	SL	Labeled BEA-b **(As Baseline)**	-	20.13	37.95
	PL	Unlabeled BEA-w	Baseline	31.93	63.04
	PLL	Labeled BEA-b+Unlabeled BEA-w	Baseline	15.93	32.16
	Inference	-	Whisper	24.32	20.15
	PL	Unlabeled BEA-w	Whisper	31.35	56.69
	PLL	Labeled BEA-b+Unlabeled BEA-w	Whisper	19.10	39.00
				Bea-eval	**CV-hu**
Mandarin (CER %)	SL	Labeled AI-1 **(As Baseline)**	-	13.00	35.87
	SL	Labeled AI-2	-	6.36	27.15
	PL	Unlabeled AI-2	Baseline	10.39	29.33
	PLL	Labeled AI-1+Unlabeled AI-2	Baseline	9.82	27.68
	Inference	-	Whisper	5.33	12.83
	PL	Unlabeled AI-2	Whisper	7.03	24.51
	PLL	Labeled AI-1 +Unlabeled AI-2	Whisper	7.11	24.24

Conclusions: Our study demonstrates that self-training significantly improves ASR accuracy for both Hungarian and Mandarin, especially when pseudo-labels are of high quality. However, limitations such as domain mismatch and pseudo-label noise−particularly in spontaneous Hungarian speech−remain critical challenges. Future research should explore more robust seed model selection, iterative self-training, and advanced domain adaptation techniques.

References:

[1] J. Kahn, *et al.*, "Libri-light: A benchmark for ASR with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 7669–7673.
[2] H. Scudder, "Probability of error for some adaptive pattern-recognition machines," *IEEE Trans. Inf. Theory*, vol. 11, no. 3, pp. 363–371, 1965. doi: 10.1109/TIT.1965.1053799.
[3] A. Radford *et al.*, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, 2023, Art. no. 1182.
[4] Z. Yao *et al.*, "WeNet: Production oriented streaming and non-streaming end-to-end speech recognition toolkit," in *Proc. Interspeech 2021*, 2021, pp. 4054–4058.
[5] P. Mihajlik *et al.*, "BEA-Base: A benchmark for ASR of spontaneous Hungarian," in *Proc. 13thLREC*, Marseille, France, Jun. 2022, pp. 1970–1977. Available: https://aclanthology.org/2022.lrec-1.211/

Improving Khalkha Mongolian ASR via Transliteration-Based Transfer Learning

D. Mengke[1], Y. Meng[1] & P. Mihajlik[1,2]

1) Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, Department of Telecommunications and Artificial Intelligence, 1111 Budapest, Hungary
2) Hungarian Research Centre for Linguistics, 1068 Budapest, Hungary

mengke@tmit.bme.hu

Background and objectives : Automatic Speech Recognition (ASR) systems have made consistent advancements, achieving notable improvements in state-of-the-art performance across various languages. However, their effectiveness often declines significantly in low-resource settings, where data and linguistic resources are limited. This paper addresses the challenges of ASR for a low-resource language, Khalkha Mongolian, by leveraging a transliteration-aided transfer learning approach. Specifically, it improves the ASR system for Khalkha Mongolian by transliterating text from a well-resourced Chakhar Mongolian (Uighur script) dataset to the Cyrillic script and then fine-tuning it with Khalkha Mongolian data. The method effectively enhances the ASR performance of Khalkha Mongolian. The effectiveness of the proposed method was validated on three popular ASR models, Wav2Vec2-BERT, Conformer-Large, and Whisper-large-v3.

Materials and methods： In this paper, we propose a transliteration method to transliterate a high-resource Mongolian dataset into the target low-resource Mongolian, utilizing a Mongolian dataset from China (the Chakhar Mongolian dialect of the Uyghur script) to help build an automatic speech recognition model that enhances the Mongolian language (the Khalkha Mongolian dialect of the Cyrillic script).

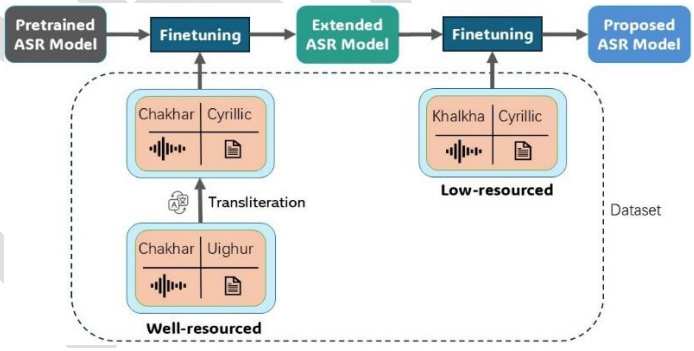


Figure 1: *This diagram illustrates the training pipeline for the proposed automatic speech recognition (ASR) model using a two-stage transfer learning approach. Initially, a well-resourced Chakhar Mongolian dataset in Uighur script is transliterated into Cyrillic script to facilitate compatibility with the target language. A pretrained ASR model is fine-tuned on this transliterated Chakhar Mongolian dataset, resulting in an extended ASR model. In the second stage, this extended model undergoes further fine-tuning on the low-resourced Khalkha Mongolian dataset in Cyrillic script to adapt to the target domain.*

Table 1: *Baseline and proposed fine-tuning results (CER%/WER%).*

Model	Zero-Shot	DFT	MDFT	EFT	TFT
Wav2Vec2-BERT	-	15.56 / 55.35	17.42 / 57.77	20.97 / 63.29	**15.33 / 52.82**
Conformer-Large	-	21.16 / 59.25	16.12 / 43.01	24.55 / 59.37	**14.50 / 39.99**
Whisper-Large-v3	45.73 / 92.60	13.96 / 43.02	16.27 / 49.95	21.96 / 61.35	**10.75 / 31.45**

Results and discussion: All test results are obtained from evaluations on the Mongolian test set (Khalkha dialect with Cyrillic script) from Common Voice v17 across all three models, and all the experimental results are shown in Table 1.

In the zero-shot setting, Whisper-large-v3 did not perform ideally, achieving a Character Error Rate (CER) of 45.73% and a Word Error Rate (WER) of 92.60%. For the direct fine-tuning stage, Whisper-Large-v3 continued to deliver the best performance, with a CER of 13.96% and a WER of 43.02%. Wav2Vec2-BERT achieved a CER of 15.56% and a WER of 55.35%, lagging behind Whisper-Large-v3. Conformer-Large had a CER of 21.16% and a WER of 59.25%, making it the least effective among the three models at this stage. Surprisingly, in the modified direct fine-tuning stage, the Conformer-Large model outperformed the others, achieving the best results with a CER of 16.12% and a WER of 43.01%. Whisper-Large-v3 followed with modest improvements, while Wav2Vec2-BERT's CER and WER increased to 17.42% and 57.77%. In the extended fine-tuning stage, Conformer-Large again achieved the best performance, with a CER of 24.55% and a WER of 59.37%. Whisper-Large-v3 reached a CER of 21.96% and a WER of 61.35%. Wav2Vec2-BERT reached a CER of 20.97% and a WER of 63.29%. However, in the final targeted fine-tuning stage, Whisper-large-v3 outperformed all models, delivering the overall best CER of 10.75% and WER of 31.45%.

Conclusions: We converted a relatively high-resource Chakhar Mongolian speech dataset (originally in the Uighur script) into Cyrillic, and then fine-tuned a pre-trained ASR model in two stages. This significantly improved the performance of the Khalkha Mongolian ASR system. We plan to expand the dataset and apply this transfer learning approach to other languages such as Uyghur and Kazakh.

References:

[1] J. A. Janhunen, Mongolian. Amsterdam, The Netherlands: John Benjamins Publishing Company, 2012.
[2] J. Burjgin and N. Bilik, "Contemporary Mongolian population distribution, migration, cultural change, and identity," in China's Minorities on the Move. London, UK: Routledge, 2015, pp. 53–68.
[3] L. Shi, F. Bao, Y. Wang, and G. Gao, "Research on Khalkha Dialect Mongolian speech recognition acoustic model based on weight transfer," in Proc. NLPCC 2019, Dunhuang, China, Oct. 9–14, 2019, pp. 519–528.
[4] T. Zhi, Y. Shi, W. Du, G. Li, and D. Wang, "M2ASR-MONGO: A free Mongolian speech database and accompanied baselines," in Proc. O-COCOSDA 2021, Singapore, Nov. 18–20, 2021, pp. 140–145.
[5] Y. Wu, Y. Wang, H. Zhang, F. Bao, and G. Gao, "MNASR: A free speech corpus for Mongolian speech recognition and accompanied baselines," in Proc. O-COCOSDA 2022, Hanoi, Vietnam, Nov. 24–26, 2022, pp. 1–6.

Speech Event Extraction: An ASR+NLP Challenge

M. Gedeon

Department of Telecommunications and Artificial Intelligence,
Budapest University of Technology and Economics, Hungary

gedeonm01@gmail.com

Background and objectives: Speech Event Extraction (SpeechEE) bridges the gap between Automatic Speech Recognition (ASR) and Natural Language Processing (NLP) by identifying events and their arguments from spoken language. Approaches to SpeechEE generally fall into two categories: pipeline-based and end-to-end. In pipeline-based approaches, the audio is transcribed, and event extraction is performed on the textual representation. Conversely, end-to-end models extract event structures directly from audio without relying on intermediate transcripts [1]. Our study explores a pipeline-based approach, leveraging a multi-step methodology that integrates Large Language Models (LLMs) with a semantic search-enhanced few-shot learning strategy.

Materials and methods: For training and evaluation, we relied on a benchmark dataset for event extraction, which was published for a shared task competition [2]. The dataset was derived from the ACE2005-EN+ dataset [3], and contains entries in the following form:

{"id": "train-6", "event": [{"trigger": "election", "type": "Elect", "arguments": [{"name": "man", "role": "Person"}]}]}

Figure 1 presents the proposed pipeline. We employed Whisper [4] and Canary [5] for ASR transcription. A classification step was then introduced to determine whether a given transcript is likely to contain an event. This was necessary, as most of the entries did not contain an event, and the LLMs returned many false positives. Three classification methods were implemented: (i) a rule-based approach that flagged instances containing trigger words identified in the training set, (ii) a BERT-based classifier trained on text embeddings, and (iii) OpenAI's o1-mini model prompted to classify event presence. A voting mechanism selected transcripts where at least two approaches agreed on event presence.

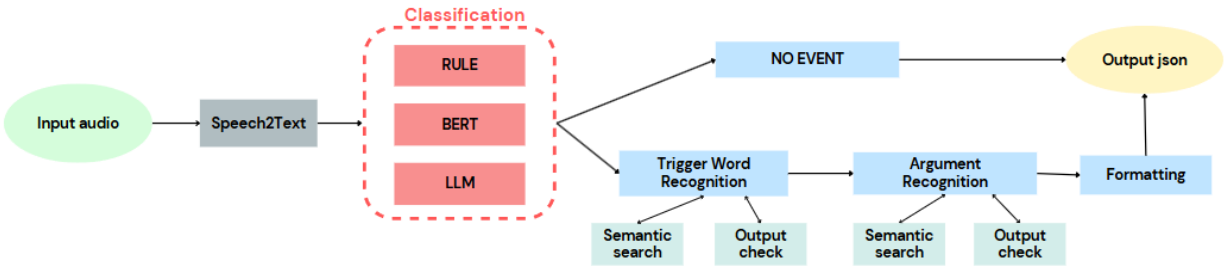


Figure 1: *Our pipeline*

In the trigger word recognition phase, an LLM was prompted to extract and classify trigger words into predefined categories. We evaluated LLaMA3-8B, GPT-4o-mini, and o1-mini for this task. Due to occasional instruction non-compliance, an automated verification step ensured output format consistency, re-executing queries when necessary. To enhance model performance, we employed a few-shot learning

approach, dynamically selecting ten semantically similar examples from the training set using embedding-based semantic search. Following trigger identification, another LLM step extracted event arguments and assigned roles, utilizing a similar semantic search and validation mechanism. Finally, a post-processing LLM call ensured uniform output formatting for JSON storage.

Results and discussion: Evaluation metrics focused on Trigger Classification (TC), assessing whether predicted event types and mentions matched reference annotations, and Argument Classification (AC), which required correct identification of event type, argument role, and argument mention. The considered baseline results were reported in [1], with an end-to-end model achieving F1-scores of 61.1 for TC and 23.3 for AC on the ACE2005-EN+ dataset. Our pipeline-based approach achieved comparable results, with an F1-score of 61.0 for TC and 24.3 for AC. It is important to note, that the used datasets are not identical, but were derived from the same dataset source.

Comparative analysis of ASR models revealed that the choice between Whisper and Canary had minimal impact on extraction performance. However, LLM selection significantly influenced results. LLaMA3-8B yielded the lowest performance (TC: 33.5, AC: 13.7), followed by GPT-4o-mini (TC: 47.1, AC: 19.5). OpenAI's o1-mini achieved the highest scores (TC: 61.0, AC: 24.3), highlighting its superior ability in event extraction tasks. This indicates, that larger models can perform this task better, and also that a reasoning model with its longer inference can be a useful tool.

Conclusions: This piece of research presents an LLM-driven pipeline for speech event extraction, combining ASR-generated transcripts with semantic search-enhanced few-shot learning. Our approach achieves performance on par with end-to-end models while offering a modular and interpretable framework. The robustness of extraction outcomes across different ASR systems suggests that transcript quality variations have limited impact. However, LLM selection plays a critical role in event extraction efficacy. Future research could explore hybrid approaches that integrate transcript-based and direct audio-based features, refine prompt engineering strategies, and investigate additional verification mechanisms to further enhance extraction accuracy.

References:

[1] Wang B, Zhang M, Fei H, Zhao Y, Li B, Wu S, et al. SpeechEE: A Novel Benchmark for Speech Event Extraction. arXiv (Cornell University). 2024 Aug 18;
[2] Github.io , "XLLM ACL 2025 Shared Task-II: Speech Event Extraction" [Online], Available: https://xllms.github.io/SpeechEE/ [Accessed on 2025 Mar 21]
[3] Lin Y, Ji H, Huang F, Wu L. A Joint Neural Model for Information Extraction with Global Features [Online]. ACLWeb. Available: https://aclanthology.org/2020.acl-main.713/
[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," in Proceedings of ICML, 2023, pp. 28492–28518.
[5] Puvvada KC, Piotr Żelasko, Huang H, Oleksii Hrinchuk, Nithin Rao Koluguri, Dhawan K, et al. Less is More: Accurate Speech Recognition & Translation without Web-Scale Data. Interspeech 2022. 2024 Sep 1;3964–8.

