1st International Seminar on the Foundations of Speech

BREATHING, PAUSING, AND THE VOICE

1st – 3rd December 2019 in Sønderborg, Denmark

Conference Proceedings

Proceedings of the

1st International Seminar on the Foundations of Speech – Pausing, Breathing and Voice

University of Southern Denmark, Sønderborg, Denmark

December 1-3, 2019

Organizers:	Oliver Niebuhr Jana Neitsch Stephanie Berger Kerstin Fischer Jan Michalsky Selina Eisenberger Matouš Jelínek

Credits:

The icons “Turkey” and “Vegetables” from the Gastronomy pack created by Smashicon were used in the statistics page (accessed November 2019 at https://www.flaticon.com/packs/gastronomy-6). The wordcloud on the statistics page was created with the wordcloud2 package in R. The country maps were created using the maps and mapdata packages in R.

Centre for Industrial Electronics

THE EUROPEAN UNION
The European Social Fund
Investing in your future

SDU

Welcome!

Dear colleagues and friends!

David Abercrombie stated in his seminal Elements of General Phonetics (1967) that speech is essentially "movements made audible" and he continues with showing, throughout his book, that "an air stream [...] is the basis of the whole of the sound, in all its variety, of human speech" (p.24).
Communication is inextricably linked with the generation of voice (in humans and other mammals). Breathing is the basis for this voice generation, and communication rests on a complex pattern of voicing, breathing and non-voicing, i.e. pausing. Therefore, breathing, pausing, and the voice together form the basis for all phonetic and phonological aspects of language.

Advancements in the analysis of interactions and dialogues (both human-human and human-machine), technological developments such as the Respiratory Inductance Plethysmograph (RIP), and the growing interest in social prosody have led to many new insights on how. This concerns, for instance, insights into how breathing and speech are coordinated, how systematically and diversely the voice is used in communication and in different speaking styles, how pauses and the perception of (dis)fluent speech are related, and how precisely pauses are prepared, indicated prosodically, and timed with turn-yielding and turn-taking.

The 1st International Seminar on the Foundations of Speech (SEFOS) will be the first event dedicated to the energetic mechanisms, states, and patterns of communication as well as to the multifaceted coloring and communicative functions of the voice. SEFOS aims to bring together researchers from different disciplines, such as phonetics, phonology, psychology, medicine, acoustics, speech technology, and computer linguistics.

It is thanks to the generous funding of the Danish Research Council (7059-00101B) that we all get together here at the University of Southern Denmark (SDU) Sønderborg for the 1st SEFOS conference. We received the funding for studying – in a Danish-Brazilian context – how breathing patterns shape and influence public speaking and its impact on listeners and, moreover, to what degree acoustic voice quality differences (represented by measures like HNR, H1-H2, Hammarberg Index, etc.) are reflected by local and global changes in speech breathing. You will see and hear about some of our results at this SEFOS. However, the main focus is of course on your contributions. We are very pleased to have attracted almost 50 junior and senior researchers from 15 different countries to Denmark at this wintery time of the year; and we hope that you will have a pleasant stay with us here and experience a bit of this typical Christmas "Hygge" in Denmark, i.e. the country where the world's happiest people are at home – well, the 2nd happiest people according to the 2019 statistics. Maybe we will hook you a bit with Danish enthusiasm and Danish agility and creativity during your stay. But that is of course not the only reason, why we are sure that there will be many more editions of SEFOS after this kick-off conference. Its interdisciplinary nature and its pre-phonological, data-driven perspective on the foundations of speech make SEFOS an ideal satellite event for many established phonetic, linguistic and acoustic conference series.

Perhaps it is no coincidence that the first SEFOS takes place at the Mads Clausen Institute (MCI) of the SDU. The SDU is both the third largest and the third oldest Danish university. Since the introduction of the ranking systems in 2012, the University of Southern Denmark has consistently been ranked as one of the top 50 young universities in the world by both the Times Higher Education World University Rankings and the QS World University Rankings. The SDU is also among the top 20 universities in Scandinavia. Within the SDU, the MCI with its 128 employees (64 professors) aims at tackling the key challenges of the 21st century by means of industry-related, applied research and development work – in the fields of nanotechnology, mechatronics, electronics, acoustics, innovation and entrepreneurship. It takes such a variety of disciplines and interests to make interdisciplinarity really work and flourish – and to allow ideas and projects to be born.

One of those "crazy" ideas was to improve the charismatic rhetorical skills of young engineers and start-up founders. Today, this idea of Acoustic Voice Profiling (AVP) has itself become a start-up business, even internationally, and with an annual growth in sales and profits of more than 50 % since 2017. AVP has been developed in Sønderborg – and with continuous support from great colleagues at the Universities of Oldenburg, Nuremberg and Prague – is now a central pillar of SDU's new Acoustic-in-Production Lab, embedded in the Centre for Industrial Electronics. The SEFOS breathing project was our first project that is concluded with the present conference. Today, we conduct applied phonetic and linguistic research not only for their own sake. Inspired by other areas of biotechnology, we are also developing innovative ideas and practical solutions for industrial sound design, noise reduction, and digital-communication tools. These developments are based on the one signal that has, like no other, shaped our human ontogenetic and phylogenetic development and that is, like no other signal, "hardwired in our brains": speech or, more specifically, speech melody. In this conceptual framework, acoustic projects are currently running to improve the sound of actuators and cello strings in a culture-specific way, to enhance the experience of call-centre customers, to measure a patient's pain level, and to reduce the noise of wind turbines in road driers.

If you are interested, we would be happy to tell you a few more details about one or another project of the Sønderborg "Acoustics-in-Production" team. In the meantime, we wish you all a great, inspiring and lively SEFOS.

In this spirit: Welcome to Sønderborg!

Oliver, Jana, Kerstin, Jan, Stephanie, Selina, and Matouš.

SEFOS 2019 Organizing Committee

Oliver Niebuhr

Associate Professor of Communication and Innovation
Centre for Industrial Electronics,
Mads Clausen Institute, SDU

Oliver Niebuhr earned his doctorate in Phonetics and Digital Speech Processing from Kiel University and worked afterwards as a post-doc researcher at linguistic and psychological institutes in Aix-en-Provence and York. In 2009, he was appointed Junior Professor of Spoken Language Analysis and returned to Kiel University, where he also headed the Kiel research center a “Speech & Emotion.” Since 2015, he is Associate Professor of Communication and Innovation at SDU and leads the Acoustics-in-Production group at the Centre of Industrial Electronics. His research focuses on the psycho-acoustic aspects of sound signals, including nonverbal persuasion and negotiation skills, noise perception, digital communication, HRI, and industrial sound design.

Jana Neitsch

Post-doctoral Researcher
Centre for Industrial Electronics,
Mads Clausen Institute, SDU

neitsch@mci.sdu.dk
www.jana-neitsch.com

Jana Neitsch earned her master's degree in General Linguistics with emphasis on phonetics and phonology from the University of Constance, based on a MA thesis about the production and mental storage of German binomials. She also earned her doctorate in Phonetics and Phonology from the University of Constance. Her thesis addressed the prosody of rhetorical questions in German in consideration of context and attitude. In 2018, she spent six months at University of Southern Denmark to advance her research with respect to the interplay between context and the prosodic realization and perception of rhetorical questions. Currently, she works as a post-doc researcher at SDU's new Centre of Industrial Electronics (CIE), where she conducts highly transdisciplinary research at the interface of engineering and speech sciences. Her main focus of research is on the production and perception of hate speech in Danish and German, a project funded by VELUX (XPEROHS project).

Kerstin Fischer

Professor (MSO) for Language and Technology Interaction
Department of Design and Communication, SDU

kerstin@sdu.dk
https://portal.findresearcher.sdu.dk/da/persons/kerstin

Kerstin Fischer Fischer is professor (MSO) for Language and Technology Interaction at the University of Southern Denmark and leads the Human-Robot Interaction Lab in Sonderborg. She received her PhD in English Linguistics from Bielefeld University in 1998, after which she did postdoctoral work at the University of Hamburg on emotion in human-computer dialog. She was assistant professor in Bremen 2000-2006 and associate professor for English Linguistics at the University of Southern Denmark 2007-2015. Her research focuses on the interface between prosody and pragmatics, on the development of smooth, seamless interaction between humans and technologies and on human-robot interaction.

Jan Michalsky

Post-doctoral Researcher
Chair of Technology Management,
FAU Erlangen-Nuremberg

jan.michalsky@fau.de
https://www.tm.rw.fau.eu/team-2/staff/jan-michalsky/

Jan Michalsky received his PhD in German linguistics from the University of Oldenburg in 2017. Since 2018 he is a post-doctoral researcher at the chair of technology management at the FAU Erlangen-Nuremberg, where he is involved in leadership research and training. His general research interests cover the entire range of pragmatic meanings in social interaction as well as their phonetic manifestations and phonological modeling. Currently, his focus is on the phonetic correlates of charisma and persuasion in negotiation and bargaining scenarios as well as on the phonetic correlates of attractiveness and likability in mating settings. Jan Michalsly is the co-founder and CEO of the consulting company "Saphire Solutions" and inventor of the the Dynamic Prosodic Adaption method.

Stephanie Berger is a PhD student of General Linguistics at Kiel University, Germany. Her research focuses on acoustic-prosodic features of charismatic speech of English-speaking YouTube Creators and how these and other features shape the perception of charismatic speech. In 2017, she finished her master’s degree in General Linguistics, investigating acoustic-prosodic features of charismatic speech as well, using methods which she is now applying, adapting and developing for her PhD thesis. In 2018, she was a temporary lecturer of the seminar “Introduction to phonetics” at Kiel University.

Stephanie Berger

PhD Student
Institute for Scandinavian Studies, Frisian Studies and General Linguistics, Kiel University

sberger@isfas.uni-kiel.de

Selina Sara Eisenberger

Master Student, Student Assistant
Department for Design and Communication, SDU

seeis16@student.sdu.dk

Selina Sara Eisenberger is student assistant at the Department for Design and Communication at the University of Southern Denmark. She is currently studying MSc in IT – Web Communication Design. Selina worked on projects concerning prosody, intonation and HRI with special focus on Danish, English and German. She is especially interested in language-teaching using robots and human-robot-interaction.

Matouš Jelínek is student assistant at the Department for Design and Communication at the University of Southern Denmark and a student of MSc in IT – Web Communication Design. Before starting studies at the SDU Matouš worked as a research assistant at the Bangor University, UK, and in the Czech Republic on projects focusing on influence of social media and customer engagement. He is now focusing on HRI and is also interested in usability and user experience.

Matouš Jelínek

Master Student, Student Assistant
Department for Design and Communication, SDU

majel8@student.sdu.dk

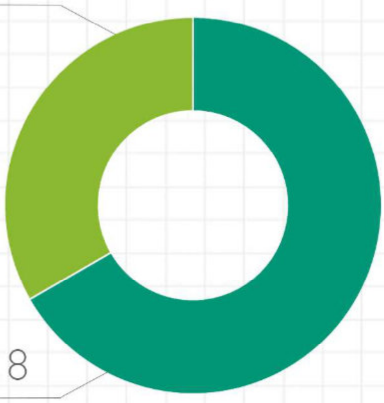
SEFOS Statistics

veggie & vegan^{10}

carnivores^{32}

42 attendees

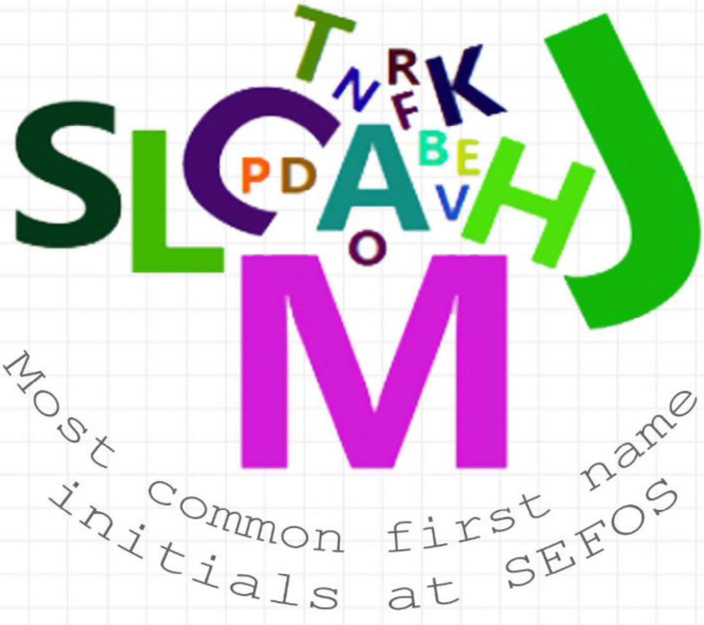
75 authors



male first names^{14}

female first names^{28}

15 countries...



Most common first name initials at SEFOS

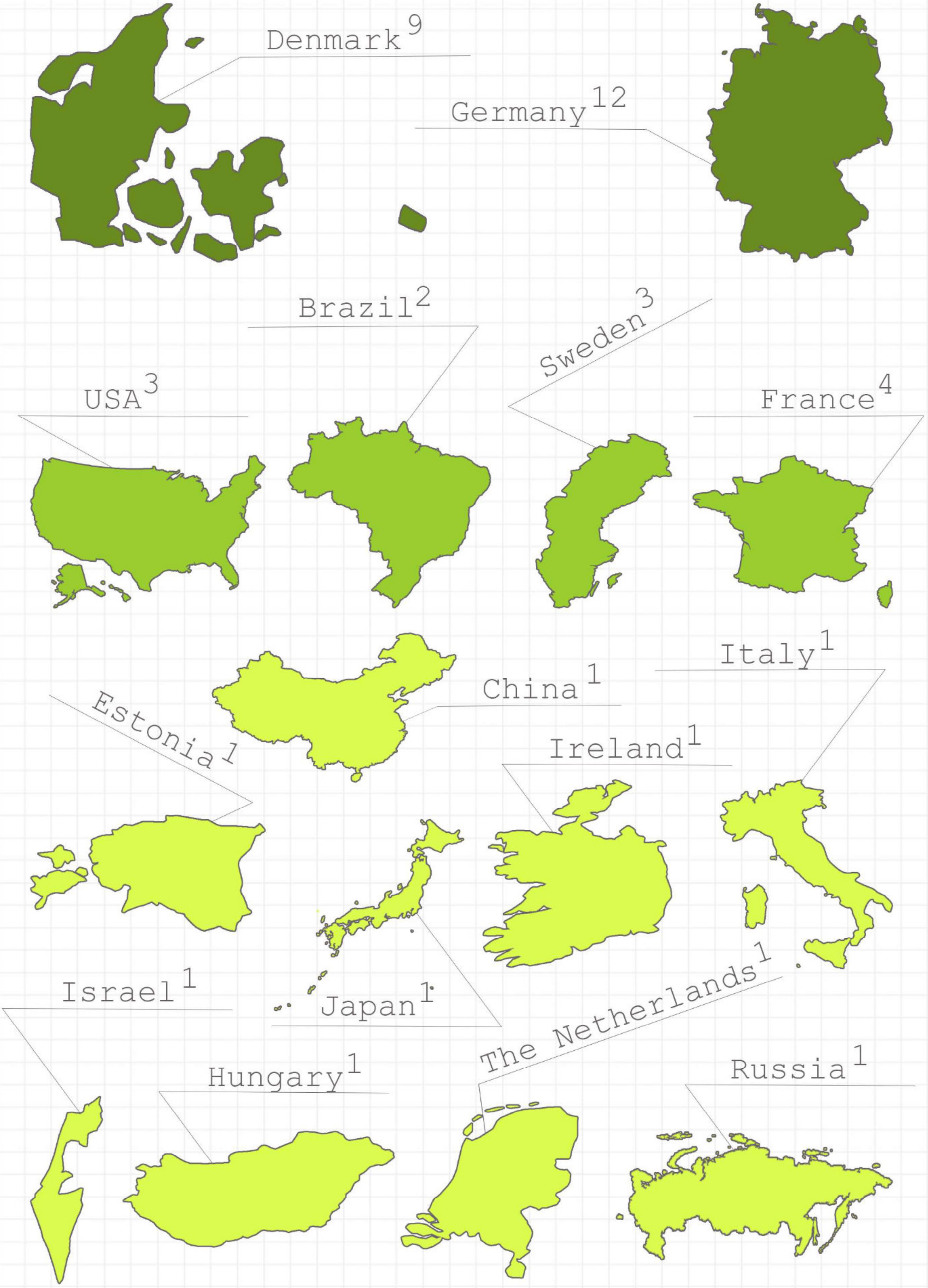
40 papers

27 talks

13 posters

4 keynotes

Where in the world do the SEFOS attendees come from?



Keynote Speakers

Jens Edlund

KTH Royal Institute of Technology, Dept. of Speech, Music and Hearing
www.speech.kth.se/~edlun

Keynote: Breathing in interaction between humans and between humans, machines and robots

Monday 02 December 2019 09:00 - 10:00, Alsion

Summary

In typical everyday life situations, few things require as little thought as breath and breathing. Breathing is a constant, life-sustainable function that we rarely pay any attention to unless we're short of breath. In the typical case, we likewise pay little explicit attention to the breath of others, nless they show signs of having difficulties with it. Still, there is widespread evidence that we react to breathing, even though we may not be aware of it. And an argument could be made that speech without breathing, such as most synthetic speech in machines and robots, will elicit different reactions from human interlocutors (and possibly from other machines) than speech with breathing.

I will present a high-level walkthrough of findings, well-grounded or impressionistic, concerning the role of breath and breathing in spoken interaction and perhaps touch on some related phenomena that behave in similar ways.. Examples and comparisons come from face-to-face interaction between humans or humans and robots, as well as from human conversations with disembodied talking machines.

Affiliation

Jens Edlund is an Assistant Professor at the Department of Speech, Music and Hearing at KTH Royal Institute of Technology in Stockholm. He has studied most aspects of speech technology and spoken interaction, but attempts to maintain a focus on the analysis of human behaviours in face-to-face interactions. Edlund's work is interdisciplinary, i.e. he received his Master's in linguistics and phonetics, followed by a PhD in speech communication and a Docent in speech technology. He has paid special attention to the collections of multimodal speech corpora, capturing, for example, voice, video, motion, and bio-signals such as breath. He is currently involved in research projects investigating gaze, breath and gesture signals in spoken interaction and their coordination with the speech signal. Edlund divides his work between the analysis of human behaviours in interpersonal (human) interactions, and the investigation of the effects of implementing such behaviours in machines.

Donna Erickson

Kanazawa Medical University, Solific Sophia University, Haskins Laboratories

Keynote: Voice: a multifaceted finely-tuned instrument for any occasion and culture

Monday 02 December 2019 13:45- 14:45, Alsion

Summary

Voice actors/actresses can change their voice in amazing ways, to give the impression of the character they are portraying. Non-professionals can too, and we do it all the time, for the most part unconsciously, in every day conversation. This talk addresses the voice as a multifaceted finely-tuned instrument, which we can change at will, depending on the situation. I discuss examples of "voice changes" appropriate and inappropriate for different cultures and social settings; for instance, what is a successful cake seller voice in Japan, and how does this voice sound to Americans or Chinese listeners? Also, what is an appropriate seductive or irritated or polite voice in the U.S., Japan, France, and Brazil? What are some characteristics of negative or positive "voices" in different cultures? I also discuss how we tune our voice instrument: what are some things we do with our anatomical production tools to bring about these voice changes in terms of adjustments at the vocal folds and vocal tract.

Affiliation

Donna Erickson has been professor for the past 20 years at Gifu City Women's College, Gifu Japan, and Showa University of Music, Kawasaki, Japan. Currently she is retired but actively pursues her research interests through her affiliations with Sophia University, Kanazawa Medical University, and Haskins Laboratories. Dr. Erickson received her Ph.D. from the University of Connecticut, where she researched laryngeal muscle activity underlying F0 contours of Thai, under the guidance of Professor Arthur Abramson. Her articulatory research spread to XRMB/EMA studies of prominence and rhythm of first and second language speakers in various languages, as well as research into source-filter contributions of voice quality changes, especially related to cross-cultural/cross-linguistic expression of social affects and emotion.

Jeremy Day-O'Connell

PhD / Associate Professor, Skidmore College, Music Department, Saratoga Springs, USA

Keynote: Voice, Ear, and Mind: The Foundations of Speech and Song

Tuesday 03 December 2019 08:30 - 09:30, Alsion

Summary

This paper aims to contextualize our collective work at SEFOS, through a comparison of language and music, the two main modes of human auditory communication. The common foundation of language and music--at least in their paradigmatic forms as speech and song--is, of course, the voice. A miracle of bioengineering, the human voice is capable of a rapid control of spectral change (enabling phonological structure) while simultaneously producing variations in fundamental frequency, timing, and loudness (enabling prosodic structure). The latter set of elements offer the clearest parallels between language and music--prosody is, in effect, the music of speech--but other parallels exist as well, highlighting the role of not only voice, but ear and mind as well.

I will discuss certain comparable features of language and music in the realms of phonology, prosody, syntax, and discourse, before turning to a deeper consideration of pitch itself. A particularly rich locus common to linguistic pitch and musical pitch is the idiosyncratic vocal form known as "stylized intonation," an apparent intermixture of spoken and sung vocal production. As I will show, stylized intonation provides a unique linguistic perspective on musical systems, even pointing to the possible vocal/linguistic foundation of an important musical universal--the pentatonic scale. I have developed a novel elicitation paradigm to systematically investigate stylized intonation cross-culturally, which I will introduce along with the 12-language corpus that it has spawned (the "Fa-Fa Corpus"). I will then present preliminary data from this corpus relevant to the universalist hypothesis.

Affiliation

Jeremy Day-O'Connell is an Associate Professor of Music at Skidmore College, where he teaches music theory and recently served as Chair. He is the author of Pentatonicism from the 18th Century to Debussy as well as articles and essays on 19th-century music, scales and harmony, and music and language. His contributions have appeared in Music Theory Spectrum, Journal of Music Theory, Music Perception, and The New Grove Dictionary of Music and Musicians. His current research focuses on the commonalities of musical and linguistic structures.

Plínio A. Barbosa

Associate Professor, University of Campinas, Brazil

Keynote: Stylistic and cross-linguistic differences in the prosodic organization of breathing, stressing, and pausing

Tuesday 03 December 2019 13:45- 14:45, Alsion

Summary

Speech breathing is very different from tidal breathing both with respect to its overall amplitude and the duration of inspiration and expiration phases. In terms of these two phases, speech breathing is strongly asymmetrical, with the expiration phase easily exceeding 2 to 4 seconds. This extended expiration time window is the foundation on which pauses, syllables and stress groups are built. In the first part of this lecture, experimental data from read and spontaneous speech in different speaking styles is presented to show how duration patterns of pauses, syllables and stress groups change cross-stylistically and how breathing patterns are adjusted by speakers to meet these stylistic changes and to support the realization of prominences and phrase boundaries in each style. For example, recent contrastive analyses of breathing patterns in read and narrated speech in Brazilian Portuguese revealed style-specific changes in inspiration amplitude and duration as well as in breath group duration and, moreover, gender-specific differences within these stylistic changes. Reading is also characterized by a higher frequency of inspiration, whereas narration shows a higher amplitude of inspiration. In the second part of this lecture, it is demonstrated that the relatively consistent and style-specific pause and prominence patterns allowed developing an algorithm for the automatic detection of terminal boundaries (of utterances in speech-act terms) in read and narrated speech. The algorithm was successfully tested for Brazilian Portuguese (BP), European Portuguese (EP), French and German. The audio file alone is sufficient for the algorithm to work, i.e. it does not require any previous transcription or linguistic-analysis efforts. The algorithm operates in two stages: first, it detects vowel onsets (VO) and then, it normalizes VO-to-VO interval durations in order to obtain smoothed z-scores; z-score peaks that exceed the threshold value of 2.5 are considered terminal boundaries. Proportion of hits for reading (circa 70 %) are better than for narration (circa 60 %) for all languages, with performance levels generally being higher for EP and French.

Affiliation

Plínio A. Barbosa obtained his PhD degree in 1994 from the Institut National Polytechnique de Grenoble, Grenoble, France. He is currently Associate Professor of the Dept. of Linguistics of the Instituto de Estudos da Linguagem at the State University of Campinas (Brazil), and responsible for the Speech Prosody Studies Group, a team of researchers and students devoted to the analysis and modeling of speech prosody. He has published more than 140 papers in pier-reviewed journals and congress proceedings and is the author of three books: Incursões em torno do ritmo da fala (2006), Manual de Fonética Acústica Experimental with Sandra Madureira (2015) and Prosódia para o Ensino Superior (2019). His interests are directed to Experimental Phonetics and Prosody. He is the editor of the Journal of Speech Sciences, a member of the editorial board of Phonetica, the International Journal of Speech Technology, and of six Brazilian journals. He is current member the IPA, ISCA, Abralin and IEEE Associated and a board member of the SproSIG.

Abstracts

Breathing in conversation — what we've learned

Marcin Wlodarczak[*], Mattias Heldner

Department of Linguistics, Stockholm University, Sweden

*wlodarczak@ling.su.se

In this paper, we provide an overview of selected findings on interactional aspects of breathing in multiparty conversation, accumulated largely over the course of a four-year research project *Breathing in conversation*, carried out at the Department of Linguistics, Stockholm University. In particular, we focus on results demonstrating the contribution of the respiratory signal to prediction of imminent speech activity, as well as on turn-holding and turn-yielding cues.

INTRODUCTION

In this paper, we present a brief overview of an investigation into interactional functions of breathing done at the Department of Linguistics, Stockholm University. The research has been primarily carried out within the project *Breathing in conversation*, funded by the Swedish Research Council (VR) in the years 2015-2018.

The project has attempted to fill a gap in phonetic literature, pertaining to pragmatic functions of respiration with particular emphasis on respiratory turn-taking cues. While recent years have seen an increased interest in such functions of breathing [e.g. 1,2], around the year 2013, when we were entering the field, the only study which addressed functions of breathing related to turn-taking was [3]. The study looked at selected respiratory patterns linked to speaker change but was based on relatively small (and partly scripted) material.

Within the project, we have built a recording setup for capturing respiratory activity in multiparty conversation using Respiratory Inductance Plethysmography (RIP). The setup was then used to collect a corpus of respiratory recordings in spontaneous conversations in Swedish and Estonian[1]. Below, we provide a short overview of the setup, the dataset as well as some of the results. In particular, we will present evidence that the respiratory signal is useful for prediction of speech in multiparty conversation. We will also discuss the respiratory cues accompanying turn holding and turn yielding.

METHOD

All results reported on below are based on a material collected with Respiratory Inductance Plethysmography (RIP) in a sound treated room in the Phonetics Laboratory at the Department of Linguistics, Stockholm University. The details of the setup, including the software and hardware developed in-house for collection and analysis of the respiratory signal is described in greater detail in a separate contribution to the present workshop.

The analyzed material includes three-party conversations in Swedish and Estonian. These were subsequently segmented into talkspurts based on manually corrected, intensity-based automatic segmentations. Turn holds and speaker changes were identified automatically using a mechanistic definition of between- and within-speaker silences, depending on presence of speaker change around the silent interval [4].

Inhalations and exhalations in the respiratory signal were labelled automatically using a method similar to that subsequently implemented in RespInPeace [5], a Python toolkit for processing and analysis of RIP data, also presented in a separate contribution.

RESULTS

To test whether the respiratory signal is useful for prediction of speech in multiparty conversation, in [6], we carried out a series of prediction experiments using stochastic turn modelling [7]. Briefly, the method represents conversation as a chronogram, that is as a series of speech and silence intervals, quantized into fixed-sized frames (in our case, 100-ms frames were used). A model can then be trained to predict whether a target speaker will produce speech or remain silent in the next frame, based on a speech

[1] The Estonian data was collected by Kätlin Aare (University of Tartu / Stockholm University).

activity history of specified duration (here, 10 frames or 1 second), and optionally, on additional features cast into the chronograph format. Prediction error, expressed on a cross-entropy scale in bits per 100-ms frames, for a range of models is presented in Figure 1, where models 1-6 are trained on joint speech activity and respiratory histories, compared to two baseline models trained on speech activity history of the target speaker alone, BL(MI), or speech activity history of all interlocutors, BL(CI). As can be seen, models trained on respiratory features achieve much lower cross-entropy values than speech activity-only models. For the best-performing respiratory feature (z-normalized respiratory slope over the 100-ms frame for the target speaker only, (5a) the benefit over the target-speaker only speech activity baseline, BL(MI), is more than twice that of including speech activity history of all interlocutors, BL(CI).

We have also carried out a number of studies identifying specific respiratory features related to turn holding and turn yielding. Similar to other studies [2], we have found that decreased duration of the inhalation was the single strongest feature related to turn-holding [8]. In addition, in [9], we looked into features of the exhalation, which had been generally overlooked before. We found that, in line with Local and Kelly [10], turn holding is often associated with near-zero exhalatory slope, likely corresponding to breath holds. Utterances followed by speaker change were also found to be completed higher in speaker's respiratory range than those followed by more speech by the same participant.

Curiously, we found a number of speaker changes in which the original speaker continues speaking after a short pause which does not involve an inhalation. Since these intervals were quite similar to exhalation found in turn keeping, we hypothesized that these instances correspond to pause interruptions, in which the original speaker planned to continue their turn but was cut short by an interlocutor. We follow Shriberg et al. [11] in referring to such events as *hidden*, and we are currently investigating other types of hidden events which are possible to identify with the respiratory signal.

A separate strand of research was devoted to respiratory properties of short feedback expressions. In [12], we demonstrated that these utterances are often not directly preceded by an inhalation but are produced on residual air instead, a finding we interpreted in the light of economy principle, whereby speakers minimize the respiratory effort associated with the inhalation. Additional evidence for this interpretation was provided by head nods, which were predominantly started towards the end of the exhalation. As production of nods is not constrained by respiratory needs, they may be preferred over verbal expressions late in the respiratory cycle.

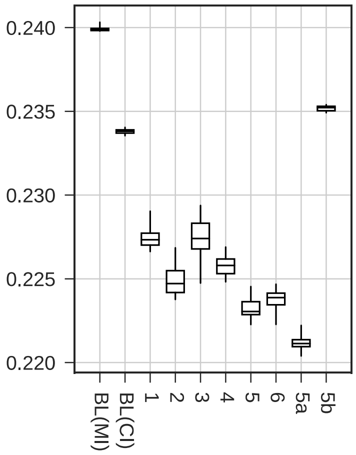


Figure 1. Cross entropies (in bits per 100-ms frames) for the baseline speech activity-only models, BL(MI) and BL(CI), as well as for respiration-augmented prediction models, 1-6. Reproduced from [6].

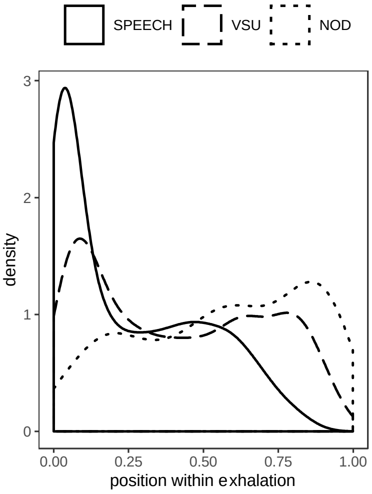


Figure 2. Timing of speech, very short utterances (VSU) and head nod onsets within the exhalation. Based on [12].

CONCLUSIONS

Even the cursory presentation of the results above should demonstrate that the respiratory signal is a rich resource for an enquiry into mechanics of conversation. While we have focused predominantly on respiratory functions related to turn-taking, breathing is also potentially useful for studying other aspects of spontaneous conversations, such as emotion expression, disfluencies or speech planning.

ACKNOWLEDGMENTS

This work was partly funded by the Swedish Research Council (VR) project 2014-1072 *Breathing in conversation* Marcin Włodarczak and by the Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to Mattias Heldner.

REFERENCES

[1] Ishii, R., Otsuka, K., Kumano, S., & Yamato, J. (2016). Using Respiration to Predict Who Will Speak Next and When in Multiparty Meetings. *ACM Transactions on Interactive Intelligent Systems, 6*(2), 1-20. doi: 10.1145/2946838

[2] Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. *Philosophical Transactions of the Royal Society Biology, 369*(1658), 20130399. doi: 10.1098/rstb.2013.0399

[3] McFarland, D. H. (2001). Respiratory markers of conversational interaction. *Journal of Speech, Language and Hearing Research, 44*(1), 128–143. doi: 10.1044/1092-4388(2001/012)

[4] Jaffe, J., & Feldstein, S. (1970). *Rhythms of dialogue*. New York, NY, USA: Academic Press.

[5] Włodarczak, M. (2019). RespInPeace: Toolkit for Processing Respiratory Belt Data. In *Proceedings from FONETIK 2019 Stockholm, June 10–12, 2019*).

[6] Włodarczak, M., Laskowski, K., Heldner, M., & Aare, K. (2017). Improving Prediction of Speech Activity Using Multi-Participant Respiratory State. In Proceedings of Interspeech 2017 (pp. 1666-1670). Stockholm, Sweden: International Speech Communication Association. doi: 10.21437/Interspeech.2017-1176

[7] Laskowski, K. (2011). *Predicting, Detecting and Explaining the Occurrence of Vocal Activity in Multi-Party Conversation.* (PhD), Carnegie Mellon University, Pittsburgh PA, USA.

[8] Włodarczak, M., & Heldner, M. (2016). Respiratory turn-taking cues. In Proceedings Interspeech 2016 (pp. 1275-1279). San Francisco, USA: ISCA. doi: 10.21437/Interspeech.2016-346

[9] Włodarczak, M., & Heldner, M. (2018). Exhalatory markers of turn completion. In Proceedings Speech Prosody 2018 (pp. 334-338). Poznań, Poland: ISCA. doi: 10.21437/SpeechProsody.2018-68

[10] Local, J., & Kelly, J. (1986). Projection and 'silences': Notes on phonetic and conversational structure. *Human Studies, 9*(2-3), 185–204. doi: 10.1007/BF00148126

[11] Shriberg, E., Stolcke, A., & Baron, D. (2001). Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In *Proceedings of EUROSPEECH* (pp. 1359–1362).

[12] Włodarczak, M., & Heldner, M. (2017). Respiratory Constraints in Verbal and Non-verbal Communication. *Frontiers in Psychology, 8*(708). doi: 10.3389/fpsyg.2017.00708

The RespTrack system

Mattias Heldner*, Marcin Włodarczak, Peter Branderud, and Johan Stark

Phonetics Laboratory, Department of Linguistics, Stockholm University, Stockholm, Sweden
*heldner@ling.su.se

This paper describes the RespTrack system for measuring and real-time monitoring of respiratory movements. RespTrack was developed in the Phonetics Laboratory at Stockholm University and the authors have been using it extensively for research for the past five years. Here, we describe briefly the underlying techniques, calibration, digitization as well as recent developments of the system. The presentation at SEFOS 2019 will also include a live demonstration of the system.

INTRODUCTION

RespTrack is a system for measuring and monitoring respiratory movements using the Respiratory Inductance Plethysmography (RIP) method [1,2].

Very briefly, the RIP method uses two elastic inductive belts placed around the rib cage and around the abdomen. Respiratory movements—inhalations and exhalations—alter the inductance of the belts. The belts are connected to electronics that convert the varying inductance into an analog signal with an amplitude that is proportional to the changes in lung volume. This signal can be digitized and recorded and/or displayed in real-time. There is general consensus that RIP is the most frequently used, established and accurate plethysmography method to monitor respiratory movements [3,4].

The RespTrack system is a recent incarnation of the RIP method developed in the Stockholm University Phonetics Laboratory by Peter Branderud and Johan Stark. RespTrack was commissioned by Mattias Heldner and it has been used extensively for research by Marcin Włodarczak and Mattias Heldner (with colleagues) for the last five years [5-23]. About ten RespTrack systems have also been made available to colleagues outside Stockholm University and are being used for research in several different disciplines (e.g. phonetics, voice science, speech and language pathology, ENT medicine). A second batch of RespTrack systems will be available during the fall 2019.

This paper gives a brief description of the the system, including recent developments and describes how we calibrate and digitize the signals from the system. The presentation at SEFOS 2019 will also include a live demonstration of the system.

SYSTEM DESCRIPTION

RespTrack was designed for ease of use, and optimized for low noise and low interference recordings of respiratory movements in speech and singing. The system consists of (i) two elastic inductive belts, (ii) a main unit, and (iii) a cable to connect the belts to the main unit. Fig. 1 shows the components of the system.

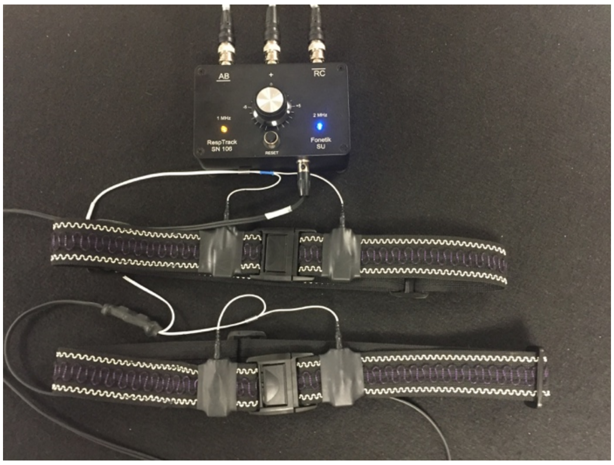


Fig. 1. RespTrack main unit, inductance belts and cable.

The main unit has one combined input for the rib cage (RC) and abdomen (AB) belts connection cable and three analog signal outputs: RC, AB, and "+" a weighted sum of the RC and AB signals allowing a direct estimation of lung volume change. All connections are made with mechanically robust connectors (input TA4F mini XLR, outputs BNC). The range of the output signals is ±2V. The main unit automatically turns on when it is connected to an inductive belt. This is indicated by LEDs (one per belt) that also serve as battery indicators.

One important ease-of-use aspect of RespTrack is the possibility to control the contribution of the individual belts to the weighted sum, which is done by means of a potentiometer knob on the main unit. This is used to ensure that a given volume of air in the lungs produce the same summed output irrespective of how that volume is distributed between the rib cage and abdominal compartments. A slightly higher weight to the RC belt is usually required [24], we have found that a setting of about +2.5 on the RespTrack potentiometer knob is a good starting point for most speakers. Correct weighting of the belts is obtained by instructing the subject to close the glottis, and then to repeatedly contract and relax the abdominal wall, while the experimenter adjusts the potentiometer so that the summed signal remains flat when air is moved from the belly to the rib cage. This is the so called *isovolume manouver* [25].

Another ease-of-use aspect is the possibility of correcting DC offset simultaneously for the RC and AB belts using a “reset” button on the main unit. This function is very practical to get signals within range to be displayed on screen, but it is also used to obtain a meaningful “zero line” in respiratory recordings. Typically, the subject is instructed to produce a relaxed sigh. The experimenter then presses the reset button just after the sigh. A zero reading will then correspond to the resting point of the elastic rib cage and lungs. This resting point is usually referred to as the resting expiratory level or REL [26]. The REL point is relevant in breathing studies as people generally seem to avoid going below REL and many inhalations are initiated around REL.

Furthermore, the reset button alleviates the need for the high-pass filter (HPF) for direct current (DC) offset removal found in many other RIP systems. More importantly, the *absence* of a HPF for DC removal in RespTrack permits the study of certain types of interesting breathing phenomena that cannot be captured in many other systems. For example, periods of breath-holding can be distinguished from slow exhalations with RespTrack, whereas they are indistinguishable in systems including such a HPF for DC removal. This can easily be tested by verifying that the signal remains constant during a breath-hold.

If needed, the weighted sum channel in the RIP recordings can be calibrated in liters using either a “spirobag” with a known volume, or a digital spirometer (e.g. CareFusion MicroLoop). In the latter case, we typically ask the subject to perform a vital capacity maneuver and use the inspirational capacity (i.e. REL to maximum inhalation) reported by the spirometer for calibration.

INDUCTIVE BELTS

In our previous work, we have used commercially available RIP transducer belts (Ambu RIP-mate, pediatric size). We are currently evaluating prototypes of new belts developed by Peter Branderud and Johan Stark which promise several advantages. From a practical point of view, they are much easier to mount and to adjust in size. This is due to a different wiring of the coils in the belts and a simplified cable connection. The new belts are also considerably less susceptible to interference from other systems used in parallel (e.g. other RIP systems or EGG).

SAFETY

The RespTrack system is electrically very safe to use due to its design. Firstly, the main unit is only supplied by low voltage power sources, two AA 1.5V batteries, or an USB connection. Secondly the inductive belts that are worn by the subjects are connected to the electronics via isolation transformers and high impedance resistors in the connection cable.

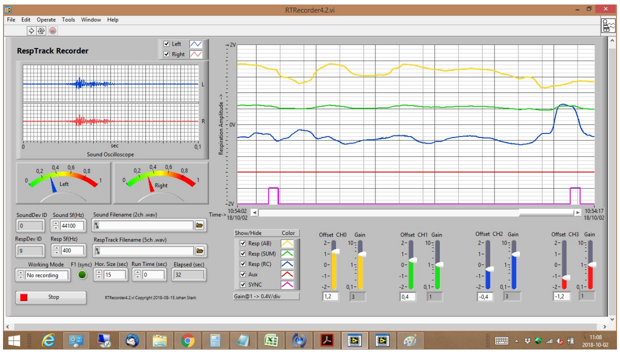
DIGITIZING THE RIP SIGNALS

The hardware requirements for digitizing slowly varying signals such as respiratory movements captured by RIP exclude the use of normal soundcards that typically have a frequency response of 20 Hz to 20 kHz. We have used different combinations of hardware and software.