Towards Creating a Carinthian Slovenian Spoken Language Resource

A. Žgank[1], G. Donaj[1], T. Koren-Zwitter[2], U. Sereinig[3], M. Piko-Rustia[3], S. Boto[2], S. Zwitter-Grilc[4] & D. Verdonik[1]

1) University of Maribor, Faculty of Electrical Engineering and Computer Science, Maribor, Slovenia
2) Mohorjeva Hermagoras, Klagenfurt, Austria
3) Slovenian Ethnographic Institute Urban Jarnik, Klagenfurt, Austria
4) ORF Kärnten, Klagenfurt, Austria

andrej.zgank@um.si

Background and objectives: Recent advancements in transformer-based speech technology, developed by global technology companies, have enabled support for an increasing number of languages [1]. Slovenian and Austrian German, both Central European languages, have undergone some development in general speech recognition [2, 3]. However, they still lag behind major languages, particularly regarding the quantity and diversity of available spoken language resources [4, 5]. Among the key factors affecting speech recognition accuracy is linguistic variation, with dialects playing a significant role. Carinthian Slovenian is spoken in the northernmost part of the Slovenian language area, encompassing not only Slovenia but also the southern region of Austrian Carinthia, where a Slovenian minority resides. This linguistic variety comprises four dialects, with the Gail Valley dialect representing the most northwestern variant. Due to its limited number of speakers, the ongoing process of assimilation with Austrian German, and its geographical position at the extreme northwestern periphery of the Slovenian language area, the Gail Valley dialect is among the most endangered Slovenian dialects. The Zila project aims to preserve the cultural heritage of this dialect in the digital era. As a first step, 100 hours of manually transcribed recordings of the Slovenian Gail Valley dialect will be created. A freely available automatic speech recognition system for this dialect will be developed in the subsequent phase.

Materials and methods: The Slovenian Gail Valley dialect is primarily spoken in the municipalities of Feistritz an der Gail/Bistrica na Zilji, Hohenthurn/Straja vas and previously independent Egg/Brdo in Austria. According to the 2001 Austrian census [6], between 5% and 10% of the local population in these municipalities belongs to the Slovenian minority. Given the limited number of speakers, all available sources were utilized to collect speech data, including recordings from ethnographic archives, media content, and newly gathered field recordings. This diversity is also reflected in the range of recording devices used, which include analogue magnetic tapes, video recorders, CDs, digital voice recorders, and smartphones. All speech recordings were converted into WAV format with a single-channel configuration, a 16 kHz sampling rate, and 16-bit resolution. The manual transcriptions were prepared by two native Carinthian Slovenian speakers. As a baseline, the transcription guidelines developed for the Slovenian Artur speech database [5] were adopted and subsequently modified to capture the Slovenian Gail Valley dialect's distinctive pronunciation, grammar, and vocabulary.

Results and discussion: A sample transcription of the Slovenian Gail Valley dialect is presented in Table 1. Currently, the database includes recordings from 44 speakers, amounting to 42 hours of speech data. The speakers range from 26 to 94 years, with an average age of 69.6 years. This demographic distribution

highlights the endangered status of the Slovenian minority in the Gail Valley. A key factor monitored in the dataset was the quality of the Slovenian Gail Valley dialect spoken by the participants. In most cases, the dialect was remarkably well preserved, which is crucial for ensuring the linguistic integrity and overall quality of the spoken language resource. Two primary challenges have been identified in the development of the Zila Spoken Language Resource. The first challenge is securing enough proficient speakers of the Gail Valley dialect. The second is associated with the transcription rules. High-quality transcriptions are essential for preserving cultural heritage when creating a spoken language resource. At the same time, the transcriptions must be consistent and structured in a way that supports the development of an automatic speech recognition system. Thus, it is crucial to balance between these two points.

Table 1: *Slovenian Gail Valley dialect transcriptions from Mijalca majalca fable [7].*

	Transcription
Gail Valley	V tistah časah sa sə vəčbart na goərah, pər Zilə bəl pər patokə pa na Ogə čudnə rečə zgadilə.
Std. Slovenian	V tistih časih so se večkrat na gorah, pri Zilji, ob potokih in na Logu godile čudne reči.
German	In jener Zeit ereigneten sich auf den Bergen, an der Gail oder an Bächen und in der Au recht seltsame Dinge.
English	In those days, strange things often happened in the mountains, at Zilja, by the streams and at Log.

Conclusions: This paper has outlined the initial framework and key challenges in creating a spoken language resource for the Carinthian Slovenian Gail Valley dialect. Once finalized, this resource will be used to develop a corresponding automatic speech recognition system.

Acknowledgment: This work is part of the LINGUA Project, funded by the EU Interreg Programme Slovenia-Austria 2021–2027. The authors wish to thank members of the Slovenian minority in the Gail Valley, Austria, who contributed their speech to the Zila Spoken Language Resource.

References:

[1] Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust speech recognition via large-scale weak supervision. International conference on machine learning 2023, 28492-28518. PMLR.
[2] Žgank A, Vitez AZ, Verdonik D. The Slovene BNSI Broadcast News database and reference speech corpus GOS: Towards the uniform guidelines for future work. LREC 2014 May 26 (pp. 2644-2647).
[3] Linke J, Geiger BC, Kubin G, Schuppler B. What's so complex about conversational speech? A comparison of HMM-based and transformer-based ASR architectures. Computer Speech & Language. 2025; 90:101738.
[4] Verdonik D, Bizjak A, Žgank A, Maučec MS, Trojar M, Gros JŽ, Bajec M, Bajec IL, Dobrišek S. Strategies for managing time and costs in speech corpus creation: insights from the Slovenian ARTUR corpus. Language Resources and Evaluation. 2024 Nov 30:1-26.
[5] Schuppler B, Kelterer A, Hagmüller M. 10 Years of GRASS development: Experiences from annotating a large corpus of conversational Austrian German. Austrian Meeting on Digital Linguistics: Recent Developments in Austria 2023.
[6] Austrian Census 2001 statistics, [Online], Available: https://www.statistik.at/en/statistics/population-and-society/population/population-stock/historic-censuses
[7] Bartoloth M. Mijalca majalca: Zilska basən / Ziljska basen / Ein Gailtaler Märchen. Klagenfurt, Vienna and Ljubljana: Mohorjeva Hermagoras, 2021.

How to Use SLICER for Stimuli Extraction from Large Speech Corpora

L. Eckert, S. Wepner & B. Schuppler

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

eckert@student.tugraz.at

Motivation: This practical session introduces a tool for stimuli search and extraction from a large speech corpus, allowing to extract and manipulate both the audio and the annotation file simultaneously. SLICER [1] is designed to complement existing software like Praat and ELAN by offering additional functionalities rather than replacing it. SLICER is designed for efficient production of natural sounding stimuli for perception experiments. It can also be used to change data types in the audio or save only a subset of the tiers in the annotation. In this session, we will explore the workflow of SLICER using examples from the audience or our own work.

Data and tools: SLICER works with input corpus data annotated in Praat TextGrid format. The key functionalities of the tool include: 1) An advanced label search, allowing users to locate specific segments based on annotations. 2) Slices can subsequently be manipulated. 3) These slices can further be filed and exported, creating a set of stimuli. When manipulating segments (e.g., phones, words), the tool offers the possibility to insert noise with configurable signal-to-noise ratios and apply smooth attack and decay transitions to ensure natural-sounding stimuli. 4) When exporting the set of stimuli, users can choose which annotation tiers to include, set audio sample rates and formats, and normalize audio output for consistency. 5) The integrated file-naming conventions allow to locate the stimulus in the original corpus file.

Learning goals: Participants will learn how they can apply SLICER on their own data to quickly and cleanly create stimuli for perception experiments, or to change their annotation tiers and audio file data types. During the session, feedback is very welcome as also we hope to learn from the participants how to improve future versions of SLICER.

Preparations: Please bring your own notebook or tablet and headphones to the session. The tool can be found in the repository [2]. If you are comfortable with python, you can directly install with pip as described in the repository. During the practical session, also an .exe will be shared with the participants, for those encountering troubles with the pip installation. The use of python is preferred to avoid shortage of time. If your notebook comes with iOS or linux, you will need to use the python version, since only a .exe will be supplied. Participants are strongly encouraged to bring their own data to work on. The tool is compatible with .wav-files for audio and TextGrid (Praat) files for annotation.

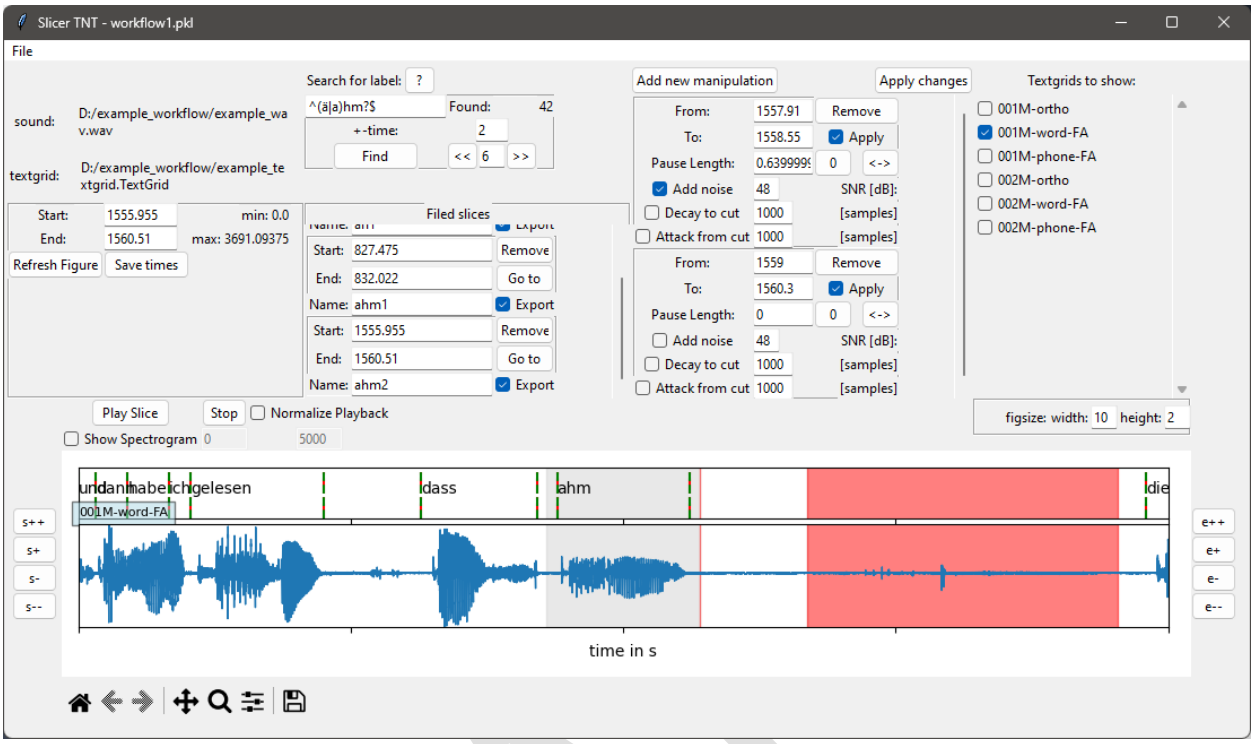


Figure 1: *GUI of SLICER*

Acknowledgements: This research was funded by the Austrian Science Fund (FWF) [10.55776/P32700].

References:

[1] Eckert, Lucas and Wepner, Saskia and Schuppler, Barbara, "Slicer – A Tool for Efficient Stimuli Extraction from Large Speech Corpora," *In Proceedings of Forum Acusticum 2025*, Malaga, June 2025

[2] Eckert, Lucas and Wepner, Saskia and Schuppler, Barbara, "Slicer – A Tool for Efficient Stimuli Extraction from Large Speech Corpora," [Github], Available: https://github.com/SPSC-TUGraz/SLICER

Annotating Creak in Healthy and Pathological Voices

A. Viehhauser

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

anna.viehhauser@student.tugraz.at

Motivation: Creak is a voice quality which is frequently occurring in healthy speech and sometimes carrying linguistic or paralinguistic information, like emotions, intentions or attitude [1]. In pathological speech creak can occur independently of its usual functions, due to the lack of precision in voice control. Therefore, the timing of creak could be an indicator for distinguishing between healthy and pathological speech [2]. To analyze this, I use *creapy*, a tool initially developed for automatically annotating creak in conversational healthy speech, which I applied to pathological speech [3]. Since pathological speech usually has different characteristics than healthy speech, false positive intervals are annotated by *creapy* [4]. To improve this, I will investigate the similarities and differences of "healthy creak" and "pathological creak".

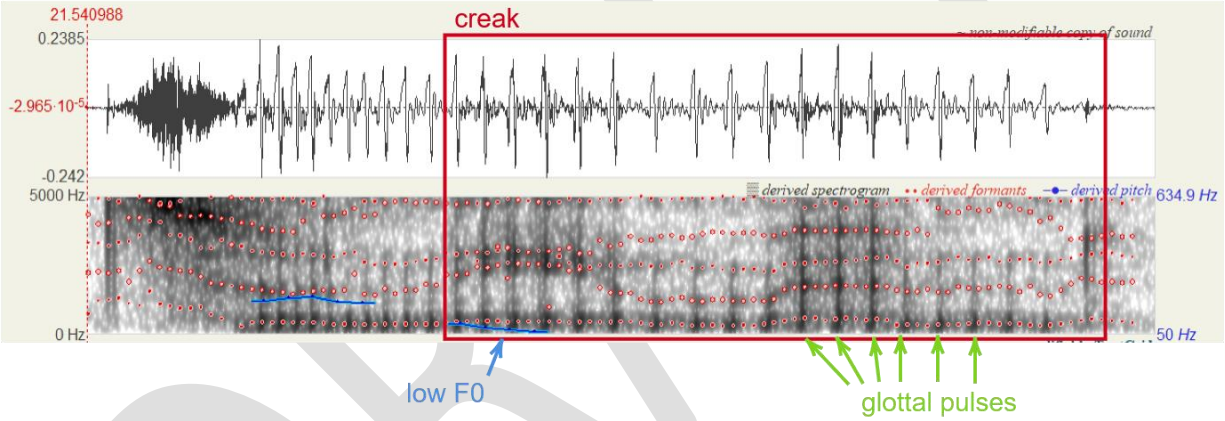


Figure 1: *Example of a recording of an unfrequently diplophonic (pathological) female speaker with creaky voice quality (red box). The signal corresponds to the word fragment "...zunehmen..". The top plot shows the waveform, the plot below the spectrogram. The glottal pulses (green) are clearly visible as well as the low fundamental frequency. Due to the irregularities of creak, the pitch detection algorithm cannot compute a fundamental frequency for the rest of the creak interval.*

Data and tools: I will bring audio files (in .wav format) with recordings from pathological speakers reading the German standard text "Der Nordwind und die Sonne". The audio files were recorded at the Medical University of Vienna. For each audio file, there will be a corresponding TextGrid and each TextGrid will have a tier called "creapy" with automatically annotated creak intervals using *creapy*. During the practical session, we will analyze and correct these creak annotations using Praat.

Learning goals: After this session you will have learned that creak cannot always be clearly defined in pathological voices. I would like to discuss the acoustic characteristics of creak in pathological voices and compare them to the characteristics of creak in healthy voices. Do you perceive/hear differences or similarities? I will use the corrected annotations from the audience to get new insights in creak in

pathological voices and revise my own annotations. With these annotations as ground-truth I will in the future then conduct a syllable-based analysis and (re-) train *creapy* on the pathological data.

Preparations: Please bring your own laptop, headphones and install Praat on your laptop. Additionally, your laptop should have an USB port or bring an adapter for it.

References:

[1] C. T. Ishi, K.-I. Sakakibara, H. Ishiguro, and N. Hagita, "A method for automatic detection of vocal fry," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 1, pp. 47–56, Jan. 2008, issn: 1558-7916. doi: 10.1109/TASL.2007.910791.

[2] V. Devaraj, I. Roesner, F. Wendt, J. Schoentgen, and P. Aichinger, "Auditory perception of impulsiveness and tonality in vocal fry," Applied Sciences, vol. 13, no. 7, p. 4186, Mar. 2023, issn: 2076-3417. doi: 10.3390/app13074186.

[3] M. Paierl, T. Röck, S. Wepner, A. Kelterer, and B. Schuppler, "Creapy: A python-based tool for the detection of creak in conversational speech," in *20th International Congress on Phonetic Sciences*, 2023.

[4] A. Viehhauser, "Creak detection in pathological voices," Bachelor's project at Graz University of Technology, Graz, Austria, 2024.

SpeechScape: An Easy-to-Use Tool for Clustering Speech Data Based on Self-Supervised Representations

J. Linke

Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria

linke@tugraz.at

Motivation: SpeechScape provides a method for investigating newly recorded speech data already before detailed annotations are available, potentially accelerating the research process and guiding future annotation efforts. The tool utilizes self-supervised speech representations to estimate acoustic similarity/distance between categories. SpeechScape builds upon the insights gained from earlier studies [1, 2], in which it was demonstrated that self-supervised speech representations encode information about languages, varieties, speaking styles, and individual speakers. By leveraging these findings, SpeechScape aims to provide researchers with a practical tool to analyze and compare speech data across different dimensions.

Data and Tools: SpeechScape leverages the wav2vec 2.0 framework [3], utilizing the XLSR-53 model for multilingual speech representation [4]. The tool processes audio recordings and generates similarity matrices based on the speaker-wise frequency usage of shared discrete speech representations (cf. Figure 1). While initially designed for language comparison, SpeechScape's versatility extends to various acoustic analyses. Researchers can use it to identify patterns in language varieties (e.g., Northern German vs. Austrian German), speaking styles (e.g., read vs. conversational speech), examine prominence levels, or investigate other extra-linguistic properties of interest. This flexibility makes SpeechScape a valuable tool for a wide range of speech-related studies beyond language documentation.

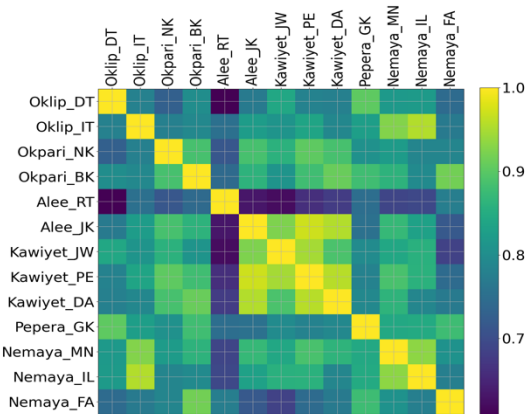


Figure 1: *Example of a similarity matrix for Papuan speakers resulting from the codebook usage of each speaker indicated by variety and speaker ID (i.e., variety ID Oklip_DT with variety Oklip and speaker ID DT). The similarities between all speakers refer not only to language/variety or individual speaker characteristics but also more dominant background noise (i.a., background sounds from other speakers, children, animals, traffic, ...). As an example, the lower similarities observed for variety ID Alee_RT may be attributed to the more dominant traffic noise present in the recordings.*

Learning Goals: After this workshop, participants will understand the concept of self-supervised speech representations and their application in spoken language analysis. They will learn how to apply SpeechScape to their analyses of acoustic similarities/distances between different linguistic and extra-linguistic properties (e.g., languages, language varieties, speaking styles, prominence levels, …) and how to interpret the results and plots provided by the tool.

Preparations: Participants should bring a laptop with Python (version 3.8 or higher) installed. They should prepare short audio samples (e.g., approx. 1 - 10 seconds each) from different speech categories they wish to analyze. Audio files should be in 16 kHz sampling rate and mono format. File names should follow the pattern: **_CATEGORY_SPKID.wav*, where CATEGORY is a consistent label for each group (e.g., language, language variety, speaking style, prominence level, another linguistic label…) and SPKID is a unique identifier for each speaker within a category.

For example, a suitable list of audio files could look like this:

- *recording001_AustrianGermanConversationalSpeech_Speaker01.wav*
- *recording002_AustrianGermanConversationalSpeech_Speaker01.wav*
- …
- *recording001_AustrianGermanConversationalSpeech_Speaker02.wav*
- *recording002_AustrianGermanConversationalSpeech_Speaker02.wav*
- …
- *recording001_AustrianGermanReadSpeech_Speaker01.wav*
- *recording002_AustrianGermanReadSpeech_Speaker01.wav*
- …

It is recommended to have at least 20 samples per category for robust statistical representation and to define shorter labels for CATEGORY (e.g., *AGCS*, *AGRS*, ...) and SPKID (e.g., *SP01*, *SP02,* …). The more samples per category and speaker you bring, the more reliable the analysis will be. Note that SpeechScape is still in development, with plans for a user-friendly GUI and automated dependency installation in future versions.

References:

[1] Linke et al., “What do self-supervised speech representations encode? An analysis of languages, varieties, speaking styles and speakers,” in Proc. Interspeech, 2023, 2023, pp. 5371-5375 [Online]. Available: https://www.isca-archive.org/interspeech_2023/linke23_interspeech.pdf
[2] SPSC-TUGraz. (2023). "SpeechCodebookAnalysis" [Online]. Available: https://github.com/SPSC-TUGraz/SpeechCodebookAnalysis
[3] Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in Advances in Neural Information Processing Systems, vol. 33, 2020, pp. 12449-12460 [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf
[4] Conneau et al., "Unsupervised Cross-Lingual Representation Learning for Speech Recognition," in Proc. Interspeech 2021, 2021, pp. 2426-2430 [Online]. Available: https://www.isca-archive.org/interspeech_2021/conneau21_interspeech.pdf

Speech Pauses in Alzheimer Disease: Exploring Challenges in Training Automatic Speech Recognition Systems

S. Lafenthaler

Department of Neurology, Neurointensive Care, and Neurorehabilitation, Christian Doppler University Hospital, Paracelsus Medical University, Salzburg, Austria, Member of EpiCARE
Department of Linguistics, Paris Lodron University, Salzburg, Austria

sa.lafenthaler@salk.at

Motivation: Spontaneous speech in individuals with Alzheimer disease (AD) is often characterized by an increased frequency of speech pauses, reflecting difficulties in word retrieval and the formulation of coherent, complete thoughts [1]. While manual transcription is time-consuming, automatic speech recognition (ASR) tools offer a promising alternative. Current ASR systems perform well with unfilled pauses but struggle with filled pauses [2], which limits accurate analysis of speech pattern crucial for understanding language decline in AD. For example, when "and" is used as a conjunction, it serves a syntactic function by linking clauses or propositions. In contrast, when used as a filler particle, it may indicate difficulties with lexical retrieval or the formulation of the next utterance. Therefore, distinguishing between these uses provides valuable insights into language processing, particularly in neurodegenerative diseases like AD.

Data and tools: In this session, we will transcribe various utterances from selected audio recordings to address the challenge of annotating filled pauses.

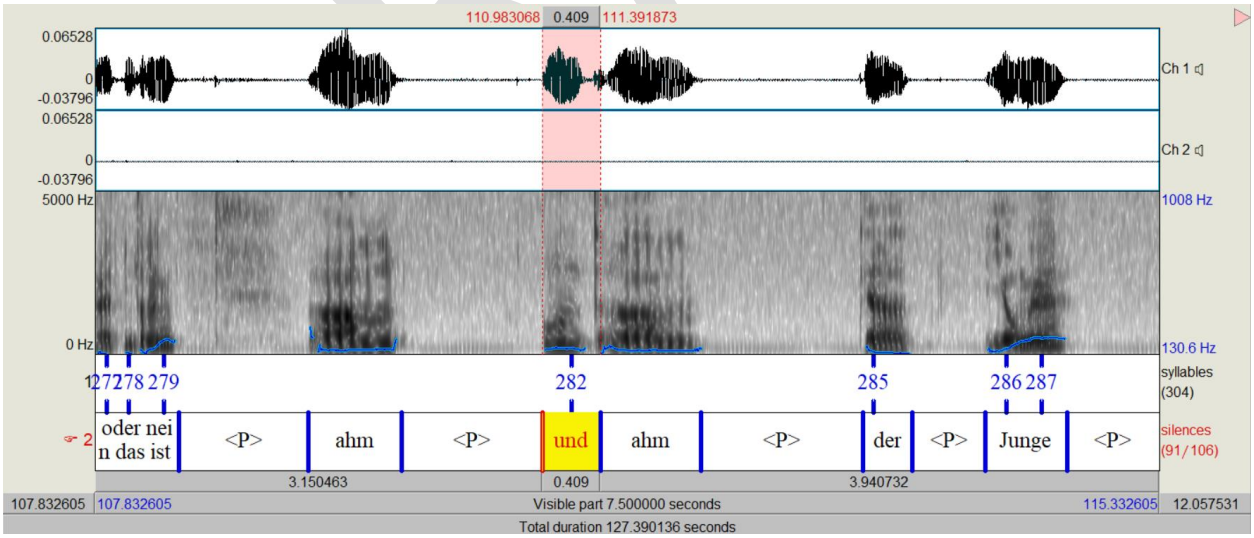


Figure 1: *A screenshot from Praat of a patient's spontaneous speech excerpt during a picture description. It highlights empty and filled pauses, as well as the word "und," which is often difficult to interpret as either a conjunction or a filled pause.*

The spoken language corpus is based on picture descriptions and comes from individuals with mild AD, who were recruited ate the Memory Clinic of the Christian Doppler University Hospital Salzburg as part of an ongoing project (Ethics Commission Salzburg; number 1044/2021). The patients are Austrian German native speakers with different Salzburg dialect backgrounds. In this session, we will work with Praat software [3].

Learning goals: In this session, participants will explore the limitations and challenges of ASR systems and gain insights into the requirements for integrating ASR into a clinical context using spoken language data from patients. We will primarily discuss the standard approach to identifying pauses, as well as the definition and reliable detection of filled pauses.

Preparations: Note that this task can be difficult even for native German speakers! Kindly bring your own notebook and wired headphones, and make sure that you have installed the Praat software on your notebook. The software is available as an open resource at the following link: https://www.fon.hum.uva.nl/praat/.

References:

[1] M. Lofgren, W. Hinzen. (2022, May-Jun). "Breaking the flow of thought: Increase of empty pauses in the connected speech of people with mild and moderate Alzheimer's disease." *J. Commun. Disord.* [Online]. vol. 97. Available: https://doi.org/10.1016/j.jcomdis.2022.106214. Erratum in: *J. Commun. Disord.* [Online]. vol. 101, Jan-Feb 2023. Available: https://doi.org/10.1016/j.jcomdis.2022.106299.
[2] S. O. C. Russell, I. Gessinger, A. Krason, G. Vigliocco, and N. Harte. (2024, Mar). "What automatic speech recognition can and cannot do for conversational speech transcription." Research Methods in Applied Linguistics [Online]. vol. 3, no. 3. Available: https://doi.org/10.1016/j.rmal.2024.100163.
[3] P. Boersma and D. Weenink. (2001, Dec). "PRAAT, a system for doing phonetics by computer." Glot International [Online]. vol. 5, pp. 341-345. Available: https://www.glottopedia.org/index.php/Glot_international.

From Medical Data to Models: Own Experiences and Challenges

M. Fleischer & D. Mürbe

Department of Audiology and Phoniatrics, Charité—Universitätsmedizin Berlin,
Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Germany

mario.fleischer@charite.de

Motivation: A detailed analysis of the acoustic transfer characteristics of the complex three-dimensional vocal tract and the sound propagation around the head is essential for voice and speech research [1-3]. We could show that finite element analysis is a valuable tool to address these research questions. This method requires not only the implementations of the governing equations (wave equation, Helmholtz equation, etc.) but also three-dimensional meshes.

Data and tools: Starting with DICOM data (MRI image stacks, optionally CT image stacks), we use a cascade of 3D Slicer, ITK-SNAP, Blender, Meshlab, and GMSH to prepare the DICOM data, segment the vocal tract, manipulate the surface meshes, and create 3D meshes.

Learning goals: This article aims to give an impression of the detailed steps required to obtain a numerically calculable 3D model from MRI images. We show that plenty of tools and several individual steps are currently necessary. How can this process be accelerated? How can the degree of automation be increased?