In most of our published studies, we have used an integrated data acquisition system for recording the RIP signals (PowerLab hardware and LabChart software by ADInstruments). This system was quite expensive but allowed us to simultaneously record three RespTrack systems in parallel as well as to get started quickly.

Johan Stark has also developed a LabVIEW application—RTRecorder—for use with data acquisition hardware from National Instruments (e.g. USB-6000). RTRecorder records 3 RIP signals, stereo audio signals and a sync signal. See Fig. 2 for a screen shot of RTRecorder.

Most recently, we have been experimenting with hardware intended for the modular synthesizer world (Expert Sleepers ES-8, ES-7 and ES-6). These Eurorack modules in combination provide a USB audio interface with 12 input channels that can be used both for audio and RIP signals due to its DC coupled inputs. A distinct advantage is that the cost of this system is a fraction of the integrated system we have been using in the past. Recording can be done with standard digital audio workstation (DAW) software.



SoundDev ID: 0
Sound Sf(Hz): 44100
Sound Filename (2ch .wav):
RespDev ID: 9
Resp Sf(Hz): 400
RespTrack Filename (5ch .wav):
Working Mode: No recording
F1 (sync)
Hor. Size (sec): 15
Run Time (sec): 0
Elapsed (sec): 32

☑ Left
☑ Right

☑ Resp (AB)
☑ Resp (SUM)
☑ Resp (RC)
☑ Aux
☑ SYNC

Gain@1 -> 0.4V/div

Offset CH0: 1,2 Gain: 3
Offset CH1: 0,4 Gain: 1
Offset CH2: -0,4 Gain: 3
Offset CH3: -1,2 Gain: 1

RTRecorder4.2.vi Copyright 2018-08-15 Johan Stark

Fig. 2. Screen shot of RTRecorder.

ANALYZING THE RIP SIGNALS

With respect to analysis of RIP signals, the second author has developed RespInPeace, a Python toolkit for pre-processing and analysis of the respiratory signal [27]. The library implements procedures related to, among others, signal normalization and calibration (drift removal, estimation of respiratory range, etc.), respiratory cycle segmentation, identification of respiratory holds, and extraction of selected respiratory features. The code has been released under a free software license and is available online [28].

ACKNOWLEDGMENTS

This work was partly funded by the Swedish Research Council (VR) project 2014-1072 *Breathing in conversation* to Marcin Włodarczak and by the Stiftelsen Marcus och Amalia Wallenbergs Minnesfond project MAW 2017.0034 *Hidden events in turn-taking* to Mattias Heldner.

REFERENCES

[1] Cohn, M. A., Watson, H., Weisshaut, R., Stott, F., & Sackner, M. A. (1978). A transducer for non-invasive monitoring of respiration. In *Proceedings of the Second International Symposium on Ambulatory Monitoring* (pp. 119-128).

[2] Watson, H. (1980). The technology of respiratory inductive plethysmography. In *Proceeding of the Third International Symposium on Ambulatory Monitoring* (pp. 537–538).

[3] Berry, R. B., Budhiraja, R., Gottlieb, D. J., Gozal, D., Iber, C., Kapur, V. K., . . . American Academy of Sleep, M. (2012). Rules for scoring respiratory events in sleep: update of the 2007 AASM Manual for the Scoring of Sleep and Associated Events. Deliberations of the Sleep Apnea Definitions Task Force of the American Academy of Sleep Medicine. *Journal of Clinical Sleep Medicine, 8*(5), 597-619. doi: 10.5664/jcsm.2172

[4] Wikipedia. (13 January 2019). Respiratory inductance plethysmography. Retrieved October 1, 2019, from https://en.wikipedia.org/wiki/Respiratory_inductance_plethysmography

[5] Aare, K., Włodarczak, M., & Heldner, M. (forthcoming). Breath holds in spontaneous speech. *Journal of Estonian and Finno-Ugric Linguistics*.

[6] Heldner, M., Carlsson, D., & Włodarczak, M. (2019). Does lung volume size affect respiratory rate and utterance duration? In *Proceedings from FONETIK 2019 Stockholm, June 10–12, 2019* (pp. 97–102).

[7] Włodarczak, M., & Heldner, M. (2018). Exhalatory markers of turn completion. In Proceedings Speech Prosody 2018 (pp. 334-338). Poznań, Poland: ISCA. doi: 10.21437/SpeechProsody.2018-68

[8] Aare, K., Lippus, P., Włodarczak, M., & Heldner, M. (2018). Creak in the respiratory cycle. In *Proc Interspeech 2018* (pp. 1408-1412).

[9] Włodarczak, M., Laskowski, K., Heldner, M., & Aare, K. (2017). Improving Prediction of Speech Activity Using Multi-Participant Respiratory State. In Proceedings of Interspeech 2017 (pp. 1666-1670). Stockholm, Sweden: International Speech Communication Association. doi: 10.21437/Interspeech.2017-1176

[10] Włodarczak, M., & Heldner, M. (2017). Respiratory Constraints in Verbal and Non-verbal Communication. *Frontiers in Psychology, 8*(708). doi: 10.3389/fpsyg.2017.00708

[11] Włodarczak, M., & Heldner, M. (2017). Capturing respiratory sounds with throat microphones. In *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016* (pp. 181-190).

[12] Šimko, J., Włodarczak, M., Suni, A., Heldner, M., & Vanio, M. (2017). Coordination between f0, intensity and breathing signals. In *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016* (pp. 147-156).

[13] Ćwiek, A., Włodarczak, M., Heldner, M., & Wagner, P. (2017). Acoustics and discourse function of two types of breathing signals. In *Nordic Prosody: Proceedings of the XIIth Conference, Trondheim 2016* (pp. 83-91).

[14] Włodarczak, M., & Heldner, M. (2016). Respiratory turn-taking cues. In Proceedings Interspeech 2016 (pp. 1275-1279). San Francisco, USA: ISCA. doi: 10.21437/Interspeech.2016-346

[15] Włodarczak, M., & Heldner, M. (2016). Respiratory belts and whistles: A preliminary study of breathing acoustics for turn-taking. In Proceedings Interspeech 2016 (pp. 510-514). San Francisco, USA: ISCA. doi: 10.21437/Interspeech.2016-344

[16] Heldner, M., & Włodarczak, M. (2016). Is breathing silence? In *Proceedings Fonetik 2016* (pp. 35-38).

[17] Włodarczak, M., Heldner, M., & Edlund, J. (2015). Breathing in conversation: An unwritten history. In *Proceedings of the 2nd European and the 5th Nordic Symposium on Multimodal Communication, August 6-8, 2014, Tartu, Estonia* (pp. 107-112).

[18] Włodarczak, M., Heldner, M., & Edlund, J. (2015). Communicative needs and respiratory constraints. In *Proceedings Interspeech 2015* (pp. 3051-3055).

[19] Włodarczak, M., & Heldner, M. (2015). Respiratory properties of backchannels in spontaneous multiparty conversation. In *Proceedings of the 18th International Congress of Phonetic Sciences*).

[20] Aare, K., Włodarczak, M., & Heldner, M. (2015). Inhalation amplitude and turn-taking in spontaneous Estonian conversations. In *Proceedings Fonetik 2015*).

[21] Edlund, J., Heldner, M., & Włodarczak, M. (2014). *Is breathing prosody?* Paper presented at the International Symposium on Prosody to Commemorate Gösta Bruce, Lund, Sweden.

[22] Edlund, J., Heldner, M., & Włodarczak, M. (2014). Catching wind of multiparty conversation. In *Proceedings of Multimodal Corpora: Combining applied and basic research targets (MMC 2014)*).

[23] Aare, K., Włodarczak, M., & Heldner, M. (2014). Backchannels and breathing. In *Proceedings from FONETIK 2014 Stockholm, June 9-11, 2014* (pp. 47-52).

[24] Banzett, R. B., Mahan, S. T., Garner, D. M., Brughera, A., & Loring, S. H. (1995). A simple and reliable method to calibrate respiratory magnetometers and Respitrace. *Journal of Applied Physiology, 79*(6), 2169–2176. doi: 10.1152/jappl.1995.79.6.2169

[25] Konno, K., & Mead, J. (1967). Measurement of the separate volume changes of rib cage and abdomen during breathing. *Journal of Applied Physiology, 22*(3), 407-422. doi: 10.1152/jappl.1967.22.3.407

[26] Cleveland, T. F. (1998). A Comparison of breath management strategies in classical and nonclassical singers: Part 1. *Journal of Singing, 54*(5), 47-49.

[27] Włodarczak, M. (2019). RespInPeace: Toolkit for Processing Respiratory Belt Data. In *Proceedings from FONETIK 2019 Stockholm, June 10–12, 2019*).

[28] Włodarczak, M. (2019). RespInPeace — Process and analyse breathing belt (RIP) data. Retrieved from https://gitlab.com/mwlodarczak/RespInPeace

Hesitation Markers and Audience Design: Position Matters

Kerstin Fischer[1*] and Nathalie Schümchen[2]

[1]*Department of Design and Communication, University of Southern Denmark, Sonderborg* kerstin@sdu.dk
[2]*Department of Design and Communication, University of Southern Denmark, Kolding* nats@sdu.dk

In this study, we investigated the effect of framing a speech event as an instance of teaching in comparison to delivering a talk on the perception of hesitation markers. We expected that, given the interpersonal and discourse structuring functions of hesitation markers and their role in marking important information for the listener, people would judge hesitation markers more positively if they occurred in pedagogical contexts. That is, when a speaker ostensibly designs their utterances for the particular communication partner, hesitation markers should be evaluated more positively. The findings of our questionnaire study of the speech of six TED talkers support these hypotheses, but significantly more so for hesitation markers before important words inside the clause. Thus, position has an influence on the effects of hesitation markers with respect to audience design.

INTRODUCTION

While hesitation markers are often treated as an unwanted speech behavior that should be avoided as much as possible [1], they actually fulfill numerous useful functions in interactions. In particular, they function to indicate ongoing thought processes [2], which helps structure the information and make one's thought processes transparent and thus accessible to the partner; furthermore, hesitation markers indicate ad hoc production in comparison to canned, prefabricated speech, which, in turn, serves a social-interactive function.

We therefore hypothesized that if listeners focus on the degree to which a speaker designs their utterance for the respective audience, for instance, in order to teach important information, listeners would appreciate the functions of hesitation markers and not associate them with nervousness or lack of knowledge [3].

METHOD

The questionnaire study was carried out using a between-subject design. In order to study the effect of the framing of the speech event, we developed a questionnaire in which people were either told that they would hear excerpts from 'great teachers' or from 'great speakers'. Then, participants heard short audio files of 11-21secs extracted from six TED talks, which they had to rate according to the speaker's perceived traits. Subsequently, participants had to answer comprehension questions concerning the sentences they had heard before, since previous work suggests that hesitation markers contribute to better comprehension and memory [4].

Stimuli Creation

We selected three male and three female TED talkers from a variety of disciplines, where topics range from robotics to sociology. We selected three initial hesitation markers in discourse structuring functions and three medial ones that occur before important words. In one condition, participants heard the original version of the stimulus (utterance including 'uh'), in another condition, the hesitation marker was edited out, and in a third condition, the hesitation marker was edited out and replaced by silence. Each participant was presented with six stimuli, each uttered by a different speaker: two original stimuli including 'uh', two without hesitation, and two with silence instead of 'uh'.

Questionnaire

All stimuli were integrated into an online survey using LimeSurvey. The survey started out with a welcome text followed by demographic questions. In order to test our hypothesis that hesitation markers serve an important function regarding addressee orientation, the audio stimuli were framed in one of two ways: participants were told to listen either to "great speakers" or to "great teachers". The framing was reinforced by asking the participants to rank their expectations toward what either a good speaker or teacher is supposed to be good at, such as *Speaks fluently*, *Preparedness*, *Intelligence*, *Friendliness*, *High education*, *Focuses on current task*. These attributes were selected on the basis of known preconceptions about hesitation markers.

Thus, two differently framed questionnaires that each included two utterances of each condition in random order were designed (resulting in 90 possible combinations). The dependent measures are participants' responses to questions about pragmatic function, but also about the suspected degree of audience design exhibited by the speaker. Therefore, the participants had to answer two sets of questions after each audio clip. The first set of questions addressed to which extent the speaker is perceived as trying to get something across, is involved, wishes the listener to really understand, and is perceived as friendly, likeable and polite (among other categories). The second set of questions concerned participants' expectations about good speakers and good teachers.

The survey was sent out via the crowdsourcing platform "Prolific". We decided to only recruit native speakers of English.

RESULTS

In total, 223 participants filled out our survey. 47 participants were excluded due to incomplete surveys, completion durations under minimum time, or L1 other than English, which left us with 176 participants, whose mean age is 37 years (range 16-73), and evenly distributed across gender (81 female, 94 male, 1 other). 19 are students, 36 hold an MA degree or higher, 42 have completed high school, 49 professional training, and 30 replied 'other'.

A first analysis of the data shows significantly more positive ratings for utterances with hesitation markers marking important words as well as more prefaced important words with 'uh' were rated as significantly less nervous and unconcentrated. With regard to the framings, analyses reveal significant differences in the ranking of important attributes. 35.8 % of all participants in condition 1 put *speaks fluently* on the first rank while the majority in condition 2 (45.1 %) prioritized *knowledge about topic* highest. In both conditions, *high education* ranks lowest (condition 1: 74.4 %; condition 2: 48.8).

These different expectations correlate significantly with participants' ratings of the speakers as honest and friendly when hesitation markers occurred clause initially, and as significantly more educated, honest, knowledgeable, nervous and wanting to get their point across when they used hesitation markers before important words. However, while these correlations are all significant, correlations are low, ranging between r=0.09 to 0.25.

Because of the considerable differences based on position, an independent-samples t-test was conducted to compare the subjective ratings of hesitation markers dependent on their positioning. There are consistent highly significant differences in the scores for initially positioned hesitation markers and hesitation markers prefacing important words; in particular, speakers who use hesitation markers in medial position before important words are rated significantly higher with respect to how important it is to them whether the listener understands them, how important it is to them to get their point across, the degree with which they take their partner into account and with respect to how intelligent, educated, knowledgeable and prepared they are, but also concerning nervousness and lack of concentration.

A chi-square test of the relation between hesitation markers' position and the participants' performance at comprehension questions shows that significantly more participants gave the correct answers to comprehension questions when the hesitation marker occurred before the important word (62.8%). Only half of the participants (48.5%) answered correctly when hesitation markers were in initial position ($p=0.003068$; $X^2=8.7665$).

DISCUSSION

The results on the effects of the framing show that people's expectations have an influence on their evaluation of hesitation marker occurrences. However, these effects are much more pronounced for utterance-medial occurrences before important words; the results show consistently more positive evaluations for 'uh' in this position, even though it is probably much more salient there [5]. The analysis of comprehension effects also favors hesitation markers in medial position.

ACKNOWLEDGMENTS

The authors would like to thank the Danish Council of Independent Research for their support of the project *Improving Second Language Pedagogy at the Prosody-Grammar-Pragmatics Interface Using Human-Robot Interaction.*

REFERENCES

[1] Clark, Herbert H. & Fox Tree, Jean E. (2002): Using *uh* and *um* in Spontaneous Speaking. *Cognition* 84: 73-111.

[2] Fischer, K. 2000. From Cognitive Semantics to Lexical Pragmatics: The Functional Polysemy of Discourse Markers. Mouton de Gruyter.

[3] Fox Tree, J. E. (2001): Listeners' uses of *um* and *uh* in speech comprehension. Memory and Cognition 29: 320-326.

[4] Fox Tree, Jean E. (2002): Interpreting Pauses and *Ums* at Turn Exchanges. *Discourse Processes* 34,1: 37-55.

[5] Niebuhr, O. & Fischer, K. (2019): Do not hesitate – unless you do it shortly or nasally. Proceedings of Interspeech 2019.

Producing and Perceiving Prosody in Autism Spectrum Disorder

Charlotte Bellinghausen[1*], Bernhard Schröder[1], Thomas Fangmeier[2], Andreas Riedel[2], Ludger Tebartz van Elst[2]

[1]*Institute of German Studies, University of Duisburg-Essen*
[2]*Department of Psychiatry and Psychotherapy Medical Center –*
University of Freiburg, Faculty of Medicine, University of Freiburg
*charlotte.bellinghausen@uni-due.de

In this paper we present an overview of studies investigating the production and perception of prosody in Autism Spectrum Disorder (ASD). There has been an increasing number of studies in this field in the last years. At the beginning we define the term ASD according to the classification of ICD-10 [1] and DSM-5 [2] and further focus on high-functioning autism. Afterwards we report on studies in the area of informational structural prosody and ASD. Also, we discuss the role of emotion recognition in speech perception in ASD. In a next step we describe studies on both the production and perception of disfluencies in ASD, finally we refer to our own work on the modelling of uncertainty by means of disfluent speech and its role for speech perception in non-autistic listeners and listeners with ASD.

INTRODUCTION

In the last years there has been an increasing number of studies on the role of prosody production and perception in Autism Spectrum Disorder (ASD). The goal of the current contribution is to give an overview of studies investigating the role of prosody for both speech production and perception in ASD. First, we give an introduction to ASD. Afterwards we present previous studies on both prosody production and perception in ASD. Finally, we present experimental studies on the role of disfluent speech in ASD.

BACKGROUND

ASD is – according to the diagnostic criteria of ICD-10 [1] and DSM-5 [2] – classified as a neurodevelopmental disease characterized by three main symptoms: i) difficulties in social communication, ii) unusually restricted, repetitive behavior and interests, and iii) specific abnormalities in perception.

It is mainly accompanied by qualitative deviations in mutual interactions and patterns of communication. With respect to the prevalence it is reported that autism occurs in approximately 1 child of 100 children (cf. [3: 107] referring to [4]) and that it is more often found in boys than in girls (3,5:1) (cf. [5: 500] referring to [6]). According to Poustka [7] possible reasons for autism are genetics, environment and birth complications.

In this paper we focus on the area of high functioning autism, i.e. autistic persons with IQ > 80 with linguistic abilities which are intact. Syntax and semantics are generally processed without problems, but differences in pragmatic processing are often found like problems with understanding of metaphors [8]. The aim of our paper is to give an overview of studies experimentally investigating the role of prosody for both speech production and perception in hearers with high-functioning autism. For these purposes, we refer to studies from different disciplines, i.e. linguistics, medicine, and psychology in order to take multiple perspectives into account.

SPEECH PRODUCTION AND ASD

As Peppé [9] points out, it is not clear from the literature how speech of autistic speakers is characterized. From her review on relevant literature it is concluded that "… speech can sound exaggerated *and* monotonous, slow *and* fast [9: 1353]."

INFORMATION STRUCTURAL PROSODY AND ASD

In the context of information status the results of DePape et al. [10] suggested that prosodic focus marking in adults depends on the degree of language functioning in autism. For autistic subjects with high language functioning a larger pitch range was observed opposed to the typically developed (TD) control group, but marking of information status was missing. On the contrary, autistic subjects with moderate language functioning used lower pitch range but mark information status.

The study of Diehl and Paul [11] showed evidence for problems in the perception and imitation of prosodic patterns in autistic children when compared to a control group, i.e. longer duration of utterances. In contrast, the results of Diehl et al. [12] suggested that children and adolescents with ASD used prosody for resolving syntactic ambiguity similarly as the control group. However, when presenting a different cue in subsequent trials the younger ASD group showed stronger preference for the initial prosodic cue than for the new one.

Grice et al. [13] found for perceiving information status, that adult hearers with Asperger syndrome made less use of prosody than the control group and relied more on lexical information like word frequency and semantic information.

EMOTIONAL PROSODY AND ASD

Doi et al. [14] tested whether adults with Asperger syndrome are able to recognize *anger, happiness*, and *sadness* as basic emotions with different intensity in facial expressions and also in acoustic utterances. Results showed in the case of the ASD group poorer performance for *anger* and *sadness* with respect to facial expression and vocal information. For low or immediate emotional intensity faces were less accurately recognized in the ASD group than in the TD group. With respect to high intensity of emotional expression in the voice the subjects with ASD showed lower accuracy compared to the control persons.

For investigating a relatively wide range of recognition abilities, Globerson et al. [15] conducted a set of perception tasks in high-functioning adults including affective and pragmatic prosody recognition tasks, psychoacoustics tasks and a facial recognition task. For the subjects with ASD and also for TD subjects showed strong associations between psychoacoustics tasks and the prosody recognition tasks, i.e. between the pragmatic and affective test. These associations were more pronounced in the ASD group. Furthermore, it was only in the ADS group that vocal emotion recognition was predicted by facial emotion recognition. With respect to pragmatic prosody no group differences were found.

Schelinski & Kriegstein [16] tested the relation between vocal emotion and vocal pitch perception abilities in high-functioning adults. They reported for the ASD group less accurate vocal emotion perception compared to the TD group. The results also suggest a correlation between vocal pitch and vocal emotion recognition only for TD subjects.

In another approach, Hsu and Xu [17] tested emotional prosody perception based on a bio-informational dimensions theory of emotion expressions in high-functioning adults. For variation of voice quality and other acoustic parameters synthetic utterances were generated by using articulatory speech synthesis [18]. Results suggested that the ASD group was less sensitive when perceiving emotional prosody compared to the TD group.

DISFLUENT SPEECH AND ASD

Individuals with ASD have problems executing theory of mind tasks [cf. 19: 43]. In autistic children deficits passing the false-believe-task (FBT) [20] can be observed [21]. However, a study by Tager-Flusberg and Sullivan [22] suggests that autistic children and mentally retarded children, who pass the first order FBT, also pass a second order FBT.

In order to investigate whether disfluencies such as ‘um’ or ‘uh’ serve as listener-oriented or speaker-oriented aspect of speech, Lake et al. [23] conducted a production study with high-functioning adults. They assumed that the production of disfluent speech in ASD is not motivated by listener-oriented behavior. They interpreted this finding as a consequence of theory of mind deficits. The experimental data showed that individuals with ASD produced less *filled pauses* but more *silent pauses* compared to the TD group and also less *revisions* and more *repetitions*. The authors explained the higher production of *silent pauses* by possible challenges in understanding perspective.

MacFarlane et al. [24] proposed a schema for coding disfluency type in children with ASD and showed that autistic children have different patterns of disfluency when compared to peers with TD.

In the study of Bellinghausen et al. [25] the role of disfluent speech characterized by different combinations of the three cues “intonation”, “pause” and “filled pause” was investigated with respect to the perception of uncertainty in articulatory speech synthesis [26]. The material was tested for a TD group presenting different levels of uncertain answers to TD subjects. Results showed generally evidence for an additive principle, i.e. the more prosodic cues of uncertainty are added to the synthetic signal the higher the perceived degree of uncertainty. The material will also be tested in a group of high-functioning autistic adults to see whether there are differences in prosody perception. It is expected that prosodic indicators of uncertainty have a weaker influence on uncertainty perception in high-functioning autistic hearers as compared to TD hearers.

REFERENCES

[1] Dilling, H., Mombour, W., Schmidt, M.H. (eds.). *Internationale Klassifikation psychischer Störungen: ICD-10 Kapitel V (F) - Klinisch-diagnostische Leitlinien* (10. Auflage). Bern: Hans Huber, 2015.

[2] Falkai, P., Wittchen, H.-U. (eds.). *Diagnostisches und statistisches Manual psychischer Störungen: DSM-5*. Göttingen: Hogrefe, 2015.

[3] Freitag, C. Diagnostik und Therapie von autistischen Störungen im Kleinkindes- und Vorschulalter. In: *Kinder- und Jugendpsychiatrie*, 10(02), pp. 106-114, 2010, doi: 10.1055/s-0038-1629070.

[4] Baird, G., Simonoff, E., Pickles, A., Chandler, S., … Charman, T. Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: the Special Needs and Autism Project (SNAP). In: *Lancet*, 368 (9531), pp. 210-215, 2006.

[5] Biscaldi, M., Rauh, R., Tebartz van Elst, L., Riedel, A. Autismus-Spektrum-Störungen vom Kindes- bis ins Erwachsenenalter. Klinische Aspekte, Differenzialdiagnose und Therapie. In: *Nervenheilkunde*, 31(7-8), pp. 498-507, 2012.

[6] Fombonne, E. The changing epidemology of autism. In: *J Appl Res Intellect*, 18(4), pp. 281-94, 2005.
[7] Poustka, F. *Autistische Störungen (Leitfaden Kinder- und Jugendpsychiatrie)*. Hogrefe: Göttingen, 2003.
[8] Riedel, A., Suh, H., Haser, V., Hermann, I., Ebert, D., Riemann, D., Bubl, E., van Elst, L.T., Hölzel, L.P. Freiburg Questionnaire of linguistic pragmatics (FQLP): psychometric properties based on psychiatric sample. In: *BMC Psychiatry*, 14(1): 14, 2014.
[9] Peppé, S. Prosodic development in atypical populations. In: Prieto, P., Esteve-Gibert, N. (eds.): *The Development of Prosody in First Language Acquisition*, pp. 343-363, Amsterdam / Philadelphia: John Benjamins Publishing Company, 2018.
[10] DePape, A., Chen, A., Hall, G., Trainor, L. Use of prosody and information structure in high functioning adults with Autism in relation to language ability. In: *Frontiers in Psychology*, 3. DOI: 10.3389/fpsyg.2012.00072, 2012.
[11] Diehl, J.J., Paul, R. Acoustic differences in the imitation of prosodic patterns in children with autism spectrum disorder. In: *Research in Autism Spectrum Disorders*, 6(1), pp. 123-134, 2011.
[12] Diehl, J.J., Friedberg, C., Paul, R., Snedeker, J. The use of prosody during syntactic processing in children and adolescents with autism spectrum disorders. In: *Development and Psychopathology*, 27(3), pp. 867-884, 2015.
[13] Grice, M., Krüger, M., Vogeley, K. Adults with Asperger syndrome are less sensitive to intonation than control persons when listening to speech. In: *Culture and Brain*, pp. 1-13, 2016.
[14] Doi, H., Fujisawa, T. X., Kanai, C., Ohta, H., Yokoi, H., Iwanami, A., . . . Shinohara, K. Recognition of facial expressions and prosodic cues with graded emotional intensities in adults with Asperger syndrome. In: *J Autism Dev Disord*, 43(9), pp. 2099-2113, 2013. doi:10.1007/s10803-013-1760-8, 2013.
[15] Globerson, E., Amir, N., Kishon-Rabin, L., Golan, O. Prosody recognition in adults with high-functioning autism spectrum disorders: from psychoacoustics to cognition. In: *Autism Res.*, 8(2), pp. 153-163, 2015.
[16] Schelinski, S., von Kriegstein, K. The Relation Between Vocal Pitch and Vocal Emotion Recognition Abilities in People with Autism Spectrum Disorder and Typical Development. In: *Journal of Autism and Developmental Disorders*, 49, pp. 66-82, 2019.
[17] Hsu, C., Xu, Y. Can adolescents with autism perceive emotional prosody? In: *Proceedings of Interspeech 2014*, Singapore, pp. 1924-1928, 2014.
[18] Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C. Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In: *Proceedings of Interspeech 2011*, pp. 2681–2684, 2011.
[19] Kamp-Becker, I., Bölte, S. *Autismus*. München: Ernst Reinhardt Verlag, 2011.
[20] Wimmer, H., Perner, J. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children‘s understanding of deception. In: *Cognition*, 13(1), pp. 103-128, 1983.
[21] Baron, Cohen, S., Leslie, A.M., Frith, U. Does the autistic child have a ‘theory of mind’? In: *Cognition*, 21, pp. 37-46, 1985.
[22] Tager-Flusberg H., Sullivan K. A second look at second-order belief attribution in autism. In: *Journal of Autism and Developmental Disorders*, 24(5), pp. 577-586, 1994.
[23] Lake J. K., Humphreys K. R., Cardy S. Listener vs. speaker-oriented aspects of speech: studying the disfluencies of individuals with autism spectrum disorders. In: *Psychonomic Bulletin & Review*, 18(1), pp. 135-40, 2011.
[24] Mac Farlane, H., Gorman, K., Ingham, R., ...van Santen, J. Quantitative analysis of disfluency in children with autism spectrum disorder or language impairment. In: *PloS One*, 12, 3, e0173936, 2017.
[25] Bellinghausen. C., Fangmeier, T., Schröder, B., Keller, J., Drechsel, S., Birkholz, P., van Elst, L.T., Riedel, A. Modelling prosodic cues of uncertainty in articulatory speech synthesis – perception in neurotypical and autistic hearers. In: *Proceedings of the 9th workshop on Disfluency in Spontaneous Speech*, Budapest, Hungary, pp. 39-42, 2019. ISBN: 978-963-489-063-8
[26] Vocal Tract Lab - Website: ww.vocaltractlab.de, 2017.

Perception Breakdown Recovery in Computer-directed Dialogues

Maria Di Maro[1*], Jana Voße[2,3], Francesco Cutugno[1], and Petra Wagner[2,3]

[1]*Università degli Studi di Napoli 'Federico II', Italy*
[2]*Phonetics and Phonology Work Group, Bielefeld University, Germany*
[3]*Center of Cognitive Interaction Technology (CITEC), Bielefeld University, Germany*
*maria.dimaro2@unina.it

The present paper examines acoustic-phonetic phenomena in users' speech directed to conversational agents, when misperception of the acoustic signal occurred. Here, we investigate whether typical Lombard speech-related parameters of hyper-articulation are observable as means of error resolution mechanisms. The promising results, although not statistically significant, show interesting tendencies, which are worth of further researches, with an additional focus on different recovery strategies caused by diverse communicative breakdowns.

INTRODUCTION

Although technical devices are more and more involved in the everyday tasks, communicating to a machine as an interlocutor is still perceived as an error-prone communicative situation by most users. This assessment becomes visible by the fact that within human-machine interaction (HMI) a high number of error resolution mechanisms, such as hyper-articulation, take place [1, 2]. These mechanisms, however, are usually detrimental, as they impair the performance of speech recognition systems [3, 4] or do not significantly improve error rates [5, 6], unless the recognition system is specifically trained on hyper-articulated data [7].

To contribute to the improvement of speech recognition systems, this study marks the starting point of a fine-grained analysis of the correspondence between error classes and error resolution mechanisms in computer-directed speech. The present study focuses on one particular error class in computer-directed communication, concretely errors that are caused by a misperception of the acoustic signal [8]. Within this error class, we expect to find a specific kind of hyper-articulation as an error resolution mechanism, namely Lombard speech effects. These effects are usually known for restoring communication in noisy environments. However, the closeness of Lombard speech to hyper-articulated speech [9],[10], which is a frequently observed error resolution mechanisms in HMI [1], lets us hypothesize to observe Lombard speech effects also in erroneous computer-directed communication. Findings from previous studies on hyper-articulation as a correction mechanism in HMI support this assumption, as they report increased amplitudes [2], increased pitch levels [2] and higher durations [2],[11] in corrective phrases, which are also typical parameters of Lombard speech [12].

The following section will specify our hypothesis regarding Lombard speech effects in computer-directed communication. The section named *Ok Google* Corpus, describes the corpus collection and its structure, before presenting the methodology and the results of our analysis. The conclusion summarizes and discusses the results of our study and gives an outlook on further research in this area.

HYPOTHESIS

In the literature, several acoustic-phonetic features are described to be relevant for a Lombard effect. These include – among others - duration (on utterance and word level), speaking rate, f0 (mean level) and intensity (mean level), whose tendencies are described in [12]. In Tab. 1 we report the phenomena that we analyzed in this work. Specifically, we are interested in whether we observe a Lombard effect expressed by the aforementioned features in human-machine interaction with differing levels of erroneous turns. More specifically, we consider two conditions, for which we hypothesize the named acoustic-phonetic features to differ significantly: condition *ne* ('no error') and condition *se* ('second error'), which are described in detail in the next section.

OK GOOGLE CORPUS

In order to analyze the user's acoustic-phonetic patterns when interacting with a dialogue system, we collected computer-directed interactions in a Wizard-of-Oz experiment. Firstly, an app was created in Android Studio, an integrated development environment for Google's Android operating system[1]. The app was built to simulate interactions with the Google Assistant in Italian, English, and German. The responding written behavior of the app was pre-coded, so we were able to simulate specific errors during the interaction, independent of the actual form of users' spoken inputs.

[1] https://developer.android.com/studio

During the experimental session, the users were asked to first activate the app by saying 'Ok Google' and then to collect specific pieces of information (i.e. movie showtime in X, recipe of X, path to X, information about X) from it via spoken requests. These requests could vary in form but not in content or order. This was important to ensure that users could follow a specific schema during the interaction, therefore assuring the collection of comparable data in pre-programmed error resolution scenarios, as reported in Fig. 1. The app responds to both the activation phrase and the request either with a confirming feedback, or with an error message (see Fig. 1)

In this work, we present the analysis of the German corpus, whose collection included the participation of 25 users, among which 14 female and 11 male participants are counted. We selected the repetition of the activation phrase *Ok Google* in three different conditions: i) *ne*, no error, when the user employs it to activate the system to proceed with a third request, after two interactions were successfully completed; ii) *fe*, first error, the second repetition of the activation phrase, after not being understood for the first time; iii) *se*, second error, the third repetition of the activation phrase, after not being understood for the second time. For our analysis, we used the two extremes of the continuum, namely the *ne* and *se* conditions.

METHODOLOGY

The data was annotated manually on utterance level, with an orthographic transcription using the software Praat [13]. For phonemic annotations on phone and syllable levels, we employed the online forced aligner WebMAUS [14] and corrected the automatically derived annotations subsequently.

The acoustic-phonetic analysis of the features listed in the hypothesis was performed automatically by means of Praat scripts. This included the analysis of the dependent variables *duration* (on word level), *speaking rate*, *f0* (median, as this is a more robust measure than the mean [15]; calculated on word level) and *intensity* (median, RMS; calculated on word level).

To figure out whether these features differ significantly in the conditions *ne* and *se*, we analyzed the results of the acoustic-phonetic analysis in a descriptive manner. Here, we referred to standardized measures in descriptive statistics that is the *center* of distribution, the *dispersion* of distribution and the *form* of distribution to ensure a comprehensive analysis of the given subject. For the center of distribution, the dependent variables' distributions are analyzed with the two-sample *t-test* and the *Wilcoxon signed-rank test*, depending on whether the dependent variable has a normal distribution or not. For the dispersion of distribution, we employed the *F-test* for equality of two variances and the *Brown-Forsythe test*. Finally, for differences regarding the form of the dependent variables' distributions, we used the *Kolmogorov-Smirnov test*.

Furthermore, we decided for unpaired statistical testing procedures, as our hypothesis demands contrasting the data in the respective conditions in an unpaired manner. The p-values are corrected with the *Holm-Bonferroni* method subsequently. All statistical analyses are performed in R.

RESULTS

Before correction, the statistical analyses yield significant differences in the center (t = -2.5074, df = 81.93, $p < 0.05$) and the form (D = 0.30952, $p < 0.05$) of distribution of the dependent variable *word duration* (*Cohen's d* = -0.5471496). After applying the Bonferroni-Holm correction, these differences disappear, so that no significant differences are observed for any dependent variable in the investigated conditions.

DISCUSSION

Finding no significant differences in the distributions of the investigated dependent variables contradicts our hypothesis. However, there is a central tendency observable in the data that supports the general idea of our hypothesis, namely that the median of most dependent variable is higher in the *se* condition than in the *ne* condition. Although the difference is not statistically significant, this observation corresponds to the Lombard effects we expected to find in the condition *se*. Therefore, we conclude that the reason for not finding significant differences might be related to the low contrast between the investigated conditions. A comparison of the condition *ne* with the participants' behavior after the fourth or fifth error, might carve out stronger differences in the investigated dependent variables. Here, it would be interesting to see whether such a higher contrast in the examined conditions will align the results to those of similar setups (e.g., [2]) or whether there will be differences, e.g. due to the higher level of smart phone-experience our participants have in contrast to [2].

Another interesting observation is that most of the dependent variables' distributions are bi-modal. This suggests the existence of an influencing factor that causes the distributions to have two main areas of data point accumulation.

We are currently preparing the analysis of spectral tilt and the formants F1, F2, F3 as further acoustic-phonetic parameters of Lombard speech in the given corpus. Moreover, the analysis of the requests' repetition would also be interesting to underline further phenomena related not only to phonetics but also to other linguistic levels to figure out further correspondences between error classes and error resolution mechanisms.

ACKNOWLEDGMENTS

This work is supported by the Italian PRIN project Cultural Heritage Resources Orienting Multimodal Experiences (CHROME, #B52F15000450001), and by the University of Bielefeld. We also thank Mattia Di Meglio for the development of the Android app used for our data collection and Dr. Annett Jorschick for her advice in the statistical analysis.

REFERENCES

[1] Orviatt, S., MacEachern, M., & Levow, G. A. (1998). Predicting hyper-articulate speech during human-computer error resolution. *Speech Communication*, vol. 24, no 2, 87-110.

[2] Pirker, H., & Loderer, G. (1999). I said "two ti-ckets": How to talk to a deaf wizard. *ESCA Tutorial and Research Workshop (ETRW) on Dialogue and Prosody*.

[3] Wade, E., Shriberg, E., & Price, P. (1992). User behaviors affecting speech recognition. *Second International Conference on Spoken Language Processing*.

[4] Soltau, H., & Waibel, A. (2000). Specialized acoustic models for hyperarticulated speech. *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Proceedings (Cat. No. 00CH37100), Vol. 3,1779-1782.

[5] Vertanen, K. (2006). Speech and speech recognition during dictation corrections. *Ninth International Conference on Spoken Language Processing*.

[6] Heracleous, P., Ishi, C. T., Sato, M., Ishiguro, H., & Hagita, N. (2013). Analysis of the visual Lombard effect and automatic recognition experiments. *Computer Speech & Language*, 27(1), 288-300.

[7] Lee, S. J., Kang, B. O., Chung, H., Park, J. G., & Lee, Y. K. (2018, September). Hypo and Hyperarticulated Speech Data Augmentation for Spontaneous Speech Recognition. *2018 26th European Signal Processing Conference (EUSIPCO)*, 2080-2084.

[8] Clark, H. H. (1996). *Using language,* Cambridge university press.

[9] Garnier, M., Bailly, L., Dohen, M., Welby, P., & Lœvenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: Global effects on the utterance, *Ninth International Conference on Spoken Language Processing*.

[10] Zhao, Yuan & Jurafsky, Dan. (2009). The effect of lexical frequency and Lombard reflex on tone hyperarticulation. Journal of Phonetics. 37. 231-247. 10.1016/j.wocn.2009.03.002.

[11] Levow, G. A. (1998, August). Characterizing and recognizing spoken corrections in human-computer dialogue. In Proceedings of the 17th international conference on Computational linguistics-Volume 1 (pp. 736-742). Association for Computational Linguistics.

[12] Landgraf, R. (2018). *Die effekte kommunikationsunterstützender Systeme und natürliche Dialogsprache im Auto – Eine phonetisch-linguistische Analyse* (Ph.D Thesis), Christian-Albrechts Universität Kiel.

[13] Boersma, P. (2006). Praat: doing phonetics by computer. Retrieved from *http://www.praat.org* (accessed 01/10/2019)

[14] Kisler, T., Schiel, F., Reichel, U. D., & Draxler, C. (2015). Phonetic/linguistic web services at bas, *ISCA*.

[15] Hatzinger, R., Hornik, K., & Nagel, H. (2011). R: Einführung durch angewandte Statistik. Hallbergmoos, Germany: Pearson Deutschland GmbH.

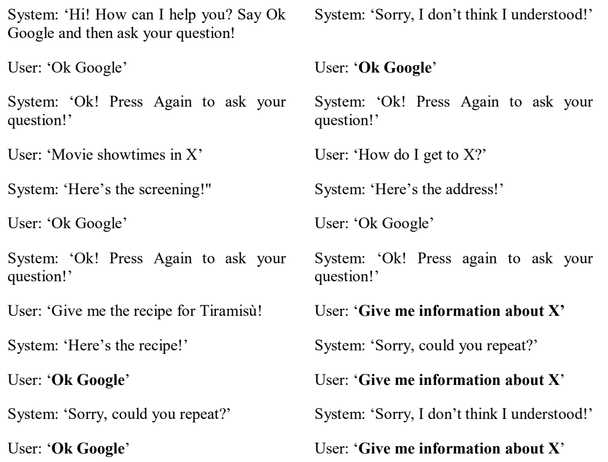
TABLES

Tab. 1: Selected Parameters Modification in Lombard Effect (modified from [11])

Parameter	Expectation	Significance
Intensity	↑	No
Fundamental Frequency	↑	No
Word Duration	↑	No
Speaking Rate	(↑) (↓)	No

FIGURES

Fig. 1. Interaction scheme in the 'OK Google corpus'.



System: 'Hi! How can I help you? Say Ok Google and then ask your question!	System: 'Sorry, I don't think I understood!'
User: 'Ok Google'	User: '**Ok Google**'
System: 'Ok! Press Again to ask your question!'	System: 'Ok! Press Again to ask your question!'
User: 'Movie showtimes in X'	User: 'How do I get to X?'
System: 'Here's the screening!"	System: 'Here's the address!'
User: 'Ok Google'	User: 'Ok Google'
System: 'Ok! Press Again to ask your question!'	System: 'Ok! Press again to ask your question!'
User: 'Give me the recipe for Tiramisù!	User: '**Give me information about X**'
System: 'Here's the recipe!'	System: 'Sorry, could you repeat?'
User: '**Ok Google**'	User: '**Give me information about X**'
System: 'Sorry, could you repeat?'	System: 'Sorry, I don't think I understood!'
User: '**Ok Google**'	User: '**Give me information about X**'

Differences in gender-specific charismatic speech across countries and languages

Anni Gutnyk[1*], Oliver Niebuhr[2], Wentao Gu[1]

[1]*Institute of Linguistics and Applied Linguistics, Nanjing Normal University, China*
[2]*Centre for Industrial Electronics, University of Southern Denmark, Sønderborg*
*****gutnyk@163.com

The present paper is the first step into determining and understanding what charismatic public-speaking means across different countries and languages. Special emphasis is placed on the factor gender, concerning both the speaker and his/her audience. Initial evidence suggests that men and women differ across languages/countries in how they use prosody in public speaking and that these differences are also reflected in how male and female audiences are addressed by speakers.

INTRODUCTION

Charisma has been a popular topic of discussion and research when it comes to defining, training, and practicing leadership [1,2,3]. A big portion of the charisma research addressed political leadership [4,5,6,7,8], but the scope of studies has been expanded in recent years to explain remarkable business success [9,10,11,2].

More fluent-, confident- and enthusiastic-sounding speakers are perceived to be more charismatic. Given the physiological and biomechanical differences of their speech production apparatus, male and female speakers obviously have different starting points when it comes to creating the acoustic-prosodic patterns of a (more) charismatic tone of voice, and there is evidence from research in linguistics, management, and psychology that also listeners set different standards and focus on different verbal and non-verbal features of a speaker when rating the charisma of men and women [12]. For example, prosody-induced charisma could even be more important for women than for men, and women have to outperform men in their charisma-inducing prosody settings to sound equally charismatic in the ears of listeners [14].

This is the basis of our large-scale cross-linguistic and cross-cultural project. Our overarching goals are, on the project's speech-production branch, to determine the prosodic characteristics that make men and women around the globe sound more charismatic in oral presentations and to check to what degree the magnitude of gender-specific prosodic differences is correlated with the gender-equality level of the country. In this context, we also test, if and to what degree male and female speakers change their presentation prosody when talking to an audience of the respective other gender.

On the project's speech-perception branch, we determine how the charisma-related prosodic patterns of speakers affect perceived charisma and whether these effects are gender-specific and differ around the globe; and if speakers indeed adjust their presentation prosody to the audience gender, we also ask on project's speech-perception branch if this means an advantage or a disadvantage for the speaker in terms of perceived charisma.

The languages that are currently involved in this project are: (Shanghai) Mandarin Chinese, German, Danish, Kenyan English, Ugandan English, Brazilian Portuguese, Ukrainian, Spanish, Russian, and Turkish. We are collaborating in this project with Unicamp Brazil, E4Impact Italy, FAU Germany, Sibirian State University of Telecom. & Inform. Sciences, and Istanbul University.

For each country, we have recorded a set of 10 male and 10 female speakers with similar age and education backgrounds and a similar public-speaking experience. All live in major cities. Thus, each language or country is represented by 20 speakers. Since 10 countries are involved, the total speech corpus recorded here includes 200 speakers. They are all recorded in silent meeting or lecture rooms (i,e. rooms meant for public speeches); and each recording includes two elicitation conditions, a spontaneous-speech presentation condition and a condition in which the speaker holds a prepared sales pitch for a newly developed smart phone application that can track employees' work time ('pitch' is a technical term in business for a concise product- or idea-oriented speech). Note that this sales pitch was the same for all 200 speakers. It is an award-winning pitch taken from the e-learning course on *"How to write a killer elevator pitch"* [13]. The sales pitch was translated by professional interpreters into the 10 involved languages. They were native speakers of the target language and got the English version of the text as common starting point for translation.

In addition to the two elicitation conditions (spontaneous speech vs. prepared sales pitch), the recordings included a bipartite language condition, i.e. both speaking tasks were conducted once in the speaker's native language and once in L2 English, as well as, a bipartite audience conditions. That is, all presentations were given once while addressing an imagined male and once while addressing an imagined female audience. The male/female audience order was balanced across speakers and elicitation conditions to avoid any order-related artifacts.

QUESTIONS AND METHOD

The present paper presents a pilot dataset from the project's speech-production branch. Acoustic-prosodic results are compared for four countries/languages: Spanish, German, Ukrainian, and Mandarin Chinese. The research questions we address here are:

- (1) Do speakers adjust their the tone of voice when addressing an (imagined) male or female audience?
  - (a) If so, in what way? Do we find a more male/female-colored prosody when addressing men/women?
  - (b) If so, does the magnitude of this prosodic adjustment depend on the country‘s gender-equality status?
- (2) If and how does the presentation prosody differ as a function of language or country?

Note that we used the current gender-gap report that is published annually by the World Economic Forum as the basis for defining and ranking the gender-equality status of the countries involved [18].

The present pilot dataset only refers to those presentations given in the speakers' native languages and, moreover, only to those that concern the prepared sales-pitch presentation. The acoustic analysis included the following empirically-selected prosodic parameters [14]. They were measured automatically by means of PRAAT scripts, but checked manually for outliers [15,16,17]:

- Pitch level (F0 median based on 90th percentile),
- Pitch range (num. of octaves, 90th percentile),
- Pitch variability (F0 standard deviation),
- Tempo (speaking rate in syll/s, excluding pauses),
- Number of silent pauses (> 200ms),
- Utterance duration (s),
- Voice-quality jitter (ppq5),
- Voice-quality shimmer (ppq5),
- Harmonics-to-noise ratio (HNR, dB).

RESULTS

Firstly, independently of the four languages compared here, we found that the pitch level (F0 median) is higher for women than for men ($p<0.001$). The sample applies to the pitch variability (F0 standard deviation, $p<0.001$). Moreover, women spoke generally at a slower tempo than men, and their sales-pitch presentation voices were characterized by lower jitter and HNR values ($p<0.01$ for both parameters). That is, female voice qualities were breathier but less "shaky" or irregular in terms of pitch. Finally, women also made fewer pauses in their sales pitch presentations than men ($p<0.05$). Note that these gender-specific differences were all greater for the sub-sample of Ukrainian speakers ($p<0.05$), see Figure 1.

With respect to differences between languages/countries, we found that pitch level and variability were highest in German and lowest in Spanish. Mandarin Chinese and Ukrainian had intermediate levels for these two parameters. However, compared to German and Spanish, presentation tempo was highest in Ukrainian and lowest in Mandarin Chinese. Furthermore, again compared to German and Spanish, HNR values were largest (i.e. presentation voices were least breathy) for Mandarin Chinese and smallest (i.e. most breathy) for Ukrainian speakers. Finally, compared to German speakers, both and Ukrainian and Spanish speakers inserted significantly more silent pauses in their sales-pitch presentations.

Regarding audience gender, German and Mandarin Chinese speakers considerably ($p<0.001$) raised their pitch level when speaking to an imagined female audience compared to an imagined male audience (this applies to both male and female speakers alike). Additionally, when addressing a female audience, Ukrainian and German speakers had a larger ($p<0.05$) and Spanish speakers a smaller pitch variability ($p<0.001$). German speakers also showed a lower HNR level. Speakers of all four countries talked faster to an imagined female audience ($p<0.01$ in all cases) and they made fewer silent pauses ($p<0.05$ in all cases). Finally, the jitter level was lower for all except the Ukrainian speakers ($p<0.001$).

DISCUSSION

Regarding our research question (1), we indeed found initial empirical evidence that speakers, across all four tested languages/countries, adjust their prosodic-parameter settings when addressing an audience of entirely male or entirely female listeners. This is all the more noteworthy as the addressed audience was not physically present. It only existed in the head of the speaker. What is also noteworthy in the context is that the prosodic changes that speakers made when talking to an imagined male or female audience reflect the gender-specific prosodic differences in the speaker's country. Thus, the empirical answer to question (1a) is also positive. Speakers used a more male/female-colored prosody when addressing men/women. As no specific prosodic instructions were given, this means that speakers have some kind of mental presentation of what the male/female prosody settings (of oral presentations) are in their country, and they seem to automatically make adjustments in this direction when presenting a sales pitch. The apparent exception of tempo can be explained with reference to gender-specific tempo levels [19]. We cannot give a positive answer to question (1b), though, as Ukrainian speakers showed the smallest and both German and Mandarin Chinese speakers the biggest adjustment of prosody to the audience gender. So there is no evidence, at least from the small subset of four languages, that the gender-equality status of a society or country affects how prosodic presentation settings are chosen by men and women and how (diversely) audiences of a specific gender are addressed. Besides these gender-related aspects we also found clear differences between the speakers of different countries/languages. They concern the entire range of prosodic parameters and, thus, suggest a positive answer to question (2): Public-speaking styles are probably not the same across all countries/languages.

ACKNOWLEDGMENTS

We would like to thank Alexander Brem, Silke Tegtmeier, Jan Michalsky, Plinio Barbosa and Alisa Sydow for their helps in collecting speech data and discussing findings. This research was partly supported by the National Social Science Fund of China (13&ZD189) and the Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences (2017STD006).

REFERENCES

[1] Beyer, J. M. (1999). Taming and promoting charisma to change organizations. The Leadership Quarterly, 10(2), 307-330.

[2] House, R. J., & Aditya, R. N. (1997). The social scientific study of leadership: Quo vadis?. Journal of management, 23(3), 409- 473.

[3] Turner, S. (2003). Charisma reconsidered. Journal of Classical Sociology, 3(1), 5-26.

[4] Boss, G. P . (1976). Essential attributes of the concept of charisma. Southern Journal of Communication, 41(3), 300-313.

[6] Hirschberg, J. B., & Rosenberg, A. (2005). Acoustic/ prosodic and lexical correlates of charismatic speech.

[7] Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. Speech Communication, 51(7), 640-655.

[8] Weber, M. (1947). The theory of social and economic organization (A.M. Henderson & T. Parsons, Eds. and Trans.). Glencoe: Ill: The Free Press.

[9] Bass, B. M. (2000). On the taming of charisma: A reply to Janice Beyer. The Leadership Quarterly, 10(4), 541-553.

[10] Den Hartog, D. N., & Verburg, R. M. (1998). Charisma and rhetoric: Communicative techniques of international business leaders. The Leadership Quarterly, 8(4), 355-391.

[11] Haslam, S. A., & Reicher, S. D. (2012). In search of charisma. Scientific American Mind, 23(3), 42-49.

[12] Niebuhr, O. & Wrzeczsz, S. (2019). A woman's gotta do what a woman's gotta do, and a man's gotta say what a man's gotta say - Sex-specific differences in the production and perception of persuasive power. Proc. 16th International Pragmatics Conference, Hong Kong, China, 1-2.

[13] https://theinterviewguys.com/write-elevator-pitch/

[14] Niebuhr, O., Tegtmeier, S., & Schweisfurth, T. (2019). Female speakers benefit more than male speakers from prosodic charisma training – A before-after analysis of 12-week and 4-hour courses. Frontiers in Communication, vol. 4, p. 12.

[15]De Jong, N.H. & T. Wempe (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods 41, 385-390.

[16] De Looze, C. & Hirst, D.J. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. Proc. 4th International Conference of Speech Prosody, Campinas, Brazil.

[17] Boersma, P. (2001). Praat, a system for doing phonetics by computer. Glot International 5, 341-345.

[18] http://www3.weforum.org/docs/WEF_GGGR_2018.pdf

[19] Weirich, M. & Simpson, A.P. (2014). Differences in acoustic vowel space and the perception of speech tempo. Journal of Phonetics 43, 1-10.

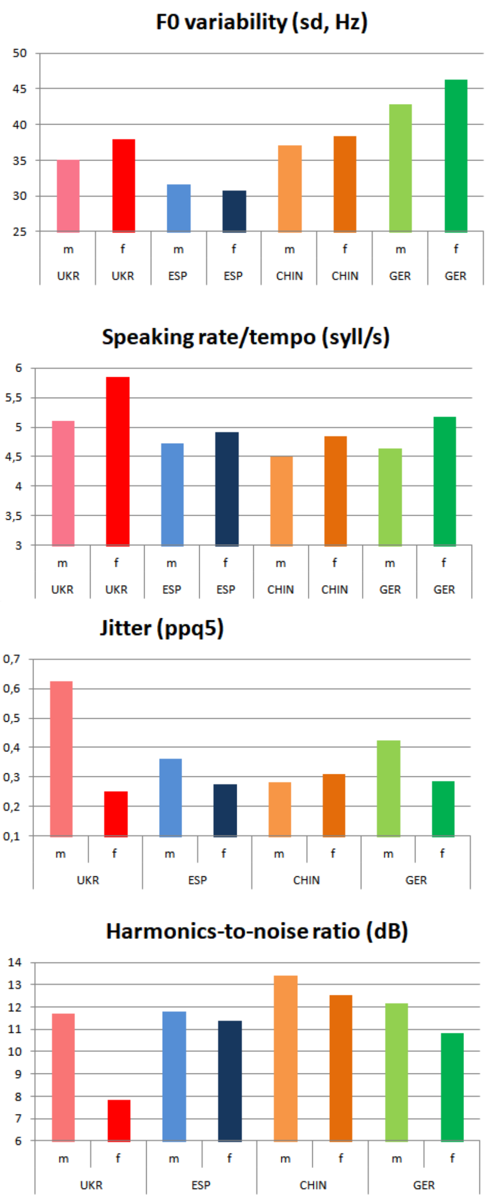


Fig. 1.: Prosodic profiles of men and women in four different countries giving an oral presentation

Age effects on voice and rate in French according to sex

Cécile Fougeron[1*], Angélina Bourbon[1] and Véronique Delvaux[2]

[1] *Laboratoire de Phonétique et Phonologie (UMR7018, CNRS – Sorbonne Nouvelle, France)*
[2] *FNRS & UMONS, Belgique*
*cecile.fougeron@sorbonne-nouvelle.fr

The current study aims at further documenting life-span changes in the speech of French adult speakers with a special focus on the differential effect of aging between men and women. Measurements related to voice quality, speaking f0, maximum phonation time and speech rate are documented in a cross-sectional study of 384 speakers (196 women and 188 men) aged 25 to 93 years.

INTRODUCTION

A better understanding of the evolution of speech throughout adulthood is critical for clinical research where data have to be sex and age-standardized. It is also crucial for our general understanding of the complexity of the speech production system since age-related changes can originate from various sources: anatomical and physiological changes in the speech apparatus affecting pulmonary function, laryngeal structure and/or vocal tract length, hormonal changes, neurological changes affecting speech motor control and/or cognitive functions [1, 2, 3].

The current study aims at further documenting life-span changes in speech and voice over adulthood with a special focus on the differential effect of aging between men and women. This is done within an on-going investigation of a recently collected cross-sectional database, the MonPaGe_HA (for Healthy Adults) [4]. To date, our knowledge of life-span changes in the speech of adults is quite sparse. A decrease in speech rate and segment duration for older speakers is the most robust (and documented) age-related effect [5, 6, 4]. The aging voice has also been wildly studied. Older voices are described as (a) less stable, with an increased jitter [7, 8, 9] or shimmer [10] and (b) noisier, with a decrease HNR [11, 12]. Finally, change in fundamental frequency with age is often reported, but with inconsistencies. Older men are found to have a higher f0 than younger ones, while older female show either no change or an opposite trend with a decrease of f0 with age [13, 14, 10]. However, these trends are not always found (e.g. 10 reported not change for the male speakers but a lowering of f0 for older female groups). Among the various acoustic parameters found to vary with sex, pitch is often considered as the major acoustic difference between male and female voices [15]. Since higher pitch is associated with increased vocal fold tonus, it can lead to more regular vibration patterns. As a consequence, perturbation measures such as jitter and shimmer may also vary with sex, although reported results vary in the literature [16]. Since age also may have an effect on f0 height, the distinction between men and women in f0 height and in perturbation measures could also reduce with age.

The goal of the present study is to explore further the effect of age on male and on female voices based on cross sectional data organized over a wider range of ages and more age groups than traditionally done. A secondary goal, relates to the complementarity of the methods and metrics used for the characterization of voice properties. Local and global voice perturbations have traditionally been measured on sustained vowels either with cycle-to-cycle variability measures (jitter & shimmer) or measures of the distribution of f0 (mean and standard deviation of f0). More recently, innovative analyses techniques such as *cepstral peak prominence smooth* (CPPS) [17] have been introduced in order to measure voice in connected speech, a more naturalistic speech setting. CPPS has the advantage of not relying on the detection of glottal cycles, which can be difficult for some voices, especially in pathologies. If CPPS has proven useful for detecting dysphonia [18, 19, 20] little is known on its variation across age and sex. This study therefore also provide normative references for CPPS in French.

METHOD

384 French speakers (196 women & 188 men) aged 25 to 93 years were selected from the MonPaGe_HA database ([4]) and further split into 4 age groups [25-44], [45-54], [55-69], [70-93] with a balance between male and female speakers across groups as shown in Table I.
Speakers are recorded following the MonPaGe speech screening protocol, which aims at characterizing a speaker's speech profile on multiple speech dimensions and tasks. Of interest here are the following tasks that relates to voice and rate parameters. On a sustained production for 2-3 seconds of the vowel /a/ at a comfortable height and loudness, we computed with Praat: *jitter (PPQ5*), *shimmer (APQ11)*, *f0_standard deviation (SD-af0), HNR,* and *CPPS (aCPPS)*. On a read sentence (composed of 7 syllables, 14 phonemes, all voiced), we computed: *speech rate* in phoneme per s., *speaking f0* (*Spkf0*: mean f0 on the sentence) and its modulation in terms of *standard deviation (SD-Spkf0)* and *coefficient of variation (Varco-Spkf0)*. A measure of CPPS was also taken on the sentence (*SpkCPPS*). A measure of *maximum phonation time (MPT)* was also taken on the best performance of a sustained vowel /a/

produced as long as possible at comfortable pitch and loudness.

RESULTS & DISCUSSION

An effect of sex is found for all voice measurements except the *Varco-Spkf0*, as shown in Table II (a). Unsurprisingly, women show a higher speaking f0 than men. Along with this high pitch, they also show a more stable voice with a lower jitter, shimmer and lower CPPS (both on the sustained vowel and sentence). Modulation of f0 over the sentence is larger for the higher female voices when expressed in Hz (*SD-Spkf0*) but this sex effect disappears when it is expressed relative to f0 height (*Varco-Spkf0*). Female voices are also found to have a higher HNR. For the temporal dimensions, a slower speech rate is found in the reading task for the female compared to the male speaker, as well as a shorter maximum phonation time.

Regarding the effect of age, the data have been first explored by looking at the relationship between chronological age (taken as a continuous variable) and the different measurements for the female and male populations. As shown in Table II (b), weak correlations over this large range of age are found. More interestingly, the relationship between chronological age and the acoustic measures is not always the same for the female and male population. On the one hand, for male as well as female speakers, speech rate tends to decrease with age (r2 ~ -.3) as can be seen on the scatterplot presented in Figure 1.(b) and the instability of f0 measured by *SD-af0* also tend to increase with age (r2=.3), which is also visible on Figure 1.(c). On the other hand, sex-dependent trends are found for other measures: a decrease in MPT and an increase of shimmer is more associated with age for the female group than the male group, while a decrease of SpkF0 and aCCPS is more correlated with age for the male population. Moreover, as shown in Figures 1.(a-d), the relationship between chronological age and the acoustic values is not linear between 25 and 93 years of age, but the distribution rather show a tilt after a certain age, with different cut-off age according to the measurement or to the sex group.

In order to capture these differences, a second analysis is done in an age groups comparison. Results are presented Table II (c). An age group effect is found on most of the acoustic dimensions and with some differences across sex. Jitter, shimmer, *SD-Spkf0* and *Varco-Spkf0* are stable across age groups for both sex groups. Reduced speech rate and f0 stability over the sustained vowel (*aCPPS*) is found for both sex groups after 70 years of age. For the other measures, age-group effects differ across sex, in terms of the groups differentiated or in terms of the presence of an effect. For instance, MPT, Shimmer and *SpkCPPS* vary across age groups only for the female speakers, while age-dependent variation in *Spkf0* is significant only in the male population.

Regarding the differences between specific age groups, they always reflect a change in voice and speech between the oldest group (70-93) and groups below 70 years of age. The CPPS analysis, either on the sustained /a/ or the sentence, show an interesting differentiation between the oldest women (70-93) and the second middle aged women group (55-69) but no other age group differences. This difference seems to reflect a tendency for CPPS in female voices to increase up to the [55-69] group and then to lower, as shown on Figure 1.(d). This trend would follow hormonal changes in women and needs to be further explored. Nonetheless, it suggests that CPPS should indeed be applied in conjunction with the other voice measures. This complementarity of the measurements is also shown by the fact that the increase of speaking f0 for men above age 70, is not accompanied by an increase in the f0 perturbation measures (jitter, shimmer, SDspkf0).

CONCLUSION

Normative data on various speech and voice measures are increasingly needed in various languages for clinical purposes. Our results confirm that age related changes are overtly sex-dependent and that one must account carefully for these two variables when assessing speech and voice. The major strength of this normative study is the large number of participants and the wide range of ages covered. Since most effects are found between the oldest groups of speakers and the speakers below 70 years of age, further investigations should be conducted to increase our knowledge of speech and voice in the old age.

ACKNOWLEDGEMENTS

This work was partially supported by the Swiss FNS Grant N. CRSII5_173711, the Belgian Fonds National de la Recherche Scientifique, and the French program "Investissements d'Avenir" ANR-10-LABX-0083 (Labex EFL).

REFERENCES

[1] Linville, S. E. (2001). Vocal aging. San Diego, CA, Singular.

[2] Torre, P., & Barlow, J.A. (2009). Age-related changes in acoustic characteristics of adult speech. *Journal of Commununication Disorders*, 42, 324-333.

[3] Bilodeau-Mecure, M., & Tremblay, P. (2016). Age Differences in Sequential Speech Production- Articulatory and Physiological Factors. *Journal of the American Geriatrics Society*, 64(11), 177-182.

[4] Fougeron C., Delvaux V., Menard L., Laganaro M. The MonPaGe_HA database for the documentation of spoken French throughout adulthood. Proceedings of the 11th LREC; 2018, Myazaki, Japan.

[5] Ramig, L.A., & Ringel, R.L. (1983). Effects of physiological aging on selected acoustic characteristics of voice. *Journal of Speech, Language and Hearing Research*, 26, 22-30. DOI:10.1044/jshr.2601.22.

[6] Verhoeven, J., De Pauw, G., & Kloots H. (2004). Speech rate in a pluricentric language: a comparison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47, 297-308. DOI:10.1177/00238309040470030401

[7] Wilcox, K.A., & Horii, Y. (1980). Age and changes in vocal jitter. *Journal of Gerontology*, 35(2), 194-198.

[8] Linville, S.E. & Fisher, H.B. (1985). Acoustic characteristics of perceived versus actual vocal age in controlled phonation by adult females. *Journal of the Acoustical Society of America*, 78(1), 40-48. DOI:10.1121/1.392452

[9] Dehqan, A., Scherer, R., Dashti, G., Ansari-Moghaddam, A., & Fanaie, S. (2013). The Effects of Aging on Acoustic Parameters of Voice. *Folia phoniatrica et logopaedica*, 64, 265-270. DOI:10.1159/000343998.

[10] Goy, H., Fernandes, D., Pichora-Fuller, M., & Van Lieshout, P. (2013). Normative Voice Data for Younger and Older Adults. *Journal of voice*, 27, 545-555.

[11] Decoster, W., & Debruyne F. (1997). The ageing voice: changes in fundamental frequency, waveform stability and spectrum. *Acta Otorhinolaryngologica*, 51,105-112.

[12] Ferrand, C. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of Voice*, 16, 480-487.

[13] Honjo, I., & Isshiki, N. (1980). Laryngoscopic and voice characteristics of aged persons. *Arch Otolaryngol*, 86, 149-150.

[14] Brown, W.S., Morris, R.J.& Michel, J.F. (1989). Vocal jitter in young adult and aged female voices. *Journal of Voice*, 3, 113-119.

[15] Pausewang, G. M., & Mikos, V. (2005). The relative contributions of speaking fundamental frequency and formant frequencies to gender identification based on isolated vowels. *Journal of Voice*, 19, 544-554.

[16] Brockmann, M., Storck, C., Carding, P. N., & Drinnan, M. J. (2008). Voice Loudness and Gender Effects on Jitter and Shimmer in Healthy Adults. *Journal of Speech Language and Hearing Research,* 51(5), 1152-1160.

[17] Hillebrand, J., & Houde, R.A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of speech and hearing research,* 39, 311–321.

[18] Hillenbrand, J., Cleveland, R.A. & Erickson, R.L. (1994). Acoustic correlates of breathy vocal quality, *Journal of Speech Language and Hearing Research,* 37, 769-778.

[19] Maryn, Y., Corthals, P., Van Cauwenberge, P., Roy, N., & De Bodt, M. (2010). Toward improved ecological validity in the acoustic measurement of overall voice quality: combining continuous speech and sustained vowels, *Journal of Voice*, 24, 540-555.

[20] Maryn, Y. & Weenik, D. (2015). Objective dysphonia measures in the program Praat: smoothed cepstral peak prominence and acoustic voice quality index. *Journal of Voice,* 29(1), 35-43.

ADDITIONAL TABLE

Table I. Distribution of the female (F) and male (M) speakers according to the 4 age groups. (N: number of speakers, Mean & SD standard deviation) of the speakers age in each group).

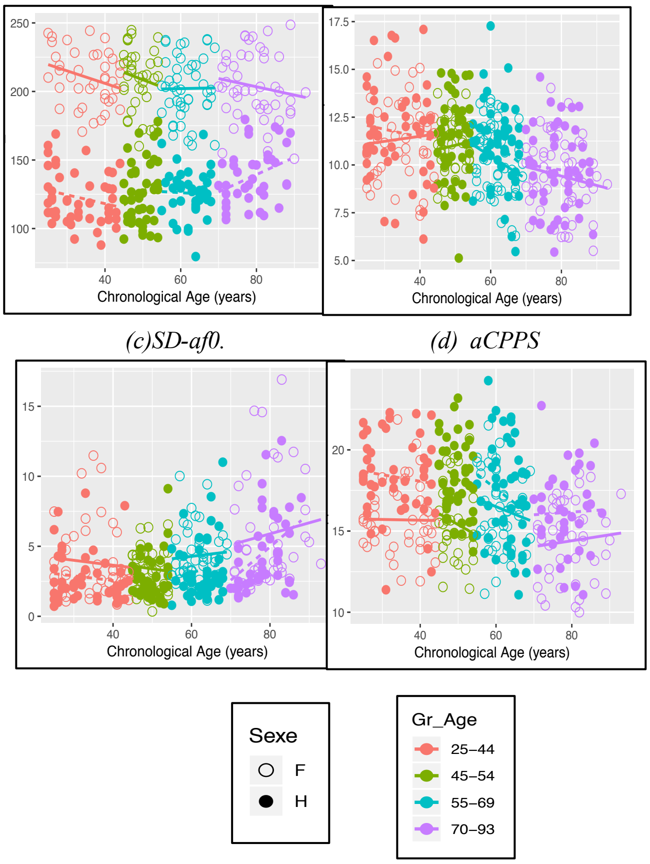
	N	Mean	SD
F	**196**	**57.65**	**14.42**
[25-44]	47	35.02	6.07
[45-54]	40	48.85	2.93
[55-69]	53	61.42	4.00
[70-93]	56	79.36	6.00
M	**188**	**56.26**	**17.71**
[25-44]	48	33.42	6.72
[45-54]	46	49.87	2.83
[55-69]	47	62.57	3.86
[70-93]	47	79.53	5.22

Table II. (a) summary of the sex effect; (b) correlation between chronological age (taken as a continuous variable across all age groups) and the various dimensions for the female (F) and male (M) populations; (c) summary of the age effect in the age-group comparison.

	(a) Sex effect	**(b) Corr with chronol. age**		**(c) Age effect**
		F	M	
speech rate	F <* M	-0.38	-0.26	70-93 <* all & 55-69<*25-44
MPT	F <* M	-0.28	-0.08	F: 55 to 93 <* 45-54 M: Ø
Spkf0	F >* M	-0.13	0.22	F: Ø M: 70-93 >* all
SD-Spkf0	F >* M	-0.02	0.08	Ø
Varco-Spkf0	Ø	0.01	-0.01	Ø
SpkCPPS	F <* M	-0.05	-0.1	F: 70-93 <* 55-69 M: Ø
aCPPS	F <* M	-0.15	-0.28	F: 70-93 <* 55-69 M: 70-93 <* 25to54
jitter	F <* M	0.19	0.11	Ø
shimmer	F <* M	0.31	0.15	F: 70-93 >* all M: Ø
SD-af0	F >* M	0.28	0.32	70-93 >* all
HNR	F >* M	-0.15	-0.06	Ø

Figure 1. For male (H, filled circles) and female (F, empty circles) speakers according to age groups :

(a) Spkf0 *(b) SpeechRate*



(c)SD-af0. *(d) aCPPS*

Voice Quality Comparison Between MPB Singing and Speech

Alexsandro R. Meireles[1], Beatriz Raposo de Medeiros[2], and João P. Cabral[3]

[1]*Federal University of Espírito Santo, CAPES, Brazil*
[2]*University of São Paulo, Brazil*
[3]*Trinity College Dublin, Ireland*
meirelesalex@gmail.com, biarm@usp.br, cabralj@scss.tcd.ie

This paper aims to compare the voice quality in MPB singing and speech. The database was composed by four texts, which are extracts from a famous Brazilian novel, and their respective recordings both in spoken and sung versions. The acoustic analyses were performed using the VoiceSauce tool to automatically extract thirteen voice quality parameters on isolated vowels. We have focused here on the parameters H1*, H1*H2*, CPP, and HNR. The results show a greater prominence of the first harmonic in singing compared to speech, allowing other parameters such as H1H2 to be different. Also, singing presents higher values of HNR, and higher cepstral peak prominence (CPP). These results suggest a more well-defined harmonic structure for MPB singing in comparison to speech.

INTRODUCTION

The voice quality is the object of study in many fields of research such as phonetics [1, 2, 3] or singing [4, 5, 6, 7]. This paper aims to compare singing and speech voice quality, in order to verify whether these two types of voice qualities can be distinguished, or at least, contrasted by voice quality parameters. Adopting similar methods used to investigate the heavy metal voice quality [8], the present work focus on gathering voice quality information for a very distinct voice from the hard rock style: the soft singing voice present in various Brazilian songs that constitutes what we call MBP ("*Música Popular Brasileira",* Brazilian popular Music).

In MPB, one can say that singers, for over five decades have been largely influenced by a special way of singing, the *bossa nova* style. In terms of phonation mode, the relative lower voice energy is one of the characteristics of *bossa nova* style [9]. We suggest here that this Brazilian singing style can be aligned with the fact that some MPB genres (eg. Bossa Nova) would basically require singing production properties that are similar to speech production.

METHOD

Three actresses and three female singers were recorded in an acoustical isolated booth (WhisperRoom MDL 4872 S) equipped with a dynamic microphone (AKG D7) and a Boss digital recorder (BR 800), with the following settings: a sample rate of 44.1 kHz, mono sound signal, and 24-bit resolution.

The database was the same used in [10] where f0 behavior was analyzed in segmented vowels, and it includes four texts, which are extracts from a famous Brazilian novel ("Macunaima") and their respective recordings both in spoken and sung versions. The sung versions are four MPB songs composed by Iara Rennó. The songs are similar between each other in terms of tempo (96 to 120 beats per minute), and also in terms of pitch range (E3 (165 Hz) to A4 (440 Hz)). While singers followed the original key and tempo to record the songs, the corresponding texts were obtained from the recording of the performance of actresses.

Actresses uttered the spoken excerpts because "Macunaima" by Mario de Andrade is also known for its dramaturgy either on stage or in a film version.

All four texts together made up 25 sentences, and in total 150 recorded sentences (3 singers x 25 sentences + 3 actresses x 25 sentences).

The acoustic analyses were performed using the VoiceSauce tool to automatically extract thirteen parameters on isolated vowels [11, 12]: H1*[1], H1*H2*, H1*A3*, CPP, Energy, HNR5, HNR15, HNR25, HNR35, F1, F2, B1, B2 (for a detailed description of these parameters, consult Meireles and Raposo de Medeiros [8]). However, the discussion here will be based on the parameters H1*, H*H2*, CPP, and HNR.

RESULTS

The results show a greater prominence of the first harmonic (see figure 1) in singing compared to speech, allowing other parameters such as H1H2 to be different. Also, singing presents higher values of HNR, and higher cepstral peak prominence (CPP).

Statistical analyses of these four parameters were run in RStudio (version 1.0.153) to test whether the differences between speech and singing were statistically significant. Mann-Whitney and/or Wilcoxon

[1] In statistics, "*" is substituted by "c" (eg. H1* = H1c), so as to ease the computation in R.

tests were used for non-parametric data and Welch's t-tests were run for checking the normality of the data. Results were significant for all tests.

In comparison to speech, the H1* parameter is higher in singing, despite the variation depicted in the boxplot of Figure 1. According to the discussion in Meireles and Raposo de Medeiros [8], higher H1* values are correlated to breathy voice. Therefore, this result suggests a more breathy voice to singing in our data.

For the H1H2* difference, lower values were found for singing, and are associated with tense phonation[1] [8], which in speech can be related to lexical stress. Regarding singing, it seems more productive to look at this parameter association to the opening of the vocal folds. Accordingly to [12], a long open-phase is more easily aligned with pitch period, enhancing the fundamental component. It seems desirable in singing that a stronger first harmonic is attained.

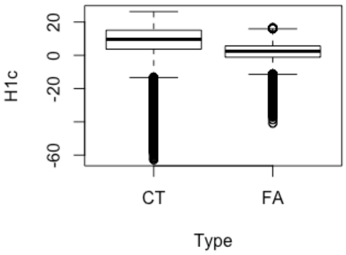


Fig. 1. Boxplot of H1 for singing (CT) and speech (FA).*

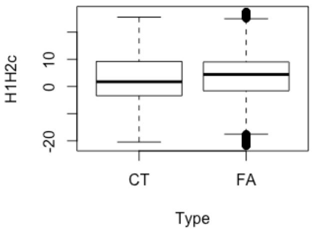


*Fig. 2. Boxplot of H1*H2* for singing (CT) and speech (FA).*

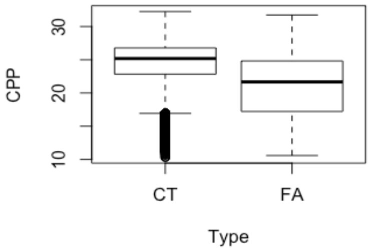


Fig.3. Boxplot of CPP for singing (CT) and speech (FA)

In the singing data, higher values were found for CPP (see figure 3), and for HNR at the ranges of 0-500 Hz, 0-1.500 Hz, and 0-3.500 Hz (see table 1). These two parameters indicate a more well-defined harmonic structure for singing in comparison to speech.

Tab. 1: HNR values in Hz for singing (CT) and speech (FA). HNR at 0-2.500 Hz was not significant. HNR 0-1500 Hz has equal mean, and significance is due to a huge amount of data.

Style/HNR	0-500 Hz	0-1500 Hz	0-3500 Hz
CT	49	50	51
FA	45	50	52

DISCUSSION

Our results suggest that MPB singing share similar voice qualities with certain styles of jazz and pop music [7]. Obviously, we need to collect extra data to confirm this hypothesis. Also, parameters related with signal periodicity (CPP and HNR) allow us to think of the tube uniformity in singing opposed to a more constricted one in speech [13]. Formant frequencies and their bandwidths deserve to be analyzed in a future study to shed light on the articulatory maneuvers or settings that allow more energy and tuning for singing. Normally this setting is related to Flow, a classical singing voice technique.

Finally, general results point out to a more harmonic signal for the singing voice. In the case of speech, it could be characterized more susceptible to voice perturbations than MPB singing.

ACKNOWLEDGMENTS

The first author is supported by the Coordination for the Improvement of Higher Education Personnel (CAPES, grant 88881.337524/2019-01). The second author is indebted to FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo - Brasil, grant 2015/06283-0).

[1] It is important to remind that most of the voice qualities measurements found on the literature refer to speech. Therefore, we reinforce that a tense breathy voice is possible in singing. Consider, for example, the voices in the blues style.

REFERENCES

[1] Ladefoged, P. (1983). The linguistic use of different phonation types. In D. Bless & J. Abbs (Eds.), Vocal fold physiology: Contemporary research and clinical issues (pp. 351-360). San Diego: College Hill Press.

[2] Laver, J. (1980). The Phonetic Description of Voice Quality. Cambridge Studies in Linguistics.

[3] Gordon, M. and Ladefoged, P. (2001). Phonation types: A cross-linguistic overview. *Journal of Phonetics*, 29(4), 383-406. DOI: 10.1006/jpho.2001.0147

[4] Sundberg, J. (1987). The science of the singing voice. Northern Illinois University Press.

[5] Broch, D. & Sundberg, J. (2010). Some Phonatory and Resonatory Characteristics of the Rock, Pop, Soul, and Swedish Dance Band Styles of Singing. *Journal of voice: official journal of the Voice Foundation* 25(5), 532-7. DOI: 10.1016/j.jvoice.2010.07.014

[6] Proutskova, P., Rhodes, C. Crawford, T. and Wiggins, G. (2013). Breathy, Resonant, Pressed – Automatic Detection of Phonation Mode from Audio Recordings of Singing. *Journal of New Music Research*, 42(2), 171-186.

[7] Rouas, J.L. & Ioannidis, L. Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings. Interspeech, Sep 2016, San francisco, United States. 2016, 150 - 154, 2016, <10.21437/Interspeech.2016-1135>. <hal-01392305>.

[8] Meireles, A. & Raposo de Medeiros, B. (2018) Acoustic Analysis of Voice Quality in Iron Maiden's Songs. Proc. 9th International Conference on Speech Prosody 2018, 20-24, DOI: 10.21437/SpeechProsody.2018-4.

[9] Lockheart, P. (2003). A History of Early Microphone Singing, 1925–1939: American Mainstream Popular Singing at the Advent of Electronic Microphone Amplification. Popular Music and Society, 26(3), 367-385. https://doi.org/10.1080/0300776032000117003

[10] Raposo de Medeiros, B &, Cabral, J. (2018) Acoustic distinctions between speech and singing: Is singing acoustically more stable than speech?. Proc. 9th International Conference on Speech Prosody 2018, 542-546, DOI: 10.21437/SpeechProsody.2018-110.

[11] Shue, Y-L. (2010). *The Voice Source in Speech Production: Data, Analysis and Models,* PhD Thesis, UCLA,.

[12] Y.-L. Shue, P. Keating, C. Vicenik and K. Yu.
VoiceSauce: A program for voice analysis", *Proceedings of the ICPhS XVII*,1846-1849,.

[13] Raposo de Medeiros, B. Formants and musical harmonics matching in Brazilian lied. J. Acoust. Soc. Am., Vol. 115. No. 5, Pt. 2, May 2004.

Menstrual Cycle Effects on Phonatory Aspects of Sustained Vowels: It's also about the Offset (but not the Onset)

Míša Hejná[1*]

[1]*Aarhus University, Aarhus, Denmark*
*misa.hejna@cc.au.dk

Numerous studies have shown that phonation can show variation depending on the phase in the menstrual cycle [e.g. 1-3]. [4] has however suggested that, rather than depending on the global phonation, it may be the onset of sustained vowels that presents the locus of this variation. The results presented here show vowel-initial glottalisation is not conditioned by menstrual phase, but the duration of vowel-final aspiration is: longer durations are associated with ovulation.

INTRODUCTION

Research has shown that voice quality can be subject to the effects of the hormonal changes related to the menstrual cycle [e.g. 1-3]. Interestingly, using the CPP measure, [4] has found that ovulation was associated with lower values (interpreted as corresponding to higher breathiness) than the follicular and luteal phases, although only the difference between the follicular and the luteal phases was statistically significant, with the luteal phase reaching the highest values (and thus the least breathy phonation). However, unlike other such studies, [4] also noted that the vowels analysed contain non-modal onsets and offsets, which contain creakiness and aspiration, respectively. The analyses were therefore also conducted only on the main part of the vocalic intervals (i.e. excluding creaky onsets and voiceless glottal friction offsets). Although the visual results were the same as when vowel onsets and offsets were included, the statistical analysis did not reveal any differences between the phases. In [4], CPP was selected because it has been identified as a robust acoustic correlate of perceived breathiness [5]. However, acoustically it does not distinguish well between breathy and creaky phonations [6]. These facts suggest that it may be the onsets and offsets, rather than (only) global phonation, that are sensitive to the hormonal changes associated with the menstrual cycle.

This motivates the following research questions:

- Do the frequency of occurrence and the duration of creaky onsets correlate with menstrual phases?
- Do the frequency of occurrence and the duration of offsets realised as voiceless glottal friction correlate with menstrual phases?
- Are any of these patterns sensitive to vowel categories? In other words, are there interactions between ovarian phases and vowel phonemes?

METHOD

We use the same data which was analysed in [4]. Here, the main aspects of the dataset are summarised, but the reader is referred to [4] for more details. The data comes from a single female subject who met a range of criteria related to the menstrual cycle as well as other important aspects (such as smoking and alcohol consumption). The subject was recorded on a daily basis, and the analyses presented here are based on the production of the following:

- phonologically short vowels of Czech twice during each session ([a], [ɛ], [ɪ], [o ~ ɔ ~ ɒ], [u])
- phonologically long vowels of Czech sustained for 5s and for maximum phonation ([aː], [ɛː], [iː], [o ~ ɔ ~ ɒː], [uː])

This resulted in 2,606 short vowels, 1,283 maximally sustained long vowels, and 1,293 vowels sustained for 5 ms.