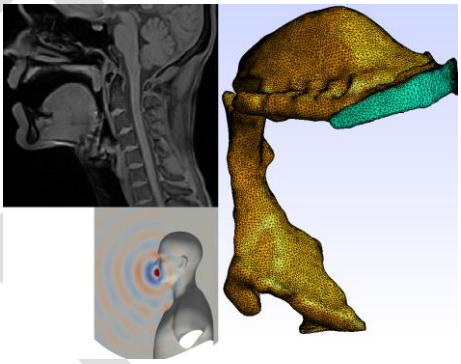


Figure 1: *Based on MRI images, 3D models are generated and analyzed numerically (for example, sound propagation around the head [4])*

References:

[1] P. Birkholz, S. Kürbis, S. Stone, P. Häsner, R. Blandin, and M. Fleischer, ‘‘Printable 3D vocal tract shapes from MRI data and their acoustic and aerodynamic properties’’, Scientific Data, vol. 7, no. 1, pp. 1–16, 2020.
[2] M. Fleischer, A. Mainka, S. Kürbis, and P. Birkholz, ‘‘How to precisely measure the volume velocity transfer function of physical vocal tract models by external excitation’’, PLoS ONE, vol. 13, no. 3, pp. 1–16, 2018.
[3] M. Fleischer, S. Rummel, F. Stritt, J. Fischer, M. Bock, M. Echternach, B. Richter, and L. Traser “Voice efficiency for different voice qualities combining experimentally derived sound signals and numerical modeling of the vocal tract”, Frontiers in Physiology, vol. 13, 1081622, 2022
[4] P. Birkholz, S. Ossmann, R. Blandin, A. Wilbrandt, P. K. Krug, and M. Fleischer, “Modeling Speech Sound Radiation With Different Degrees of Realism for Articulatory Synthesis” *IEEE Access*, vol. 10, pp. 95008-95019, 2022

Evidencing Physiological and Acoustic Outcomes of Clinical Voice Interventions Using Voice Maps

S. Ternström

Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Stockholm, Sweden

stern@kth.se

Motivation: Conventional acoustical and physiological measurements are time-consuming and/or weak in terms of providing evidence for effects of treatment and/or training of voices. In this workshop we will see that typically, it is not the metrics themselves that are inappropriate, but rather the legacy approaches to how data from voice signals are collected and collated. In voice mapping, multiple measurements are made continuously in parallel, collecting data from practically all phonated cycles, and providing real-time visual feedback during voice production tasks.

Data: Volunteering participants will analyze their own voices with real-time voice mapping. Also, pre-recorded examples of pre- and post-intervention recordings of patients and singers will be examined. *Tools:* Hardware: audio interface, microphone, electroglottograph with analog output, laptop computer, loudspeaker. Software: SuperCollider 3.13.0 [1], FonaDyn 3.3.2 [2], and any spreadsheet app.

Learning goals: To understand the differences between the phonetogram/VRP and the voice map; to understand why SPL and f_0 must be matched when comparing measurements across clinical interventions; to understand how to make and interpret pre-, post-, and difference voice maps. The benefits of considering the (de-noised) EGG signal jointly with the acoustic signal will be demonstrated.

Participants will experience with their ears, eyes and hands-on how the public-domain FonaDyn system can be used to map voice attributes over the range, and assess the effects of interventions. We will talk also about drawing substantiated conclusions from voice maps, and the role of statistical testing of differences in maps. Finally, the participants' own ideas for pre-post assessments will be welcomed and discussed. For best results, voice production tasks should exercise all relevant parts of the informant's voice range – but not more.

Preparations: Participants do not need to bring anything, but please read at least references [3] and [4]. You will find further references therein. Those wanting to try running FonaDyn on their own computer are recommended to download the software in advance [1, 2], and bring a microphone and headphones.

Acknowledgments: Several of the examples were collected in cooperation with phoniatrician Silvia Capobianco during her visit to Stockholm in 2024. Thanks to Gunnar Björck, M.D., at the Karolinska University Hospital for letting us observe patients pre- and post- intervention with injection laryngoplasty. Thanks also to baritone Joris Grouwels for his supra-normal voice.

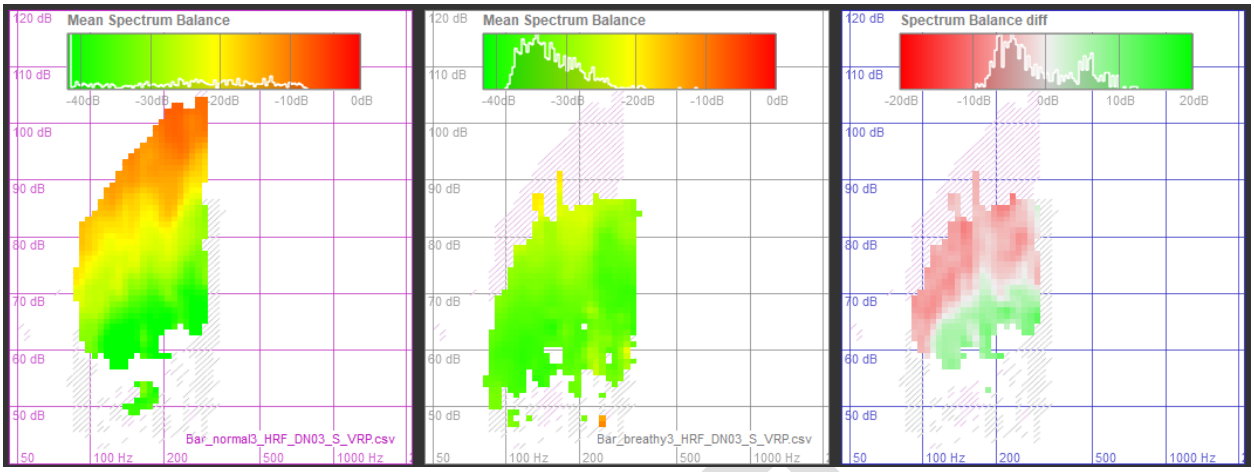


Figure 1: *Voice maps of a highly trained baritone covering his full modal voice range. The vertical axis is SPL (dB) at 30 cm, the horizontal axis is fundamental frequency (Hz). The metric shown here is the spectrum balance, in normal voice (left), intentionally breathy voice (centre), and the difference (right). The spectrum balance is the power ratio between high frequencies (>2 kHz) and low frequencies (<1.5 kHz), expressed in dB. The spectrum balance is usually negative and increases (becomes less negative) with increasing vocal effort. The difference map shows a reduced spectrum balance (red) in normal-to-loud voice and an increased one (green) in soft voice. The transition aligns with the threshold of vocal fold contacting. This can be seen by comparing with other layers in the same voice map.*

References:

[1] SuperCollider download: https://supercollider.github.io/downloads . Download also the "sc3-plugins".
[2] FonaDyn download: https://www.kth.se/profile/stern
[3] (short read) Ternström, S. "Vocal Function and Range", *NCVS Insights*, vol. 2, no. 2, 2024. https://doi.org/10.62736/ncvs183832
[4] (overview of FonaDyn) Ternström, S. "FonaDyn Quick Look", online at [2]. (3 March 2025).
[5] (long read) Ternström, S., and Pabon, P. "Voice Maps as a Tool for Understanding and Dealing with Variability in the Voice," *Applied Sciences*, 12, 11353, 2022. https://doi.org/10.3390/app122211353
[6] (very long read) Pabon, P. *Mapping Individual Voice Quality over the Voice Range : The Measurement Paradigm of the Voice Range Profile*. PhD thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2018.
https://kth.diva-portal.org/smash/get/diva2:1253755/FULLTEXT01.pdf

Optopalatography for Phonetic Research

J. Yun, J. Menezes, A.-L. Fietkau, P. Birkholz

Institute of Acoustics and Speech Communication, Dresden University of Technology, Germany

jeehyun.yoon@tu-dresden.de

Motivation: We introduce the optopalatography (OPG) system OPG2023 and its application to research in phonetics and speech science. OPG2023 is the current development at the TU Dresden for measuring the movement of the lips and the tongue, which is preceded by a series of different versions [1]. With a compact and lightweight control unit (Fig. 1), the system is portable and easy to set up. It is suitable not only for laboratory experiments with multi-channel data acquisition but also for field research with unspecified participants.

Data and tools: The OPG2023 is available with personalized- or non-personalized artificial palates. A total of 15 light source emitter and detector pairs provide the measurements related to the distances between the tongue surface and the palate and between the lips. The signals are recorded at a sampling rate of 100 Hz with 18-bit quantization. The normalized signal strength, although with large variability depending on the condition, could be roughly estimated to be maximum at contact, decreasing as the cube of the distance and converging to 0 at 3 cm. Articulatory Data Recorder 2 (ADR2) is a dedicated recording software for simultaneous recording of OPG and acoustic signals and their automatic synchronization. The intuitive display panel in ADR2 facilitates real-time and offline monitoring of the optical sensors, and thus of lip and tongue gestures.

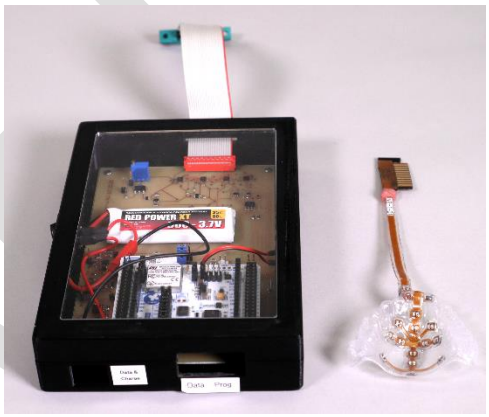


Figure 1: *OPG2023 control unit (left) and artificial palate (right)*

Learning goals: The configuration of artificial palates and common data acquisition procedures will be explained. A step-by-step demonstration will show how to record articulatory data using a pseudopalate, control unit, ADR2 software, and microphone. The characteristics of the OPG signal and the advantages and disadvantages of the device will be discussed. You will be able to understand how the system works and develop ideas on how to use the equipment for your own research topics.

Preparations: During the presentation, a non-personalized artificial palate could be available for one volunteer to participate in the demonstration. Those interested in wearing the pseudopalate may want to rinse their mouth before the session.

References:

[1] P. Birkholz, S. Stone, C. Wagner, S. Kürbis, A. Wilbrandt, M. Bosshammer. "A review of palatographic measurement devices developed at the TU Dresden from 2011 to 2022," In Conf. International Congress of Phonetic Sciences (ICPhS), Prague, Czech Republic, 2023, pp. 883-887.

Using Ultrasound Tongue Imaging for Phonetic Research

N. Elsässer, J. Luttenberger, H. Behrens-Zemek & E. Reinisch

Acoustics Research Institute, Austrian Academy of Sciences, Vienna, Austria

nathalie.elsaesser@oeaw.ac.at

Motivation: In this practical session, we explore the application of Ultrasound Tongue Imaging (UTI) in phonetic research. UTI is a widely used method for examining tongue motion during speech, as it is non-invasive (unlike electropalatography or electromagnetic articulography) and safe for humans (unlike X-Ray) [1]. Sagittal UTI offers the possibility of objectively determining different places of articulation and subtle articulatory differences. It is also used as a tool in speech therapy [2]. UTI studies allow for a wide range of methods in tongue contour identification (manual vs. (semi-)automatic methods), including the use of reference recordings and data normalization. We present our approach of performing UTI studies and invite discussion on various topics, especially various approaches to data analysis.

Data and tools: This session includes a demonstration of the UTI data recording process, covering participant preparation, optimizing ultrasound settings, reference recordings (e.g., palate tracing, bite plane) and an overview of the software tools used for data acquisition and analysis.

We collected data in two different projects concerning Austrian German /r/ and /l/, which serve as exemplary cases in this session. We discuss these two highly variable phonemes and propose solutions to common challenges, such as the limited visibility of the tongue tip in phonemes where the tongue tip is raised (e.g. alveolar trill or retroflex /l/), as ultrasound does not accurately capture very steep or retroflex tongue shapes (see Figure 1).

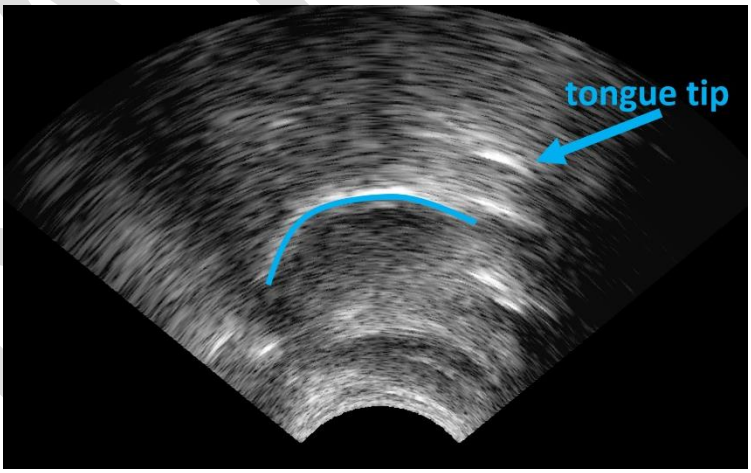


Figure 1: *Example picture of an ultrasound image of the tongue during the production of an alveolar trill [r] with the contour of the tongue body traced in blue. The steep tongue tip on the right is not depicted correctly but creates a white reflection further up instead (where the arrow is pointed).*

For UTI recordings, we use Articulate Assistant Advanced (AAA) [3], a software frequently utilized in articulation research. We demonstrate key steps in AAA, including data recording, processing, and extraction, as well as fitting splines that represent tongue contours using both manual and automatic methods provided by AAA, including DeepLabCut, a machine learning tool trained to identify tongue contours in ultrasound data [4]. Additionally, we present a method for normalizing tongue splines based on previously recorded reference data. We outline a schematic evaluation process for the collected data and provide a brief introduction to our planned statistical analysis using R [5]. We also demonstrate the visualization of tongue splines as polar and cartesian coordinates and discuss possible analyses using the R package rticulate [6].

Learning goals: Participants new to UTI data recording will get a first impression of setting up UTI experiments and analyzing tongue contours based on recorded images. They will also gain an understanding of potential challenges and different methods for data normalization. Additionally, we aim to engage more experienced participants in discussions on common challenges related to data extraction and analysis such as using AAA's analysis values system, exploring alternative R packages or applying different statistical models.

Preparations: No preparation is needed to join this session, as it is not necessary for attendees to work along step by step. However, for those who are interested and have access to a AAA license key they are welcome to work along on their own laptops as we show the key steps in AAA. If attendees are interested in replicating the proposed analysis steps in R, they should install the package rticulate before the session. We will provide an exemplary dataset and analysis code for interested participants.