Tab. 1: Vowel categories per task and menstrual phase (#).

Vowel	Menstrual Phase		
	Follicular	Luteal	Ovulation
/aː, a/	474	536	32
/eː, e/	477	532	33
/iː, ɪ/	474	531	32
/oː, o/	471	531	32
/uː, u/	478	528	32

Creakiness or glottalisation was defined either as irregularly timed glottal pulses in the acoustic signal, or as an interval of a sudden drop in f0. Voiceless glottal friction was identified by the lack of periodicity in the waveform and the presence of glottal friction in the spectrogram. These two phenomena are illustrated in Fig. 1. The *lmerTest* [7], *lme4* [8], and *emmeans* packages [9] were used with RStudio [10] for the analyses.

RESULTS

The frequency of word-initial glottalisation is not conditioned by the ovarian phase. No differences between the three phases are found (Mixed Effects Models: dependent variable = glottalisation frequency; independent variables = ovarian phase * vowel phoneme;

p > 0.05). See also Fig. 2. The same is the case for vowel-final aspiration.

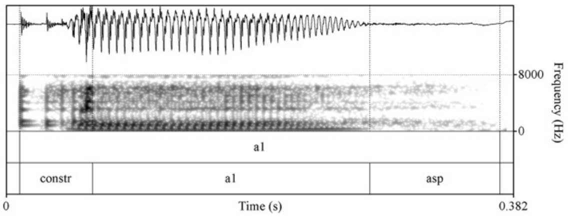


Fig. 1. Identification of laryngeally constricted onsets 'constr' (or occasionally also offsets), and offsets realised as voiceless glottal friction 'asp' (or occasionally also onsets).

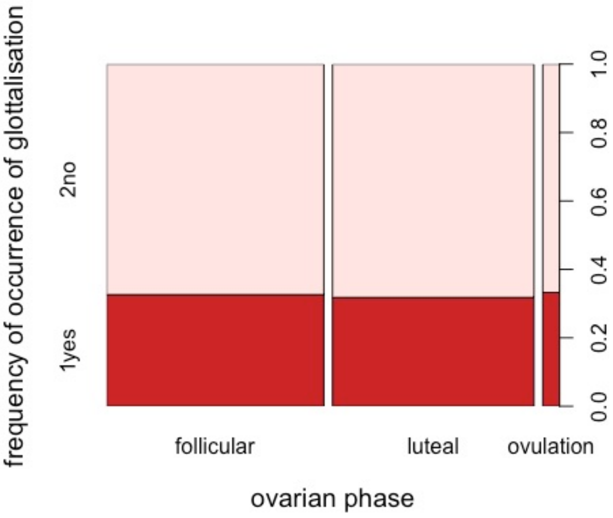


Fig. 2. Frequency of vowel-initial glottalisation by ovarian phase. The results for the frequency of vowel-final voiceless aspiration are comparable.

The duration of word-initial glottalisation duration is not conditioned by the ovarian phase (p > 0.05; no significant differences in post-hoc comparisons). However, the duration of vowel-final voiceless aspiration is. The Mixed Effects Model analyses point to a significant difference between the luteal and the follicular phases (p < 0.01). Post-hoc test comparisons (*emmeans*) nevertheless do not confirm this, pointing to significant differences between ovulation and the follicular phase (p < 0.0001), and ovulation and the luteal phase instead (p < 0.01); Fig. 3.

CONCLUSION

This study aimed to tap into three questions. Firstly, we conclude that glottalised vowel onsets are not conditioned by ovarian phase either regarding their frequency of occurrence or duration. Secondly, the frequency of vowel-final voiceless aspiration is not conditioned by ovarian phase either. However, ovulation is associated with longer vowel-final aspiration intervals than the follicular and the luteal phases. No interactions are found between any of these parameters and vowel phonemes, suggesting that any potential menstrual effects are not vowel-phoneme specific. On the whole then, it seems that the findings reported in [4] are indeed due to breathiness/aspiration levels rather than glottalisation levels.

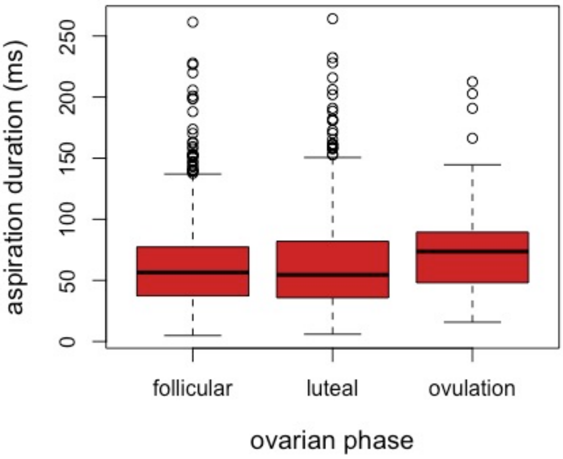


Fig. 3. Duration of vowel-final aspiration (ms) by ovarian phase.

REFERENCES

[1] Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice*, 13(3), 424-446.

[2] Gunjawate, D. R., Aithal, V. U., Ravi, R., & Venkatesh, B. T. (2017). The effect of menstrual cycle on singing voice: a systematic review. *Journal of Voice*, 31(2), 188-194.

[3] Raj, A., Gupta, B., Chowdhury, A., & Chadha, S. (2008). A study of voice changes in various phases of menstrual cycle and in postmenopausal women. *Journal of Voice*, 24(3), 363-368.

[4] Hejná, Míša. (2019). A case study of menstrual cycle effects: global phonation or also local phonatory phenomena? *19th International Congress of Phonetics Sciences, Melbourne*, 2630-2634.

[5] Klatt, D., & Klatt, L. C. (1994). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *JASA*, 87(2), 820-857.

[6] Kelterer, A., & Schuppler, B. (2019). Acoustic correlates of phonation in Chichimec, *Proc. Interspeech, Gratz*, 1981-1985.

[7] Kuznetsova, A. (2015). lmerTest: Tests in Linear Mixed Effects Models. Version 2.0-25, http://cran.r-project.org/web/ packages/lmerTest/index.html.

[8] Bates, D., Maechler, M., Bolker, B., Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7, http://CRAN.R-project.org/package=lme4.

[9] Lenth, R., Singmann, H., Love, J., Buerkner, P., Herve, M. (2019). Emmeans: estimated marginal means, aka least-squares means.

[10] RStudio. 2009-2019. Version 1.2.1335.

Effect of breathing on reaction time in a simple naming experiment: Evidence from a pilot experiment

Christine Mooshammer[1*], Oksana Rasskazova[1,2], Alina Zöllner[1] and Susanne Fuchs[2]

[1]Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin, Berlin, Germany
[2]Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany
*christine.mooshammer@hu-berlin.de

In this pilot study we investigate whether and how breathing patterns affect acoustically measured reaction time. It adapted Sternberg's seminal experiment [1], varying utterance length in a simple naming experiment of ascending and descending number sequences. Nine native speakers of German were recorded acoustically and by means of inductive plethysmography. On average reaction time was 80 ms longer if the stimulus trigger occurred prior or during the inspiration phase. The results indicate that breathing is an integral part of speech planning during initiation and leads to substantial delays in speech onset. It also suggests that parts of the large variability found in reaction time experiments can be explained by the observed breathing patterns.

Introduction

Naming experiments are often used to track speech planning by measuring planning time. In speech research, planning time has been operationalized as the interval from the stimulus onset to the acoustical onset of the first segment. Longer planning or reaction time has been associated with planning strategies and stages on several linguistic levels (e.g. syntax [2], lexicon [3, 4]). Phonological factors also affect planning time, e.g. syllable frequency [5], syllable structure [6], and unit size [1]. However, physiological and phonetic factors have been rarely investigated. Furthermore, planning time is acoustically based on the assumption that the interval from articulatory initiation to acoustic onset is fixed for all segments [7]. However, there is ample evidence against this assumption. Stops, for example, are acoustically detected at the burst noise, indicating the release of the constriction, whereas fricatives are detected at the onset of constriction (see e.g. [6, 8, 9]).

Even though, the discrepancy between acoustic and articulatory onset of speech has recently attracted attention in the field, the interval between the response trigger and movement onset towards the first segment does not take the respiratory system into account. Breathing is not only essential for the oxygen exchange and therefore vital; it also provides the airflow that is essential for speaking. Since most speech acts are performed during expiration the respiratory activity has to be planned well in advance to the utterance onset. Evidence for the role of respiration in speech planning has been found from pre-speech phases and inter-speech pauses in read speech [10, 11] as well as turn-taking in conversations [12, 13]. For example, prior to speaking, speakers take in more air for longer utterances [11]. In conversations breathing of a speaker responding to the question of an interlocutor is initiated well before the end of a question [13], indicating that breathing is an integral part of speech planning. Moreover, respiratory coordination among interlocutors of a dialogue varies with turn taking type (interruption, smooth turn, butting in) showing a fine tuning of advanced planning [12].

The aim of this study is to investigate whether and how respiratory activity affects the planning time during a simple naming experiment, adapted to German from the seminal work of Sternberg et al. (1978). In this delayed naming experiment speakers were presented with sequences of 1-5 digits in ascending order. Then auditory and visual signals were presented after 4 seconds in 85% of the trials. 4 female speakers were instructed to utter the list as soon as possible. Sternberg et al. found that planning time increased linearly with the number of digits which was interpreted as an effect of locating, unpacking and activating of a larger number of subprograms. In our current pilot experiment, we applied a simple naming experiment, so speakers could not prepare their breathing activity to the response trigger. We expect that the measured planning time will increase if the response trigger is applied during or before breathing in (e.g. [14]).

METHOD

Ten native speakers of German (9 f, 1m) were recorded at 16 kHz with inductance plethysmography (Respitrace), simultaneously with the audio signal. Due to technical problems one female speaker had to be excluded. Two flexible bands were wrapped around the torso of the speaker, one around the thorax and the other around the abdomen. Via amplifiers changes in rib and abdomen volume are registered. The task consisted of reading ordered sequences of 1 to 5 digits. Ascending sequences started with numbers from 1 to 5, descending sequences from 9 to 2. The stimuli were presented as numbers on a screen at the same time as an acoustic beep and a change of color on the screen.

The acoustic onset of the responses (speech) and the maximum amplitude of the beep signal were detected using Praat software [16]. The planning time RT_ac was defined as the interval from the beep to the acoustic onset of the response. The respiratory data of the thorax and the abdomen were labelled using Praat and EmuR ([16,17]). The long recordings were cut into single trials with respect to thoracic change from the onset of inspiration to the end of expiration. Only results from the thorax signal are reported here because most subjects showed larger movements for the thorax. 12 out of 1980 trials were without inspiration phase and were therefore excluded. The inspiration phase is defined as the interval from the thoracic minimum prior to speech (=onset of inspiration) to the maximum of the thorax signal and the expiration phase from the maximum (=onset of expiration) to the next minimum (see Fig. A1). Based on the timing of the respiratory activity and the beep, the trials were categorized for *phase:* <I if the inspiration onset occurred before the beep, I if the beep occurred during inspiration and E if the beep occurred during expiration.

Statistical analysis was carried out in R with the linear mixed model package [18]. Sequence order (ascending vs. descending) is ignored because there was no effect on RT.

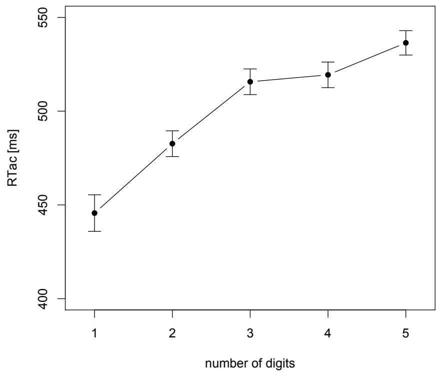


Fig. 1. Acoustic reaction time, averaged over all speakers, for number of sequences with varying numbers of digits.

RESULTS

In a first step we want to know whether we could replicate Sternberg et al.'s findings with our method and data. As can be seen in Figure 1 the planning time increased for longer sequences with less increase between 3 and 4 numbers. Sternberg et al. report a regression equation of 263 + 12.6x. The equation, fitted to our data, 460.9+20.3x. shows a larger intercept and steeper slope. This can be explained by a difference in experimental methods: in simple naming experiments (our experiment) additional planning steps that are more time-consuming are involved than in delayed naming experiments (Sternberg et al.'s experiment).

As explained in the Methods section we categorized the data according to the timing of the trigger signal during respiration. Table A1 (see appendix) shows that, overall, in the majority of cases (58%) the beep occurred during the inspiration phase. However, there is also extensive individual variability. Figure 2 shows a clear effect of respiratory activity on the acoustically measured planning time: if the beep occurs before the speakers initiated inspiration, the planning time is longest ($\bar{x}$=600 ms), shorter if it is during inspiration ($\bar{x}$=510 ms) and shortest during expiration ($\bar{x}$=450 ms). A linear mixed model with number of digits and phase as fixed factors and speaker as random effect showed significant main effects for *number of digits* (β=16.4, df=1959, t=2.8, p<0.01) and for *phase* (I: β=-72, df=1961, t=3.0, p<0.01, E=-110, df=1961, t=-4.4, p<0.001) but no significant interaction.

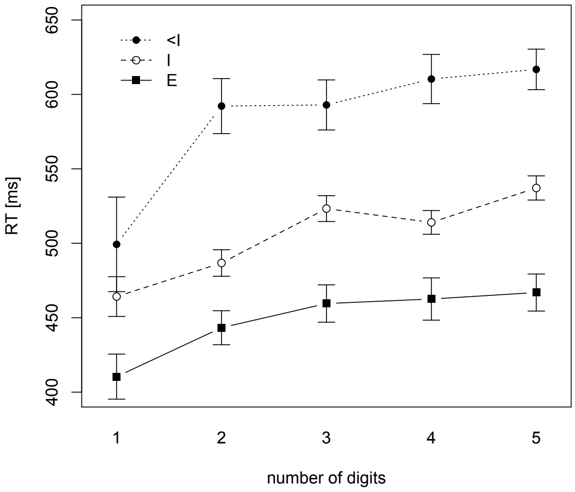


Fig. 2. Acoustic reaction times split by phase in which the beep triggers the reaction: <I = beep before inspiration, I = during inspiration, E = during expiration.

DISCUSSION

This pilot study gives evidence that the planning time is strongly affected by the respiratory phase during which the trigger signal occurs. The effect is quite large (72 ms shorter if expiration was already initiated) compared to linguistic effects (e.g. number of digits here 20 ms, or between 5 and 20 ms in several studies on syllable frequency, [5]). This implies that motor execution is planned with regard to the breathing cycle, i.e. speakers finish inspiration before starting to speak. Recent work by Perl et al. (2019) [15] also showed a strong phase-locking for spontaneous inhalation at task onset. The authors provided evidence for better task performance during inhalation than exhalation and discuss this in light of "general brain information-processing mode triggered by inhalation" (p. 1).

In a follow-up study we will compare these strategies with a delayed naming task during which the speakers can to control their breathing behavior before the trigger signal.

ACKNOWLEDGMENTS

The authors would like to thank Jörg Dreyer and Jane Mertens for help with the recordings. This work was supported by the Leibniz Association and a grant from the BMBF.

REFERENCES

[1] Sternberg, S., Monsell, S., Knoll, R. L., & Wright, C. E. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In Information processing in motor control and learning (pp. 117-152). Academic Press.

[2] Ferreira, V. S. (1996). Is it better to give than to donate? Syntactic flexibility in language production. Journal of memory and language, 35(5), 724-755.

[3] Jescheniak, J. D., & Levelt, W. J. M. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. Journal of Experimental Psychology: Learning, Memory and Cognition, 20, 824–843.

[4] Vitevitch, M. (2002). The influence of phonological similarity neighborhoods on speech production. Journal of Experimental Psychology: Learning, Memory and Cognition, 28, 735–747.

[5] Levelt, W. J. M., & Wheeldon, L. (1994). Do speakers have access to amental syllabary? Cognition, 50, 239–269.

[6] Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E., & Tiede, M. (2012). Bridging planning and execution: Temporal planning of syllables. Journal of Phonetics, 40(3), 374–389.

[7] Kawamoto, A. H., Liu, Q., Mura, K., & Sanchez, A. (2008). Articulatory preparation in the delayed naming task. Journal of Memory and Language, 58(2), 347–365.

[8] Rastle, K., Harrington, J. M., Croot, K. P., & Coltheart, M. (2005). Characterizing the motor execution stage of speech production: Consonantal effects on delayed naming latency and onset duration. Journal of Experimental Psychology: Human Perception and Performance, 31(5),1083-1095.

[9] Schaeffler S., Scobbie J. M., Schaeffer F. (2014). Measuring reaction times: vocalisation vs. articulation, Proc. of the 10th ISSP Cologne.

[10] Rasskazova, O. Mooshammer, C. & Fuchs, S. (2018) Articulatory settings during inter-speech pauses," Proceedings of the Conference on Phonetics Phonology in German-speaking countries (PI&P13), p. 161-165.

[11] Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. Journal of Phonetics, 41(1), 29-47.

[12] Rochet-Capellan, A., & Fuchs, S. (2014). Take a breath and take the turn: how breathing meets turns in spontaneous dialogue. Philosophical Transactions of the Royal Society B: Biological Sciences, 369(1658), 20130399.

[13] Torreira F., Bögels S., Levinson S. C. (2015). Breathing for answering: the time course of response planning in conversation, Frontiers in Psychology, 6, 135-145.

[14] Izdebski, K., & Shipp, T. (1978). Minimal reaction times for phonatory initiation. Journal of Speech and Hearing Research, 21(4), 638-651.

[15] Perl, O., Ravia, A., Rubinson, M., Eisen, A., Soroka, T., Mor, N., ... & Sobel, N. (2019). Human non-olfactory cognition phase-locked with inhalation. Nature Human Behaviour, 3(5), 501.

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

[17] Winkelmann, R., Harrington, J., & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. Computer Speech & Language, 45, 392-410.

[18] Bates, D., Maechler, M., Bolker, B., Walker S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software 67(1), 1-48.

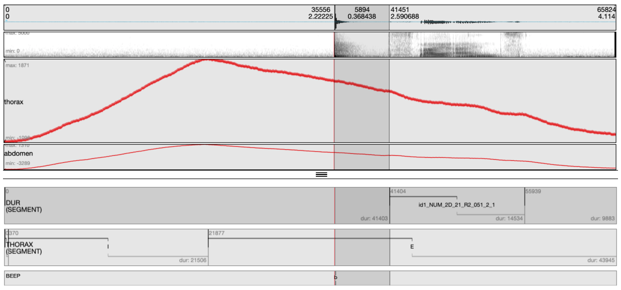


Fig. A1. Screenshot of EmuWeb-App from speaker F01, uttering "zwei eins", showing from top to bottom the audio signal, the spectrogram, the thoracic and the abdominal activities and the labels for the acoustic signal, thoracic activity and the beep. The dark grey area marks the planning time RT_ac from beep to acoustic onset of speech.

Tab. A1: Frequencies of occurrences for <I (beep prior to inspiration), I (beep during inspiration) and E (beep during expiration) in percent per subject and overall.

Participant	Beep-respiration		
	<I	I	E
F01	7.69	53.39	38.91
F04	5.96	75.23	18.81
F05	2.83	79.25	17.92
F07	23.08	47.51	29.41
F08	45.70	42.99	11.31
F09	19.27	40.83	39.91
F10	13.33	54.67	32.00
F11	1.40	73.36	25.23
M02	8.68	54.79	36.59
All	14.22	58.00	27.78

Proposal for a novel analysis of the organization of breath pauses, silent pauses and speech intervals in spontaneous speech

Heather Weston

Leibniz-Centre General Linguistics, Berlin, Germany
weston@leibniz-zas.de

As speech unfolds over time, talk is continually interspersed with breath and non-breath pauses. While there is some evidence that breath-pausing is sensitive to linguistic structure, most studies focus on read speech or turn-taking, and less is known about how breath vs. non-breath pauses interact with the grammatical structuring of an utterance. To better understand how physiological and linguistic demands are coordinated, this work aims to develop an annotation method to examine the distribution of pause types across utterances. In a first step, reported here, pause types are annotated in spontaneous monologues as the speech–breathing system is perturbed by increasing respiratory load via aerobic exercise. The preliminary results unexpectedly show that with heavy respiratory load, non-breath pauses all but disappear. The next step is to annotate the grammaticality of each pause type's location and to identify patterns of distribution.

INTRODUCTION

Breath control makes speech possible, but the principles that govern how breaths are interwoven with speech and pauses are not entirely clear. Investigations of read speech have revealed that almost all breath pauses occur at syntactic/grammatical breaks [1] and that the depth and duration of inspiration is correlated with the length of the upcoming utterance [2]. While few studies have investigated spontaneous speech, the results are similar [3], though more inhalations are reported at non-grammatical breaks [4]; this is attributed to increased cognitive load due to online speech planning.

The present study pursues the idea that breath pauses may serve planning needs as speech unfolds (microplanning), not just before an utterance is executed (macroplanning). This is one departure from the current literature. The other is that the literature to date tends toward a "speech-first" perspective, asking: How does breathing adapt to syntax? How does breathing serve upcoming speech needs? The alternative proposed here is to consider speech and breathing on equal footing: How do speech and breathing respond to each other over time? And with respect to speech planning: Are breath pauses fundamentally different from non-breath (silent) pauses in their interaction with grammatical/utterance structures?

While breath pauses clearly serve physiological needs, there is some evidence that they may also be involved with cognitive aspects of speech planning. For example, speakers can actively vary the type of hesitation strategy used for cognitive planning [5] (e.g., silent vs. filled pause), suggesting that breath and silent pauses could be used interchangeably in some contexts. On a more conceptual level, work in the embodied cognition framework reports that respiration can have a positive effect on cognition [review: 6, 7]; thus, breathing could, e.g., facilitate word-finding. In this case, breath pauses might be used in different contexts than silent pauses.

One way to explore these questions is to introduce additional breath pauses through aerobic activity and to compare pause distribution in different conditions. To do this, it is necessary to annotate pause type and whether the pause occurs at a grammatical juncture. Following the literature, it is hypothesized that breath and non-breath pauses will generally be in complementary distribution (e.g., breath pauses are planned and tend to occur at non-grammatical junctures; silent pauses arise in periods of hesitation/planning and thus at non-grammatical junctures). In a second step, pauses in non-grammatical junctures will be analyzed in relation to the adjacent speech material.

METHOD

An existing dataset [8] is being used to develop the annotation method: acoustic and respiratory data were collected from 11 (1 male) adult speakers of German while performing aerobic activity in a laboratory setting. A stationary bicycle (Daum electronic) was used to perturb respiratory load; respiratory inductance plethysmography (Respitrace; RIP) was used to record breathing. Spontaneous speech was elicited with an item-choice task in three conditions: 1) sitting still; 2) cycling with a moderate load of 70W; and 3) cycling with a heavy load of 140W. The acoustic data (≈8 min./participant for cond. 1 and 2; ≈2 min. for cond. 3) were transcribed and segmented into speech intervals and pauses (>130ms) using phonetic analysis software [9].

The current analysis used the RIP data (combined abdominal and rib cage signals) to further label pause intervals as inhalations, silent pauses or audible exhalations. Inhalations show a sharp positive slope in the RIP signal; silent pauses have no noise in the acoustic signal and no change in the RIP signal and thus no measured respiratory action; and audible exhalations show acoustic noise and a negative RIP slope. The following disfluency phenomena were also annotated:

fillers (e.g., *um*), false starts, partial words, repetitions and prolongations (drawing out sounds within a word). Timestamps were extracted for all intervals and the durations calculated to find the percentage of breath/silent pauses and speech per condition.

The method of grammatical and breath/pause annotation is still a work in progress. The aim is to develop an annotation system that can test whether a systematic relationship exists between pause type (i.e., breath-related and silent pause) and the surrounding speech material. A central question is the extent to which breath pauses can also be used for cognitive/planning purposes.

One idea is to examine pauses at the micro level, considering duration and depth of inspiration in relation to the surrounding material (e.g., annotating for disfluency/hesitation phenomena, parts of speech). Another idea is to construct a larger-scale overview at the macro level (e.g., 30s–60s) to pursue the (not uncontested) notion that speech consists of alternating phases of speech planning and fluent execution. Here one could examine the distribution and duration of breath vs. silent pauses in relation to the topic structure of the greater utterance. The overarching goal is to investigate whether breath planning in spontaneous speech is sensitive to speech planning at different linguistic levels.

PRELIMINARY RESULTS

Preliminary results from a subset of 5 speakers were surprising: as respiratory load grows, and the percentage of breath-related pauses increases, non-breath pauses almost disappear. Data from a representative speaker are shown in Figure 1 (a within-speaker comparison was used due to differences in trial times and individual disfluency behavior across all conditions).

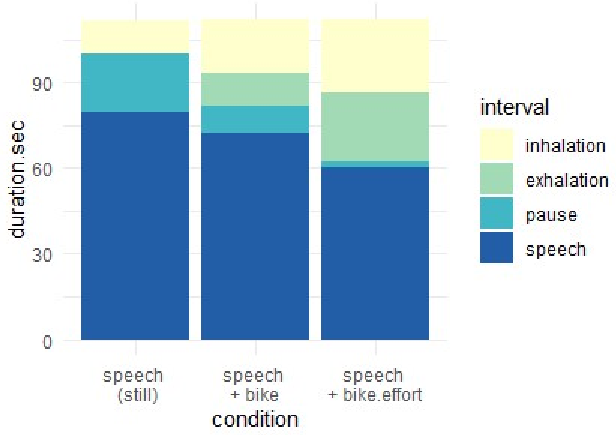


Fig. 1. Ratio of speech to non-speech intervals for a representative speaker (Subject 1).

The percentage of speech also decreased with higher respiratory load. Table 1 shows total time in seconds and percentage for interval type per condition.

Tab. 1: Total time of speech vs. non-speech intervals for a representative speaker (Subject 1).

Interval type	Condition (total time in sec and %)					
	speech (still)		speech + bike		speech + bike.effort	
inhalation	11.5	10%	18.8	17%	25.9	23%
exhalation	0	0%	11.5	10%	24.2	22%
pause	20.7	19%	9.9	9%	2.1	2%
speech	79.6	71%	72.1	64%	60.3	53%
TOTAL	111.8s		112.3s		112.5s	

DISCUSSION

Unexpectedly, silent pauses – characterized by no noise in the acoustic signal and no change in thoracoabdominal kinematics – all but disappeared under heavy respiratory load. If some silent pauses in the speech-only condition serve cognitive planning needs, the preliminary data suggest that breath-related pauses may take on this function. In other words, it is possible that breath pauses do not serve only physiological needs but can also be recruited to some extent for cognitive goals. This suggests a close and flexible mechanism of coordination for the speech and breathing systems.

To pursue this idea, the next steps will be 1) to annotate whether each pause occurs at a grammatical juncture (criteria still to be determined) and 2) to annotate the *type* of inhalation. One surprising observation in the data is that speakers regularly appear to make "small" inhalations (ca. half the range of the fullest inhalation in the trial) in the middle of utterances. Presently it is not clear how to quantify these "partial" inhalations or whether they appear in conjunction with certain speech phenomena (e.g., disfluencies/hesitation, before certain parts of speech). A rough hypothesis is that these partial inhalations may have a word-finding/planning function.

A further surprising observation in the data is that both breath and silent pauses appear in unexpected positions, such as between an article and noun (*das* <230ms> *Zelt* 'the tent') or preceding a verb in final position (*auch ein bisschen vor den Außentemperaturen nachts* <breath.500ms> *geschützt bin* 'I have some protection from the temperature outside at night'). This raises questions of whether certain pause types (e.g., breath/non-breath or partial/full inspiration) are more likely to occur predominantly before or within utterances.

It is hoped that examining breath and silent pauses in the context of their immediate speech environments will both shed light on the coordination of physiological and linguistic aspects of speech planning and also better capture the dynamic nature of online language production.

REFERENCES

[1] Winkworth, A. L., Davis, P. J., Ellis, E., & Adams, R. D. (1994). Variability and consistency in speech breathing during reading: lung volumes, speech intensity, and linguistic factors. *Journal of Speech, Language, and Hearing Research* 37, 535-556.

[2] Whalen, D. H., & Kinsella-Shaw, J. M. (1997). Exploring the relationship of inspiration duration to utterance duration. *Phonetica* 54, 138-152.

[3] Rochet-Capellan, A., & Fuchs, S. (2013). The interplay of linguistic structure and breathing in German spontaneous speech. In *Proceedings of Interspeech* 2013 (2014-2018). Retrieved from https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_2014.pdf (accessed 09/12/2019).

[4] Henderson, A., Goldman-Eisler, F., & Skarbek, A. (1965). Temporal patterns of cognitive activity and breath control in speech. *Language and Speech* 8(4), 236-242.

[5] Beattie, G. W., & Bradbury, R. J. (1979). An experimental investigation of the modifiability of the temporal structure of spontaneous speech. *Journal of Psycholinguistic Research* 8(3), 225-248.

[6] Varga, S., & Heck, D. H. (2017). Rhythms of the body, rhythms of the brain: Respiration, neural oscillations, and embodied cognition. *Consciousness and Cognition* 56, 77-90.

[7] Grassmann, M., Vlemincx, E., von Leupoldt, A., Mittelstädt, J., & Van den Bergh, O. (2016). Respiratory changes in response to cognitive load: a systematic review. *Neural Plasticity* 2016:8146809.

[8] Fuchs, S., Reichel, U., & Rochet-Capellan, A. (2015). Changes in speech and breathing rate while speaking and biking. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Paper no. 1005). Glasgow: University of Glasgow. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS1005.pdf (accessed 07/03/2019).

[9] Boersma, P., & Weenink, D. (2019). Praat: doing phonetics by computer [Computer program]. Version 6.0.43, retrieved from http://www.praat.org/ (accessed 09/08/2018).

Disfluencies in German adult- and infant-directed speech

Charlotte Bellinghausen[1*], Simon Betz[2*], Katharina Zahner[3*], Alina Sasdrich[1], Marin Schröer[2], Bernhard Schröder[1]

[1]Department of German Studies, University of Duisburg-Essen
[2]Faculty of Linguistics and Literary Studies, Bielefeld University
[3]Department of Linguistics, University of Konstanz
**shared first authorship (alphabetical order); all corresponding*
charlotte.bellinghausen@uni-due.de, simon.betz@uni-bielefeld.de,
katharina.zahner@uni-konstanz.de

We investigate the occurrence of disfluencies (lengthening, filled pauses, silent pauses, abandoned utterances, and repairs) in German infant-directed speech (IDS), as compared to German adult-directed speech (ADS). The corpus consists of speech of nine mothers talking to their toddler (IDS condition), or to an adult experimenter (once in a task with low cognitive load, ADS1; once with higher cognitive load, ADS2). In line with previous studies, ADS contains more instances of disfluencies than IDS. Also, with increasing cognitive load in ADS, the number of silent pauses and lengthened instances increases. Overall, our results corroborate earlier findings on IDS in other languages, and on disfluencies and cognitive load. From an explorative perspective, our results also allow to derive hypotheses for future experiments – both for studies regarding IDS and infant speech processing.

INTRODUCTION

Spontaneous speech among adults (adult-directed speech, ADS) contains large amounts of disfluencies and hesitations. Disfluencies occur when speakers are involved in tasks with increased cognitive load, e.g., map tasks [1] and also depend on listeners feedback [2]. The actual frequencies of disfluencies also vary depending on language status ((non)native), as well as psychological or developmental factors like stuttering (cf. [3], Ch. 2).

Infant-directed speech (IDS), on the other hand, is described as highly fluent ([4, 5] on English), with disfluencies becoming slightly more frequent as children grow older (0.58 disfl./100 words in IDS to 7-12 mo-olds vs. 1.03 disfl / 100 words in IDS to 13-24 mo-olds, [6] on Swedish). [5] report that only 5-10% of the prosodic breaks are disfluent in IDS to 6-10 mo-old infants. Similarly, in [6] only 19 filled pauses ("uh"), occurred in a Swedish IDS corpus (35500 words), compared to 249 in ADS (24100 words). Yet, recent processing studies suggest that toddlers discriminate between fluent and disfluent speech ([5, 7] for 22 mo-old children), and also benefit from filled pauses when recognizing novel words ([8] for 24-month-old American children; [9] for 32-mo-old French, English, and English-French children).

In this paper, we compare the occurrence of the disfluencies *lengthening, filled pauses*, *silent pauses*, *abandoned utterances*, and *repairs* in a) German IDS, as compared to ADS and b) in different ADS conditions that differ in cognitive load. Based on the reviewed background, we expect more disfluencies in German ADS compared to IDS, H(ypothesis) 1. We also predict disfluencies to increase with increasing cognitive demands, i.e., increasing task difficulty (H2).

METHODS

The corpus consists of audio-recordings of 9 German-speaking mothers (between 32 and 39 years; currently living in Konstanz and surroundings) in three conditions: one IDS condition, and two ADS conditions (differing in cognitive load). In the IDS condition (86.5 minutes and 6709 words in total), mothers were recorded when unpacking a treasure chest (containing toys) together with their children (all female, between 19 and 24 months). In the ADS1 condition (low cognitive load; 30.0 minutes and 3868 words in total), mothers unpacked the same objects from the treasure chest and talked about them with an adult female experimenter. Finally, in the ADS2 condition (high(er) cognitive load; 9.7 minutes and 1079 words in total), mothers performed a map task (cf. [1]) with the experimenter as the interlocutor. Specifically, the mothers' maps contained a path leading through the depicted objects and their task was to guide the experimenter's way to the destination, see [10] for a more detailed description. These two ADS conditions differ in cognitive load, since ADS1 allows free spontaneous speech on everyday-objects (e.g., a cat, a cup etc.), while ADS2 the speaker needs to find her way through the map in order to guide the interlocutor.

The corpus was annotated following the DUEL guideline, which allows for on-the-fly markup of spontaneous speech elements (DisflUencies, Exclamations and Laughter [11]). Four labelers (authors) received an introduction to the DUEL manual and segmented utterances based on syntactic and pausal criteria in Praat [12]. For each utterance, they marked *lengthening* of words, *filled pauses* (e.g., 'ähm'), *silent pauses*, *abandoned utterances* ('I am leaving - what was the task again?')), and *repairs* (('and the-+ and then) to the left')).

A fifth expert annotator (author) checked the files for compliance with DUEL.

To test the hypotheses outlined in the introduction, we calculated Chi-Square tests to compare two (of the three) conditions in regard to one hypothesis. With respect to H1 (more disfluencies in ADS than in IDS), we chose ADS1 as speech register since in IDS and ADS1 the communicative situation is similar: In IDS the mother talks to the child about object in the treasure chest, in ADS1 she talks to the interviewer about objects in the treasure chest. With respect to H2 (more disfluencies when cognitive load is higher), we compared ADS1 and ADS2 (map task) which had the same register, but the map task required a higher cognitive effort.

RESULTS

Tab. 1 shows the occurrences of the different types of disfluencies per 1000 words in the three conditions. Chi-Square tests assessed the difference in occurrences of disfluencies across conditions. P-values were adjusted using the Benjamini-Hochberg correction [13] to account for the fact that multiple comparisons were run (10 altogether), one for each type of disfluency. As level of significance we chose alpha=0.05. We report the original p-values, along with the adjusted p-values (p_{adjust}).

Tab. 1: Number of occurrences (per 1000 words) for the different types of disfluencies. The values in brackets indicate the absolute numbers of occurrence. N gives the total number of words in each condition.

Condition	**Type of disfluencies**				
	Lengthening	Filled pause	Silent pauses	Aband. utts	Repairs
ADS1 (Treasure Chest) (N=3869 words)	35 (134)	23 (88)	109 (422)	11 (42)	19 (73)
ADS2 (Map Task) (N=1079 words)	117 (126)	27 (29)	151 (163)	15 (16)	14 (15)
IDS (N=6709 words)	56 (378)	2 (16)	67 (447)	11 (77)	8 (52)

H1 (IDS vs. ADS1). As predicted, there were more disfluencies in German ADS than in IDS, Tab. 2a for an overview of results of the Chi-Square Test. Specifically, there were more occurrences of filled pauses (χ^2=16.5, df=1, p<0.0001, p_{adjust}<0.001), and silent pauses (χ^2= 10.3, df=1, p=0.001, p_{adjust}=0.005), and more repairs (χ^2=4.6, df=1, p=0.03, p_{adjust}=0.05). Interestingly, instances of lengthening were more frequent in IDS than in ADS1 (χ^2=5.2, df=1, p=0.02, p_{adjust}=0.05). A preliminary analysis of the lengthened instances in IDS reveals that they were not primarily due to hesitations but used for accentuation purposes (for a distinction of different types of lengthening, see [14]).

H2 (ADS1 vs. ASD2). We predicted generally more disfluencies in ADS2 (map task with high cognitive load) than in ADS1 (treasure chest, low cognitive load). This prediction held for the occurrences of *lengthening* (χ^2= 25.1, df=1, p<0.0001, p_{adjust}<0.0001) and *silent pauses* (χ^2= 6.8, df=1, p=0.009, p_{adjust}=0.02). For the other disfluency types there was only a trend, see Tab. 1 for proportions, Tab. 2b for results of the Chi-Square Test.

Tab. 2: Overview of results of Chi-Square Tests for the different types of disfluencies. "Yes" indicates that there is a difference between conditions (direction in brackets); "No" that there is no evidence to assume a difference in the distribution of disfluent occurrences across conditions.

Chi-Square Test	**a) Comparison IDS vs. ADS1 (treasure chest)**				
	Lengthening	Filled pause	Silent pauses	Aband. utts	Repairs
Difference (yes / no)	**Yes** (IDS > ADS1)	**Yes** (ADS1 > IDS)	**Yes** (ADS1 > IDS)	No	**Yes** (ADS1 > IDS)
	b) Comparison ADS1 (treasure chest) vs. ADS2 (map task)				
	Lengthening	Filled pause	Silent pauses	Aband. utts	Repairs
Difference (yes / no)	**Yes** ADS2 > ASD1	No	**Yes** ADS2 > ASD1	No	No

DISCUSSION

We compared the occurrence of disfluencies in a) different speech registers (IDS vs. ADS) and b) ADS under different amounts of cognitive load (low vs. high). Regarding a) IDS shows only few instances of disfluencies while the ADS condition with low cognitive load shows a higher number of *filled pauses*, *silent pauses* and *repairs*. Hence, our findings on German fit previous studies on fluency in IDS in other languages [4-6]. Yet, there are more instances of *lengthening* in IDS than in ADS, which, based on a preliminary analysis, are not indicative of hesitations but mainly used for highlighting. German IDS has been shown to exhibit many accents [15], more than one would expect in ADS. The difference in lengthening may thus be due to a larger number of accents in IDS than in ADS.

Regarding b), we expected the map task to generally elicit more disfluencies, but this was only the case for *silent pauses* and most strongly for *lengthening.* We see two explanations that may account for the strong effect of lengthening in the map task: First, it might be a strategy of keeping pace with the interlocutor who finds her way through the map. Second, it could be an indication of an iconic relation between the task object and the speech phenomena describing it (path as a continuous line transferred to speech). Yet, it is possible that a more demanding task would further increase the number of disfluencies and hence increase the difference between the ADS conditions differing in cognitive load.

Possibly, disfluencies would also increase in IDS when cognitive load gets higher, e.g., when caretakers are mentally distracted, which provides an interesting hypothesis for future research. As mentioned above, infants distinguish fluent from disfluent speech [5, 7] and use disfluencies for word recognition [8, 9]. In future work, we plan to test whether disfluencies may serve as indicators for different speech registers in German and to what extent children might benefit from this.

ACKNOWLEDGMENTS

The authors would like to thank Katrin Busse for data collection (as part of her bachelor thesis in the Babylab in Konstanz). We are further grateful to Bettina Braun, the head of the Babyspeech Lab in Konstanz, for making the recordings available. Finally, we want to say thank you to the mothers and their children for participating in the study.

References

[1] Belz, M. and Klapi, M., "Pauses following fillers in L1 and L2 German map task dialogues," in *Proceedings of the Disfluency in Spontaneous Speech Workshop (DiSS-2013)*, 2013, pp. 9-12.

[2] Nicholson, H. B. M., Bard, E. G., Anderson, A. H., Flecha-Garcia, M. L., Kenicer, D., Smallwood, L*., et al.*, "Disfluency under feedback and time-pressure," in *Proceedings of EUROSPEECH-2003,* 2003, pp. 205-208.

[3] Eklund, R., "Disfluency in Swedish human–human and human–machine travel booking dialogues," PhD thesis, Department of Computer and Information Science, Linköping University, Sweden, 2004.

[4] Newport, E. L., Gleitman, H., and Gleitman, L. R., "Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style," in *Talking to children: Language input and acquisition*, Ferguson, S., Ed., Cambridge, UK: Cambridge University Press, 1977, pp. 109-149.

[5] Soderstrom, M. and Morgan, J. L., "Disfluency in speech input to infants? The interaction of mother and child to create error-free speech input for language acquisition," in *Proceedings of the Disfluency in Spontaneous Speech Workshop (DiSS'05)*, Aix-en-Provence, France, 2005, pp. 157-162.

[6] Björkenstam, K. N., Wiren, M., and Eklund, R., "Disfluency in child-directed speech," in *Proceedings of Fonetik 2013, the XXVIth Swedish Phonetics Conference, Studies in Language and Culture*, Linköping, Sweden, 2013.

[7] Soderstrom, M. and Morgan, J. L., "Twenty-two-month-olds discriminate fluent from disfluent adult-directed speech," *Developmental Science,* vol. 10, pp. 641-653, 2007.

[8] Kidd, C., White, K. S., and Aslin, R. N., "Toddlers use speech disfluencies to predict speakers' referential intentions," *Developmental Science,* vol. 14, pp. 925-934, 2011.

[9] Morin-Lessard, E. and Byers-Heinlein, K., "Uh and euh signal novelty for monolinguals and bilinguals: evidence from children and adults," *Journal of Child Language,* vol. 46, pp. 522-545, 2019.

[10] Busse, K., *Are there differences between German IDS and ADS compared to Japanese?*, Bachelor Thesis, Department of Linguistics, University of Konstanz, 2019.

[11] Hough, J., de Ruiter, L., Betz, S., and Schlangen, D., "Disfluency and laughter annotation in a light-weight dialogue mark-up protocol," in *Proceedings of the Disfluency in Spontaneous Speech Workshop (DiSS-2015)*, Edinburgh, UK 2015.

[12] Boersma, P. and Weenink, D. (2016). *Praat: doing phonetics by computer. Version 6.0.23 (version depended on labeller) [Computer program].*

[13] Benjamini, Y. and Hochberg, Y., "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society,* vol. 57, pp. 289-300, 1995.

[14] Betz, S., Wagner, P., and Voße, J., "Deriving a strategy for synthesizing lengthening disfluencies based on spontaneous conversational speech data," in *Proceedings of the 12. Tagung Phonetik und Phonologie im deutschsprachigen Raum*, München: LMU, 2016, pp. 19-22.

[15] Zahner, K., Schönhuber, M., Grijzenhout, J., and Braun, B., "Konstanz prosodically annotated infant-directed speech corpus (KIDS corpus)," in *Proceedsings of the 8th International Conference on Speech Prosody*, Boston, USA, 2016, pp. 562-566.

On the Multifunctionality and Multimodality of Silent Pauses in Native and Non-native Interactions

Loulou Kosmala

PRISMES EA 4398/SeSyLiA, Sorbonne Nouvelle University, Paris, France
loulou.kosmala@sorbonne-nouvelle.fr

This paper explores the multifunctional uses of silent pauses, adopting a multimodal perspective. Pauses are said to serve numerous functions in speech, and this small comparative study looks at their distribution across L1 and L2 speakers in face-to-face interactions. While overall results showed a higher rate of pauses in L2 than in L1, a lot of individual differences were also found. The visual-gestural features of discourse (gesture, gaze, facial expressions) were found to be essential when analyzing pauses as they can reflect their multifunctionality in a different modality, which allows for a deeper understanding of their internal processes in L1 and L2.

INTRODUCTION

When we speak spontaneously, we constantly need to produce pauses for a variety of reasons. From a strictly physiological point of view, pausing is necessary to allow us to breathe and inhale between cycles of speech; but pauses can also serve several functions in discourse, such as planning and structuring speech, marking a word, holding the floor, or allowing other people to take the turn. This paper focuses on the use of silent pauses, i.e. "silent periods between vocalizations" [1].

Silent pauses have been studied from several perspectives: they can be viewed as hesitation or disfluency markers [2, 3] which suspend the speech flow and occur when speakers are uncertain or thinking what to say next. In non-native speech, L2 speakers tend to produce more pauses of longer duration and in mid-clause position before low-frequency words [4,5] which can be explained by their limited proficiency of the language [1]. However, pauses can also be viewed as important components of interaction, as they serve several communicative functions. In *Conversation Analysis*, a delay in speech can be interpreted as meaningful as it can be a sign of a dispreferred answer in assessments [6], and pauses can also be used to manage turn-taking [7]. The use of pauses is also affected by several factors, such as individual variation, speaking style, and the type of speech produced [8].

Grounded in a functional, dynamic, and interactional approach to grammar [9], this paper aims to explore the multifunctionality of silent pauses in native and non-native interactions of French and English, following a multimodal perspective [10]. A few studies have looked at the production of gestures during pauses in native and non-native conditions [11, 12], but it is still an aspect that has not been widely explored in pausing phenomena. Since pauses are known to serve a variety of functions, the visual-gestural modality of discourse (eye gaze, facial expression, manual gestures) can provide a deeper understanding of their different internal processes. The main hypotheses for this study are that (1) silent pauses occur more frequently in L2 speech than in L1, but they appear in similar functional contexts, (2) L2 speakers will make use of more gestures during pauses than L1 speakers and these gestures can reflect their multifunctionality.

METHOD

This small study of pauses was conducted on a sample of the SITAF Corpus [13] which includes 12 video recordings of dyadic interactions between French and American students (undergraduate level) engaged in semi-directed conversations (6 pairs). The speakers interacted respectively in English and in French, alternating from L1 to L2. The methodology used for this analysis is adapted from previous studies conducted on the use of (dis)fluencies [14, 15] which looked at the distribution of several (dis)fluency markers (filled pauses, silent pauses, repetitions, repairs etc.). The silent pauses were extracted from the data and were coded manually according to their (a) duration in ms (400ms minimum duration threshold, following [4]) (b) level of combination, whether they appeared isolated, or combined with other (dis)fluency markers, and their (c) functional contexts. The term "function" here is replaced by "context" as it is difficult to tell whether a single silent pause serves one attributed function as it can always be affected by its co-occurring markers, its context of use, and the accompanying manual gestures. Four functional contexts were defined: (1) structuring—contexts in which the speaker is currently planning, structuring, emphasizing parts of their speech, which can be indicated by the presence of discourse markers (and, but, or); (2) uncertainty—contexts in which the speaker shows signs of uncertainty (e.g. frown); (3) interactive—contexts in which the speaker is engaged in the interaction and expresses their stance (e.g. gaze towards the interlocutor, interactional gesture), (4) undefined. Since pauses can also be used for physiological reasons or other reasons that are unknown or very difficult to detect, this category was used when pauses occurred in contexts which did not apply to those described above. Lastly, the gestures which occurred during pauses were also annotated, by looking at the "gesture phrase" [10] whether the gestures remained in rest position (on the lap), were held in the same position, or if they were fully produced. The functional types of gestures were also annotated, mainly (1) *referential gestures*, which are related to the meaning conveyed in the content of speech, (2) *deictic-anaphoric* gestures, which place referents in time and space through pointing, (3) *parsing gestures*, used for emphasis or marking speech segments (production-oriented), (4) *thinking gestures*, metapragmatic gestures produced during communication breakdown, (5) *interactional gestures*, which enact a communicative action (speech act, interactional move-interaction-oriented).

RESULTS

A total of 468 silent pauses were found in the data. On average, L2 speakers produced significantly more silent pauses than L1 speakers (p=0.0004). They produced 6,3 silent pauses per hundred words (N=288) as opposed to L1 speakers who produced 3,7 (N=180), which corroborates previous findings. Less significant differences were found in the duration of the silent pauses, as they lasted on average 739ms in L1 and 795ms in L2 (STDV 428 and 458) but a lot of individual differences were found. For example, one American speaker produced pauses of an average duration of 1197ms in his L1 and of 846ms in his L2, which does not support previous predictions [5], so no clear-cut conclusions can be made at this point. Silent pauses were also found to occur more frequently in co-occurrence with other (dis)fluency markers in L2 (54% N=156) than in L1 (38% N=69), which shows that silent pauses often combine with other markers, such as filled pauses [16,17,18], but the fact that L2 speakers produced more clusters of pauses and (dis)fluency markers may indicate that they resort to different stalling mechanisms and need to fill the silence with other vocal markers.

Results also show that the pauses behaved very similarly in L1 and L2 as no significant differences (p=0.08) were found in their distribution according to the functional contexts: 44% for undefined (N=79/180 in L1; N=128/288 in L2), 7% for uncertainty (N=11/180 in L1; N=19/288 in L2), 9% for interactive (N=16/180 in L1; N=21/288 in L2), and 40% for structuring (N=71/180 in L1; N=114/288 in L2). This finding shows that a lot of pauses tend to be produced in a semi-automatic way for no clear pragmatic reasons other than just buying time in speech (44%) and for structuring and segmenting speech (40%), but it does not seem to be affected by language proficiency as the native and non-native conditions show similar behavioral patterns. But once again, a lot of individual differences were found, as some speakers produced much more pauses in contexts of uncertainty in their L2 than in their L1. Additionally, more gestures occurred with pauses in L2 speech (56% N=163) than in L1 speech (36% N=65) (p=0.01), which partially confirms the view that L2 speakers produce more gestures than L1 speakers [18], but this needs to be confirmed by looking at all the gestures in the data. L2 speakers also produced more thinking gestures in pauses (27% of the completed gestures) than L1 speakers (17%) while L1 speakers produced more parsing gestures in pauses (46% of their completed gestures, as opposed to 38% for L1 speakers). These findings point out the gestural activity found during pauses, and this will be illustrated in the following qualitative analyses.

Qualitative analyses: example from a pair

As it has often been mentioned in the literature, the use of (dis)fluency markers varies strongly per speakers [20], so it is essential to look at the individual profiles of speakers when investigating the use of silent pauses. This paper will now focus on two qualitative analyses taken from one pair of the corpus. In these examples, the two participants are interacting in French and are discussing the differences between being a traveler and being a tourist.

In the first example, the non-native speaker (NNS) is talking about travelers who get to stay in a foreign country for a longer time, but she mispronounces the noun "longer" *(plus longtemps*). She quickly realizes her mistake, and as Fig. 1 shows, she then produces a silent pause of 580 ms, but she also holds her hands in the same position, and with her gaze fixed on her interlocutor she slightly moves her head in the direction of her partner, to elicit help.

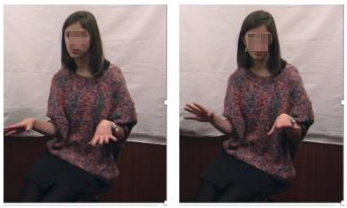


Fig 1. Multimodal activity during a silent pause: implicitly seeking help from the interlocutor

The native speaker (NS) quickly understands her partner's request, and gives her the right pronunciation of the word; then NNS repeats the target word, this time with the right pronunciation, and finishes her sentence. This example shows that NNS did not verbally seek help from her interlocutor, but she instead relied on multimodal resources (pausing, holding her hands, and gazing) to ask for the right pronunciation and to give her partner the floor, which stresses out the interactional dimension of pauses.

In the second example, NS is talking about how tourists tend to see trips as 'to-do lists' as they aim to go to one place and see different monuments. When she starts saying that tourists like to go to one place, she produces a silent pause before the noun "place" and produces a vertical movement with her right hand towards her left open palm (Fig. 2). She then repeats the same gesture after saying the word 'place' and repeats it again as she produces a lexical repetition: “this monument, this monument, this monument”. The type of gesture produced in the pause is different from the one in Fig. 1 as in this case she is rather “marking the words” (see [12]), in other words, she is emphasizing key words in her utterance, as to draw attention to them. When she first initiated this gesture, it was during the silent pause, which stresses out its pragmatic use. She may have produced it to make her speech clearer for her partner (interaction-oriented) or to mark distinct speech segments (production-oriented). In any case, the gesture accompanying the silent pause was found to be informative of the ongoing processes associated with her pausing.

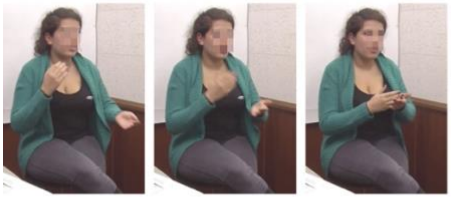


Fig 2. Initiation of a parsing gesture during a pause: marking the words

CONCLUSION

To conclude, this small comparative study has stressed out the multifunctionality and multimodality of silent pauses in native and non-native speech. Silent pauses can be used as strategies to resolve speech difficulties, to emphasize parts of speech, to make an interactional move, or simply to buy time. While some of them appear to be semi-automatic, it is important to take individual differences into account, as speakers make use of them in different ways to achieve different things. The multimodal approach used in this study was found to be essential when analyzing the interactional dimension of pauses in the qualitative analyses. In this view, pauses are no longer ‘silent’ or ‘unfilled’ in the multimodal sense as they can be filled with other rich semiotic resources.

REFERENCES

[1] Cenoz, J. (1998). Pauses and Communication Strategies in Second Language Speech. (ERIC Document ED 426630).

[2] Goldman-Eisler, F. (1958). The predictability of words in context and the length of pauses in speech. *Language and Speech, 1*(3), 226–231

[3] Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word, 15*(1), 19–44.

[4] Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers. *ELT Journal, 65*(1), 71–79.

[5] De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching, 54*(2), 113-132.

[6] Roberts, F., Francis, A. L., & Morgan, M. (2006). The interaction of inter-turn silence with prosodic cues in listener perceptions of “trouble” in conversation. *Speech Communication, 48*(9), 1079–1093.

[7] Schegloff, E., Jefferson, G., & Sacks, H. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696–735.

[8] Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech, 25*(1), 11–28.

[9] Mondada, L. (2001). Pour une linguistique interactionnelle. *Marges Linguistiques, 1*, 142–162.

[10] Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press

[11] Tellier, M., Stam, G., & Bigi, B. (2013). Gesturing while pausing in conversation: Self-oriented or partner-oriented?”. *The Combined Meeting of the 10th International Gesture Workshop and the 3rd Gesture and Speech in Interaction Conference.*

[12] Stam, G., & Tellier, M. (2017). The sound of silence. *Why Gesture?: How the Hands Function in Speaking, Thinking and Communicating, 7*, 353.

[13] Horgues, C., & Scheuer, S. (2015). Why some things are better done in tandem. In *Investigating English Pronunciation* (47–82).

[14] Kosmala, L., & Morgenstern, A. (2017). A preliminary study of hesitation phenomena in L1 and L2 productions: A multimodal approach. *TMH-QPSR*, 37.

[15] Kosmala, L., Candea, M., & Morgenstern, A. (2019). Synchronization of (Dis)fluent Speech and Gesture: A Multimodal Approach to (Dis)fluency. *Gesture and Speech in Interaction.* Presented at the Paderborn University. Paderborn University.

[16] Betz, S., & Kosmala, L. (2019). Fill the silence! Basics for modeling hesitation. *The 9th Workshop on Disfluency in Spontaneous Speech*, 11.

[17] Grosjean, F., & Deschamps, A. (1972). Analyse des variables temporelles du français spontané. *Phonetica, 26*(3), 129–156.

[18] Merlo, S., & Mansur, L. L. (2004). Descriptive discourse: Topic familiarity and disfluencies. *Journal of Communication Disorders, 37(*6), 489–503.

[19] Graziano, M., & Gullberg, M. (2018). When speech stops, gesture stops: Evidence from developmental and crosslinguistic comparisons. *Frontiers in Psychology, 9*, 879.

[20] Betz, S., & Gambino, S. L. (2016). Are we all disfluent in our own special way and should dialogue systems also be? Elektronische *Sprachsignalverarbeitung (ESSV) 2016*, 81.

Development of models for the automatic detection of prosodic boundaries in spontaneous speech

Barbara Heloha

Speech is segmented into intonational units marked by prosodic boundaries for cognitive and linguistic reasons. This work aims to: investigate the phonetic-acoustic parameters that guide the production and perception of prosodic boundaries; to develop models for automatic detection of prosodic boundaries in spontaneous speech. Two samples of male spontaneous speech excerpts were segmented into intonational units by two groups of trained annotators. The prosodic boundaries perceived by them were noted as or non-terminal or terminals. A script was used to automatically extract phonetic-acoustic parameters along speech signal, positions at which at least 50% of the annotators indicated a boundary of the same type were considered as prosodic boundary. A training of models composed by the combination of multiple parameters designed to the automatic identification of boundaries marked by the annotators was developed using Linear Discriminant Analysis (LDA). The automatic terminal boundary model shows a convergence of 80% in relation to the terminal boundaries identified by the annotators in sample I and 75% in sample II. For the non-terminal boundaries, three statistical classification models were obtained. Together, the three models show a 98% convergence in relation to the non-terminal boundaries indicated by annotators in sample I and 88% in sample II.

INTRODUCTION

Speech is prosodically segmented into intonation units determined by prosodic boundaries for cognitive and linguistic reasons. These units can be functionally analyzed according to different theoretical perspectives like syntactic, pragmatic or cognitive [1-3]. However, prosodic boundaries can be studied *per se*, independently of the theoretical perspective from which the units are observed, since these boundaries are clearly perceivable by listers. Prosodic boundaries are normally associated with the perception of conclusion or continuation of the intonation unit. In general, the first type is called terminal boundary (TB), the second one is called non-terminal (NTB). They are also signalled in the speech flow through acoustic parameters. Some acoustic parameters are commonly considered in the literature as acoustic correlates of prosodic boundaries, some of them are silent pause, pre-boundary lenghtening, reset of the fundamental frequency (f0) and change in speech rate. However, the acoustic distinction between terminal and non-terminal boundaries is not so clear until these moment. Therefore, tools for automatic detect terminal and non-terminal prosodic boundaries are not currently avaible.

OBJECTIVES

This work aims: to investigate the phonetic-acoustic parameters that guide the production and perception of prosodic boundaries; to develop automatic models for detect prosodic boundary in Brazilian Portuguese spontaneous speech.

DATA AND DATA TREATMENT

Full data includes 14 excerpts of monologic spontaneous speech extrated from the three sections of the C-ORAL-BRASIL corpora [4, 5], namely the informal and the formal sections of the natural context and the media section. The excerpts have on average 190 words and they are organized into two different samples composed by seven excerpts (Samples I and II). The samples were segmented into intonational units by two groups of trained annotators. The first group includes 14 annotators and segmented the sample I, the second one includes 19 annotators and segmented the sample II. The annotators received the audio file and the orthographic transcription without any annotation of prosodic boundaries, their task was to annotate the two main types of boundaries following their perception using a simple slash symbol (/) to indicate a non-terminal boundary and a double slash one to indicate a terminal boundary (//). The agreement among the annotators was evaluated through the Fleiss kappa coefficient. For the excerpts in first sample, the general interrater agreement was 0.80 for the annotation of terminal boundaries and 0.75 for non-terminal ones. For the excerpts in second sample, the general interrater agreement was 0.73 for the annotation of terminal boundaries and 0.72 for non-terminal ones.

The audio files are annotated with Praat TextGrid objects [6] with five tiers as follows: 1) Vowel-to-Vowel interval tier accompanied by a broad phonetic transcription (Vowel-to-Vowel interval units were delimited by two consecutive vowel onsets); 2) Phonological word point tier with points at every phonological word boundary. In each point tier, it was informed how many annotators signalled that point as a non-terminal prosodic boundary; 3) Phonological word point tier with points at every phonological word boundary. In each point tier, it was informed how many annotators signalled that point as a terminal prosodic boundary; 4) Interval tier delimiting silent pauses. 5) Textual transcription of utterances.

In order to generate models for the automatic detection of prosodic boundaries, the Praat script BreakDescriptor [7] was used to automatically extract phonetic-acoustic parameters along speech signal. This Praat script extracts the acoustic-phonetic parameters for all the V-

V units in a window centered in all the boundaries between phonological words. The windows scanned by the BreakDescriptor includes ten V-V units to the left and ten V-V units to the right of each analyzed V-V unit (those at the boundary of phonological words).

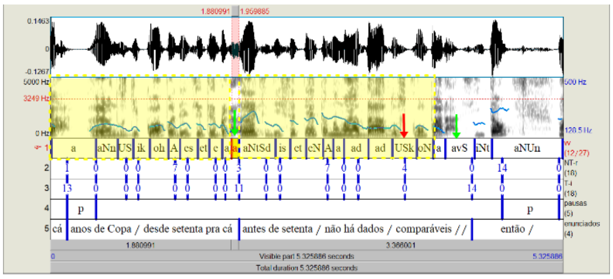


Fig. 1. Starting at the top: wave form, broad-band spectrogram, and the different tiers in a Praat TextGrid. The position of terminal boundary used here for the analysis is highlighted in yellow; it constitutes the central point of the analyzed window.

The Praat script BreakDescriptor consider as prosodic boundary positions with at least 50% of agreement among annotators. Thus, in first group, positions at which at least seven annotators indicated a boundary of the same type were considered as prosodic boundary. In second one, positions at which at least ten annotators indicated a boundary of the same type were considered as prosodic boundary. The remaining positions located at phonological word boundaries were considered by the script as non-boundary. The Praat script BreakDescriptor extract five groups of acoustic parameters for the positions perceived as terminal, non-terminal, non-boundary and their respective windows: 1) Measures for speech rate and rythm (6 measurements); 2) Measures for normalized duration of V-V units (34 measurements); 3) Measures for fundamental frequency (65 measurements); 4) Measures for intensity (4 measurements); 5) Measures for silent pause (4 measurements). In total, it evaluates 111 measurements.

STATISTICAL ANALYSIS

The Linear Discriminant Analysis (LDA) algorithm was used to develop models composed by the combination of multiple parameters destined to the automatic identification of boundaries marked by the annotators. The process of training models was heuristic and evaluated all measures extracted by BreakDescriptor. In general, the main idea of this process is improving the performance of the algorithm by looking for a more accurate prediction of segmentation performed by annotators.

The 111 parameters, used as predictors, had their number reduced following two different heuristics. Firstly, we gradually eliminated the parameters ranked as less relevant in the hierarchy of the algorithm, following the weight assigned by the model. Secondly, we reintroduced or eliminated some parameters based on the findings available in the literature and not only on the weight attributed by the algorithm. For non-terminal boundaries a third step was needed, since the two first phases did not yield a satisfying result. In order to have a better prediction of non-terminal boundaries, we eliminated positions of TB because NTB model did not recognize the disctintion between TB and NTB and we looked for models that could explain sub-groups of boundaries characterized by different configurations of acoustic parameters. After the best result obtained with the first model, we eliminated the boundaries that this model correctly detected and developed a new model for the remaining boundaries still not identified automatically.

RESULTS

The terminal model shows a convergence of 80% in relation to the terminal boundaries marked by the annotators in sample I. For the non-terminal boundaries, three statistical classification models were obtained. Together, the three models show a 98% convergence in relation to the non-terminal boundaries indicated by annotators in sample I. The models were validated later in sample II. The results of the validation indicate that the performance of the model dedicated to the terminal boundaries is 75% correct in the second database. The models for non-terminal boundaries identify 88% of the non-terminal boundaries marked in sample II. At least in the full database used in this study, terminal boundaries seems to be more typified, while non-terminal appears to be more stratified.

The models found also presents a hierarchy of acoustic parameters and describes their relative importance in each model. Up to this point, the results of the research shows that pauses and f0 measurements are very important for terminal boundaries and that this is not true for non-terminal ones. On the contrary, measurements of speech rate and of normalized duration of the pre-boundary segments seens to be very important for non-terminal boundaries and less important for terminal ones. Measurements of intensity do not show much weight in both type of prosodic boundaries. The first model for detect NTB is mainly marked by measurements of pause and lengthening immediately next to the boundary. The second model model for detect NTB is mainly characterized by measurements of speech rate, articulation rate and some f0 measurements, the thrid one is mainly characterized by measurements of peak rate of smoothed z-score.

REFERENCES

[1] SELKIRK, E. (2005) Comments on Intonational Phrasing in English. In: Frota, S.; Vigário, M. & Freitas, M.J. *Prosodies*, Berlin, Mouton de Gruyter, 11-58.
[2] Cresti, E. (2000) *Corpus di Italiano parlato*. v. 1. Firenze, Accademia della Crusca.

[3] Chafe, W. (1994) Discourse, consciousness anda time: The Flow and dsiplacement of Conscious Experience in Speaking and writing. Chicago, University of Chicago Press.

[4] Raso, T.; Mello, H. (2012) *C-ORAL-BRASIL I: corpus de referência do português brasileiro falado informa*l. 1ed. Belo Horizonte, UFMG.

[5] Raso, T.; Mello, H.; Ferrari, L. (in press). *C-ORAL-BRASIL II: corpus de referência do português brasileiro*.

[6] Boersma, P.; Weenink, D. (2015) *Praat: doing phonetics by computer*. Software.

[7] Barbosa, P. BreakDescriptor (2016/ 2019). Script for Praat. Available with the author.

How prosody and context shape the acoustic nature of rhetorical questions in German

Jana Neitsch[1*] and Oliver Niebuhr[1]

[1]*Centre for Industrial Electronics, Mads Clausen Institute, University of Southern Denmark, Sønderborg/DK*
* neitsch@mci.sdu.dk

In contrast to what is known as information-seeking questions (ISQs), rhetorical questions (RQs) usually occur in non-neutral contexts since they are frequently used to challenge, criticize, or persuade the addressee. This pilot study investigating the prosodic realization of two attitudinally loaded *wh*-RQs (*disgust* and *mockery*; e.g., *Who likes lavender?*) in German relative to a string-identical sincere interest control condition (ISQs) indicates that RQs show context-specific acoustic and phonological differences. Overall, the way in which German RQs and ISQs differ from each other as well as from statements suggests that questions represent a coherent bundle of prosodic parameters (i.e., prosodic construction) consisting rather of gradual phonetic parameters than of categorical phonological parameters.

INTRODUCTION

Information-seeking questions (ISQs) provide the listener with information that is usually given by the addressee [1,2]. In contrast, rhetorical questions (RQs) imply answers that are known to both the speaker and the listener [3] and hence seek the addressee's commitment with respect to the underlying proposition [4]. For the correct differentiation between RQs and ISQs (especially if they are string-identical), context plays a crucial part.

Isolated from context, *Who likes playing soccer?* could be interpreted as both an ISQ or an RQ, as in German [5, 6]. Regarding RQs, context has been considered as most salient determiner [4, 5, 6] since it facilitates the "understanding of the question as not doing questioning" [7, p.55]. The findings presented by [8] suggest that RQs are, just like ISQs, multi-parametric prosodic entities ([9]).

Prosodically, RQs are not closer to statements than ISQs, even though RQs are often perceived as stating. Both RQs and ISQs differ from statements along the same prosodic parameters, and RQs even more from statements than ISQs with respect to duration and voice quality [8, 9]. A potential reason for this might be the ironic flavor that characterizes many RQs, since irony is a major function of RQs [5,10,11,12]. This ironic flavor might be expressed by an exaggerated prosody. This ironic overtone enables speakers to criticize and persuade the addressee, and even to create humor [12,13,14,15,16,17,18,19].

This pilot study focuses on RQ realizations across attitudinal contexts and respective ISQ realizations and the question whether which prosodic parameters (acoustic-phonetic and phonological) vary as a function of the question-statement difference itself and which parameters show a separate pattern that might be contextually/attitudinally driven.

The multi-parametric prosodic variation of string-identical *wh*-questions is particularly investigated in different contexts triggering *sincere interest* (ISQs as contextual reference condition) or the two RQ types *disgust* and *mockery*. Both disgust and mockery represent typical and at the same time very different attitudinal/expressive realizations for RQs. The selection of acoustic and phonological prosodic parameters is based on those that turned out to be crucial for both the production and the identification of German ISQ and RQ in previous empirical investigations (e.g., [8,9]).

METHOD

We designed 6 target *wh*-questions (e.g., *Wer mag denn Lavendel?* 'Who likes PRT lavender?'; *wh*-word, verb, modal particle *denn* – occurring in both illocution types in German [20] – sonorous sentence-final object noun) that could be interpreted as both RQs and ISQs. Each of the questions was paired with short contexts as texts triggering a *disgust*, *mockery* and *sincere interest* attitude on the part of the speaker (Tab. 1). Each context introduced the object noun to make sure that it was prosodically realized as given information in the subsequently realized *wh*-question.

Tab. 1: *Examplar contexts for RQ/ISQ elicitations.*

Mockery	Disgust	Sincere interest
Your mother tells you that her neighbor read in the newspaper that **lavender** can be eaten. The other day she observed him sit-ting in the garden and eating the blossoms which **you both find extremely funny.** You say:	You and your friend walk into a perfumery where a woman instantly offers you a new scent with **lavender**. But you and your friend find the scent so gross that you quickly continue walking because **you are feeling nauseous.** You say:	You and your roommates want to plant a small flowerbed in your garden, and you have always dreamt of **lavender**. **You are very interested** in whether the others agree with that. You say:
Who likes lavender?		

In total, 90 questions were recorded (5 speakers x 6 questions x 3 attitudes) by 5 voluntary monolingual native speakers of German, who have been recorded in a sound attenuated booth (ø = 22.6 years, 2 male). They were presented with the context-question pairs in a Power Point presentation and were asked to read the given context silently followed by the realization of the respective question. Occurrences of the same question were separated as far as possible from one another. Participants had to realize the question aloud as naturally as possible fitting into the situational context.

RESULTS

Fig. 1 summarizes (and proportionally for each parameter) to what extend and in which way the question realizations in the three context conditions differ from each other.

Nuclear accents: The results showed that sentence-final object nouns in *disgust* contexts were exclusively realized with rising pitch patterns: L*+H (73%) or L+H* (23%). In *mockery* contexts, more than every fourth object noun had non-rising pitch patterns like H* (10%) and L* (10%). Rising L*+H was also the most frequent pattern in *mockery*, though (63%, L+H*: 10%). The variation in nuclear pitch accent patterns was largest in the *sincere-interest* contexts, were the low or falling patterns prevailed (L*+H: 37%, L+H*: 10%, L*: 47%, H*: 3%).

Final boundary tone: In *disgust* and *mockery* contexts, questions were predominantly realized with a low boundary tone (L%: 90% and 97%), whereas questions in *sincere-interest* contexts mainly showed a high final boundary tone (H-^H%: 47%, L-H%: 37%).

Initial pitch level: In *mockery* contexts (207 Hz), the level was higher than in *disgust* contexts (195 Hz). Questions realized in *sincere-interest* contexts started on average at an intermediate pitch level (203 Hz).

Overall question duration: Questions realized in *disgust* contexts were on average longer (1384 ms) than those in *mockery* contexts (1351 ms). Both of them had longer overall durations than those realized in *sincere-interest* contexts (1141 ms). Similar to the findings in [8], the noun was longest in *disgust* contexts (807 ms), with an intermediate duration in *mockery* contexts (772 ms), and the shortest duration in *sincere-interest* contexts (620 ms).

Voice quality (VQ: based on vowel midpoints): *Wh*-words were realized with a breathier VQ in the *sincere interest* (13.3 dB) and *disgust* contexts (13.5 dB) than in the *mockery* contexts (14.1 dB). Regarding the verb, however, VQ was breathier in the *disgust* contexts (11.6 dB) than in the *mockery* (13.3 dB) and the *sincere interest* contexts (13.6 dB). This tripartite difference became stronger within the sentence-final object noun (*disgust*: 14.5 dB, *mockery*: 15.1 dB, *sincere interest*: 16.4 dB).

F0-shape: The tonal-target labels L and H were taken to determine and compare the f0 shapes of the rising and falling slopes in the nuclear tunes. Only those nuclear tunes were taken into account that consisted of tritonal LHL sequences. The three f0 values of L, H, and L were measured as well as the two f0 values halfway in between the three tones. On this basis, the range proportion measure (R_{prop}) was determined, following [21]. The results showed that nuclear question tunes in *disgust* and *mockery* contexts were characterized by clearly concavely shaped f0 rises with R_{prop} values well below 0.5 (0.40 and 0.36), followed by more convexly shaped f0 falls (0.69 and 0.76). The opposite was true for the nuclear question tune in the *sincere-interest* context showing a convex shape with a R_{prop} value of 0.59, followed by a strongly concavely shaped f0 fall of R_{prop} = 0.96.

DISCUSSION

Regarding the general differences between *mockery/disgust* RQs and *sincere interest* ISQs, our results are consistent with [8]. Phonologically, we found, like [8], that the nuclear pitch accent is mainly rising in RQs but falling in ISQs. Regarding the final boundary tone, RQs had a low and ISQs mainly a high boundary tone. This supports what was stressed by [9]: Final boundary tones have a specific meaning, and this is not simply 'question'. Hence, there is no straightforward link between the final boundary tone and sentence mode. This already implies that the prosodic differences between RQs and ISQs do not originate from signaling illocution type, but from different attitudinal stances. Phonetically, both the total duration and the duration of the object noun was greater in RQs than in ISQs. It also accords with [8] that breathiness dominated in RQs as opposed to ISQs. RQ breathiness was strongest on the verb, but the difference to ISQs was strongest on the final object noun.

Our results show that the realization of German RQs is context-specific and that there is not a single stable prosodic RQ profile. Yet, RQs still differ from ISQs along the entire prosodic profile that we analyzed; and the differences between RQs and ISQs were larger than those between the two RQ types *mockery* and *disgust*. Thus, RQs are also not simply "real" questions (ISQs) with prosodic differences, but rather prosodically an entirely different phenomenon. As for the specific nature of this phenomenon, note that, firstly, RQs differ from ISQs along the same prosodic parameters that also distinguish ISQs from regular statements in German [9]. Compared to the latter, ISQs are realized longer, breathier, and with stronger concave rises and convex falls in the LHL sequences of nuclear tunes [22, 23]. RQs seem to exaggerate these differences in terms of being even more question-like than ISQs. This exaggeratedness is also observed in ironic utterances [10-13] where people tend to use a "borrowed voice" instead of their own voice. A similar mechanism could be at work here, e.g. to attenuate the expressed *mockery* or *disgust*. That (esp. *mockery*) RQs started with a higher initial pitch than ISQs supports the idea of an exaggerated RQ prosody (see [9]).

Secondly, note with respect to *mockery* and *disgust* that the prosodic profiles of the realized RQs are not simply one-to-one reflections of the respective emotions, (e.g. [24]). Instead, the corresponding emotion profiles seem to be modified and translated into gradual parallel changes along a parameter profile that distinguishes questions from statements in German.

Overall, RQs and ISQs clearly differ from each other and so do contextual subtypes of RQs. These differences manifest themselves as parallel gradual parameter shifts along the same prosodic profile, which can also turn questions into statements in German.

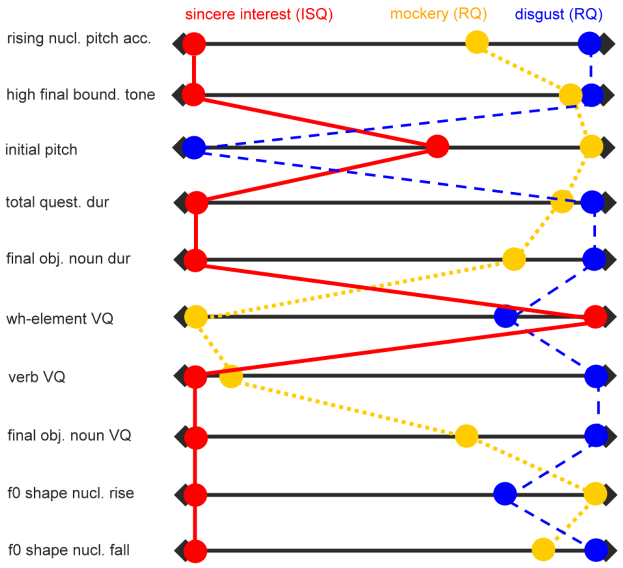


Fig. 1. Value ranges of all parameters from the smallest level or frequency (left) to the highest level or frequency (right). The intermediate level or frequency is displayed proportionally within each value range.

REFERENCES

[1] Groenendijk, J., Stokhof, M., 1984. *Studies on the semantics of questions and the pragmatics of answers.* Amsterdam: Universiteit van Amsterdam PhD Thesis.

[2] Meibauer, J., 1986. *Rhetorische Fragen* (Lingu-istische Arbeiten 167). Berlin: De Gruyter.

[3] Caponigro, I., Sprouse, J. 2007. Rhetorical Ques-tions as Questions, In: Puig-Waldmüller, E. (ed.), *Proc. of Sinn und Bedeutung* 11, Barcelona, Spain: Universitat Pompeu Fabra, 121-133.

[4] Biezma, M., Rawlins, K. 2016. Rhetorical Ques-tions: Severing asking from questioning. In *Semantics and Linguistic Theory* 27, 302-322.

[5] Frank, J. 1990. You call that a rhetorical question?: Forms and functions of rhetorical questions in conversation. *Journal of Pragmatics, 14*(5), 723-738.

[6] Jung, V., Schrott, A. 2002. A question of time? Question types and speech act shifts from a historical-contrastive perspective: Some examples from Old Spanish and Middle English. In: *PRAGMATICS AND BEYOND NEW SERIES*, 345-372.

[7] Koshik, I. 2003. Wh-questions used as challenges. *Discourse Studies, 5*(1), 51-77.

[8] Braun, B., Dehé, N., Neitsch, J., Wochner, D., Zahner, K. accepted. The prosody of rhetorical and information-seeking questions in German. *Language and Speech*.

[9] Niebuhr, O., Bergherr, J., Huth, S., Lill, C., Neuschulz, J. 2010. Intonationsfragen hinterfragt – Die Vielschichtigkeit der prosodischen Unterschiede zwischen Aussage- und Fragesätzen mit deklarativer Syntax. *Zeitschrift für Dialektologie und Linguistik, 77*(3), 304-346.

[10] Freed, A. F. 1994. The form and function of questions in informal dyadic conversation. *Journal of Pragmatics, 21*(6), 621-644.

[11] Hudson, R. A. 1975. The meaning of questions. *Language*, 1-31.

[12] Oraby, S., Harrison, V., Misra, A., Riloff, E., Walker, M. 2017. Are you serious?: Rhetorical Questions and Sarcasm in Social Media Dialog.

[13] Cacioppo, J. T., Petty, R. E. 1982. Language variables, attitudes, and persuasion. *Attitudes towards language variation*, 189-207.

[14] Cantor, J. R. (1979). Grammatical variations in persuasion: Effectiveness of four forms of request in door - to - door solicitations for funds. *Communications Monographs, 46*(4), 296-305.

[15] Gibbs, R. W. 2000. Irony in talk among friends. *Metaphor and Symbol, 15*(1-2), 5-27.

[16] Ilie, C. 2015. Questions and questioning. *The International Encyclopedia of Language and Social Interaction*. 1-15.

[17] Petty, R. E., Cacioppo, J. T., Heesacker, M. 1981. Effects of rhetorical questions on persuasion: A cognitive response analysis. *Journal of Personality and Social Psychology, 40*(3), 432-440.

[18] Swasy, J. L., Munch, J. M. 1985. Examining the target of receiver elaborations: Rhetorical question effects on source processing and persuasion. *Journal of consumer research, 11*(4), 877-886.

[19] Schaffer, D. 2005. Can rhetorical questions function as retorts?: Is the Pope Catholic? *Journal of Pragmatics, 37*(4), 433-460.

[20] Thurmair, M. 1991. *Zum Gebrauch der Modalpartikel 'denn' in Fragesätzen. Eine korpusbasierte Untersuchung.* In: E. Klein (ed.), *Betriebslinguistik und Linguistikbetrieb: Akten des 24. Linguistischen Kolloquiums* vol. 1. Linguistische Arbeiten 260. Tübingen: Niemeyer, 377-387.

[21] Dombrowski, E., Niebuhr, O. 2005. Acoustic patterns and communicative functions of phrase-final F0 rises in German: Activating and restricting contours. *Phonetica*, *62*(2-4), 176-195.

[22] Petrone, C., Niebuhr, O. 2014. On the intonation of German intonation questions: The role of the prenuclear region. *Language and Speech*, *57*(1), 108-146.

[23] Niebuhr, O. 2012. Das ist (k)eine Frage: Phonetische Merkmale in der Identifikation standarddeutscher Deklarativfragen. In: I L. Anderwald (eds.), *Sprachmythen: Fiktion oder Wirklichkeit*. Frankfurt: Peter Lang, 203-222.

[24] Murray, I. R., Arnott, J. L. 1993. Towards the Simulation of Emotion in Synthetic Speech: A Review of the Literature of Human Vocal Emotion. *Journal of Acourstic Society of America 93* (2), 1097-1198.

Temporal characteristics of silent pauses and breathing: the effects of speakers' age

Gyarmathy Dorottya[1*], Krepsz Valéria[1**]

[1] *Research Institute for Linguistics, Budapest, Hungary*
**gyarmathy.dorottya@nytud.hu*
***krepsz.valeria@nytud.hu*

The present study investigates the relationship between the pattern of the audible respiration and the occurrence and the duration of silent pauses in two age groups according to their position and function in two speech genres (narrative and conversation). We analyzed how silent pauses and respiration patterns are influenced by their position in spontaneous speech and the age of the speakers based on 20 Hungarian conversations and narratives from young and older adult speakers. Results showed that a. the strategies of pausing are determined by their functions and speech genre. We empirically and statistically confirmed that the physiological necessity of breathing is subordinated to the speech planning processes.

INTRODUCTION

Speech is occasionally interrupted by silent pauses of various length. Pauses serve various functions in speech, like breathing, grammatical function, marking syntactic boundaries, providing time for speech planning processes, for self-repair and for perception as well [1]. The realization of pauses depends on various factors, e.g. the speaker's age, the length and the complexity of the utterance or the speech style [2, 3]. Researches revealed connection between the speech situation and the pauses. The more complex a speech task was – the greater cognitive effort it required – the longer and more frequent the pauses became.

Pause in a conversation has also various functions: it plays an important role in turn-taking system, can be connected with pragmatic or social meanings or with cognitive reasons. Furthermore, conversations can have pauses for thinking or for dramatic effect, the speaker can use them to highlight new information, and they can also be used to structure the discourse [4].