References:

[1] M. Stone, "A guide to analysing tongue motion from ultrasound images," *Clinical linguistics & phonetics*, vol. 19, 6-7, pp. 455–501, 2005, doi: 10.1080/02699200500113558.
[2] L. S. Da Barberena, B. d. C. Brasil, R. M. Melo, C. L. Mezzomo, H. B. Mota, and M. Keske-Soares, "Ultrasound applicability in Speech Language Pathology and Audiology," *CoDAS*, vol. 26, no. 6, pp. 520–530, 2014, doi: 10.1590/2317-1782/20142013086.
[3] Articulate Instruments Ltd, *Articulate Assistant Advanced User Guide: Version 2.14*. Edinburgh, UK: Articulate Instruments Ltd, 2012.
[4] A. Wrench and J. Balch-Tomes, "Beyond the Edge: Markerless Pose Estimation of Speech Articulators from Ultrasound and Camera Images Using DeepLabCut," *Sensors (Basel, Switzerland)*, vol. 22, no. 3, 2022, doi: 10.3390/s22031133.
[5] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria. [Online]. Available: https://www.r-project.org/
[6] Stefano Coretta, *rticulate: Articulatory Data Processing in R*. [Online]. Available: https://github.com/stefanocoretta/rticulate

Analysis of Uni- and Multi-Dimensional Contours with GAMs and Functional PCA: an Application to Ultrasound Tongue Imaging

M. Gubian

Institute for Phonetics and Speech Processing at LMU, Munich, Germany

m.gubian@phonetik.uni-muenchen.de

Motivation: This session illustrates two alternative statistical methods to analyze ultrasound tongue imaging (UTI) data. One method is based on generalized additive (mixed) models (GAMs), the other on a combination of functional principal components analysis (FPCA) followed by linear (mixed-effects) regression (LMER). Both approaches are suitable for UTI analysis represented in the polar domain, offering different views on the same data set. Each approach will be applied both on a static and on a dynamic scenario, where static means a single point in time is included in the analysis (e.g. the onset or the point in time where the articulatory target is reached), while dynamic means that each sound realization is represented by a sequence of tongue contours along time. The geometrical setup in the dynamic scenario is the same as in [1].

Data and tools: The session is entirely based on artificial data. The code is based on the following R libraries. GAMs: *mgcv, itsadug*; FPCA: *funData, MFPCA, fda, landmarkregUtils*, the last one available at https://github.com/uasolo/landmarkregUtils; data processing and plotting: *tidyverse*. The code used in the session will be uploaded at https://github.com/uasolo/FPCA-phonetics-workshop, where didactic material and example scripts on GAMs and FPCA from past tutorials are available.

Learning goals: To be able to set up a UTI static/dynamic analysis pipeline based on GAMs/FPCA+LMER, to understand how to read results, to appreciate differences between the two approaches. Participants who do not engage with UTI data will still profit from the session, as the concepts applied here can be reused with several other types of speech data.

The use of artificial data has the purpose to concentrate the learning on the concepts rather than on the specifics of a single dataset/problem. Special attention is given to the geometry of data representation and modelling, which gets particularly involved in the dynamic case, where time, angle and radius dimensions intertwine with principal component directions.

The session is meant to be an entry point to introduce FPCA and GAMs as general frameworks for curve data analysis. For a tutorial introduction, see [2] for FPCA and [3] for GAMs.

Preparations: The session is organized as a 'hands-off' illustration of the methods. Participants can follow along while the code is illustrated and executed and do not need to pre-install anything on their laptop.

The exposition will focus on GAMs and FPCA, which will be introduced in some detail, while familiarity with linear regression as well as with the *tidyverse* ecosystem (essentially *dplyr*, *tidyr* and *ggplot2*) will be assumed.

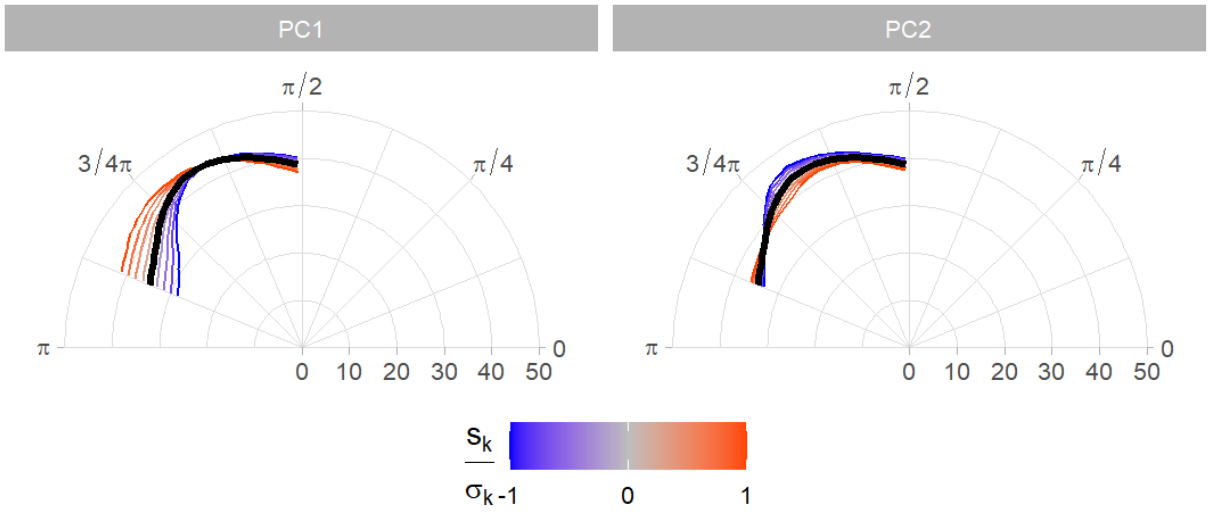


Figure 1: *An example of static FPCA on an artificial dataset of tongue contours. Each PCk captures a different shape variation in the data set, parametrized by score s_k, (k=1 left, k=2 right). The black contour is the mean shape, color-coded contours show the relation between score s_k (normalized by its standard deviation σ_k) and the shape modelled by PCk.*

References:

[1] Al-Tamimi, Jalal, and Pertti Palo. "Dynamics of the tongue contour in the production of guttural consonants in Levantine Arabic." *20th International Congress of Phonetic Sciences (ICPhS)*. 2023.
[2] Michele Gubian, Francisco Torreira, and Lou Boves. "Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts." *Journal of Phonetics* 49 (2015): 16-40.
[3] Martijn Wieling. "Analyzing dynamic phonetic data using generalized additive mixed modeling: A tutorial focusing on articulatory differences between L1 and L2 speakers of English." *Journal of Phonetics* 70 (2018): 86-116.

Voice Conversion in Pathological Speech: Applications and Challenges

B. Mayrhofer[1,3], M. Hagmüller[1,3] & P. Aichinger[2,3]

1) Signal Processing and Speech Communication Laboratory, Graz University of Technology
2) Medical University of Vienna, Department of Otorhinolaryngology, Div. Phoniatrics-Logopedics
3) Medical University of Vienna, Comprehensive Centre for AI in Medicine

benedikt.mayrhofer@tugraz.at

Background and objectives: Voice disorders can significantly impair human communication, affecting social participation, self-esteem, and quality of life [1]. Traditional solutions like the electro-larynx (EL) device often produce unnatural and robotic speech, making communication difficult [2]. Recent advancements in artificial intelligence, particularly in voice conversion (VC), promise improvements by converting pathological speech into more natural-sounding speech [3, 4, 5]. This research evaluates the effectiveness of state-of-the-art VC models, including FreeVC [6], QuickVC [7], LLVC [8] and XVC [9], targeting dysphonic and EL speech, highlighting their potential for clinical applications and personalized voice rehabilitation.

Materials and methods: The models were fine-tuned on Austrian-German speech using the GRASS Corpus [10]. Objective evaluation involved intelligibility assessment using Word Error Rate (WER), Speaker Similarity and neural Mean Opinion Scores (MOS), while subjective assessments involved ratings from 93 listeners. Participants evaluated naturalness, perceived vocal health, rhythm, and intonation, and provided comparative preference ratings of each model's conversion.

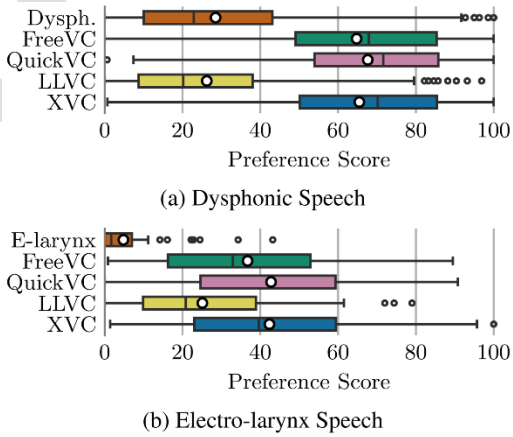


Figure 1: *Listener preference distributions with their mean (white dot) for Dysphonic (Dysph.) and Electro-larynx (E-Larynx) compared to conversions from four VC models. Higher scores indicate greater preference.*

Results and discussion: Significant improvements were observed with FreeVC, QuickVC, and XVC for dysphonic speech. QuickVC and XVC demonstrated the highest performance, achieving MOS scores up to 4.15 and significant increases in listener preference (up to 39%) compared to pathological counterparts. Conversions for EL speech showed lower but still meaningful improvements, highlighting ongoing challenges due to lack of variations in the fundamental frequency (f0) and distorted acoustic characteristics. LLVC, despite computational efficiency, scored lowest in dysphonic speech, underscoring a trade-off between model complexity and conversion quality. Yet, the model still achieved noticeable improvements in EL speech conversion. The results show VC models' effectiveness for dysphonic speech and underline ongoing challenges for EL speech conversions.

Conclusions: The presented VC models have significantly improved speech quality for individuals with dysphonic disorders, yet severe dysphonic and EL speech remain particularly challenging due to inherent constraints like heavily distorted vocal characteristics, constant f0, the absence of unvoiced segments in EL speech, limited prosodic variation, slow speaking rates, noisy qualities, and insufficient parallel datasets. Real-time VC shows additional challenges, particularly the balance between computational complexity and achievable speech quality, as illustrated by LLVC, which achieves reduced speech quality for dysphonic speech conversion. Moving forward, research should focus on increasing model robustness via expanded parallel training sets, enhanced prosodic modeling, and further optimized computational efficiency to realize clinical adoption.

References:

[1] S. M. Cohen, W. D. Dupont, and M. S. Courey, "Quality-of-Life Impact of Non-Neoplastic Voice Disorders: A Meta-Analysis," *Annals of Otology, Rhinology Laryngology*, vol. 115, no. 2, pp. 128–134, Feb. 2006.
[2] A. K. Fuchs, M. Hagmüller, and G. Kubin, "The New Bionic Electro-Larynx Speech System," I*EEE Journal on selected topics in signal processing*, vol. 10, no. 5, pp. 952–961, Aug. 2016.
[3] M. Chu, M. Yang, C. Xu, Y. Ma, J. Wang, Z. Fan, Z. Tao, and D. Wu, "E-DGAN: Encoder-Decoder Generative Adversarial Network for Pathological Voice Conversion," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 5, May 2023, pp. 2489–2500.
[4] D. Ma, L. P. Violeta, K. Kobayashi and T. Toda, "Two-Stage Training Method for Japanese Electrolaryngeal Speech Enhancement Based on Sequence-to-Sequence Voice Conversion," *IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, Jan. 2023, pp. 949-954
[5] Y. Yang, H. Zhang, Z. Cai, Y. Shi, M. Li, D. Zhang, X. Ding, J. Deng, J. Wang, "Electrolaryngeal speech enhancement based on a two stage framework with bottleneck feature refinement and voice conversion," *Biomedical Signal Processing and Control*, vol. 80, ISSN 1746-8094, Feb. 2023
[6] J. Li, W. Tu, and L. Xiao,, "FreeVC: Towards High-Quality Text-Free One-Shot Voice Conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece, June 2023, pp. 1–5.
[7] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "QuickVC: A Lightweight VITS-Based Any-to-Many Voice Conversion Model using ISTFT for Faster Conversion," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, Dec. 2023, pp. 1–7.
[8] K. Sadov, M. Hutter, and A. Near, "Low-latency Real-time Voice Conversion on CPU," arXiv, Nov. 2023.
[9] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, "Using Joint Training Speaker Encoder With Consistency Loss to Achieve Cross-Lingual Voice Conversion and Expressive Voice Conversion," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, Dec. 2023, pp. 1–8.
[10] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: the Graz corpus of Read And Spontaneous Speech," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May 2014, pp. 1465–1470.

Realtime Personalized Deep Learning Voice Morphing for Electrolaryngeal Speech in Polish Language

P. Cyrta, K. Zieliński

Uhura Bionics, Warsaw, Poland

pawel@cyrta.com

Background and objectives: Voice disorders resulting from laryngeal cancer, neurological conditions, or motor speech disorders significantly impact communication and quality of life. Traditional assistive devices like electrolarynx produce mechanical-sounding speech that lacks personalization and natural prosody. This research introduces VoxFlow AI, a wearable voice conversion system designed to transform atypical speech, particularly electrolaryngeal speech in Polish, into more natural-sounding output while preserving linguistic content. Our objective is to develop a voice conversion solution that operates in real-time on wearable hardware, addressing both the technological challenges and practical usability concerns that have limited previous solutions [1,2].

Materials and methods: We developed a specialized corpus of Polish speech featuring both typical speech and atypical speech from individuals with voice disorders. Data collection involved two stages: (1) recording of phonetically balanced materials including vowels, syllables, words, and sentences; and (2) collecting spontaneous speech in three formats: reading a 350-word story, emotional role-playing dialogues, and free-form conversations.

For the voice conversion system, we evaluated several architectures including LLVC [3], FreeVC [4], and QuickVC [5] before developing our custom solution with three components: feature extraction for F0 estimation, U/V detection, mel-cepstrum coefficients, and speech representation; a U-Net-like encoder-decoder for feature mapping; and vocoder implementation using LPCNet [6] and HiFi-GAN [7]. The system was optimized for deployment on ARM-based devices (Raspberry Pi, smartphones) with specialized microphone configurations to maximize real-world usability.

Results and discussion: Our approach successfully addresses the real-time conversion requirements with minimal latency, crucial for maintaining natural conversation flow. The system effectively transforms electrolaryngeal speech in Polish to more natural-sounding output while preserving the speaker's intended linguistic content. Hardware optimization through model compression, quantization, and efficient feature extraction allows the system to run on wearable devices with reasonable battery life. User testing with individuals after laryngectomy demonstrated significant improvement in speech intelligibility and naturalness compared to unprocessed electrolaryngeal speech, with particular success in preserving Polish-specific phonetic features. The wearable form factor with close-mouth microphone configurations proved effective in noisy environments, addressing a key limitation of previous systems.

Conclusions: VoxFlow AI demonstrates that real-time, personalized voice conversion for electrolaryngeal speech in Polish is feasible on wearable hardware. The solution offers people with voice disorders a more natural communication experience while maintaining usability in real-world situations. Future work will focus on expanding language support beyond Polish, reducing hardware size, extending battery life, and integrating the system more seamlessly into speech rehabilitation protocols. We believe this approach

represents a significant step toward restoring more natural communication for individuals with atypical speech.