It is known that silent pauses caused by speech planning difficulties differ in their occurrence and duration from pauses occurring in syntactic boundaries.

Breathing is one of the vital functions. There are still many controversies about the relationship between silent pause and breath-taking. Not all silent pauses serve as breathing points: the length of the preceding and following speech units, speech genre, age of the speaker, current physical and mental condition, etc. you can also determine whether the speaker take to breathe there or not.

Although several studies have previously been conducted in different languages (such as English and German) on the relationship between silent pauses and breathing, it may be of interest because of the different prosodic features of the Hungarian language: i. Hungarian is an agglutinating language, that is, a different type of language compared to English and German. ii. Hungarian is a left-headed head/edge-prominence language, while German and English are right-headed head-prominence languages, as well as Hungarian prosody is typologically different from the prosody of Germanic languages. The information structure in Hungarian is primarily expressed by word order, i.e. logical functions are linked to certain sentence positions. The position of focus is defined syntactically (it is immediately pre-verbal), while prosodic prominence marking plays only a secondary role and is partly optional. Consequently, it can be assumed that due to the different language planning and implementation, both the pause pattern and their realization can differ between the Hungarian and the aforementioned German and English that raises the question of these features in the case of the Hungarian language as well.

A recent study confirmed differences in young adults' speech among various types of silent pauses in terms of occurrences and duration, and the relation of pausing strategies and audible breathing. It is a well-known fact that there are many changes with age which may affect, among other things, the temporal characteristics of utterances. Therefore, the aim of this study is to analyze the relationship between the pattern of the audible silent pauses and the occurrence and duration of silent pauses in two age groups according to their function and position in narratives and conversations. The research question is the following: how pauses and breathing patterns are influenced by their position in spontaneous speech and the age of the speakers.

Our hypotheses were that: (i) Silent pauses realize with different patterns according to age groups. (ii) The duration of silent pauses is determined by their position. (iii) There would be proportional and/or durational differences among pause categories depending on breath-taking, (iv) Breathing pattern of silent pauses is different depending on speech genre.

METHOD

20 conversations and narratives from two age groups (20-35-year-olds ($n_{narratives}$=10, $n_{conversations}$=10) and 40-55-year-olds ($n_{narratives}$=10, $n_{conversations}$=10)) were selected from the Hungarian Spontaneous Speech Database, BEA [6]. Three speakers participated in each conversation; the interviewer and one speaker were colleagues; the third participant was the subject. Silent pauses were categorized based on the system developed by [7]. The first distinction was whether the pause was related to disfluency (in these cases, the time span between the interruption of articulation and the beginning of correction was taken into account, as part of the editing phase – E), or it had a syntactical function (S). Pauses with editing function (E) were further categorized based on whether the disfluency phenomena were due to the speaker's uncertainty or errors. Silent pauses with a syntactical function (S) were distinguished based on their position. Utterance onset pauses (S_Uo) occur when a speaker claims the turn; here the pause may only be preceded by a filler word or a discourse marker. Silent pauses at phrase boundaries (S_PhrB) are found between semantic and syntactic units, often before or after a conjunction. Within phrase pauses (S_PhrW) are found within a grammatical unit ('phrase'). End of phrase pauses (S_PhrE) are silent pauses at the end of a syntactic unit, after which the speaker starts another semantic and syntactic unit that often represents a new thought unit. The frequency and the duration of silent pauses was also analyzed with regard to these categories.

The annotation was carried out in Praat [8], statistical analysis in R. Linear mixed-effects models were run in the lme4 [18] in R [17]. The dependent variable was duration, the fixed factors were the age, the breathing, the main and minor categories of the pauses, the speech style and their interaction. The speaker was set as random factor. Both a random intercept model and a random slope model were created for each dependent variables. These were compared by the means of ANOVA in lmerTest package [9]. The AIC (Akaike information criterion) numbers [10] were compared, and also ANOVA was run in [9]. The model with lower AIC number was kept for data explanation. As the linear mixed-effects models do not compute p-values [11], p-values were computed with Satterthwaite approximation in lmerTest package by ANOVA [9].

RESULTS

The occurrence of the silent pauses depended on the speech genre; in narratives there were more silent pauses than in the dialogues. The syntactical silent pauses were more frequent than the editing phases in each speech genres and age groups. The occurrence of the types of syntactical silent pauses differed from each other: the most frequent type is the phrase boundary pauses independent from the speakers' age and speech genre. The duration of the syntactical silent pauses was greater than the editing phase in each speech genres and age groups. The duration of syntactical silent pauses depends on the type or position of syntactical silent pause: the longest types are the utterance onset pauses and the end of phrase pauses, and the shortest are the within phrase pauses.

In case of the editing phases the pauses between the two parts of uncertainty phenomena (e.g. repeated words) took more time than the repairing process of the errors.

The frequency of audible breathing depended on the type of pauses; speakers make audible breathing more frequent in syntactical silent pauses than in editing phases. The pauses with audible breathing are significantly longer than the other types of silent pauses.

Audible breathing appeared most frequently at the end of phrase and these pauses were the longest.

There was more audible breathing in the narratives, than in conversations. Older speakers showed higher occurrence of audible breaths than young speakers.

DISCUSSION

Results showed that the strategies of pausing are mostly determined by their functions and speech genre. There was a difference in the acoustic realization of the silent pauses according to their position. It is probably due to the different function of these silent pauses. The differences between the speakers were remarkable which refers to individual pause strategies in addition to general patterns. Pauses occurred in a grammatically justified position in a greater ratio without breaking the unity of the utterance. Like silent pauses, respiratory patterns differed significantly between the two groups. It can be explained by the physiological differences between the younger and older speakers. The speech genre differed even within each speaker: presumably in the conversations, the occurrences of the turn-taking points strongly influenced the respiratory pattern as opposed to the narratives.

We empirically and statistically corroborated that the physiological necessity of breathing is subordinated to the speech planning processes.

ACKNOWLEDGMENTS

The research was supported by the NKFI-K-128810 and the Bolyai János Research Scholarship.

REFERENCES

[1] Levelt, W. J. M. (1989). Speaking: From intention to articulation. A Bradford Book. Cambridge (Massachusetts)–London, The MIT Press.

[2] Duez, D. (1982). Silent and non-silent pauses in three speech styles. Language and Speech 25(1):11–25.

[3] Krivokapic, J. (2007). Prosodic planning: Effects of phrasal length and complexity on pause duration. Journal of Phonetics 35(2): 162–179.

[4] Esposito, A., Stejskal, V., Smékal, Z. & Bourbakis, N. (2007). The significance of empty speech pauses: Cognitive and algorithmic issues. Advances in Brain, Vision, and Artificial Intelligence. Berlin Heidelberg, Springer, 542–554.

[5] Mády, K., Kleber, F., Uwe D. R. & Szalontai, Á. (2016). The interplay of prominence and boundary strength: a comparative study. http://real.mtak.hu/45641/1/mady_etal_basket_pundp2016.pdf (accessed: 11/17/2017)

[6] Neuberger, T., Gyarmathy D., Gráczi, T. E., Horváth, V., Gósy, M. & Beke, A. 2014. Development of a large spontaneous speech database of agglutinative Hungarian language. In: Sojka, P., Horák, A., Kopeček I. & Pala, K. (eds.), Proceedings of TSD 2014. Berlin: Heidelberg; New York: Springer. 424–431.

[7] Gyarmathy, D. (2018). The functions of silent pauses in spontaneous Hungarian speech. Phonetician, 115. 53-71.

[8] Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program] www.praat.org (accessed 08/14/2018)

[9] Kuznetsova, A. Brockhoff, P. B. and Christensen, R. H. B. (2017) LmerTest Package: Tests in Linear Mixed Effects Models, Journal of Statistical Software, 82. 1-26.

[10] Akaike,H.| (1974.) A new look at the statistical model identification, IEEE Transactions on Automatic Control, 19(6): 716–723.

[11] Lawrence, M. A. (2013) Ez: Easy Analysis and Visualization of Factorial Experiments. R Package Version 4.

In Support of the Laryngeal Articulator Model? A Case Study of Vowel Height and Glottalisation in Czech

Míša Hejná[1]

[1]*Aarhus University, Aarhus, Denmark*
*misa.hejna@cc.au.dk

[1] have shown that the presence of glottalisation is correlated with vowel height, as attested by a range of phonological processes. These findings support and extend the predictions of the Laryngeal Articulator Model [2-3], which acknowledges interactions between the laryngeal and the supralaryngeal components of the vocal tract. This study analyses vowel production data from Czech, a language with obligatory glottal onsets [4], in order to bring evidence from a Slavic language into the discussion. Indeed, we find that /a/ and /aː/ (low vowels) are more likely to be associated with vowel-initial glottalisation than the high vowels, and they are also associated with the longest duration of glottalisation in contrast to all the other Czech vowel phonemes.

INTRODUCTION

The Source Filter Theory of speech production assumes a separation between the source and the filter [e.g. 6]. A simplistic reformulation would be that sounds of speech can be generated in the larynx and subsequently modified above the larynx, but crucially what takes place in the larynx is not coupled with what takes place above the larynx. Yet, many scholars have demonstrated important connections between the larynx and the rest of the vocal tract [e.g. 4, 7, p. 179]. More specifically, [1] show connections between laryngeal constriction and vowel height: low vowels are more likely to go hand in hand with glottalisation. [2] argues that the cardinal vowel /ɑ/ is not just a low back vowel in the sense that it is primarily the laryngeal work that results in its quality, rather than the tongue gesture alone.

In this paper, I examine the links between vowel height and phonatory settings to test the predictions of the Laryngeal Articulator Model with data from Czech, a Slavic language. Czech has the following short and long vowels: /a/, /ɛ/, /ɪ/, /o/, /u/, and /aː/, /ɛː/, /iː/, /oː/, /uː/ [8]. Czech is also a language known to exhibit obligatory glottal attack onsets in vowel-initial words [4], such as *abrakadabraka* "abracadabra" and *a* "and". The following questions are addressed here:

1. Do low vowels show a stronger tendency for the presence of glottalisation? Do the different vowel height categories (1. /a/, /aː/; vs. 2. /ɛ/, /ɛː/, /o/, /oː/; vs. 3. /ɪ/, /iː/, /u/, /uː/) show a gradually lower frequency of glottalisation occurrence with increasing vowel height category?
2. Is the duration of glottalisation shorter as we move from low to high in these three broader vowel height categories?

METHOD

I use the data originally collected for work presented in [9]. Here, the main aspects of the dataset are summarised, but the reader is referred to [9] for more details. The data comes from a single female subject. She was recorded on a daily basis, and the analyses presented here are based on the production of the following:

- phonologically short vowels of Czech twice during each session ([a], [ɛ], [ɪ], [o ~ ɔ ~ ɒ], [u])
- phonologically long vowels of Czech sustained for 5s and for maximum phonation ([aː], [ɛː], [iː], [o ~ ɔ ~ ɒː], [uː])

This procedure resulted in 2,606 short vowels, 1,283 maximally sustained long vowels, and 1,293 vowels sustained for 5 ms.

Laryngeal constriction was defined either as irregular vocal fold vibration (i.e. irregularly timed glottal pulses), or as an interval of a sudden drop in f0. Voiceless glottal friction could also occur, and it is identified by the lack of periodicity in the waveform and the presence of glottal friction in the spectrogram. These two phenomena are illustrated in Fig. 1.

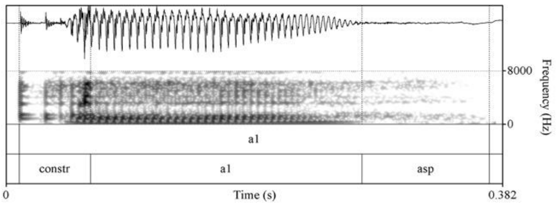


Fig. 1. Identification of laryngeally constricted onsets 'constr' (or occasionally also offsets), and offsets realised as voiceless glottal friction 'asp' (or occasionally also onsets).

RESULTS

Firstly, as shown in Fig. 2, high vowels show a mild dispreference for vowel-initial glottalisation. The statistical analysis confirms this: /a/ is associated with more instances of vowel-initial glottalisation than /iː/, /u/, and /uː/ (Mixed Effects Models: dependent variable = glottalisation frequency of occurrence; independent variable = vowel phoneme with /a/ as the reference level; $p < 0.01$-0.0001).

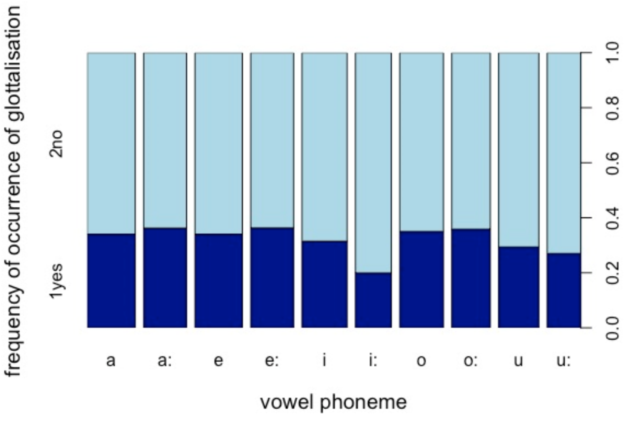


Fig. 2. Frequency of occurrence of vowel-initial glottalisation by vowel phoneme.

Interestingly, the difference between /a/ and /ɪ/ is not significant, which may be due to the fact that /u, uː, iː/ are *phonetically* higher than /ɪ/. In other words, this suggests that we are indeed dealing with a phonetic and not a phonological effect. Post-hoc analyses (conducted with the *emmeans* package in RStudio) further point to the highest vowels being the ones dispreferring glottalisation: /a/, /aː/, and /ɛ/, /ɛː/, /o/, /oː/ show significant differences in being associated with higher rates of glottalisation than the three phonetically highest vowels, /u, uː, iː/ (p = 0.5-0.0001).

As shown in Fig. 3, the highest vowels are associated with the shortest durations of glottalisation as well. However, visually there is practically no difference between low vowels and mid vowels. The statistical analysis nevertheless suggests that all comparisons of vowel phonemes with /a/ are significant (Mixed Effects Models: dependent variable = glottalisation duration (ms); independent variable = vowel phoneme with /a/ as the reference level; p < 0.01-0.0001). On the whole, post-hoc tests point to /iː, u, uː/ consistently contrasting with the other vowels in having the shortest glottalisation durations, but the same does not consistently hold for the other vowel pair comparisons.

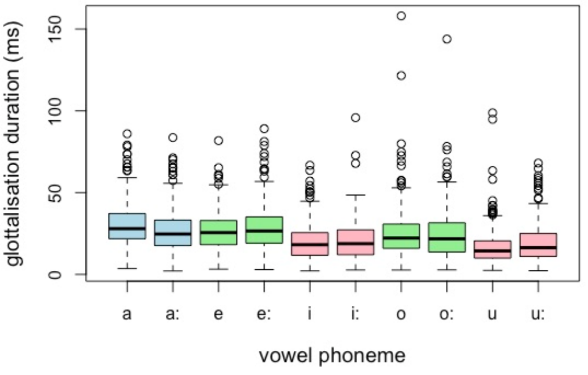


Fig. 3. Duration of vowel-initial glottalisation (ms) by vowel phoneme.

Although there is a possibility that this pattern is due to inherent *vowel duration* differences conditioned by vowel height rather than vowel height itself (or, rather, the way in which the larynx is involved in the vowel production), the long vowels were produced once for 5-6 seconds irrespective of height and once for as long as possible. If the results were solely due to correlations with overall vowel duration due to vowel height, the long phonemes should pattern differently from the short phonemes. Because they do not, I conclude that the results reflect a relationship between vowel height and glottalisation, rather than (solely) vowel duration and glottalisation.

CONCLUSIONS

The results presented here reveal that vowel height does indeed condition the frequency of vowel-initial glottalisation in the Czech data analysed here. /a/ is associated with more instances of vowel-initial glottalisation than /u, uː, iː/, three of the four phonologically high vowels, and the three phonetically highest vowels. Secondly, /a/ is also associated with longer durations of vowel-initial glottalisation than all the other vowel phonemes. Considering /a/ and /aː/ are the most likely vowels to be associated with laryngeal involvement resulting in glottalisation, the findings presented here are in line with the previous reports of low vowels patterning with more glottalisation, and more generally with the Laryngeal Articulator Model. Further research should investigate whether there is a positive correlation between F1, the acoustic correlate of vowel height, and the duration of glottalisation. In addition, glottalisation duration should be normalised as a percentage of the overall vowel duration.

ACKNOWLEDGMENTS

I would like to thank Scott Moisik, Marzena Żygis, and John Esling for their kind encouragement to finally poke into this sort of analysis of my Czech vowel materials.

REFERENCES

[1] Brunner, J., & Żygis, M. (2011). Why do glottal stops and low vowels like each other? Proceedings of the 17th International Congress of Phonetic Sciences, Hong Kong, 376-379.

[2] Esling, J. H. (2005). There are no back vowels: The laryngeal articulator model. *Canadian Journal of Linguistics*, 50, 13-44.

[3] Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality. The Laryngeal Articulator Model*. CUP, Cambridge.

[4] Skarnitzl, R. (2004). Acoustic categories of nonmodal phonation in the context of the Czech conjunction ”a”. *Phonetica Pragensia*, X, 57-68.

[5] Scheer, T. 2009. Representational and procedural sandhi killers: Diagnostics, distribution, behaviour. In Dočekal, M., & Ziková, M. (eds.), *Czech in Formal Grammar*, 155-174. München: Lincom.Google Scholar

[6] Fant, G. (1960). *Acoustic Theory of Speech Production. With Calculations based on X-Ray Studies of Russian Articulation*. Hague: Mouton.

[7] Thurgood, G. (1999). *From Ancient Cham to Modern dialects. Two thousand years of language contact and change*. Oceanic Linguistics Special Publication, 28. University of Hawaʻi Press, USA.

[8] Dankovičová, J. (1999). Czech. In *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*, 70-74. Cambridge: Cambridge University Press.

[9] Míša Hejná. A case study of menstrual cycle effects: global phonation or also local phonatory phenomena? *International Congress of Phonetic Sciences, Melbourne*, 2630-2634.

Silent pause duration and distribution in Older-Women’s Speech: A case study with 4 Within-Speaker Comparison

Julie Kairet

Potsdam Universität, Potsdam, Germany
*kairet@uni-potsdam.de

This paper presents a longitudinal case study aiming at within-speaker comparison of duration and distribution of silent pauses produced by four older women (ranging from 71 to 78 years old). The data analysed are ca. 40 minutes retrieved from two sets of narrative interviews recorded in 2005 and 2015. The results show no significative lengthening of the mean pause duration. However, the modality of the distribution of pauses bring interesting new hypotheses into light: while three speakers use only one category of pauses, the fourth speaker uses two categories of pauses. Furthermore, 49's distribution of pauses changes between 2005 and 2015. This case study provides promising assumptions that will need to be tested through larger scale study.

INTRODUCTION

The study presented in this abstract is part of a PhD project focusing on temporal features (speech rate, articulation rate and pauses) of the speech of older people. While the thesis focuses on the interaction between silent pauses, filled pauses, syllable duration and vocal lengthening, the main goal, here, is to investigate within-speaker differences in silent pauses duration and distribution.

If the age-effect on the speech rate is well documented in the literature [1,2,3], the longitudinal studies focusing on healthy older speakers of French are rare, if not inexistent. The slow-down described in the literature is explained by several factors including longer syllable duration and more frequent and longer pauses [4,5,6]. The present study focuses on silent pauses which are defined by Duez [7] as: "any interval of the oscillographic trace where the amplitude is indistinguishable from that of the background noise." On the first hand, the study aims to test the following hypothesis: After 10 years, are the studied speaker using longer silent pauses?

On the other hand, numerous studies analysed the pause distribution as a combination of two or three normal curves. [8,9,10]. In other words, the distribution of pauses may be unimodal, bimodal or trimodal. Previous studies showed that spontaneous speech was characterized by a trimodal distribution [11] and, on a speech style comparable with our data, Goldman et al. observed a bimodal distribution of pauses in narrative conversation [12]. In line with those findings, our second hypothesis is that we expect to find a bimodal or a trimodal distribution in our data.

METHOD

Our corpus consists of ca. 40 minutes of speech retrieved from the LangAge corpus [13] which comprised narrative interviews with older speakers of French living in the city of Orléans (France). The 4 participants included in the subsample are healthy older women aged, in 2005, between 71 and 78 years old (mean age: 74). Their education level and their professions are comparable. The sample consists in the first five minutes of each monologue retrieved from two interviews for each speaker. The first one was recorded in 2005 and the second one in 2015.

A total of 767 silent pauses (>120 ms [14]) have been manually annotated with Praat [15]. The articulatory pauses (e.g. pre-occlusive short pauses) have been discarded during the annotation phase. The extracted duration of pauses was log-transformed as it is advised in the literature [9,10,16]. Mclust [17] was used for clustering and density estimation.

Tab. 1: Description of the corpora

Speaker	Age in 2005	Silent pauses (2005 sample)	Silent pauses (2015 sample)
15	78	97	91
16	72	98	93
46	71	97	104
49	74	94	93
Total		386	381

RESULTS

The median and mean duration of silent pauses are reported in the Table 2. The within-speaker comparison of silent pauses duration outlines two opposite tendencies. Two speakers (15 and 49) produce, on average longer pauses in 2015: for example, 15’s mean duration of silent pause increases from 562 ms, in 2005, to 708 ms, in 2015. On the contrary, the two other speakers (16 and 46) produce, on average, shorter pauses in 2015: 16’s mean duration of silent pause drops from 617 ms to 550 ms. However, the differences between the means are not statistically significant.

Consequently, our first hypothesis, which states that the speakers will produce longer pauses in 2015, cannot be validated. More information concerning the distribution of the silent pauses are included in the figure 2 (p.3).

Tab. 2: Medians and Means of silent pauses duration for each speaker in 2005 and 2015

	Year	Median ($\log_{10}$ ms)	Median (ms)	Mean ($\log_{10}$ ms)	Mean (ms)
15	2005	2.79	617	2.75	562
	2015	2.82	661	2.85	708
16	2005	2.83	676	2.79	617
	2015	2.79	617	2.74	550
46	2005	2.70	501	2.66	457
	2015	2.63	427	2.61	407
49	2005	2.76	575	2.68	479
	2015	2.72	525	2.69	490

Modality of the distribution of silent pauses

When analysing the density plot produced by Mclust (see figure 3, p.3), the distribution of silent pauses of three speakers (15, 16 and 46) is unimodal. On the contrary, the silent pause distribution of 49 is bimodal. In other words, her speech contains the co-occurrence of two categories of pauses. A clustering of the pauses produced by 49 for each sample has been operated. When comparing the density of distribution in 2005 and in 2015, a change is observable (see figure1).

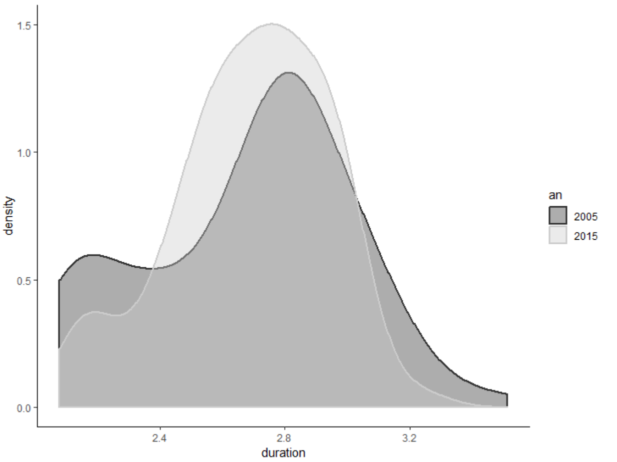


Fig. 1. Density plot of speaker 49's silent pauses ($\log_{10}$ ms)

The first category (C1) groups fewer pauses: 19 out of 187. When compared when becoming older, the number of pauses in the first cluster drops from 12 in 2005 to 7 in 2015. Most of the pauses are part of the second category (C2): 168 pauses out of 187 have been categorized in the second cluster. However, the number of pauses belonging to the second category slightly increases: 82 (out of 94) pauses in 2005 and 86 (out of 93) in 2015.

Concerning the mean duration of pauses in each category, the C1 silent pauses lengthen: the mean duration increases from 129 ms in 2005 to 145 ms in 2015. The C2 silent pauses are, in average, shorter: 575 ms in 2005 and 537 ms in 2015. In other words, 49, in contrast to the three other speakers, uses two categories of pauses. The first category she uses counts less and longer pauses in 2015 and the second category counts more and shorter pauses in 2015.

DISCUSSION

Our first hypothesis assuming that the silent pauses would be longer in 2015 for each speaker is not met. While it is descriptively true for two speakers out of four, the difference in means is not statistically significant. This observation could be explained by several factors: individual variation, that seems to predominate research linked to the process of ageing [6,18], or, since numerous studies show significant duration differences between younger and older speakers in different languages [1,2,19], by the selection of speakers who might be already too old to show any within-speaker variation. Further studies should also consider the interaction between silent pauses and other features usually labelled as disfluency in order to investigate potential temporal compensation.

Our second hypothesis is also denied: three speakers use one category of pauses, while 49 do use two categories of pauses. Although literature lacks description on the modality of the distribution of pauses produced by older speakers, this last observation remains intriguing because it contradicts the results of the existing study using the same methodology.

Another puzzling finding is the observable change in the density of the two clusters of pauses used by 49: in 2015, there are fewer and longer pauses in the first cluster and, more and shorter pauses in the second cluster. It might indicate a centralization of the silent pause distribution.

To conclude, our results, in our opinion, highlight the advantage of study case: such small samples can put into light promising hypotheses that need to be tested on large-scale study. Undoubtedly, it is not possible to confirm the importance of our findings based on a case study: more data of a representative sample of the whole population is needed to confirm our theory. A complete lifespan perspective, including more than one repeated measure across ten years and considering younger speakers would certainly reveal interesting observations with regards to the modality of pause distribution in older speaker discourse.

REFERENCES

[1] Ramig, L. A., & Ringel, R. L. (1983). Effects pf Physiological Aging on Selected Acoustic Characteristics of Voice. *Journal of Speech and Hearing Research, 26*, 22–30.

[2] Bilodeau-Mercure, M., & Tremblay, P. (2016). Age Differences in Sequential Speech Production: Articulatory and Physiological Factors. *Journal of the American Geriatrics Society*.

[3] Bona, J. (2014). Temporal characteristics of speech: The effect of age and speech style. *The Journal of the Acoustical Society of America*, *136*(2).

[4] Gerstenberg, A., Fuchs, S., Kairet, J., Frankenberg, C., & Schröder, J. (2018). A cross-linguistic, longitudinal case study of pauses and interpausal units in spontaneous speech corpora of older speakers of German and French. *Proceedings of the 9th International Conference on Speech*, 211–215. Poznan, Poland.

[5] Linville, S. E. (2001). *Vocal Aging*. San Diego: Singular Thomson Learning.

[6] Fougeron, C., Delvaux, V., Ménard, L., & Maganaro, M. (2018, May 7). *The MonPaGe_HA Database for the Documentation of Spoken French Throughout Adulthood*. 4301–4306. Miyazaki, Japon.

[7] Duez, D. (1982). Silent and non-silent pauses in three speech styles. *Language and Speech*, (25), 11–28.

[8] De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390.

[9] Grosman, I., Simon, A. C., & Degand, L. (2018). Variation de la durée des pauses silencieuses: Impact de la syntaxe, du style de parole et des disfluences. *Langages*, *2018/3*(111), 13–40.

[10] Auchlin, A., Simon, A. C., Goldman, J.-P., & Avanzi, M. (2018). Pauses avec et sans prise de souffle. Typologie acoustique et fonctionnelle. In E. Richard (Ed.), *Des organisations dynamiques de l'oral* (pp. 57–72). Bernes: Peter Lang

[11] Campione, E., & Véronis, J. (2002, April 11). *A Large-scale Multilingual Study of Silent Pause Duration*. Presented at the Speech Prosody 2002, Aix-en-Provence, France.

[12] Goldman, J.-P., François, T., Roekhaut, S., & Simon, A. C. (2010). Etude statistique de la durée pausale dans différents styles de parole. *Actes Des 23èmes Journées d'étude Sur La Parole*. Presented at the Mons, Belgique, 25-28 mai 2010. Mons, Belgique, 25-28 mai 2010.

[13] Gerstenberg, A. (2011). *Generation und Sprachprofile im höheren Lebensalter. Untersuchungen zum Französischen auf der Basis eines Korpus biographischer Interviews*. Frankfurt: Klostermann.

[14] Heldner, M. (2011). Detection thresholds for gaps and overlaps and no-gap-no-overlaps. *The Journal of the Acoustical Society of America*, 130(1), 508–513.

[15] Boersma, P., & Weenink, D. (2015). *Praat* [Windows]. Retrieved from http://www.fon.hum.uva.nl/praat/

[16] Oehmen, R., Kirsner, K., & Fay, N. (2010). Reliability of the Manual Segmentation of Pauses in Natural Speech. *NLP 2010: Advances in Natural Language Processing*, 263–268.

[17] Fraley, C., E. Raftery, A., Murphy, T. B., & Scrucca, L. (2012). *Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation* (Technical Report No. 597). USA: University of Washington.

[18] Ringel, R., & Wojtek J., C.-Z. (1987). Vocal Indices of Biological Age. *Journal of Voice*, *1*(1), 31–37.

[19] Dimitrova, S., Andreeva, B., Gabriel, C., & Grünke, J. (2018). Speaker Age Effects on Prosodic Patterns in Bulgarian. *Proceedings of the 9th International Conference on Speech*, 709–713. Poznan, Poland.

ADDITIONAL FIGURES

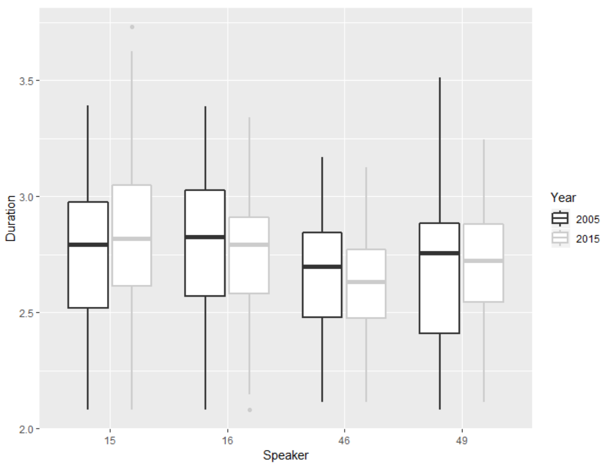


Fig. 2. Boxplot: pause duration (log$_{10}$ *ms) for each speaker.*

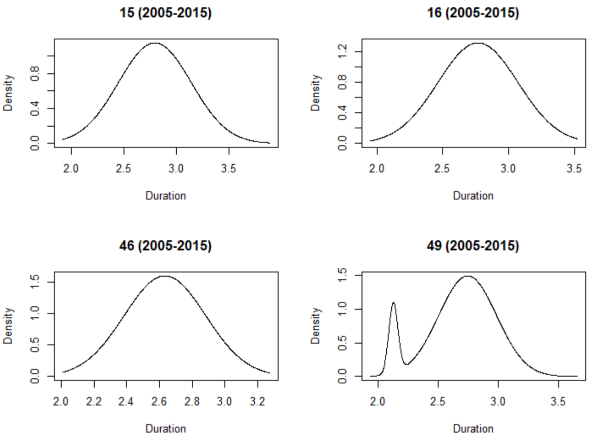


Fig. 3. Density plot of silent pauses (log$_{10}$ *ms) for each speaker.*

Final Lengthening in Danish

Maria Alm[1*]

[1]*Dept. of Design and Communication, University of Southern Denmark*
*mhalm@sdu.dk

Final lengthening means that speakers slow down their speech rate in the last syllables of an intonation phrase in order to signal its upcoming end. However, according to Grønnum [1], final lengthening is relatively modest in Danish. She suggests that this might cause problems for speakers of languages with final lengthening to recognize when a Danish utterance is finished. In an ongoing investigation of intonation and prosody in Danish, the use of final lengthening in Danish talk-in-interaction is also investigated. The sample is small, and it is for several reasons a difficult to calculate the speech rate, but it is suggested that there is a tendency for the conversation partners to use of speech rate differences at least in turn-final position.

INTRODUCTION

In many languages, final lengthening is used at the end of intonation phrases for signaling an upcoming major boundary. However, in Danish, final lengthening is moderate. Grønnum [1] suggests that this causes problems for speakers of other languages, for example English and Swedish, in recognizing the end of a Danish utterance. Recognizing completion points is an important cue for a smooth turn-taking [2].

In an ongoing investigation of the prosodic design of questions in Danish conversations, speech rate is one of the analyzed parameters. It was then observed that that the beginning of a question is on average spoken faster than at the end of the question, albeit with a great variation.

Since Grønnum's findings are based on speech from laboratory settings [1; 3], the question asked here is if the special affordances of authentic talk-in-interaction, putting the conversations partners under a strong pressure to be clear in signaling their intentions, might influence their conversational behavior.

METHOD

The speech data comes from recordings from the Danish part of the Talkbank.org. The sound quality of this kind of recordings is not always the best for prosodic analysis, but authentic talk-in-interaction is difficult to document in a laboratory.

Two different sets of data are analyzed: First, conversational questions were extracted from three of the recordings, following the definition of conversational questions in Selting [4]. Second, a small case study of indefinite pronoun *noget* ("something") was conducted.

The prosodic analysis was conducted using the freely available computer program PRAAT [5].

At this point, there are four different methods for measuring final lengthening known to me, all with their advantages and disadvantages:

1) The simplest method is to measure the speech rate of the syllables in the beginning vs. in the end the intonation phrase. The syllables should have the same status with respect to accentuation, because one of the prosodic correlates of accented syllables in Danish is a longer duration. Thus, I have compared the unstressed syllables before the first accent with unstressed syllables following the last of intonation phrases.

2) The cleanest method is to measure the very same lexical word in an early/medial vs. late position of the intonation phrase. Again, you have to control for accentuation [1]. Constructed sentences are probably the easiest way to obtain this kind of data for comparison. However, in the Danish data the pronoun *noget* ("something") was used repeatedly both in utterance-final and utterance-medial position, and often in an unaccented use, and so I made a case study of *noget*.

3) Another method consists in measuring the duration variation of the single phones by a speaker and comparing it to the mean phone duration of that same speaker [6]. This makes sense. For example, the phone [s] in my data always has a long duration in comparison to other consonant phones; long enough to make a difference in a comparison based on single syllables. However, this method is laborious, requiring the segmentation and labelling of a fairly large amount of speech data. Since I, like Hansson [7], do not have a speech recognizer for automatic segmentation, I will not use this method.

4) Hansson [7] instead uses a method developed by Dankovičová [8], which phonological consists in measuring and comparing the individual phonological words within the intonation phrases. This method requires you to define what a phonological word is. In Swedish, a phonological word contains an accented word [7]. Therefore, the Danish so-called "stress group" (*trykgruppe*), which is the basic rhythmic and tonal unit [1; 3], seems to suggest itself as a correspondence. Pre-accent syllables and words with hesitation prolongation were discarded from the analysis.

Hansson [7] found that the first prosodic word was always very fast (maybe a “final shortening” phenomenon). The reliable, significant difference was found between the penultima and the ultima phonological word. I have thus calculated the speech rate for these phonological words.

Speech rate is usually measured by dividing the number of spoken syllables with their total duration in seconds. This turned out to be a problem, primarily due to the frequent use of *schwa*-assimilation in Danish [1; 3]: Unstressed endings often contain a *schwa*-vowel (written as <e>), for example *stor-e* (“big” + inflection ending). In spontaneous speech, the *schwa*-ending is often assimilated and fused together with a preceding or following sound. The result is often a prolongation of the fused sound Sometimes the syllabicity remains, sometimes it is lost. I often found it difficult to decide if I was dealing with one or two syllables. However, as I am a native speaker of Southern Swedish, which is close related to Danish, I decided to trust my judgement and count only the syllables that I could hear. This is still problematic, because the loss of syllabicity but addition of prolongation will make the syllable longer than an “ordinary” syllable; but still shorter than the same syllable followed by an intact ending.

For illustration, the utterance in (1) contains four *schwa*-assimilations/reductions in the last phonological word (capitals mark accented syllables). The speech rate is then based on five syllables, while the corresponding standard forms of the words contain nine syllables (2):

(1) *har i wEEkend-Arbejd non gang os*
“do you have weekend work sometimes, too”

(2) har i wEEkend-Arbejd-**e** nogl-**e** gang-**e** ogs-**å**.

I have not yet found a solid solution to the syllable count.

RESULTS

Table (1) shows the speech rate in syllables/second for the question types yes/no question with inversion (V1), yes/no question without inversion (V2) and question-word questions (WH). The questions are extracted from the conversations *Anne og Beate*, *Gamledage* and *Kartofler & Broccoli*:

Tab. 1: Question type and the speech rate in syl/sec

Quest type	Pre-accent	Post-accent	First Phon Word	Penultima Phon Word	Ultima Phon Word
V1 17 it.	10,4 4,7-16,7 17 items	7,2 2,2-11,7 16 items	8,1 3,8-13,0 14 items	6,0 45-8,3 5 items	6,8 3,6-11,9 12 items
V2 11 it.	8,4 5,2-12,0 10 items	7,0 4,3-11,9 6 items	5,7 4,1-7,2 5 items	6,1 1 item	4,4 2,7-6,6 5 items
Quest type	Pre-accent	Post-accent	First Phon Word	Penultima Phon Word	Ultima Phon Word
WH 11 it.	10,9 7,4-17,6 6 items	6,4 3,6-11,1 6 items	6,9 4,0-8,7 6 items	6,8 6,4-7,0 3 items	5,1 3,3-6,7 6 items
ALL 39 it.	9,9 5,8-15,5 33 items	6,9 3,7-11,6 27 items	6,9 4,0-9,6 25 items	6,3 5,5-7,7 9 items	5,4 3,2-8,4 23 items

I then did a case study of the indefinite pronoun *noget* (“something). The examples are extracted from *Kartofler & Broccoli* and *225_deller*, which contain the same three speakers. *Noget* is pronounced like anything from one syllable, one syllable with prolongation and two syllables. Since I now had the advantage of dealing with the same word, I decided to measure the duration of *noget* in milliseconds instead. In addition, I calculated the part *noget* made up of the phonological word of which it is part. Accented occurrences of *noget* were excluded. The results show that it makes sense to divide the occurrences into categories depending on the position of *noget* in the turn (turn-final = a speaker change follows); the expression *så no* “such things”, which divides into intonation phrase-final and non-phrase final occurrences; and finally the phrase-internal occurrences of *noget*.

Tab. 2: Speech rate of noget *in ms, its part of the phonogical word in percentage and categorized by position in the turn*

Position	Length	Part of phonological word,
Turn-final 7 items	254 ms 220-286 ms	45% 31-59%
så no, last word 5 items	147 ms 65-235 ms	28% 15-45%
så no, not last word 5 items	145 106-221 ms	16% 10-22%
Other positions in intonation phrase, 17 items	186 124-300 ms	36% 24-84%

DISCUSSION

Although the variation in speech rate is great within each of the examined categories, the average length of pre-accent and post-accent syllables in the questions shows that the speech rate generally seems to be faster in the beginning than in the end of an intonation phrase. In contrast to Hansson’s study of Southern Swedish [7], the speech rate of the first and penultima phonological word does not seem to differ much in the Danish data. However, the ultima phonological word seems to be slightly slower than the penultima phonological word. The question is if this change in speech rate is large enough to be discernable to listeners.

The case study of *noget* (“something”) indicates that it is really the turn-final position that is important for final lengthening, not the phrase-final position in general. Incidentally, all the questions are turn-final, most often one intonation phrase corresponding to a whole turn. This might explain the speech rate differences in the questions.

More data needs to be analyzed, but up to this point I am suggesting that there is a tendency for final lengthening in Danish talk-in-interaction, at least as a cue for turn-final phrases, making it an available cue for turn-taking.

ACKNOWLEDGMENTS

Many thanks to the Independent Research Fund Denmark.

MATERIAL

The examples are extracted from the Danish part of Talbank (https://samtalebank.talkbank.org), a freely available collection of conversation-analytical material. The questions are extracted from Sam2 *Anne og Beate* (10:08 min), Sam3 Gamledage (13:08 min) and Sam3 *Kartofler & Broccoli* (44:38 min). The occurrences of *noget* ("something") are extracted from Sam3 *Kartofler & Broccoli* and Sam3 *225_deller* (50:01 min).

REFERENCES

[1] Grønnum, N. 2005. *Fonetik og fonologi.* Viborg: Akademisk forlag, p. 191.

[2] Sacks, H., Schegloff, E.A. & Jefferson, G. 1974. A simplest systematics for the organization of turn-taking in conversation. In *Language,* 50, 696-735.

[3] Grønnum, N. 2007. *Rødgrød med fløde: En lille bog om dansk fonetik.* Copenhagen: Akademisk Forlag.

[4] Selting, M. 1995. *Prosodie im Gespräch: Aspekte einer interaktionalen Phonologie der Konversation.* Tübingen: Niemeyer.

[5] Boersma, P. 2001. PRAAT: A system for doing phonetics by computer. In *Glot International,* 5, 341-347.

[6] Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf & P. J. Price. 1992. Segmental Durations in the Vicinity of Prosodic Phrase Boundaries. In *The Journal of the Acoustical Society of America* 91, 1707-1717. Reported in Hansson, p. 52 [7].

[7] Hansson, P. 2003. *Prosodic Phrasing in Spontaneous Swedish.* Dissertation, Lund University, Sweden.

[8] Dankovičová, J. 1997. The domain of articulation rate variation in Czech. In *Journal of Phonetics* 25, 287-312.

Teaching Strategic Use of Hesitation Markers

Nathalie Schümchen[1*]

[1]*Department of Design and Communication, University of Southern Denmark, Sonderborg* nats@sdu.dk

Although handbooks on public speaking often state that hesitation markers should be avoided altogether, research has shown that some hesitation markers in the right positions can in fact aid communication while other hesitation markers should be avoided. This paper presents the results of a usability study on visual learning material about the strategic use of hesitation markers. Initial results indicate that language users are generally unaware of the difference between different phonetic formats of hesitation markers and their associated positive and negative effects. This usability study is thus a first step toward a more nuanced general understanding of the effect of hesitation markers in public speaking.

INTRODUCTION

Many guidebooks on public speaking condemn hesitation markers (such as “uh” and “uhm”) as negative and advise speakers to avoid all types of voiced hesitation markers altogether. However, research has shown that the really disturbing hesitation markers are relatively loud and open (“äh” with a long vowel as in ‘bad’) whereas closed (“uh”) and nasal sounds (“mm”) go largely unnoticed while still fulfilling discourse structuring functions. In this study, I present the process of developing and testing learning material based on the assumption that the conscious and strategic use of hesitation markers can be beneficial in public speaking.

Preliminary results of the conducted usability test indicate that language users are not consciously aware of the different functions of hesitation markers but generally open to the idea of a distinction between disturbing and useful hesitation markers. The qualitative analysis of the usability test data thus delivers promising results for future designs of learning materials on public speaking. The quantitative pre- and post-intervention analysis did not yield conclusive results regarding the effect of the learning material on the study participants’ speaking behavior. A possible reason for the inconclusiveness of the results may be found in the study design which might have included too many stimuli for the participants to pay attention to their speech production.

METHOD

The study was set up as a think-aloud usability test [1-3] in which one study participant at a time looked at a paper-based draft of learning material about hesitation markers (cf. Figure 1). The hesitation marker material was part of a larger usability testing setup which included the testing of various other learning materials (e.g. about grammatical issues).

The usability test sessions were video-taped and analyzed using usability testing analysis methods (cf. [1]) as well as conversation analysis [4] as a qualitative, data-driven approach to the analysis of interaction. Furthermore, presentation-like speech was elicited before and after showing the participants the learning material to determine whether the intervention (i.e. working with the material) had an impact on the participants’ use of hesitation markers. To this end, these presentation-like utterances were transcribed and coded with regard to the different types of hesitation markers used.

RESULTS

Preliminary results of the qualitative analysis suggest that language users do not consider hesitation markers as something that can be used strategically in public speaking contexts, but that they are able to distinguish between open and closed/nasal sounds when presented with this distinction including examples. Furthermore, the think-aloud approach taken in this study uncovers both content- and layout-related usability issues which inform future designs of the material. The quantitative analysis of the participants’ use of hesitation markers before and after working with the learning material is still in progress. A cursory look at the data indicate that there is little difference between the two presentation-like speech production events.

DISCUSSION

The finding that language users do generally not think about hesitation markers as a category that consists of formally and functionally distinct entities opens up the opportunity to spread awareness of the different forms and functions of hesitation markers in the context of public speaking. Learning material that presents different forms of hesitation markers (e.g. open vs. closed sounds) and relates them to different (communicative) functions can help to improve public speaking skills. With regard to the specific study at hand, the preliminary observation that the specific learning material does perhaps not have a significant effect on the participants’ use of hesitation markers can probably be attributed to the test design itself. Apart from the hesitation marker material, the usability study included several other learning materials on other aspects of language. This means that the participants' attention was split between different cognitively challenging aspects of language that court the participants' attention.

Design processes are iterative, and the findings of this study inform the redesign of the learning material draft, which can then be tested again in order to develop useful and user-friendly learning material on the strategic use of hesitation markers..

REFERENCES

[1] Dumas, J. S., & Redish, J. (1999). *A practical guide to usability testing:* Intellect books.

[2] Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87* (3), 215.

[3] Schriver, K. A. (1997). *Dynamics in Document Design* (1. ed.): Wiley.

[4] Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*(4), 696-735.

This is a ”read-aloud” task. Read the text out loud and verbalize your thoughts and comments while doing so.

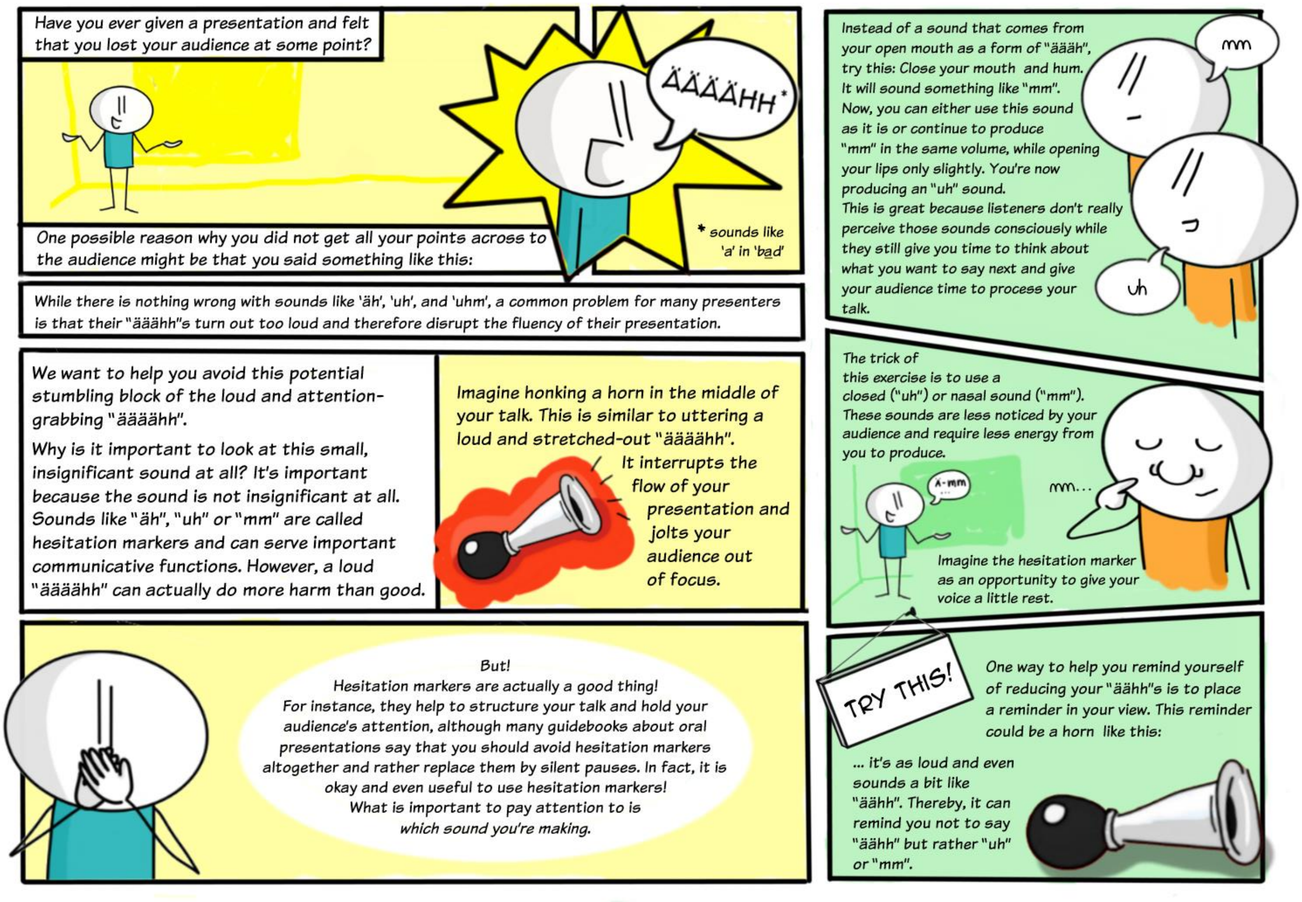

Have you ever given a presentation and felt that you lost your audience at some point?
ÄÄÄÄHH*
One possible reason why you did not get all your points across to the audience might be that you said something like this:
* sounds like 'a' in 'bad'
While there is nothing wrong with sounds like 'äh', 'uh', and 'uhm', a common problem for many presenters is that their "äääähh"s turn out too loud and therefore disrupt the fluency of their presentation.
We want to help you avoid this potential stumbling block of the loud and attention-grabbing "äääähh".
Why is it important to look at this small, insignificant sound at all? It's important because the sound is not insignificant at all. Sounds like "äh", "uh" or "mm" are called hesitation markers and can serve important communicative functions. However, a loud "äääähh" can actually do more harm than good.
Imagine honking a horn in the middle of your talk. This is similar to uttering a loud and stretched-out "äääähh".
It interrupts the flow of your presentation and jolts your audience out of focus.
But!
Hesitation markers are actually a good thing!
For instance, they help to structure your talk and hold your audience's attention, although many guidebooks about oral presentations say that you should avoid hesitation markers altogether and rather replace them by silent pauses. In fact, it is okay and even useful to use hesitation markers!
What is important to pay attention to is which sound you're making.
Instead of a sound that comes from your open mouth as a form of "äääh", try this: Close your mouth and hum. It will sound something like "mm". Now, you can either use this sound as it is or continue to produce "mm" in the same volume, while opening your lips only slightly. You're now producing an "uh" sound.
This is great because listeners don't really perceive those sounds consciously while they still give you time to think about what you want to say next and give your audience time to process your talk.
mm
uh
The trick of this exercise is to use a closed ("uh") or nasal sound ("mm"). These sounds are less noticed by your audience and require less energy from you to produce.
A-mm
mm...
Imagine the hesitation marker as an opportunity to give your voice a little rest.
TRY THIS!
One way to help you remind yourself of reducing your "äähh"s is to place a reminder in your view. This reminder could be a horn like this:
... it's as loud and even sounds a bit like "äähh". Thereby, it can remind you not to say "äähh" but rather "uh" or "mm".

Fig. 1. Paper-based learning material on the strategic use of hesitation markers

Alternative phrase boundary symbolization and its effect on pause duration

Stephanie Berger[1*], Oliver Niebuhr[2], and Margaret Zellers[1]

[1]Institute for Scandinavian Studies, Frisian Studies and General Linguistics, University of Kiel, Germany
[2]Centre for Industrial Electronics, University of Southern Denmark, Sønderborg, Denmark
*sberger@isfas.uni-kiel.de

This study aims to find alternative phrase boundary symbolizations in written text to enable consistent pause production. Shorter phrases and many pauses are beneficial for successful presentations and the communication of information structure. Texts with alternative symbolizations (dashes, bars, spaces, line breaks, hashtags, and arrows) are used to indicate minor and major phrase boundaries. Results suggest that these methods can elicit more consistent pause production than normal punctuation, especially using line breaks.

INTRODUCTION

Research concerning the effect of intended pauses in public speaking contexts has revealed that many pauses with shorter phrases are advantageous, for example for charisma perception [1]. It would therefore be helpful to develop methods that can consistently elicit many pauses of differing lengths and a system to help parse longer sentences into smaller units based on content and emphasis rather than punctuation and syntax.

In 16th century England, today's punctuation marks (, ; : . ?) were already in use, primarily for reading aloud, even though some first indications of syntactic usage were clearly apparent [2]. By the 18th century, punctuation was mostly seen as syntactic, while scholars still attested a role for pausing, suggesting that only about half of spoken pauses are represented in written symbolization (cf. [3], p. 75). At the same time, an alternative to syntactic punctuation called 'rhetorical' or elocutionary punctuation appeared, with "the purpose of regulating the pauses of the voice in reading" ([4] cited in [2], p. 60). However, this kind of punctuation never took hold [2].

Some studies have investigated the influence of sentence formatting (commas, spaces, line breaks) on the information parsing and readability of a text. [6] found that the isolation of individual constituents using large spaces is conducive for readability, while inserting larger spaces in the text at prosodically-informed places made the text less readable "because it evokes auditory processing strategies" ([6], p. 83). Nevertheless, [6] generally showed that "isolating major phrases within extra spaces facilitates reading" (p. 83). According to [5], inserting commas in ambiguous sentences can "enhance the awareness of constituent structure and phrasal boundaries, both physically, in terms of eye movements, and mentally, in terms of processing […] with the added benefit of reducing the need to reread the sentence" (p. 585). Double spaces instead of commas did not affect processing but showed "a similar effect on many eye-movement parameters regardless of processing implications" ([5], p. 586). Line breaks also have "a powerful influence on the reader's behavior" ([7], p. 70) and "cue clausal segmentation" ([5], p. 568).

The present study is part of a research initiative named INSPECT (**I**nnovating **Sp**eech **E**li**C**itation **T**echniques) which focuses on researching the influence of outside factors on the elicitation of specific phenomena (see [8; 9] for text formatting studies; see [10] for a general summary).

The aim of this study is to return to the idea of 'rhetorical punctuation' for public speaking training and to develop methods to elicit systematic minor and major phrase boundaries resulting in shorter and longer pause durations using symbols in a printed text. These methods can be used to help speakers consistently produce a minimum of two different pause durations on-line during reading aloud or presenting, leading to the main question: Can we get speakers to produce a specific, consistent style of speech?

METHOD

A reading text ("Mobile App Developer", [11], an example of a well-made elevator pitch) was slightly altered by the authors. It was set with Times New Roman, 16 pt, double line spacing and block setting.

The third author (English L1 speaker) set minor and major phrase boundaries in the text based on an auditory interpretation of a recording of her voice. There is clear evidence for more than one level of phrase boundary, often referred to as *intermediate phrase* and *intonational phrase* [12,13]. Major phrase boundaries were set at full stops or question marks in the original text (N=13). Minor phrase boundaries were set where short pauses or phrase breaks should occur to structure the text more efficiently—this was the case where the original text had commas, but also when there was no punctuation (N=22). The phrase boundaries were symbolized using six methods in addition to normal punctuation (which served as a baseline condition). Table 1 summarizes and justifies the different conditions used for recordings and analyses.

Per condition, 8 participants were recorded in an empty, quiet lecture room at the University of Southern Denmark, resulting in 48 participants total; all were first-semester students of the Master programme "Innovation & Business" and English L2 speakers. All speakers read the *Baseline* condition first and then one of the six (randomly assigned) test conditions with a one-hour break in between, and the opportunity to familiarize themselves with the text for ten minutes before each reading.

Tab. 1: Summary of the symbols used for the different conditions with justifications.

Condition	Symbols		Justification
	Minor PB	Major PB	
Baseline	,	. ?	normal punctuation
Bar	|	||	subtle, but visible break
Arrow	>	>>	might lead the eye to a new constituent
Hashtag	#	##	might indicate new constituents (inspiration: social media)
Enter	line break	empty line	inspired by poem formatting
Dashes	–	—	more subtle, normal in writing
Space	3 spaces	5 spaces	in theory; block setting resized the spaces; included as an inconsistent symbolization.

RESULTS

Measurements were analyzed using linear mixed-effects models, with pause duration as the dependent variable. Only pauses within a time window of 200-1000 ms were included in the statistical models. The lower value was chosen with respect to the perceptual detection threshold of silent pauses in speech and to avoid inclusion of plosive closures. The upper value was small enough to avoid including any disfluent, hesitation pauses in the results (cf. [14], p. 40). Therefore, this study does not deal with a difference in presence or absence of pausing at phrase boundaries—phrase boundaries marked by different prosodic parameters other than silence are not included here—and only investigates pause duration at phrase boundaries when pauses arise. Figures were created using [15].

The first model included two fixed factors, i.e. *Boundary* (minor vs major, within-subjects variable) and *Group* (1-6, between-subjects variable), and *Speaker* as a random factor. The model was run only for the baseline condition (i.e., normal punctuation) to check whether speakers produce different silent-pause durations at minor and major phrase boundaries and/or in connection with regular punctuation marks and, moreover, whether there are significant differences in the pausing habits between the six groups of participants.

Results show a significant difference of about 175 ms between the silent pauses associated with minor and major phrase boundaries (F[1,42]=422.07, p<.001, η_p^2=0.91). Minor phrase boundaries were on average shorter (514 ms) than major phrase boundaries (689 ms), yet there was considerable overlap between minor and major pause durations, see Fig.1a. Pause durations between groups differed on average by about 30 ms, but this factor was not significant (F[5,1536]=0.69, p=.625, η_p^2=0.01), nor were any interactions. Thus, any significant differences between the compared pause-marking strategies are not due to artifacts of between-group differences in pausing habits but represent genuine effects of the pause-marking strategies.

The second model had two fixed factors, *Boundary* (minor vs major, within-subjects variable) and *Marking* (six pause-marking conditions, between-subjects variable), with *Speaker* as a random factor. The model yielded significant main effects of both *Boundary* and *Marking* as well as a significant interaction between them (F[5,35]=64.01, p<.001, η_p^2=0.90). There was no significant main effect of *Speaker*, but a significant interaction arose between *Speaker* and *Boundary* (F[7,35]=2.34, p=.046, η_p^2=0.32). A subsequent t-test comparing the difference between minor and major pause durations for each speaker shows that the *Dashes*, *Arrows*, *Hashtags* and *Enter* conditions create a clearer differentiation between minor and major boundaries than the baseline.

On this basis, the key results can be summarized as follows. For some marking conditions, the elicited minor and major pause duration differences were larger and more consistent across speakers than others. This applies to the *Arrow* (see Fig.1c) and *Hashtag* markings and, in particular, to the *Enter* markings, see. Fig.1b. In contrast, the *Space* marking performed worst in these respects, see Fig.1d. The *Bar* condition performed similarly poorly and, together with the *Space* condition, led to the greatest variation (i.e. inconsistency) between speakers. *Hashtag* markings, in turn, were worse than *Arrow* and *Enter* markings as they elicited relatively long minor pause durations; the *Dashes* markings were still worse in this respect. Pauses elicited by dashes were generally the longest (743 ms), about 30 % longer than those elicited by arrows (479 ms). The *Enter* markings, in contrast, elicited pausing behavior that was on average closest to that of the regular orthographic punctuation marks (570 ms) but with less internal variability.

DISCUSSION

Overall, the results suggest that there is a hierarchy of effectiveness of different symbolization methods for separating shorter from longer pauses and eliciting natural and unmarked pause durations. The method that was most effective in this first study was using line breaks (*Enter* condition), followed by *Arrows*, *Hashtag*, *Dashes*, *Bars*, and the completely ineffective *Space* condition. Altogether, symbolizations that performed well were typical in typesetting like the *Enter* condition, except for the *Arrows* condition. A disadvantage of the *Enter* condition is the number of pages necessary, which is not convenient for reading tasks on paper. A possible application would be to mix systems with each other and with ‘normal’ orthographic punctuation. For example, dashes can be inserted in unusual places alongside normal punctuation to increase some pause durations for dramatic effect in a speech. Future studies should conduct field surveys with actual presentations and presentational training in order to improve further the consistency and efficiency of this version of a rhetorical punctuation.

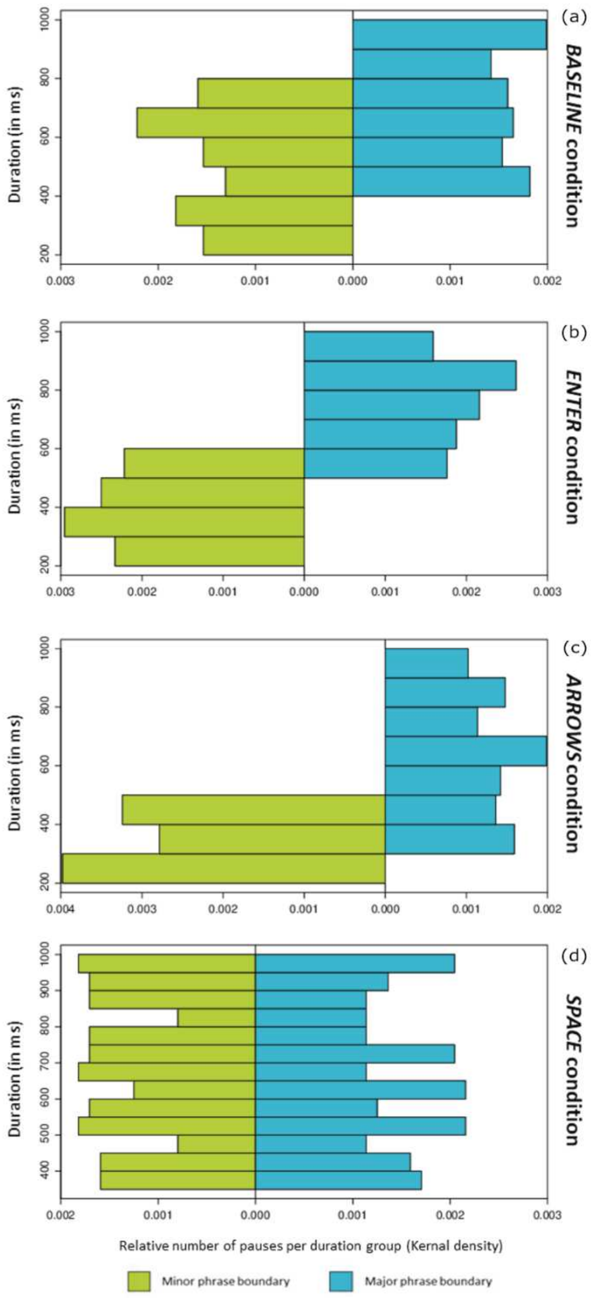


Fig. 1. Kernal Density plots for (a) the Baseline condition, as well as three test conditions: (b) Enter condition, (c) Arrows condition, and (d) Space condition. The relative number of pauses per duration group was used to account for the difference between minor phrase boundaries (N=22) and major phrase boundaries (N=13) per text (x48 in the Baseline condition; x8 in each of the test conditions). Note also that no pauses shorter than 200 ms and longer than 1000 ms are included.

ACKNOWLEDGMENTS

The authors would like to thank Nele Kiupel for her helpful comments. Thank you also to the participants of the study. The work of the first author was supported by the Federal State Grant Schleswig-Holstein of the University of Kiel.

REFERENCES

[1] Niebuhr, O., Brem, A., & Tegtmeier, S. (2017). Advancing research and practice in entrepreneurship through speech analysis – From descriptive rhetorical terms to phonetically informed acoustic charisma profiles. *Journal of Speech Sciences*, 6(1), 3-26.

[2] Cruttenden, A. (1990). Intonation and the comma. *Visible Language*, 25(1), 54-73.

[3] Robertson, J. (1785). *An essay on punctuation*. Charing-Cross, UK: J. Walter.

[4] Bell, A. (1835). *The practical elocutionist*. London: Sherwood, Gilbert and Piper.

[5] Hill, R. L., & Murray, W. S. (2000). Commas and spaces: Effects of punctuation on eye movements and sentence parsing. In A. Kennedy, R. Radach, D. Heller, & J. Pynte (eds.). *Reading as a perceptual process*. Amsterdam: Elsevier, 565-589.

[6] Bever, T. G., Jandreau, S., Burwell, R., Kaplan, R., & Zaenen, A. (1990). Spacing printed text to isolate major phrases improves readability. *Visible Language*, 25(1), 74-87.

[7] Kennedy, A., Murray, W. S., Jennings, F., & Reid, C. (1989). Parsing complements: Comments on the generality of the principle of Minimal Attachment. *Language and Cognitive Processes*, 4(3/4), 51-76.

[8] Berger, S., Marquard, C., & Niebuhr, O. (2016). INSPECTing read speech – How different typefaces affect speech prosody. *Proc. of Speech Prosody 2016, Boston, MA, USA*, 513-517.

[9] Berger, S., Niebuhr, O., & Fischer, K. (2018). Eliciting extra prominence in read-speech tasks: The effects of different text-highlighting methods on acoustic cues to perceived prominence. *Proc. of Speech Prosody 2018, Poznań, Poland*, 75-79.

[10] Niebuhr, O., & Michaud, A. (2015). Speech data acquisition – The underestimated challenge. *Kieler Arbeiten zur Linguistik und Phonetik (KALIPHO)*, 3, 1-42.

[11] Simpson, M. How to write a killer elevator pitch (examples included). *ANZFIRST – The Job Search Website*. URL (accessed 06 November 2019): https://www.anzfirst.com/post/workplace/283bfad62115b

[12] Frazier, L., Carlson, K., & Clifton Jr., C. (2006). Prosodic phrasing is central to language comprehension. *TRENDS in Cognitive Sciences*, 10(6), 244-249.

[13] Pierrehumbert, J., & Hirschberg, J. B. (1990). The meaning of intonational contours in the interpretation of discourse. *Intentions in Communication*, 271-311.

[14] Fors, K. L. (2015). *Production and perception of pauses in speech*. PhD thesis, Gothenburg, Sweden: University of Gothenburg.

[15] Wessa P. (2019). Back to Back Histogram (v1.0.7) in *Free Statistics Software* (v1.2.1), Office for Research Development and Education, URL (accessed 06 November 2019): https://www.wessa.net/rwasp_backtobackhist.wasp

Mellow the Cello! Determining Correlations between Human Voices and Instrument Voices as a New Source for Innovating Cello Strings

Kristina Diekjobst[1*] and Oliver Niebuhr[2]

[1]*Mads Clausen Institute, University of Southern Denmark, Sønderborg*
[2]*Centre for Industrial Electronics, University of Southern Denmark, Sønderborg*
**krdie16@student.sdu.dk* **oniebuhr@mci.sdu.dk*

The acoustic and physical characteristics of strings are to some extent comparable to the ones of voices, however, no intensive research in the area of string quality and preferences has been conducted so far. This research is the first proof of concept to being able to apply the extensive existing knowledge of voices and speeches to the area of innovating musical strings and opening immense opportunities for future development. For this, seven different strings from Larsen Strings A/S were recorded and transformed into voices and then a supervised questionnaire with twenty participants was conducted. In this, the seven different strings were each represented by two original string recordings, a more aggressive and a romantic one, and four different kinds of created string voices, a female and male one for each melody. The data results revealed that there is a correlation between strings and their respective string voices in specific cases. The aggressive string melody correlated strongly with the aggressive male voice, while the romantic string melody correlated strongly with the romantic female voice. Currently, a second iteration based on the outcome of this research is being conducted.

INTRODUCTION

Even though the sound of a bowed instrument is heavily influenced by the used strings, there is neither intensive conductive research on string quality, nor standard description tools. This difference in research is surprising since the characteristics, e.g. the range and distribution of acoustic energy across the frequency spectrum, are to some degree comparable between musical strings and human voices.

Can something thus be learned from the extensive empirical knowledge about the acoustics and perception of human voices for the sound design of musical strings? Meaning, if one knows why a human voice sounds better, if and in what ways is it possible to make use of this knowledge in developing better-sounding musical strings?

If such a correlation exists, it does make sense to predict how a better musical string should sound like from the solid knowledge of human voices.

METHOD

In total, seven A-Strings were recorded. All strings were Cello A-Strings with medium tension. In the next round of testing an aggressive and one romantic melody were chosen to be shown, which represent the characteristics of each string. The second aspect of testing were the string voices. These were split into four categories, both female and male versions were created for each of the two melodies, as the gender plays an important role for voice perception as it is a strong indicator for mate quality [1]. This resulted in six different categories in total.

The criteria through which musical strings are being evaluated are very similar to the ones of voices. As research concluded common descriptive words for music can be reduced to three scales that will not correlate with each other [2]. These three, as well as two scales more and an additional question the general impression were asked. See Fig. 2.

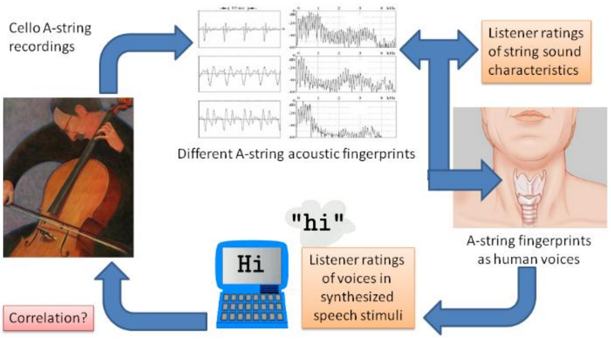

Cello A-string recordings
Different A-string acoustic fingerprints
Listener ratings of string sound characteristics
A-string fingerprints as human voices
"hi"
Hi
Listener ratings of voices in synthesized speech stimuli
Correlation?

Fig. 1. Schematic representation of the project's milestones

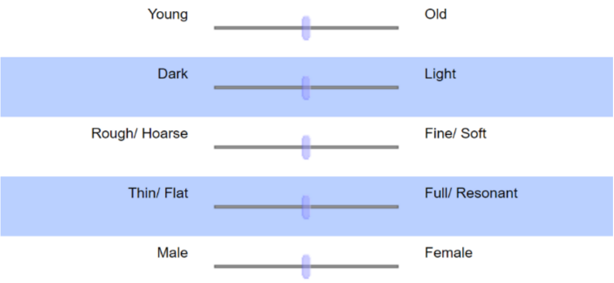


Fig. 2. Scales used in test round

In total, 20 people participated in the survey. Six of them were male and 14 female. The age ranged from 19 to 57 with 17 out of the 20 being in the age range of 19 to 25 years.

Six participants stated that they have no musical background and 13 stated they mostly spoke German between the age of eight and 18 and four were mostly exposed to Danish.

RESULTS

There are many possible correlations to be reported, one of them being the difference in ratings for the seven strings. String 7 was the one with the lowest quality and all yellow marked values are statistically significantly higher to the one of String 7, and all grey ones are lower. See Tab. 2. Especially looking at String 1 and String 4, one can see, that there is a better overall impression.

Tab. 1. Mean Values of all Values – The 7 Strings

	Young/Old	Dark/Light	Rough/Fine	Thin/Full	Male/Female	Attractiveness
String 1	47,3	64,5	59,7	54,3	63,4	55,3
String 2	49,9	59,9	58,1	57,6	60,3	54,2
String 3	50,8	59,6	59,2	55,8	60,0	54,6
String 4	47,0	65,0	62,0	57,4	63,2	58,0
String 5	49,1	60,5	61,2	55,7	63,2	51,1
String 6	51,4	59,2	54,7	57,5	61,2	51,9
String 7	52,2	59,5	57,7	54,0	59,5	52,6

The results for the main research aim, if there are statistically significant correlations between the perception of strings and their corresponding string voice show that there is a correlation between certain categories (Pearson product-moment correlations).

Tab. 2. r-values of correlations between strings and string voices (significant correlations at p<0.05, n=80)

		Aggressive	Romantic
String 1	M	0,18	0,22
	F	0,15	0,18
String 2	M	0,33	0,14
	F	0,04	0,42
String 3	M	0,16	0,13
	F	0,1	0,26
String 4	M	0,34	0,25
	F	0,3	0,48
String 5	M	0,42	0,13
	F	-0,03	0,14
String 6	M	0,32	0,08
	F	0,04	0,17
String 7	M	0,35	0,16
	F	0,01	0,3

As can be seen in Tab. 2 and Fig. 3 correlations can be observed in five out of seven strings between the romantic string and the romantic female string voice and in six out of the seven strings between the aggressive string and the aggressive male string voice.

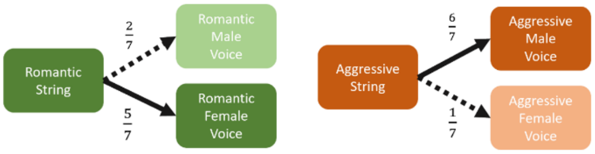


Fig. 3. Visualization of Correlations

This can be easily explained. The correlation exists because the strings and the specific string voices were rated similar and by perceiving and rating the aggressive string melody similar to the male voice and the romantic string melody closer to the female voice it is natural that a higher correlation exists in these cases. The interesting observation for this is that the participants already assigned the melody to characteristics one can stereotypically have about the voices of the different genders. This can also be formulated in a different way, such as that an aggressive melody apparently possesses more male traits and that a romantic melody possesses more female traits.

DISCUSSION

There is an extensiveness of categories and scales, which yield many opportunities for further combining the existing data to reveal interesting correlations.

The open question that needs to be asked is, which correlation can be made for a string melody that is neither romantic nor aggressive. Finding a precise rule for when strings and their equivalent voices correlate is the goal for subsequent research.

FUTURE RESEARCH

Learning from the process and outcome of this research a second iteration of this research has been started to test for further and clearer correlations.

ACKNOWLEDGMENTS

The presented research was only made possible by amazing support throughout the process, all research participants with helpful inputs, and Kristian Bach Sigvardt and Thomas Zwieg from Larsen Strings A/S.

REFERENCES

[1] Hill, A., & Puts, D. A. (2017). *Vocal Attractiveness*. (January). https://doi.org/10.1007/978-3-319-16999-6

[2} Štěpánek, J., & Moravec, O. (2005). *Barva hudebního zvuku a její slovní popis*.

There is music in speech melody! - How pitch intervals shape speaker charisma

Mikkel Ploug[1] and Oliver Niebuhr[2]*

[1]*Guitarist, Composer, Frederiksberg, Copenhagen,Denmark*
[2]*Centre for Industrial Electronics, MCI, University of Southern Denmark*
**oniebuhr@mci.sdu.dk*

The present study sheds a new light on the relevance of pitch-scaling differences, i.e. the interval sizes of the rising and falling pitch movements in speech. Based on a pilot sample of two speakers, an auditory translation of their pitch movements into musical (semitone) notations by a professional Danish musician yields two results: First, different syntactic-melodic elements like phrase onsets and pitch accents occur with different interval sizes; second, speakers generally prefer different intervals in their speech, and these individual preferences have consequences for the charismatic impact of the speaker.

INTRODUCTION

Today's view on meaningful intonational patterns is dominated by temporally unfolding melodic patterns, their syntactic concatenation, and their timing with the string of sound segments. Also the tonal H/L difference in the Autosegmental-Metrical (AM, [1]) framework is tem-poral in nature insofar as it is an interval-size insensitive binary distinction, defined in relation to the (H/L) level of preceding tonal target. In other words, with regard to pitch displays like that of Figure 1, intonational meanings are commonly assumed to be created along the x-axis (time) and not along the y-axis (fundamental frequency, f0, i.e. the primary acoustic counterpart of perceived voice pitch, [1]). Of course, this assumption is well justified and a major driving force in our discovery and understanding of melodic form-meaning links across languages since the early works of [2] and [3]. However, the time-dominated view on intonation to some extent made the frequency axis and its pitch-interval or *'scaling'* differences slip from the research agenda.

Studies still addressing pitch-scaling patterns mainly look at gradually pronounced paralinguistic phenomena. For example, if a particular rising L-H sequence in a language signals a question, then a larger scaling difference from L to H creates a stronger question signal [5]; and while smaller scaling differences between Hs and Ls can barely transport expressive traits such as certain emotions and emphatic accentuation, these features are the more strongly conveyed the larger the scaling differences become [6]. So, what scaling differences should *not* produce (for rare exceptions see [5]) are *qualitative* changes in the communicative interpretation of intonation.

PUZZLE

It was shown by [7] -- and has been replicated a number of times since then -- that the listeners' perception of speaker charisma is closely correlated with holistic prosodic characteristics of the speaker like, for example, pitch level, pitch range, tempo, loudness level, the duration of interpausal units, etc. [8-10]. The 2nd author (ON) is CEO of a public-speaking consultancy company and has, in this context, analyzed and trained the vocal charisma of more than 500 male and female speakers, based on a patent-pending acoustic-prosodic assessment algorithm [10]. Despite the great performance of the assessment algorithm, there are occasionally speakers whose differences in perceived vocal charisma are greater or smaller than their holistic prosodic characteristics alone would predict. Apart from the possibility that probably not all relevant holistic prosodic characteristics are taken into account by the assessment algorithm (not to speak of pronunciation differences and other phenomena at the level of sound segments, [11]), this suggests that there are other non-holistic prosodic features that play a role in perceiving speaker charisma.

IDEA

The first author (MP) is an internationally acknowledged Danish star guitarist and composer. He has a long-standing experience with translating the speech melodies of a speaker into musical notations. The syllable-based musical translations of MP represent a sophisticated, fine-grained analysis of the pitch-scaling patterns in speech in terms of 12 tone intervals, i.e. semitones (st). During this analysis-and-translation work, an idea emerged: There seem to be systematic differences not only in which musical st intervals speakers use for the different building blocks of a tune, but also in which musical st intervals speakers generally prefer to use in their speech.

QUESTION

Combining puzzle and idea above, the following two questions are tested here: (1) If we translate the pitch-scaling patterns of speech melodies into sequences of musical st intervals, can we find empirical evidence for the idea that speakers differ in how they make use of the musical intervals in their speech? And if so, (2) is there a link between the use of musical st intervals in speech and the perceived charisma of the corresponding speaker?

METHOD

In order to test the above questions, we chose a sample of 300-400 syllables of two speakers, a more charismatic speaker and a less charismatic speaker. The more charismatic speaker was Steve Jobs (SJ). The speech excerpt we chose of him comes from his iPhone 4 keynote given in 2010. The less charismatic speaker was Kirsten Jensen (KJ), a politician of the social-democratic party and mayor of the city of Hillerød in Denmark. The speech excerpt we chose for this speaker comes from a speech that presents the program of the party's policies in 2017. Note that the two speakers were chosen because they are similar in terms of the holistic prosodic characteristics of speaker charisma, i.e. pitch level (SJ=211 Hz, KJ=214 Hz), pitch range (SJ=22 st, KJ=20 st), tempo (SJ=4.9 syll/s, KJ=5.1 syll/s), pause frequency (SJ=every 10.7th syll, KJ=every 10.8th syll), etc. However, relative to the existing minor acoustic-prosodic charisma differences between SJ and KJ, the two speakers differ considerably in perceived charisma.

The translation of pitch-scaling patterns into sequences of musical intervals was done separately for the two speakers by MP. Note that MP was, at that point in time, not informed (by ON) about the specific research question. In order to reduce the amount of data and to increase the linguistic implications of the data, the analysis was guided by the AM model and restricted to the three main types of pitch movements within a prosodic phrase: (i) the pitch movement at the beginning of a phrase, (ii) the pitch movement at the end of a phrase, and (iii) the pitch movement associated each pre-nuclear and nuclear pitch accent inside a prosodic phrase. The musical intervals (i)-(iii) have been determined for all prosodic phrases in the samples of SJ and KJ.

RESULTS

Table 1 summarizes in % the results of the translation of pitch-scaling patterns into sequences of musical intervals. Two results are immediately obvious. First, the musical intervals that are mainly used (i) at the beginning of a phrase, (ii) at the end of a phrase, and (iii) in connection with pre-nuclear and nuclear pitch accents are not the same ($\chi^2[14]=78.7$, $p<0.001$). There are relatively more musical intervals smaller than 3 st at the beginning of a phrase (i) than at the end of a phrase (ii). Moreover, the on average largest musical intervals occur in connection with pre-nuclear and nuclear pitch accents (iii).

Secondly, SJ and KJ prefer, at each position (i)-(iii) in their tunes, different musical intervals. There is research in music psychology on the attributes and personality traits that listeners associate with musical intervals. Table 2 provides a summary of that work conducted by [12] (re-analyzed in [13]). Only attributes related to the concept of speaker charisma are listed and sorted such that the first five musical intervals support speaker charisma, whereas the second five musical intervals reduce speaker charisma. If the percentages of interval use by SJ and KJ are projected onto the 2x5 attributes of Table 2, then the results pattern emerges that is displayed in Figure 2. The two speakers SJ and KJ clearly differ in their use of charisma-supporting musical intervals ($\chi^2[5]= 57.3$, $p<0.001$). While the majority is SJ's intervals is associated by listeners with charisma-supporting attributes, the opposite applies to the majority of musical intervals used by KJ. This difference is strongest at (i), i.e. the beginning of a phrase ($z[320]=-14.5$, $p<0.001$). Additionally, SJ uses the highest relative number of charisma-supporting musical intervals for pre-nuclear and nuclear pitch-accents ($z[436]=-3.08$, $p=0.002$), i.e. those syntactic-melodic elements that occur most often in speech communication.

DISCUSSION

Our results clearly suggest positive answers to both research questions. We found empirical evidence for the idea that speakers differ in how they make use of musical intervals in their speech. A twofold difference emerged. First, different interval sizes are used for the syntactic-melodic elements of a tune, especially phrase-onsets (smallest intervals) and (pre)nuclear accents (largest intervals) differ in this respect. Second, speakers generally prefer different intervals. For example, KJ has a preference for a minor and major second as well as for a minor third, whereas SJ more often used a perfect fourth, a perfect fifth, and a major sixth in his speech. Note that these are also differently large intervals (2-4 semitones in the case of KJ and 5-8 semitones in the case of SJ), a fact that we will analyze and discuss in a follow-up paper.

Regarding question (2) about the existence of a link between the use of musical intervals by a speaker and his/her perceived charisma, we found that SJ clearly uses more charisma-supporting intervals in his speech than KJ. Given the fact that SJ and KJ are otherwise similar in all major holistic prosodic characteristics of charisma, the results suggest that the perceived charisma difference between SJ and KJ indeed relies to some extent on the pitch-scaling or musical-interval preferences of the two speakers. Follow-up studies have to examine this possibility in more detail. Not because of its limitation in terms of sample size, the present study only laid the foundation for this new line of research, which has, however, the potential to give new impetus to the study of meaningful scaling differences in intonation. Within this new line of research, it also needs to be determined (and automatized) how the translation from acoustic f0 patterns into sequences of musical intervals actually works and to what extent musical intervals like those derived here by PM precisely and consistently mirror the speech-melody perception of ordinary listeners.