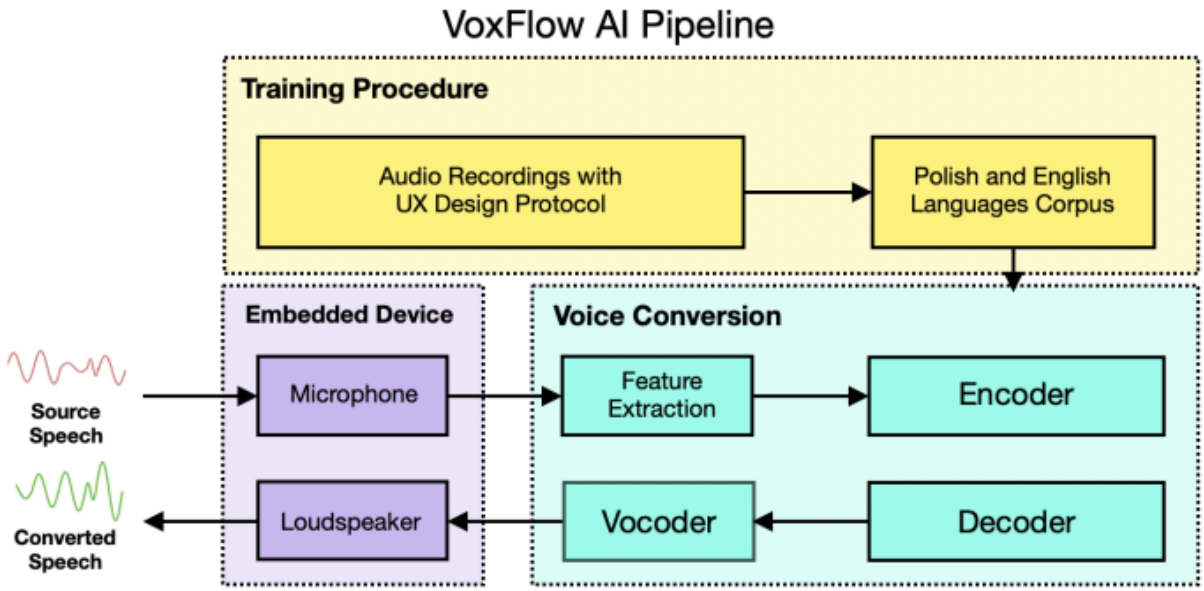


Figure 1: *System diagram*

References:

[1] G. P. Mika, K. Zieliński, P. Cyrta, M. Grzelec, "VoxFlow AI: Wearable Voice Converter for Atypical Speech," in INTERSPEECH 2024, Sep. 2024, pp. 993-994.
[2] K. Zieliński and J. Raczaszek-Leonardi, “A complex human machine coordination problem: essential constraints on interaction control in bionic communication systems,” in CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–8.
[3] K. Sadov, M. Hutter, and A. Near, “Low-latency real-time voice conversion on cpu,” arXiv preprint arXiv:2311.00873, 2023.
[4] J. Li, W. Tu, and L. Xiao, “Freevc: Towards high-quality text-free one-shot voice conversion,” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
[5] H. Guo, C. Liu, C. T. Ishi, and H. Ishiguro, “Quickvc: Any-to-many voice conversion using inverse short-time fourier transform for faster conversion,” arXiv preprint arXiv:2302.08296, 2023.
[6] J.-M. Valin and J. Skoglund, “Lpcnet: Improving neural speech synthesis through linear prediction,” in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019, pp. 5891–5895.
[7] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” Advances in neural information processing systems, vol. 33, pp. 17 022-17 033, 2020.

A Medical Application of ASR and LLM

I. Jánoki[1,2,3], Zs. Nagy[4], Á. Gasparics[4], Á. Jermendy[4], E. Zsáry[4], P. Földesy[1,2] & P. Mihajlik[5]

1) Pázmány Péter Catholic University, Budapest, Hungary
2) HUN-REN Institute for Computer Science and Control, Budapest, Hungary
3) Medicor Elektronika Zrt.
3) Semmelweis University, Budapest, Hungary
5) Budapest University of Technology and Economics, Budapest, Hungary

janoki.imre.gergely@sztaki.hun-ren.hu

Background and objectives: During a meeting with physicians from Semmelweis University (SU), the need for automated anamnesis transcription was identified. The requirements for such a system include ease of operation, accurate transcription of spoken postpartum anamnesis and medical records, immediate feedback with the option for correction, and automatic generation of a digital record conforming to a predefined structure and format. While similar tools exist with varying degrees of accuracy [1,2], the specialized medical terminology of specific fields limits the universal applicability of even high-performing AI-based solutions. These systems typically require separate training or fine-tuning for each field. Concurrently, advancements in AI architectures and models are continually improving overall performance, suggesting the potential for superior results in specialized fields following fine-tuning.

Materials and methods: At the outset of this research, we sought a suitable, state-of-the-art text-to-speech architecture with an available, commercially usable pretrained model. OpenAI's Whisper [3] best met these requirements, and it includes a pretrained model for the Hungarian language. Although Whisper was not originally designed for real-time use, Whisper.cpp [4], a fast, community-driven C++ implementation, allowed us to achieve near-real-time continuous transcription of Hungarian speech with acceptable accuracy in everyday language. The system was tested on an Intel i7-12700H CPU coupled with an Nvidia 3070 (laptop) GPU.

However, a specialized dataset was needed to fine-tune the model for postpartum anamnesis transcription. As no such voice dataset was publicly available, we created our own. We requested a set of anonymized anamneses from doctors at *Semmelweis University (SU)*. The provided text encompassed approximately 3500 words across 20 cases. Due to limited human resources, a significantly larger dataset was not feasible. Therefore, we employed various generative language models to create artificial, synthetic anamnesis texts. These synthetic texts were subsequently reviewed and validated by a medical resident and PhD candidate specializing in the field.

Next, we are using the synthetic data, combined with a portion of the original data, to create a corresponding voice dataset. To achieve this, we are compiling a specialized text containing the core text data, augmented with all the different pronunciations of each critical term and abbreviation. Human participants are asked to read this text aloud, articulating each pronunciation, while the reference "label" text contains only the standardized written form of the words. This approach may guide the AI during fine-tuning, enabling it to transcribe different pronunciations to the same written word.

Furthermore, to expand the voice dataset, we are generating synthetic voice data using text-to-speech (TTS) AI solutions. Two suitable and convincing models supporting the Hungarian language were selected for this task: XTTS-v2 [5] and ElevenLabs TTS [6].

Results and discussion: As this research is ongoing, we present our progress below. We have developed and tested an environment that will serve as the foundation for a product, incorporating real-time Whisper transcription with color-coded confidence feedback. On an Nvidia 3070 (laptop) GPU, the transcription delay was only 3 seconds. We have collected a baseline anamnesis dataset of 3500 words in text format to serve as a foundation for the dataset. Using multiple generative large language models we have expanded the text corpus to a total of 17,500 words currently under revision.

Currently, our medical resident is reviewing the text corpus and augmenting it with the various pronunciations. Concurrently, we are experimenting with XTTS-v2 and ElevenLabs TTS to prepare for synthetic voice data generation. The codebase for fine-tuning Whisper is also ready.

In the future, we will first evaluate three different Whisper pretrained models on the resulting voice dataset. The best-performing model will be fine-tuned using the voice data and subsequently evaluated on independent real data. The resulting transcription may be processed by a large language model to summarize it and format it according to a predefined structure and format.

Conclusions: In this abstract, we have outlined the steps taken, and those that lie ahead, in creating an AI model and software for real-time transcription of postpartum medical records and anamneses. The research and development are still in progress. While the results of similar approaches appear promising, a satisfactory outcome cannot be guaranteed. Because we intend to use synthetic data at multiple stages of dataset generation and training, rigorous manual checking and validation at all stages are crucial to prevent the accumulation of errors.

Acknowledgements: Project no. 2023-2.1.2-KDP-2023-00011 C2309621 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the KDP-2023 funding scheme.

References:

[1] “Speech Recognition - Belux.” en.belux.hu. Accessed: Mar. 21, 2025. [Online]. Available: https://en.belux.hu/speechrecognition
[2] “SpeechTex Jogi Beszédleíró.” SpeechTex Jogi Beszédleíró. Accessed: Mar. 21, 2025. [Online]. Available: https://speechtex.com/
[3] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, ‘Robust Speech Recognition via Large-Scale Weak Supervision’, arXiv [eess.AS]. 2022.
[4] “GitHub - ggerganov/whisper.cpp: Port of OpenAI's Whisper model in C/C++.” GitHub. Accessed: Mar. 21, 2025. [Online]. Available: https://github.com/ggerganov/whisper.cpp/tree/master
[5] “GitHub - coqui-ai/TTS: 🐸💬 - a deep learning toolkit for Text-to-Speech, battle-tested in research and production.” GitHub. Accessed: Mar. 21, 2025. [Online]. Available: https://github.com/coqui-ai/TTS
[6] “Free Text To Speech Online with Lifelike AI Voices | ElevenLabs.” ElevenLabs. Accessed: Mar. 21, 2025. [Online]. Available: https://elevenlabs.io/text-to-speech

The Mere-Measurement Effect in Patient-Reported Outcomes: A Randomized Controlled Trial with Speech Pathology Patients

P. Long[1], V. Ritschl[1,2], T. Stamm[1,2], P. Aichinger[3]

1) Medical University of Vienna, Centre for Medical Data Science, Institute of Outcomes Research, Vienna, Austria
2) Ludwig Boltzmann Institute for Arthritis and Rehabilitation, Vienna, Austria
3) Medical University of Vienna, Department of Otorhinolaryngology, Division of Phoniatrics-Logopedics, Speech and Hearing Science Lab, Vienna, Austria

preston.long@meduniwien.ac.at

Background and objectives: Patient-reported outcome measures (PROMs) are frequently used to assess patient perspective and quality of life across a variety of speech disorders. However, currently little is known about the effect of item word choice, emotional valence, or frequency of use on the responding patients. This effect is referred to as the mere-measurement effect, which is the effect of subjects having their perceptions and/or behaviours on the inquired topic affected simply through the act of being exposed to psychological measurements [1,2]. That is, 1) does a positive or negative wording of items affect the patient's perspective on the latent variable, 2) is there a degree of subliminal influence or measurement effects on their behaviour resulting from exposure to PROs, and finally, 3) is such an effect amplified with repeated exposure? This study will assess the impact that questionnaire item wording and frequency of exposure has on speech pathology patient's self-perception, speech production, and self-esteem.

Materials and methods: This study follows a 2x2 design. Participants responding to our social media adverts are randomly assigned to one of four groups, using a randomizer software [3]. Each group receives online questionnaires at different frequencies (two high and two low exposure groups), and with modifications of the wordings (two original and two positive wording groups). The questionnaires are the Voice Handicap Index (VHI) [4,5], and the Communicative Participation Item Bank (CPIB) [6]. In the high and low exposure group, a subject completes the questionnaires two and four times respectively, separated each time by one week. Lastly, each participant is asked to record online with his/her own device a circa 45 seconds audio file containing text readings, to complete the Rosenberg self-esteem scale, and three disease activity items. In addition, a healthy case control group is recruited while undergoing case matching with respect to age and gender.

In contrast to the groups completing original VHI and CPIB questionnaires, the other two groups complete questionnaires with positively modified items. In particular, while many items of the original versions of the VHI and the CPIB assume the presence of a handicap or medical condition, the positively modified versions rather ask about abilities instead. For example, for positive wording modifications, instead of asking in the CPIB "Does you condition interfere with asking questions in a conversation?" we modified the item to "Regarding your ability to speak, is it easy for you to ask questions in a conversation?". In the VHI, for example, instead of the statement "I am less outgoing because of my voice problem", we use the statement "My voice enables me to be outgoing".

Results and discussion: The study remains active. To date, 71 subjects have been enrolled and 54 have completed the trial, resulting in a dropout rate of 19 %. For a preliminary interims analysis, we randomly selected 17 subjects who had received the original VHI questionnaires. Here, the mean estimate differed between the first and final exposure by 10 of 120 points, approximately, reflecting a worsening of the PROM (First exposure mean: 76.5; Last exposure mean: 86.75). In contrast, other randomly selected 17 subjects from the positive experimental condition showed no such difference (First exposure mean: 93.71; Final exposure mean: 93.86). Both paired t-tests turned out to be non-significant, however.

Conclusions: As the study is ongoing, it would be imprudent to draw any formal conclusions at this time. Currently, neither of the hypothesis test returned a significant value, the sample sizes were small, although an early trend may have been identified. A review of the means indicates a noticeable worsening of the PROM over time in the original condition, however, there is almost no change observed in the positive. This may be preliminary evidence for the effect of wording. In addition, we see an increased drop-out rate in the negative condition compared to the positive and control, suggesting the negative wording may be a demotivation to continue. For the future, we plan to look at the influence of exposure frequency, the CPIB, the Rosenberg self-esteem and disease activity. Finally, assessment of the speech recordings may help to differentiate changes in the PROM that are triggered by worsening of speech production and changes triggered by the mere-measurement effect. Such assessment may be based on detailed manual annotation, but also automated approaches to ASR and/or pronunciation assessment.

References:

[1] P. A. Long *et al.*, “The mere-measurement effect of patient-reported outcomes: a systematic review and meta-analysis,” *Qual. Life Res.*, no. 0123456789, 2025, doi: 10.1007/s11136-025-03909-y.
[2] V. G. Morwitz and G. J. Fitzsimons, “The Mere-Measurement Effect: Why Does Measuring Intentions Change Actual Behavior?,” J. Consum. Psychol., vol. 14, no. 1–2, pp. 64–74, 2004.
[3] “MedUni Vienna Randomizer Website.” online: https://www.meduniwien.ac.at/randomizer/login. (Accessed 28.03.2025)
[4] B. Jacobson et al., “The Voice Handicap Index (VHI): Development and Validation,” J. Speech-Language Pathol., vol. 6, no. 3, pp. 66–70, 1997.
[5] G. Chow, M. Scher, G. P. Krisciunas, and L. F. Tracy, “Comprehensive Review of Multilingual Patient-Reported Outcome Measures for Dysphonia,” J. Voice, pp. 1–7, 2025, doi: 10.1016/j.jvoice.2025.01.005.
[6] C. Baylor, T. Eadie, and K. Yorkston, “The Communicative Participation Item Bank: Evaluating, and Reevaluating, Its Use across Communication Disorders in Adults,” Semin. Speech Lang., vol. 42, no. 3, pp. 225–239, 2021, doi: 10.1055/s-0041-1729947.

Speech-Based Cardiorespiratory Health Monitoring with VOICE-BIOME: a scalable voice biomarker platform

P. Pombala[1], L. Nippet[1], J. Hoxha[1] and S. O. Simons[2]

1) ZANA Technologies GmbH, Karlsruhe, Germany
2) Department of Respiratory Medicine, Institute of Nutrition and Translational Research in Metabolism (NUTRIM), Maastricht University, Netherlands

prashanth@zana.ai

Motivation: Speech carries rich physiological signals that can be used to monitor a patient's health status non-invasively. In recent years, voice-based technologies have shown promising results in detecting symptoms related to chronic conditions such as Chronic Obstructive Pulmonary Disease (COPD) and Heart Failure (HF). However, most research in this area focuses narrowly on model performance, while often neglecting the broader infrastructure needed to deploy such systems in real-world settings. Zana offers an AI-driven, device-agnostic voice biomarker platform (VOICE-BIOME) that leverages the human voice as a novel biomarker for health monitoring. The cloud-based VOICE-BIOME solution delivers objective and unbiased data on a patient's disease progression and treatment response. A key innovation is the longitudinal analysis of voice data to predict critical events, such as heart decompensation in Heart Failure patients or exacerbations in COPD/asthma patients. Our results demonstrate the effectiveness of our methods: In heart Failure detection, the VOICE-BIOME platform achieves over 81% sensitivity and 79% specificity in identifying decompensated HF using voice features alone. For COPD, 79.0% ± 3.0% of exacerbations are detected at least three days before onset, and 85.1% ± 1.7% are identified by the onset day or earlier using voice measures alone.

This session introduces VOICE-BIOME, a modular and adaptable voice biomarker platform for speech-based health monitoring that supports data collection, acoustic feature processing, continuous retraining, and results visualization—all integrated into a streamlined pipeline. Instead of centering on specific machine learning algorithms, we aim to showcase how the underlying platform enables reproducibility, cross-study comparability, and seamless integration with clinical workflows. We hope to open a discussion on how the VOICE-BIOME platform can be customized for different research contexts and clinical needs. An Overview of the modular speech processing and monitoring pipeline is given in Figure 1.

Data and tools: The session will demonstrate our VOICE-BIOME platform using an in-house dataset comprising voice recordings from patients with chronic conditions such as COPD and Heart Failure. The recordings include structured speech tasks like vowel prolongation and text reading, collected via a mobile application. To facilitate daily voice data collection, we use a Digital Companion that encourages regular voice input through structured tasks, seamlessly integrating into users' routines to promote adherence and ensure reliable data capture. The dataset consists of 204 participants, contributing a total of 298 hours of recordings across 41,034 speech samples. The participants are categorized into three cohorts: Chronic Obstructive Pulmonary Disease (COPD), Asthma, and Acute Heart Failure (AHF). The dataset includes 19,868 vowel recordings, 13,860 text readings, and 7,306 spoken answers, providing a diverse set of voice data for analysis. Participants in the COPD and Asthma cohorts were tracked with daily voice recordings

over a 3-month period, while those in the Heart Failure group were monitored during hospitalization with decompensated HF and for six months post-discharge.

Demographic distribution varies across cohorts, with a mix of male and female participants. In addition to speech recordings, the dataset includes questionnaire responses, such as EXACT, and physiological measurements, including weight, ejection fraction, and NT-proBNP levels. We will showcase how the VOICE-BIOME platform handles data ingestion, novel acoustic feature extraction, automated retraining, and visualization through a modular pipeline. All demonstrations will be conducted using preloaded data, and participants are not required to bring any datasets or tools of their own.

Learning goals: By the end of the session, participants will gain:

- An understanding of speech-based feature extraction and machine learning for health monitoring – including how acoustic features relevant to respiratory and cardiovascular conditions can be derived from recorded speech and used for exacerbation detection.
- Familiarity with a reusable end-to-end framework for speech analysis – learning how a modular pipeline handles data ingestion, feature processing, automated model retraining, prediction storage, and clinical dashboard integration.

Preparations: Participants do not need to bring anything to the session. All necessary data, tools, and demonstrations will be provided and presented live. No prior setup, software installation, or data preparation is required.

Figure 1: *End-to-End Speech-Based Health Monitoring Framework: A scalable platform for collecting, processing, and analyzing patient speech data to detect health conditions. The system integrates real-world data collection, feature extraction, biomarker analysis, and secure data management to provide actionable health insight.*

Deep Learning ASR for a Patient with Permanent Tracheostomy

D. Nadrchal

Institute of Computational Perception, Johannes Kepler University Linz, Austria

k12213656@students.jku.at

Motivation: We present an ongoing case study on the development of a personalized automatic speech recognition (ASR) system for a patient with a severe speech disability caused by a permanent tracheostomy and potential neurological damage resulting from a debilitating car accident. Conventional ASR systems fail to accurately recognize the patient's speech, highlighting the need for tailored solutions. We analyze the specific ways in which the patient's speech differs from that of healthy speakers and propose a series of adaptations to improve ASR performance. Specifically, we fine-tune Whisper [1], a state-of-the-art deep learning-based ASR model, to the patient's speech. In addition, we introduce a publicly available dataset of the patient's recordings and provide a high-level guide on adapting deep learning models to such custom data. Our fine-tuned Whisper model achieves near-healthy-voice accuracy on in-domain data. To illustrate the capabilities and limitations of the model in its current state, we will provide an interactive demonstration. Furthermore, we discuss key challenges in deploying personalized ASR systems, lessons learned from this development process, and potential future improvements to enhance speech recognition accuracy in real-world scenarios.

Data and tools: Only the data collected from our patient will be used for the session. We will use the web service Google Colab (colab.research.google.com) for their inspection, therefore no pre-installation is required.

Learning goals: On the theoretical side, the participants will learn our strategy for collecting a large volume of speech data from a patient using *artificial conversations* generated by an LLM (ChatGPT) and a voice conversion (Voice.ai). Furthermore, they will learn how to fine-tune a state-of-the-art ASR system on a custom dataset. We will provide them with our codes for these tasks. On the practical side, participants will gain hands-on experience with recognising an impeded speech using deep-learning models. They will learn about the strengths and limitations of this approach and they will see several examples.

Additionally, we will discuss the real-world applicability of our system and invite participants to share insights on how it could be effectively deployed.

Preparations: To participate in the hands-on parts of the session, a laptop with a web browser is required.

Acknowledgement: This work was done as a bachelor's thesis in cooperation with the Institute of Computational Perception at Johannes Kepler University Linz and under the supervision of Dipl.-Ing. Florian Schmid

Table 1: *This table presents word error rates (WER) for the two smallest versions of the Whisper model—tiny (39M parameters) and base (74M parameters)—under two training setups: the standard pipeline and our adapted version. These models were evaluated on three datasets:* ***(1)*** *healthy Czech speech from the Common Voice dataset [2], which we used for pre-training,* ***(2)*** *in-domain patient speech (artificial conversations), and* ***(3)*** *out-of-domain patient speech (real-world conversations).*

		Dataset		
Whisper version	**Pipeline**	Common Voice Czech [2]	Patient: In-domain	Patient: Out-of-domain
Tiny	Standard	0.45	0.44	0.98
	Adapted	0.34	0.40	0.85
Base	Standard	0.30	0.37	0.83
	Adapted	0.31	0.35	0.82

References:

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever. "Robust Speech Recognition via Large-Scale Weak Supervision," in *International Conference in Machine Learning*, Honolulu, HI, 2023, pp.28492-28518

[2] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, G. Weber. "Common Voice: A Massively-Multilingual Speech Corpus," in International Conference on Language Resources and Evaluation, Marseille, France, 2020, pp. 4218-4222

Physically-Based Machine Learning for Vocal Fold Video Data Interpretation

C. Drioli

Dept of Mathematics, Computer Science, and Physics (DMIF), University of Udine, Italy

carlo.drioli@uniud.it

Background and objectives: High-speed video data of vocal folds provides a full-field visualization of vocal fold dynamics with high time resolution, allowing for accurate and detailed inspection of vibratory patterns of the folds. Applications in clinical diagnostics include the precise assessment of pathologies and treatment evaluation, and more in general, it is a fundamental tool for the understanding of the phonation mechanisms and the validation of models of the folds. In this investigation, HSV data are used to explore the potential of a numerical biomechanical model of the folds presented recently [1], which has the capability of generating oscillatory patterns of the vocal folds along the vertical and the sagittal dimensions. The model has a low computational complexity structure, yet is able to reproduce most of the fold edge patterns commonly observed in HSV data. This makes it suitable to be combined with parametric optimization and machine learning methods, in the aim of obtaining the accurate fitting of the model to real data. Among the possible application of a visual data-matched model based on physical insights, we cite the possibility of generating realistic synthetic data, which is a desirable tool for researchers and clinicians, as well as a necessity for technological applications involving the training of machine learning tools from large amount of data. Previous studies with this respect include [2], in which a variational autoencoder is trained to enable the creation of synthetic GAWs that closely replicate real-world data. With this study we aim at extending this approach by having a biomechanical model generating the prototypical oscillation, and the generative model transforming the sequence of prototypical oscillations into photorealistic frames.

Materials and methods: The model used to generate the simplified motion of fold edges during the glottal cycle, is a lumped/distributed elements mechanical model, described in details in [1]. It is worth to recall here that it includes a 3D description of the fold cover, providing the description of fold displacements along the vertical and the sagittal axes, which in turn allows to obtain a time-varying visual representation of the folds seen from above comparable to the one of the HSV data. The aim here is to use this prototypical description of the fold oscillation as the input to a Neural Network able to process video data, which is trained to turn different configuration of the folds during the cycle into realistic images captured by the high speed video endoscope. The procedure consists into the following steps:

1. tuning of the biomechanical model parameters to provide a prototypical oscillatory behavior of vocal fold edges, consistent with the fold oscillation observed in the reference HSV data. This step necessitates of the synchronization of the oscillations.

2. build the training dataset by collecting a sequence of paired label frame-video frame data, and train the image transformer model.

3. validate the trained model by measuring the accuracy of the synthetic video sequence if compared to the training one and to comparable data (e.g., by the same subject) not used during training.

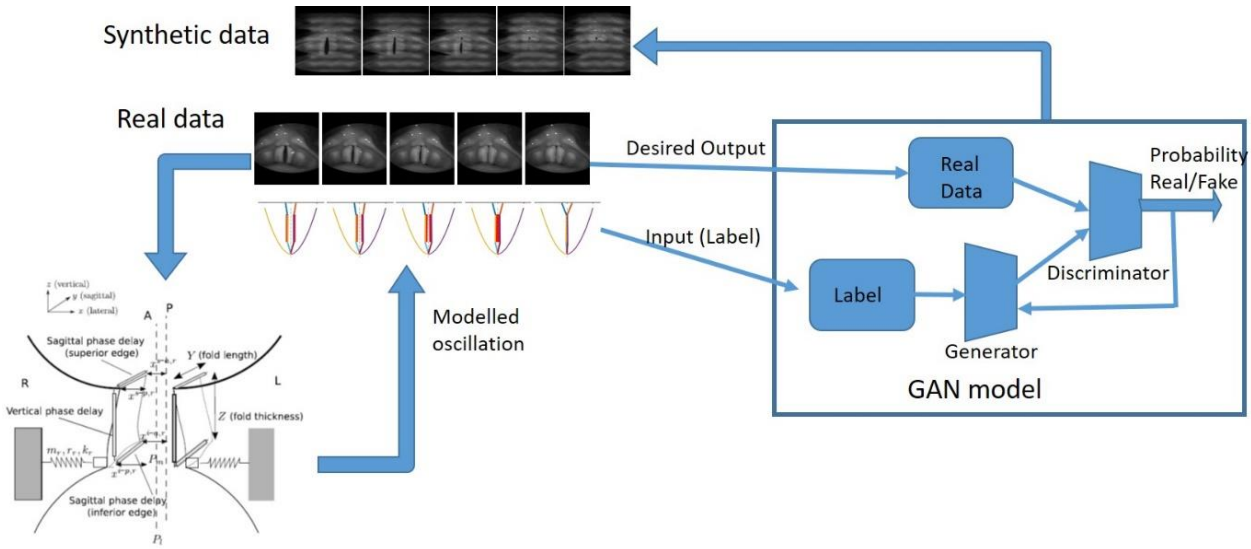


Figure 1: *Overview of the training data setup and synthetic data procedure. On the lower-left side of the picture, the glottal model to generate the input data, used in turn to train the GAN model (center-right) to imitate the real video data*

Results and discussion: Preliminary experiments conducted on a reduced set of HS video samples from BAGLS dataset, and using a conditional GAN [3] trained on the paired label-image sequences obtained as discussed, have shown the feasibility of the approach, although the quality of results is very preliminary. Much of the accuracy of the training seems to depend on the accuracy of the model configuration and tuning with respect to HSV data, which is not a trivial task.

Conclusions: To extend this preliminary study, the following steps are necessary: complete the necessary refinements to the model; perform a thorough assessment conducted on a large set of samples, to demonstrate the generality of the method, using phonatory data from healthy subjects (but uttering at different phonation settings); evaluation of the synthetic video sequences by clinicians; assessing the model in increasingly complex conditions, e.g. with asymmetrical or irregular fold oscillations, as in pathological voices.

References:

[1] Carlo Drioli, Philipp Aichinger, Modelling sagittal and vertical phase differences in a lumped and distributed elements vocal fold model, Biomedical Signal Processing and Control, Volume 64, 2021
[2] Mahdi Darvish, Andreas M. Kist, A Generative Method for a Laryngeal Biosignal, Journal of Voice, 2024,
[3] Nuha Aldausari, Arcot Sowmya, Nadine Marcus, and Gelareh Mohammadi. Video Generative Adversarial Networks: A Review. ACM Comput. Surv. 55, 2, Article 30 (February 2023)

Sampling Rate Bias of Vocal Jitter and Shimmer

J. Schoentgen[1], A. Kacha[2] & F. Grenez[1]

1) Université Libre de Bruxelles, Brussels, Belgium
2) Université de Jijel, Jijel, Algeria

jean.schoentgen@icloud.com

Background and objectives: Acoustic features that report vocal cycle length or amplitude perturbations are biased when they involve data sampled at the pace of the vocal cycles. The reason is that the features include high-pass filtering of the cycle lengths or amplitudes to separate fast (jitter or shimmer) from slow perturbations (drift, declination, physiological or neurological tremor and flutter). Indeed, the analog cut-off frequency of a digital filter depends on the sampling frequency that is equal to the vocal frequency (f_o) when the speech signal is sampled once every cycle. Jitter or shimmer are therefore segregated from slow perturbations at a frequency that varies with the vocal frequency. The reported size of jitter or shimmer consequently is token-dependent loose from any physiological or anatomical causes.

We have compared two unbiased (i.e. fixed-rate) to respectively five and six biased (i.e. variable-rate) jitter and shimmer features that are conventionally obtained by means of the software PRAAT version 6.4.43. The purpose has been to examine the correlation with f_o and the intra-corpus dispersion. Both are expected to be lower for fixed-rate than for variable-rate features. Another objective has been to estimate the expected error when replacing fixed-rate by variable-rate features. The latter are the only option available in popular analyses software such as PRAAT or MDVP. Fixed-rate and PRAAT features cannot be compared directly because they involve unequal gains and bandwidths. The feature values have therefore been replaced by their ranks. That is, two features are equivalent when they rank vowel sounds identically. The root mean square error reported below is the average disagreement in number of ranks between fixed-rate and PRAAT features.

Materials and methods: The corpus has included 2-sec vowel [a] fragments sustained by 54 normophonic speakers. The number of male and female speakers has been equal to 18 and 36. All the vowel sounds have been type I (that is, they have involved one single frequency of vibration) and the recordings have been sampled at 44100 Hz. Eighteen female and 18 male vowel sounds have been downloaded from the website of the ATIC Research Group, Dept. Ingenieria de Comunicaciones, Universidad de Malaga (ATIC Research Group, 2018). The chronological age of the speakers has been equal to 45 +/- 11 years and all the speakers have been labelled grade 0 on the GRABS scale. An additional 18 vowel sounds sustained by normophonic female speakers have been provided by the Speech Therapy Department, Faculty of Psychology, Speech Therapy and Education Sciences, University of Liège, BE.

Estimating cycle amplitudes and lengths as well as replacing variable-rate by fixed-rate features has involved the following steps. First, the up-sampled (4 x) vowel sounds have been zero-phase low-pass filtered at 1000 Hz to remove any spurious peak perturbations due to the super-imposed oscillations of higher formants. Second, the amplitudes of the main cycle peaks have been obtained via a conventional peak picking routine. Third, the observed cycle lengths and amplitudes have been re-sampled at a fixed rate equal to 800 Hz via a smooth interpolation and the averages have been subtracted to obtain the total length and amplitude perturbations. Fourth, the resampled perturbations have been high-pass filtered at a fixed

cut-off frequency equal to 20 Hz via an orthogonal discrete cosine transform. The upper bound of the high-pass filtered perturbations has remained equal to half the vocal frequency. Finally, the high-pass filtered cycle length and amplitude perturbations have been summarized via the relative standard deviations, which is a natural choice because the orthogonal discrete cosine transform conserves the energy of the total perturbations. The variable-rate features have been obtained by implementing exactly the definitions provided in the PRAAT manual.

Results and discussion: Table 1 summarizes the properties of the PRAAT and fixed-rate features. The Table reports the rank correlation with f_o, the relative inter-quartile range, the root mean square error in number of ranks as well as the rank correlation between variable-rate and fixed-rate features for jitter (top) and shimmer (bottom). The correlation with f_o and the relative inter-quartile range are lowest for the fixed-rate features, as predicted. However, the distinction between PRAAT and fixed-rate features is more evident for shimmer than for jitter. The root mean square error is indeed halved for the jitter features and the rank correlation between fixed-rate and variable-rate features is higher. The reason is that the jitter features involve a normalization by the average cycle length, which (in part) compensates for the dependence of the raw cycle length perturbations on f_o. This compensation is lacking in the shimmer features as well as the jitter feature called *local, abs*. Removing the excess dependence on the vocal frequency of variable-rate features is not the only reason for re-sampling cycle lengths and amplitudes at a fixed rate. Equally relevant motivations are being able to segregate fast from slow perturbations at a frequency that is perceptually meaningful as well as being able to process cycle amplitude, cycle length and Fourier frequency time series by means of conventional signal processing methods that assume that the data have been sampled at a constant rate.

Conclusions: The excess dependence of popular features of shimmer and jitter on the vocal frequency as well as their arbitrary segregation from slow perturbations disregarding any auditorily determined boundary between roughness and tremulousness invites a nuanced rereading of the literature. Especially, published correlations between features, between features and vocal frequency as well as between features and perceptual scores of roughness could be subject to reinterpretation.

Table 1: *Absolute value of the rank correlation with f_o, relative inter-quartile range, root mean square error in number of ranks, as well as rank correlation of the variable-rate and fixed-rate features for jitter (top) and shimmer (bottom). The nomenclature of the variable-rate features agrees with the voice report in PRAAT. The relative standard deviations of the fixed-rate features are reported in the right-most column.*

	local	local, abs	rap	ppq5		ddp	rel. std_{jit}
|corr(f_o)|	0.32	0.59	0.23	0.33		0.23	0.20
rel. IQR	0.71	1.40	0.59	0.79		0.59	0.57
RMSE	4.58	10.04	3.84	4.23		3.86	0.
corr(rel. std_{jit})	0.84	0.60	0.86	0.85		0.86	1.
	local	local, dB	apq3	apq5	apq11	dda	rel. std_{shim}
|corr(f_o)|	0.41	0.41	0.41	0.47	0.42	0.41	0.21
rel. IQR	0.80	0.80	0.86	0.90	0.86	0.86	0.64
RMSE	7.88	7.88	9.87	7.95	7.18	9.87	0.
corr(rel. std_{shim})	0.70	0.70	0.64	0.71	0.74	0.64	1.

Resonance Frequencies in Non-Rigid Compressed Waveguides

A. Eliraki, F. Vixege, X. Pelorson & A. Van Hirtum

CNRS, Grenoble INP, LEGI, University of Grenoble Alpes, France

annemie.vanhirtum@univ-grenoble-alpes.fr

Background and objectives: Voiced speech sounds are characterised by distinct resonance frequencies. The underlying physics is commonly examined using simplified waveguide geometries that replicate the constricted regions observed in human speakers. While such studies have successfully validated key aspects, such as the influence of constriction position and degree on resonance frequencies, they predominantly utilise rigid waveguides. This constraint limits the exploration of dynamic articulation and the subsequent production of varying vowels. To address this limitation, the present study investigates circular waveguides with wall properties ranging from rigid to elastic. An experimental analysis of resonance frequencies is conducted for both unconstricted waveguides and those featuring one or two constrictions, with systematic variation in the position and degree of constriction. An analytical model of the deformed waveguide geometry is employed, and critical geometrical parameters are validated against experimental data across different wall properties. The geometrical model is then employed to simulate acoustic wave propagation within the waveguides. Comparisons between modelled and experimentally measured resonance frequencies are presented, with particular attention given to the influence of waveguide wall materials on the observed acoustic phenomena.

Materials and methods: A circular uniform waveguide with length L_0 is positioned along the x-direction. The waveguide's cross-section is deformed by squeezing it locally between a pincher consisting of two rounded symmetrical bars as illustrated in Fig. 1(a). The squeezing can be expressed as a function of two parameters, namely the constriction position xc and the constriction degree P. The deformed waveguide is modelled by representing each cross-section within the deformed region as a stadium-shaped ring [1]. Three-dimensional rigid waveguides with one or two constrictions are generated based on the model considering a moderate constriction at P=77% and a severe constriction at P=90%. Three distinct constricted waveguide configurations ([A], [B] and [C]) are examined, based on the positioning and degree of one or two constrictions along the waveguide as shown in Fig. 1(b). In addition, uniform waveguides with different elasticity are obtained using both 3D printing and molding techniques. Concretely, and in order of increasing deformability, five wall materials are assessed: R (rigid, 3D printed); D (deformable, 3D printed); S (deformable commercial); DR (deformable, 3D printed); M (deformable, molded). Waveguides are mounted to the experimental setup illustrated in Fig. 1(c). The acoustic pressure is measured along its centreline. A compression chamber is connected to the waveguide inlet (at $x = -L_0$) in order to repeatedly emits a 30-second linear frequency sweep so that the frequency range of 0.1 up to 15 kHz can be studied. A rigid Plexiglas rectangular baffle is attached at the waveguide exit (at $x = 0$). Measured data (acoustic probe and source signal; wall displacement) are analysed as a function of the instantaneous acoustic source frequency by computing its amplitude and phase [1]. Acoustic waveguide resonances are identified at frequencies corresponding to peaks in the acoustic pressure amplitude (in dB). Pressure-pressure transfer functions between different probe positions are quantified.

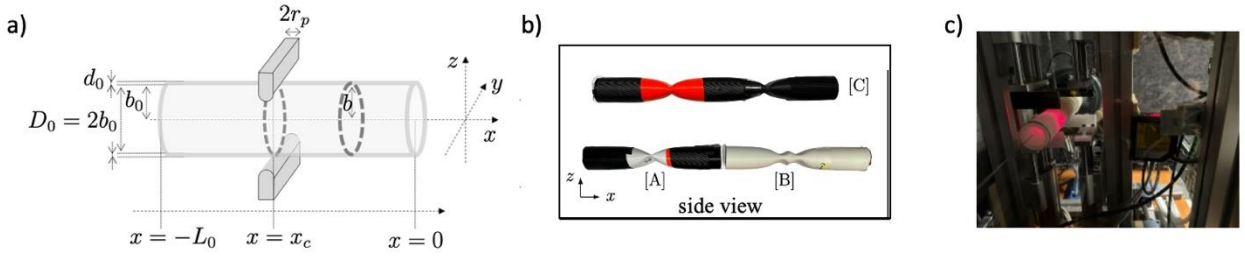


Figure 1: *a) Illustration of a cylindrical waveguide squeezed between two rounded parallel bars at constriction position x_c imposing constriction degree P; b) Prototypes of modelled squeezed waveguides; c) illustration experimental setup allowing to squeeze deformable waveguides.*

Results and discussion: The impact of various wall materials (R, D, S, DR and M) on the measured resonance frequencies for uniform, unconstricted waveguides with lengths L0 ∈ {175, 180} mm is considered. Resonance frequencies across wall materials are mostly consistent, with standard deviations below 6% (< 75 Hz) relative to the rigid wall reference (R). Nevertheless, resonances for deformable materials are generally lower, reflecting increased boundary compliance and energy absorption. However, for molded (M, most deformable) waveguides, the lowest resonance frequency decreases to 210 Hz, with an additional new resonance near 650 Hz. Measured microphone probe and wall displacement spectra indicate strong mode coupling between the acoustic quarter-wave frequency and wall vibration. The same observation holds for squeezed waveguides. This implies that mode coupling is observed for squeezed molded waveguides as well. Measured resonance frequencies are compared to modelled ones obtained using a plane wave model [1]. For frequencies up to 2.5 kHz, the model accurately estimates measured resonance frequencies across all parameter sets and waveguide configurations [A], [B] and [C]. For frequencies higher than 2.5 kHz, the accuracy declines due to the propagation of non-plane higher-order modes. The cut-on frequency of these modes for a uniform circular waveguide portion is approximately 4.4 kHz and decreases with increasing pinching effort, as the constricted section along the major axis increases. To illustrate frequency behaviour up to 10 kHz, modelled and measured pressure-pressure transfer functions are examined.

Conclusions: Acoustic resonances in deformable uniform and compressed waveguides are studied experimentally using an original setup developed to study vocal tract like waveguides dynamics. Experimental results are in close agreement with plane wave model results up to the cut-on frequency of higher order modes. Different degrees of deformability do not significantly affect resonances except for the most deformable waveguide (molded) for which vibro-acoustic coupling is observed affecting the lowest resonance frequency.

Reference:

[1] Van Hirtum, A., Blandin, R., Pelorson, X. Validation of an analytical compressed elastic tube model for acoustic wave propagation, J Appl Phys, 118, 224905, (2015).

Physical Model of Phonation with Reduced and Measurable Parameters

X. Pelorson & A. Van Hirtum

CNRS, Grenoble INP, LEGI, University of Grenoble Alpes, France

xavier.pelorson@legi.cnrs.fr

Background and objectives: While voice adjustment by a speaker or singer can be a very complex phenomenon from a physiological point of view, the physical parameters controlled are few (pulmonary pressure, ab- or adduction of the vocal cords, tension and elongation of the vocal cords, etc.). In contrast, most physical models of phonation rely on a considerably larger number of parameters. The proposal by Ishizaka and Flanagan [1], for example, requires the choice of no less than 26 parameters. While some of these parameters can be inferred from anatomy, for example the geometric dimensions of the vocal folds, others, such as the elastic constants of the springs in the two-mass model, are arbitrary and even often artificial, such as the subdivision of the vocal cords into two asymmetrical parts. In practice, the imbalance between the parameters of theoretical models and those actually controlled by a speaker makes them of little interest for clinical applications.

Materials and methods: In this communication, we present a simplified physical model of phonation, focusing on describing the most important physical phenomena and validated on the basis of experiments on mechanical replicas. Particular attention will be paid to reducing the number of control parameters of the physical model and on how these parameters can be inferred from in vivo observations.

Tissue mechanics is described using an oscillator model with a limited number of mechanical modes. It is shown that, compared to other approaches, this description allows to significantly reduce the number of mechanical parameters. Moreover, most of them can be measured or deduced from experiments on human speakers.

The fluid mechanics of the air flow through the glottis, is described using a quasi-steady viscous flow theoretical model, accounting for dissipation by turbulence downstream of the glottis. Different theoretical models of increasing complexity have been tested against measurements on mechanical replicas of the vocal folds in terms of accuracy but also in terms of their computational costs. Lastly, the acoustic coupling between the vocal folds and the sub- and supraglottal cavities is also accounted.

Results and discussion: Simulations, varying systematically the parameters of the physical model, have been performed. They compared well with reported observations on human speakers, such as the relationship between the Sound Pressure Level and the fundamental frequency of vocal folds vibration with the subglottal pressure, for example. The vibrating mass of the vocal folds, m is the only arbitrary parameter (within anatomical plausible limits, though), however its effect seems to be limited on the simulation outputs.

Further, an improved biomechanical description, based on continuum mechanics, linking the width of the vocal folds to the mechanical resonance frequency, allows to reduce the number of parameters to 7 only.

Conclusions: Compared with other physical models of phonation, this proposal has the advantage of relying on a very limited number of parameters. Further, all these parameters but one can be either measured or inferred from human voicing experiments. Further theoretical work, and experimental validation, is to be done in order to link the tissue biomechanics to the vocal fold geometry. This will lead to further reduction of the number of parameters.

Reference:

[1] Ishizaka, K. and Flanagan, J. L. Synthesis of voiced sounds from a two-mass model of vocal folds. *The Bell System Technical Journal*, 51(6) :1233–1267. (1972)