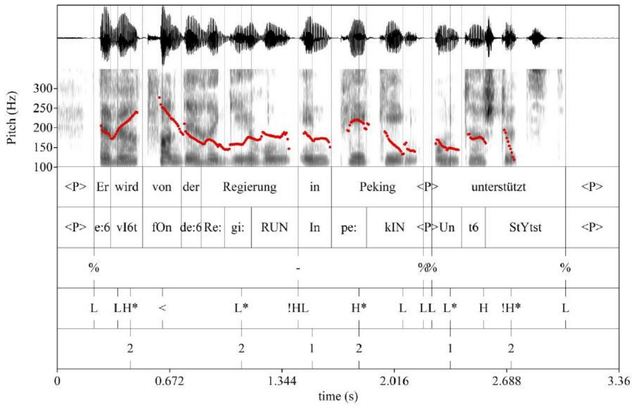


Fig. 1. Example of a tonal H/L annotation based the German DIMA version of the AM framework [4]

Tab. 1: Percentages of use of melodic intervals (st) by the two speakers in (i) phrase-initial and phrase-final (ii) position as well as (iii) on (pre)nuclear accented syllables.

	(i) onset		(ii) acc.		(iii) offset	
Mus. intveral	SJ	KJ	SJ	KJ	SJ	KJ
Unisonous (0 st)	4	11	0	0	1	3
Minor second	7	30	1	3	4	4
Major second	0	26	3	18	5	25
Minor third	12	7	14	29	7	23
Major third	9	8	15	15	21	6
Perfect fourth	13	7	28	9	11	5
Augm. fourth	3	4	6	10	16	3
Diminished fifth	2	7	6	6	7	6
Perfect fifth	10	0	12	5	15	10
Minor sixth	12	0	0	5	0	0
Major sixth	3	0	6	0	10	4
Minor seventh	9	0	0	0	0	3
Major seventh	5	0	0	0	0	0
Octave (12 st)	11	0	9	0	3	0

Tab. 2: Attributes and speaker traits associated with musical intervals according to [12] .

Charisma-supporting musical intervals	
Perfect fourth	Firmness, achievement, simplicity
Perfect fifth	Balance, love, certainty, mastery
Major sixth	Light, kindness, satisfaction, gratification
Octave	Solid, stable, courage, heroism
Charisma-reducing musical intervals	
Minor second	Fear, anger, shyness, illness
Major second	Friction, request, wish, displeasure
Minor third	Heaviness, sadness, pain, discouragement
Dimin. fifth	Instability, anxiety, doubt, uncertainty

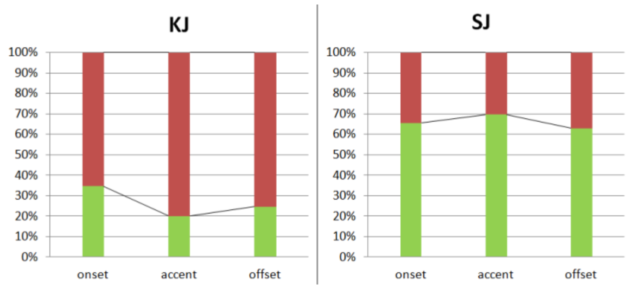


Fig. 2. Percentages of charisma-supporting (green) and charisma-reducing (red) musical intervals found in the speeches of KJ and SJ at positions (i), (ii), and (iii).

REFERENCES

[1] Moore, B. C. (2012). *An introduction to the psychology of hearing*. San Diego: Academic Press.

[2] Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation (PhD thesis)*. Massachusetts Institute of Technology, Dept. of Ling. & Phil., USA.

[3] Kohler, K. J. (1987). Categorical pitch perception. *Proc. 11th International Congress of Phonetic Sciences, Tallinn, Estonia*, 331-333.

[4] Kügler, F., Baumann, S., Andreeva, B., Braun, B., Grice, M., Neitsch, J., ... & Wagner, P. (2019). Annotation of German Intonation: DIMA compared with other annotation systems. *Proc. International Congress of Phonetic Sciences, Melbourne, Australia*, 1-5.

[5] Michalsky, J. (2015). Pitch Scaling as a Perceptual Cue for Questions in German. *Proc. 16th International Interspeech Conference, Dresden, Germany*, 924-928.

[6] Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics* 25, 313-342.

[7] Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Communication* 51, 640-655.

[8] Niebuhr, O., Voße, J., and Brem, A. (2016). What makes a charismatic speaker? A computer-based acoustic prosodic analysis of Steve Jobs tone of voice. *Computers and Human Behvior* 64, 366–382.

[9] D'Errico F., Signorello R., Demolin D., Poggi I. (2013). The perception of charisma from voice. A crosscultural Study. *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland*, 552–557.

[10] Biadsy, F., Rosenberg, A., Carlson, R., Hirschberg, J. & Strangert, E. (2008). A cross-cultural comparison of American, Palestinian, and Swedish perception of charismatic speech. *Proc. 4th International Conference of Speech Prosody, Campinas, Brazil*, 579–582.

[11] Niebuhr, O. & Gonzalez, S. (2019). Do sound segments contribute to sounding charismatic? Evidence from acoustic vowel space analyses of Steve Jobs and Mark Zuckerberg. International *Journal of Acoustics and Vibration* 24, 343–355.

[12] Willems, E. (1977). *L'oreille musicale*. Biemre: Pro Musica.

[13] Costa, M., Ricci Bitti, P. E., & Bonfiglioli, L. (2000). Psychological connotations of harmonic musical intervals. *Psychology of Music* 28, 4-22.

Speaking Intensity Potential (SIP) – A new Clinical Measure for the Voice Disordered Population

*Frederikke Dam Hansen, stud. Audiologoped, Lærke Hansen Siedentopp, stud. Audiologoped
**Ågot Møller Grøntved, MD
***Trine Printz, Audiologoped MA

Institute of Language and Communication, University of Southern Denmark, Odense M, Denmark

* *freha16@student.sdu.dk*, *lasie15@student.sdu.dk*
** *aagot.grontved@rsyd.dk*
*** *trine.printz@rsyd.dk*

Doctors and speech-language-therapists sometimes have difficulties guiding, treating, and evaluating patients with intensity decreasing voice disorders because no measures of the intensity potential of the speaking voice have ever been made. This research has in one way solved this clinical problem by investigating the Speaking Intensity Potential (SIP) of Danish adults with healthy voices and this study is the first to investigate this particular measure. In the future it will be beneficial to investigate the equivalent measure for diagnose-specific patient groups consisting of people with voice disorders known to cause intensity decrease such as vocal fold paralysis.

INTRODUCTION

The Voice Range Profile (VRP) is an objective method used by speech language pathologists to measure the extent of the voice (Sanchez, Oates, Dacakis, & Holmberg, 2014). It gives a visual view of the potentials and the capacity of the voice in the parameters of frequency (x-axis) and intensity (y-axis) (Sanchez et al., 2014) (fig. 1). The Speech Range Profile (SRP) is an equivalent visual view of the potentials of the speaking voice (fig. 1). The VRP can be reduced in multiple ways, both in the frequency range and the intensity range. Some patients with voice disorders have reduced voice intensity and experience the need to put strain on their speaking voice in order to speak in daily life. This will be further presented.

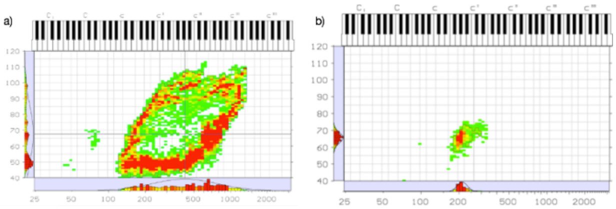


Fig. 1. a) Voice Range Profile and b) Speech Range Profile.

The Speaking Intensity Potential (SIP) is a value that describes the intensity potential from the habitual intensity in the speaking voice and to the maximal intensity in the VRP at the same frequency level (fig. 2).
Clinically, SIP is a relevant value for evaluating treatments for patients with voice intensity problems. Measuring SIP has not been done previously. Our aim was to establish a normative data material for SIP.
A low SIP-value means that the speaker has problems increasing his or her voice habitual speaking intensity.

Consequently, the speaker is forced to putting extra strain on the voice even though speaking with a normal intensity. The voice may sound normal, but the speaker is using excessive effort to produce a hearable voice; for example, in a public place with background noise. Also, this can induce problems in occupational contexts.

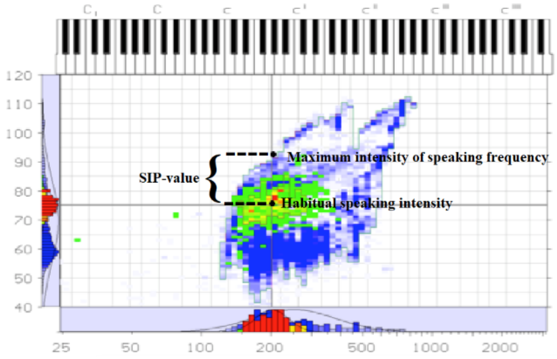


Fig. 2. Speaking Intensity Potential.

METHOD

This quantitative, retrospective study is including Danish adults ≥20 years with healthy voices (n=86). The subjects were divided into groups based on gender, age, and smoking-habits. The standard values were calculated for all groups and included medians, 5%, and 95%-percentiles.

Data were collected between July 7th, 2016 and December 31st, 2018 by Speech Language Pathologists at Odense University Hospital (OUH) (clinical setting, fig 3). Data were stored in REDCap electronic data capture tools hosted by OUH (Harris et al., 2009).

Two measures from the VRPs and SRPs were extracted in order to calculate the SIP value: (1) the habitual speaking intensity of the SRP, and (2) the maximal intensity in the VRP at the same frequency level (fig. 2).

Fig. 2. Clinical setting; recording a voice range profile.

RESULTS

Normative data will be presented. Women overall had a larger SIP value than men. Furthermore, the SIP value decreased with age unaffected by gender and smoke habit. It was unclear what effect smoking had on the SIP value, yet men's SIP values were more affected by smoking than women's SIP.

DISCUSSION

More participants are needed to provide statistically significant normative SIP-values in all age/gender/tobacco subgroups, however, most of the results can be clinically implemented. Interpretation of the SIP-values should always be done with caution and investigation of the voice should always involve additional assessment tools due to the multidimensionality of the voice (Dejonckere et al., 2001). Future studies will explore other types of normative SIP values, and also investigate the SIP-value in diagnose-specific groups of patients with intensity problems.

REFERENCES

Dejonckere, P. H., Bradley, P., Clemente, P., Cornut, G., Crevier-Buchman, L., Friedrich, G., . . . Committee on Phoniatrics of the European Laryngological, S. (2001). A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques: Guideline elaborated by the Committee on Phoniatrics of the European Laryngological Society (ELS). *European Archives of Oto-Rhino-Laryngology, 258*(2), 77-82. doi:10.1007/s004050000299

Hallin, A. E., Fröst, K., Holmberg, E. B., & Södersten, M. (2012). Voice and speech range profiles and Voice Handicap Index for males - methodological issues and data. *Logopedics Phoniatrics Vocology, 37*(2), 47-61. doi:10.3109/14015439.2011.607469

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42*(2), 377-381. doi:10.1016/j.jbi.2008.08.010

Sanchez, K., Oates, J., Dacakis, G., & Holmberg, E. B. (2014). Speech and voice range profiles of adults with untrained normal voices: Methodological implications. *Logopedics Phoniatrics Vocology, 39*(2), 62-71. doi:10.3109/14015439.2013.777109

The prosodic features of the Topic information unit in spontaneous speech from a crosslinguistic perspective

Frederico Cavalcante[1*], Plínio A. Barbosa[2], Tommaso Raso[1]

[1]*Federal University of Minas Gerais* (UFMG)
[2]*State University of Campinas* (UNICAMP)
*fredericoa4@gmail.com

Based on the *Language into Act Theory*, a pragmatic framework for the study of spontaneous speech, this paper presents a statistically based study of the prosodic features of the of *Topic information unit*, specifically understood as the unit that supplies a cognitive domain for the interpretation of the illocutionary force, carried by the Comment. We use Functional Data Analysis and Principal Component Analysis to examine the melodic curves of Topic units from spontaneous speech corpora of American English and European and Brazilian Portuguese. We show the substantial results that were obtained in an interrater agreement test to assess the identification of Topic in spontaneous speech data. In addition, we provide statistical evidence for the classification of the prosodic forms of TOP as well as a study of the duration pattern of the unit showing that the unit exhibits significantly longer syllables in its nucleus.

Introduction

This study is based on the Language into Act Theory (L-AcT), a pragmatic framework for the study of spontaneous speech [1]. We examine the Topic information unit (TOP) defined specifically as the unit that supplies a domain of identification for the interpretation of the illocutionary force, which is in turn carried by another unit, the Comment (COM).

Within L-AcT, the basic unit of speech analysis, called the *utterance*, corresponds to the smallest stretch of speech exhibiting pragmatic autonomy. Such autonomy is conveyed prosodically through terminal breaks that yield the auditory impression of conclusion. TOP, present in 9-15% of utterances, is realized by a non-autonomous prosodic unit, always preceding COM, this being the only information unit an utterance must have.

Different studies on TOP [2–6] have suggested that it is realized by prosodic forms characterized by nuclei consisting of longer syllables with higher intensity which may be classified into three types according to melodic shape into three types: Type 1 having a rising-falling nucleus, Type 2 having a rising one, and Type 3 having two frequently discontinuous semi-nuclei, the first of which usually featuring much higher f0 values than the second (see Fig. 1).

The goal of this study is to show the results of an interrater agreement test with four annotators who had to identify TOPs in spontaneous Brazilian Portuguese speech data and to use statistical techniques to examine the melodic forms and duration pattern of TOPs in American English (AE), European and Brazilian Portuguese (EP and BP). The purpose of the study is to verify the appropriateness of the descriptions of TOP proposed within L-AcT's framework.

Methods

We used the Kappa statistics [7] in the agreement test with four raters in a task where they were given 100 utterances, each containing at least two prosodic units, randomly selected from the three sections of the formal part of the C-ORAL-BRASIL corpus [8]: 40 utterances from the Media section (TV and radio programs), 40 from the Natural Context section (debates, lectures, public hearings etc.), and 20 from Telephonic section. These utterances amount to 541 prosodic units, and the raters were asked to decide, by listening to them in context, whether or not each prosodic unit constituted a TOP. The context was comprised of four utterances, two immediately before and two immediately after the target utterance. A carefully designed protocol was created for the test, and we used Praat [9] for the task itself and R [10] for computing the results.

As for the prosodic forms, we applied Functional Data Analysis (FDA) [11, 12] to 56 melodic curves of TOP from the AE minicorpus [13], 59 from the BP minicorpora [14], and 28 from the EP component of the C-ORAL-ROM [15]. FDA provided us with smoothed and time-aligned representations of the curves (Fig. 2), thus allowing the qualitative comparison as well as the use of Functional Principal Components Analysis (F-PCA) to unveil the main modes of variation in the sample and to assess the classification proposed by previous studies [2–6]. For the assessment, we used the PC scores computed for the curves and ANOVA.

F-PCA provides functional approximations *f(t)* of the curves, which are computed through the linear combination of the mean curve $\mu(t)$ and the main *PC(t)* curves multiplied by the PC scores s_i (computed for each curve individually) according to the Eq. 1.

$$f(t) \approx \mu(t) + s1 * PC1(t) + s2 * PC2(t)... (1)$$

In addition, we used the median PC scores of curves classified as belonging to the same class in order to generate reliable models for each class (Fig. 1) and to check whether such models resemble the ones proposed by others before this study.

Class membership information interfered with no computation whatsoever, as this information was only introduced after FDA and F-PCA had been applied.

For examining the temporal dimension of the unit, we used Praat to segment and annotate 78 TOPs, which add up to 475 syllables, and used the *SGDetector* script [16] to obtain the normalized durations of the syllables. We then used ANOVA to verify whether the nuclear syllables were significantly longer than non-nuclear ones.

Results

In the Kappa test, we obtained substantial *k* scores, no matter how we partition the data: 0.79 (all the utterances, n=100), 0.80 (Media, n = 40), 0.79 (Natural Context, n = 40), and 0.66 (Telephonic, n = 20).

The FDA output curves (Fig. 2) already suggested the appropriateness of the classification scheme proposed in previous studies. The Flat curves that are shown in the bottom-right panel of Fig. 2 are similar to Type 3, except for their exhibiting much less f0 variation, hence their being shown separately. We are currently trying to determine whether Flat is a subclass of Type 3 or a form on its own right.

The PC scores computed with F-PCA provide further evidence for the separability of the forms proposed in the above-mentioned studies (see Fig. 3). Kruskal-Wallis rank sum tests show that the PC scores 1, 2 and 3 differentiate the curves of our sample by type (chi sq.= 107.18, 47.5, and 16.239, df = 3, p < .01). The first 3 PCs together account for 94% of the variability in the data.

Fig. 4 shows the curves that we obtain by using the median PC scores of the 1st and 2nd components for the curves of each class. As the figure shows, the curves for Types 1, 2, and 3 are faithfully close to the ones in Fig. 1.

Finally, a Wilcoxon rank sum test showed that the durations of the nuclear and the non-nuclear syllables of the TOPs in our sample are significantly different (W=21667, p << .001), the difference being maintained even if the apparent outliers (Fig. 5) are excluded.

References

[1] Cresti, E. (2000). *Corpus di Italiano parlato*. Firenze: Accademia della Crusca.

[2] Firenzuoli, V., & Signorini, S. (2003). L'unitá informativa di topic: correlati intonativi. In G. Marotta (Ed.), *in Atti delle Giornate del Gruppo di Fonetica Sperimentale - XIII, Pisa, Novembre 2002 ETS, Pisa* (pp. 177–184).

[3] Mittmann, M. M. (2012). *O C-ORAL-BRASIL e o estudo da fala informal: um novo olhar sobre o Tópico no Português Brasileiro*. Belo Horizonte, Universidade Federal de Minas Gerais.

[4] Rocha, B. N. R. de M. (2012). *Características Prosódicas do Tópico em PE e o uso do pronome lembrete*. Belo Horizonte, Universidade Federal de Minas Gerais.

[5] Cavalcante, F. A. (2016). *The topic unit in spontaneous american english: a corpus-based study*. Belo Horizonte, Universidade Federal de Minas Gerais.

[6] Raso, T., Cavalcante, F., & Mittmann, M. M. (2017). Prosodic forms of the Topic information unit in a cross-linguistic perspective: A first survey. In A. De Meo & F. M. Dovetto (Eds.), *La comunicazione parlata/Spoken communication* (pp. 473–498). https://doi.org/10.4399/978882552064428.

[7] Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382. https://doi.org/10.1037/h0031619.

[8] Raso, T. & Mello, H. (forthcoming). *The C-ORAL-BRASIL II. Corpus de referência do português falado (formal em contexto natural, mídia e telefone)*.

[9] Boersma, P., & Weenink, D. (2001). Praat: a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

[10] R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

[11] Ramsay, James O. & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). Springer.

[12] Gubian, M., Torreira, F., & Boves, L. (2015). Using Functional Data Analysis for investigating multidimensional dynamic phonetic contrasts. *Journal of Phonetics*, *49*, 16–40. https://doi.org/10.1016/j.wocn.2014.10.001.

[13] Cavalcante, F., & Ramos, A. (2016). The American English spontaneous speech minicorpus: architecture and comparability. *CHIMERA: Romance Corpora and Linguistic Studies*, *3.2*, 99–124.

[14] Mittmann, M. M., & Raso, T. (2011). The C-ORAL-BRASIL informationally tagged minicorpus. In H. R. Mello, A. Panunzi, & T. Raso (Eds.), *Pragmatics and Prosody. Illocution,Modality, Attitude, Information Structure and Speech Annotation* (pp. 151–183).

[15] Cresti, E., & Moneglia, M. (Eds.). (2005). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*. Amsterdam / Philadelphia: John Benjamins Publishing Company.

[16] Barbosa, P. A. (2006). *Incursões em torno do ritmo da fala*. Campinas: Pontes.

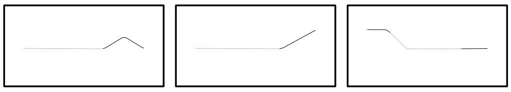


Fig. 1. Schematic representations of the TOP prosodic forms as described in previous studies. Starting from the left: Types 1, 2, and 3. Dotted lines indicate non-nuclear portions, which constitute optional syllables.

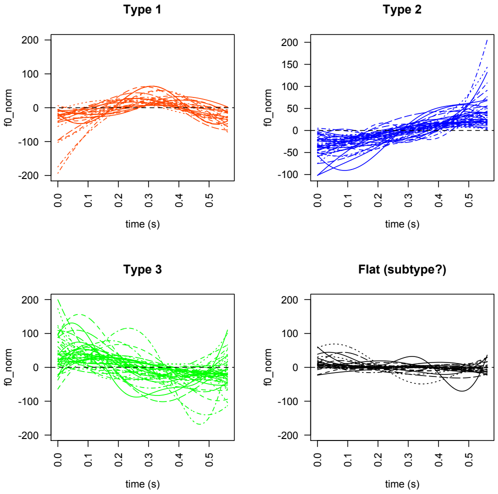


Fig. 2. Smoothed and time-aligned melodic curves obtained with FDA with no recourse to class-membership information; shown separately only to aid in the visualization.

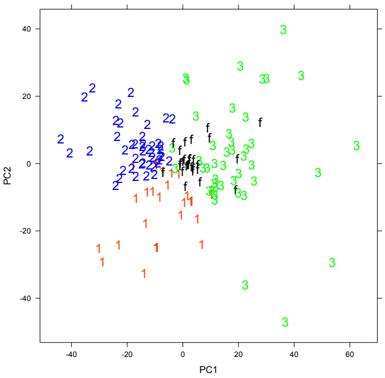


Fig. 3. PC1 and PC2 scores of the curves in the sample.

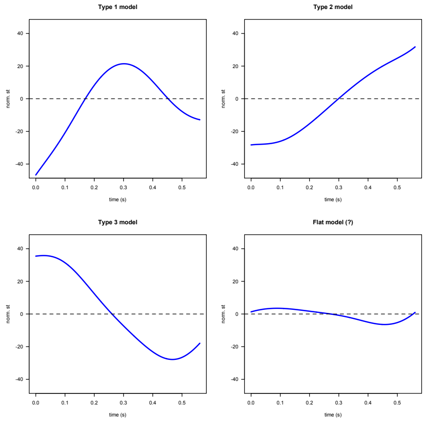


Fig. 4. Models of TOP melodic forms computed using the median PC1 and 2 values for the curves of each class.

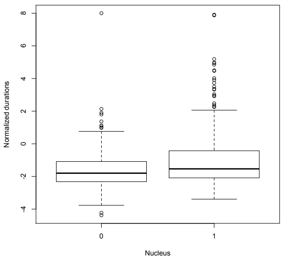


Fig. 5. Normalized z-scores of syllable durations of TOPs in our sample.

Prosodic correlates of dominance and self-assurance. Acoustic cues to testosterone related personality states of male speakers

Jan Michalsky[1*], Tobias Kordsmeyer[2], Oliver Niebuhr[3], Lars Penke[2]

[1]Institute for German studies, University of Oldenburg, Germany
[2]Department of Psychology & Leibniz Science Campus Primate Cognition,
University of Goettingen, Germany
[3]Centre of Industrial Electronics, Mads Clausen Institute, University of Southern Denmark
***** j.michalsky@uni-oldenburg.de

The steroid hormone testosterone modulates human behavior related to both mating and competition. Increased levels of testosterone in male speakers correlate with both self-reported as well as externally observed degrees of dominance and self-assurance. In this paper, we ask whether features of vocal acoustics contribute to the perception of dominance and self-assurance and hence may contribute to perceive testosterone-related changes in male behavior. We analyzed 290 recordings of short self-presentations of 145 male speakers for changes of f0, rhythm and voice quality. The acoustic cues were related to the evaluation of dominance and self-assurance by 400 female raters. Dominant and self-assured speakers show an increased speech rate. In terms of f0, dominant speakers employ a wider f0 range as well as more and deeper final falls. However, contradicting expectations from previous studies we do not find differences in f0 mean or median. In terms of voice quality, dominant speakers show smaller differences between the first harmonic and the third formant corresponding to less breathiness or more tenseness as well as higher energy in the fundamental frequency range below 250 Hz.

The full paper of this contribution can be found online by January 2020.

Types of hate speech in German and their prosodic characteristics

Jana Neitsch[1*] and Oliver Niebuhr[1]

[1] *Centre for Industrial Electronics, Mads Clausen Institute, University of Southern Denmark, Sønderborg/DK*
* neitsch@mci.sdu.dk

The present pilot study investigates spoken hate-speech items, produced by a professional German speaker after actual Twitter and Facebook posts and reflecting characteristic morpho-syntactic features of German hate speech. Acoustic-prosodic signal analyses reveal a considerable feature-specific variation in the production of hate speech. Thus, at least at the level of phonetics and spoken hate speech, no evidence was found that hate speech is a separate communicative function conveyed by a set of uniformly realized prosodic characteristics.

INTRODUCTION

Hate speech is becoming more and more of a concern in societies around the world [1]. The term “hate speech”, however, remains highly controversial. There is a lack of consensus on its definition and impact, while the motivation and justification for its criminalization and regulation are inextricably linked to allowing freedom of speech on the one hand and protecting the human rights of equality and dignity on the other [2].

Given the pressure that hate speech exerts on the pillars of modern civilization, it is striking how little is known about the linguistic and communicative mechanisms underlying the expression and perception of hate speech. The Velux-funded XPEROHS project is intended to address some of these gaps. The project is based on a large German-Danish corpus of about 3.5 million real instances (posts) of hate speech on Twitter and Facebook, see [3,4] for further details on the compiled corpus and the goals of the cross-linguistic project. The present study is concerned with the German subset of this corpus, which consists of ~1.7 million hate-speech posts.

Although hate speech is primarily associated with written language, at least in the way it is experienced and discussed in the media [5], there is, of course, also hate speech in spoken language, for example in political discourse [6] or in connection with bullying on school yards and football fields [7]. The present study takes the first step into the field of spoken hate speech by asking the question, at an acoustic-phonetic level, whether hate speech has a uniform manifestation in the sense of a context-independent phonetic core and/or how large the phonetic variation of hate speech is with respect to major lexical and morpho-syntactic features that have emerged – for the German subset of the corpus – from a preceding linguistic processing of the Twitter and Facebook data.

For providing an empirical definition and scientific guidance in identifying and handling (e.g., reporting, deleting) hate speech, relatively or absolutely constant characteristics are required. At the level of speech communication, this would mean that phonetic parameters are not entirely shaped by the morpho-syntactic features underlying the functional and semantic variants of hate speech but also by the phenomenon of hate speech itself.

METHOD

The acoustic-phonetic analysis conducted here is based on German hate speech. A set of 12 hate-speech items (posts) were selected from the German part of the Twitter and Facebook corpus. They were all similarly short. That is, they consist of less than 25 words and include between 20 and 30 syllables. Moreover, semantically they are all directed against the minority group of immigrants. Six items use the more general and relatively neutral term "Ausländer" (foreigners) to refer to this group, the other six use the more specific, religious and – in this context – negatively connoted term "Muslime" (Muslims).

These 12 original hate-speech items have been modified or supplemented in terms of major lexical and morpho-syntactic features that crystallized as being characteristic of hate-speech items posted on Twitter and Facebook [3], thus leading to 6 feature conditions in addition to the original baseline condition: (1) rhetorical questions (RQs, e.g., "who wants/needs/would ever..."), (2) irony (IRO, modal particles expressing irony in German), (3) imperatives (IMP, directed either at the writer's peer group "we must..." or at the minority group "go/stop...."), (4) metaphors (MET, e.g., "Muslims are like..."), (5) holocaust reference (HOL), and (6) indirectness (IND e.g., "I am not against foreigners/ Muslims, but..."). In this way, 6 items sets have been created in addition to the original set. Thus, the total number of items was 84 (7x12).

The 84 hate-speech items were evaluated by two independent groups of people, an expert panel of researchers working on hate speech and a panel of ordinary users of Twitter and Facebook. The 84 items passed this pre-test in that they were consistently identified as hate speech. Moreover, both experts and non-experts rated all items similarly as to the degree of expressed hate.

The 84 items were realized, in line with [8], by a male native speaker of German (BP, 47 years old). He is a professional speaker as well as an experienced public-speaking trainer with an academic education and a PhD degree in phonetics and linguistics. BP is able to control the phonetic characteristics of his speech and to deliberately choose and produce phonetic patterns in order to create specific semantic-pragmatic effects.

On this basis, BP received the instruction to familiarize himself thoroughly with the set of 84 separate items and to practice their elicitation with different phonetic realizations in order to find one that suits each individual item and makes it sound like authentic and natural spoken hate speech. Note that BP was not informed about the 7 feature conditions included in the 84 items and according to a debriefing interview, he also did not become aware of these conditions during the speech-production task.

The speech-production task itself was conducted in the sound-proof booth of the Kiel Phonetics Lab [9]. Recordings were made with a Microtech Gefell M940 microphone at a 44.1 kHz sampling rate and a 16-bit quantization. BP produced the 84 items as isolated hate-speech utterances.

The acoustic-phonetic analysis of the 84 items included two pitch (f0) parameters, i.e. f0 mean (Hz) and f0 range (semitones, st), as well as two voice-quality parameters, i.e. mean HNR (Harmonics-to-Noise Ratio measured in dB) and the Hammarberg index (dB). Loudness was included in terms of mean RMS intensity (dB). Duration and tempo parameters were excluded due to the items' different morpho-syntactic make-up. Analysis was done automatically by means of PRAAT scripts written by [10,11,12].

RESULTS

Analyzed were 84 recorded stimuli that were produced by the phonetically trained speaker. Results were analyzed for mean f0, HNR, Hammarberg Index, and f0 range. *P*-values were adjusted using the Benjamini-Hochberg correction [13] to account for the multiple variables that were tested. All statistical models showed significant interactions between *target group* (*Muslims* vs. *foreigners*) and all of the *feature conditions* (see above; all *p*-values < 0.03). Interactions were split up according to the two targets and compared against the original items. Results for *foreigners* indicate that RQ, IRO, IND, IMP, and HOL were realized with a significantly lower mean f0 than their original base items (all *p*-values < 0.04). For *Muslims*, IRO and HOL items were realized with a significantly lower mean f0 (all *p*-values < 0.0003), whereas IMP, IND and RQ were produced with a higher mean f0, see Fig.1. HNR results for *Muslims* show significantly higher values compared to the original base items for all the feature conditions except for HOL (all *p*-values < 0.03), indicating less breathiness than in the original base items. For *foreigners*, IMP, IND and RQ were also produced with higher HNR values, indicating less breathiness than in the original base items (all *p*-values < 0.03). The Hammarberg index yielded higher values for RQ and HOL and lower values for MET (all *p*-values < 0.03) when the minority group concerned were *foreigners*. For *Muslims*, all feature conditions had higher Hammarberg indices than the original base items (all *p*-values < 0.04; see Fig. 2). Finally, f0-range results indicate lower values for IMP, MET and RQ compared to original items (all *p*-values < 0.02) in the *foreigner* sub-set, but higher values in the *Muslim* subset (all *p*-values < 0.0003).

DISCUSSION

A main result of our data is that items addressing the target minority group of *foreigners* differ prosodically from those addressing target minority group of *Muslims*. In particular, the prosodic differences caused by the six feature conditions compared to the original items were a lot more pronounced when the hate speech targeted Muslims rather than foreigners. That is, the communicative functions that were involved in the tested hate speech items and embodied by the six lexical and morpho-syntactic features were conveyed more clearly and consistently, i.e. with a greater phonetic effort, in Muslim-oriented speech. The greater phonetic effort includes a greater vocal effect, see the higher HNR and Hammarberg-index values for the Muslim items. This makes sense given that Muslims are a smaller, more specific and negatively connoted target group than foreigners in general. A further implication of this finding is that the "intensity" of hate speech can vary at the phonetic level.

However, as was stated above, the nature of this variation points in the direction of a more or less strong signalling of the communicative functions (irony, rhetorical questions, etc.) conveyed by the different feature conditions. For example, rhetorical questions turned out to be the condition that differed most in its prosodic characteristics from the original base items, which is plausible insofar as this communicative function is, unlike indirectness and holocaust reference, primarily conveyed by means of prosody. The prosodic characteristics of rhetorical questions show such a coherent co-variation across speakers and contexts that researchers assume the existence of a 'prosodic construction' for this communicative function in the sense of a bundle of linked parameters and settings [14]. The parameters and settings we find here for the realization of rhetorical questions in hate speech are in accord with those reported in [14].

The conclusions we can draw from these findings with respect to our research questions is that there is obviously no uniform, context-independent phonetic core pattern that characterizes hate speech (we have also tested other prosodic parameters with the same result). The opposite is true. The contextual variation of hate speech is substantial. The nature of this variation suggests that it is externally driven, i.e. due to signalling communicative functions that can be involved in hate speech, such as rhetorical questions. Unlike for the latter, we found no evidence for something like a separate 'prosodic construction' of hate speech. Of course, this conclusion is still very preliminary since the analyzed production data only relies on a single speaker (although his skills argue in favor of the generalization of the produced items), and since we still need to compare the different hate-speech conditions to a baseline condition of no hate speech from the same speaker.

ACKNOWLEDGMENTS

The XPEROHS project (project number 95-16416) was funded by the Velux Foundation. The authors would like to thank Benno Peters from Kiel University for recording the files for the purpose of the study. Please note that the speaker was chosen since he fulfilled the criteria (gender, age) of a prototypical hate speaker and is phonetically trained. Furthermore, we thank Andrea Kleene, Nicole Baumgarten, and Klaus Geyer for their discussions and support in conducting this study.

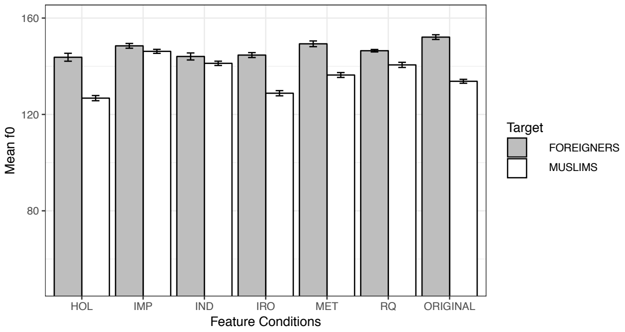


Fig. 1. Mean f0 levels across all feature conditions for the two target groups (foreigners and Muslims).

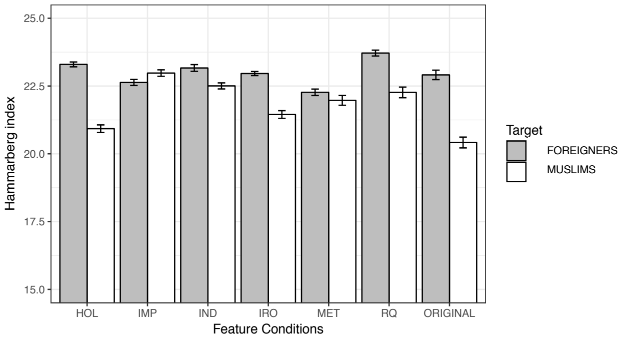


Fig. 2. Hammarberg index across all feature conditions for the two target groups (foreigners and Muslims).

REFERENCES

[1] Guterres, A. (2019). United Nations Strategy and Plan of Action on Hate Speech. Taken from: https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf

[2] Herz, M., & Molnár, P. (Eds.). (2012). *The content and context of hate speech: rethinking regulation and responses*. Cambridge University Press.

[3] Baumgarten, N., Bick, E., Geyer, K., Lund Iversen, D., Kleene, A., Lindø, A. V., Neitsch, J., Niebuhr, O., Nielsen, R. & Petersen, E. N. (Submitted). Towards Balance and Boundaries in Public Discourse: Expressing and Perceiving Online Hate Speech (XPEROHS). *RASK – International journal of language and communication*.

[4] XPEROHS Project. (2019). https://www.sdu.dk/en/om_sdu/institutter_centre/isk/forskning/forskningsprojekter/xperohs

[5] Assimakopoulos, S., Baider, F. H. & Millar, S. (2017). *Online Hate Speech in the European Union: A Discourse-Analytic Perspective.* Cham: Springer.

[6] McClay, R. (2017). Us and them: A descriptive analysis of Donald Trump's campaign speeches. *Unpublished Master Thesis. University of Birmingham. Retrieved by https://www. birmingham. ac. uk/Docum ents/college-artslaw/cels/essays/appliedlinguistics/McClay2017. Trump-Speech-Discourse-Analaysis. pdf.*

[7] Darmstadt, A., Prinz, M., Rocholl, F. & Saal, O. (2018). *Hate Speech und Fake News: Fragen und Antworten.* Amadeu Antonio Stiftung und Berliner Landeszentrale für politische Bildung, Berlin, Germany.

[8] Hrdina, M. (2016). Identity, activism and hatred: Hate speech against migrants on Facebook in the Czech Republic in 2015. Naše společnost, 14(1), 38-47.

[9] Niebuhr, O. & Peterson. J. M. (2011). *KALIPHO: Kieler Arbeiten zur Linguistik und Phonetik*. Kiel, Germany.

[10] Xu, Y. (2013). ProsodyPro – A Tool for Large-scale Systematic Prosody Analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), Aix-en-Provence, France. 7-10.

[11] De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2), 385-390.

[12] De Looze, C., & Hirst, D. (2008). Detecting changes in key and range for the automatic modelling and coding of intonation. In *Proceedings of International Conference on Speech Prosody*, 4, 6-9.

[13] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, Series B 57, 289-300.

[14] Neitsch, J., & Niebuhr, O. (2019). Questions as prosodic configurations: How prosody and context shape the multiparametric acoustic nature of rhetorical questions in German. In *Proceedings of the 19th International Congress of Phonetic Sciences*. 2425-2429.

I Shall Know you by your Voice – Melodic and Physical Dominance in the Design of Robot Voices

Kerstin Fischer[1*], Oliver Niebuhr[2], Rosalyn M. Langedijk[1] and Selina Eisenberger[1]

[1]Department of Design and Communication, University of Southern Denmark, Sonderborg kerstin@sdu.dk; rla@sdu.dk; seeis16@student.sdu.dk
[2]TEK, University of Southern Denmark, Sonderborg
olni@sdu.dk

In this paper, we present a systematic attempt at identifying parameters of synthesized voices that can be adjusted in order to index the robot to which the voice belongs. We carried out an experimental study for which we first generated stimuli by synthesizing the robot voices using a free text-to-speech synthesis system, which were then manipulated according to the parameters f0 level, f0 range and formant dispersion, associated with melodic and physical dominance. These stimuli were matched with videos of robots of different sizes and appearance. We embedded the videos in an online questionnaire, where participants' tasks were to indicate the impressions evoked by the robots and to identify the robot to which the voice was the best fit.

INTRODUCTION

Robots are envisioned to soon take over numerous roles in our daily lives, and there are currently enormous efforts to develop the technology and the appropriate social behaviors for robots to join social spaces with people. However, one of the least explored areas concerns robot voices (Crumpton & Bethel 2016). There are many reasons for this; for instance, technically it is not trivial to change the voice in a text-to-speech system without causing distortions and rendering the resulting speech hard to understand. Moreover, that robot voices might need attention is only slowly discovered in robotics engineering, where often neither the awareness nor the competence is there to pay attention to the choice of voice for a given robot. Furthermore, in humans, voices are unique, and the general parameters that cause the particular properties of human voices, especially in relation to the respective speaker's appearance, are not well understood beyond some very general correlations like body height, i.e. length of vocal tract, and formant spectrum, or between the typical female and the typical male speaker (e.g. Nass & Brave 2005).

Thus, besides the technical issues involved, it is not clear how a voice can be designed to index the speaker; that is, when we hear a typical female voice, we envision a typical female (possibly with additional characteristics), and if we are n a room with several people, we can use these inferences to identify the respective speaker. However, for robots, no such associations exist nor have they been systematically explored. Recently, McGinn & Torre (2019) have shown that most of the voices of even commercially successful robots would not be related to the respective robot if heard in isolation. If robots are to share social spaces with humans, especially if it is more than one robot at a time, robot voice should indicate who is speaking. This is especially crucial in elderly care, one of the domains in which robots are expected to support caregivers very soon (e.g. Riek 2013). Especially residents who suffer from some form of cognitive decline will have problems identifying who is speaking (cf. Fischer et al. 2020).

To address the lack of knowledge on possible associations between voice characteristics and appearance, we carry out a study to systematically explore the effects of the acoustic features of formant dispersion, which corresponds to vocal tract size in humans and which can be interpreted as signaling physical dominance, and pitch level and pitch range, which have functions concerning emotionality, engagement and submission and which therefore can be seen as expressing melodic dominance. In an in-between subject experiment, participants rated the interpersonal effects of the different voices and identified which robot relates to the different voices best. The study therefore addresses both interpersonal functions and correspondence between robot and voice.

METHOD

The study is set up as a between-subjects experiment.

Stimulus Creation

We first created four robot videos of robots which differ in size (two large, two small robots) and in the degree of anthropomorphic design (two anthropomorphic, two abstract robots). Furthermore, we created four robot voices that differ on the two dimensions melodic and physical dominance by varying formant dispersion on the one hand and pitch range and pitch level on the other. That these features are indeed related to physical and melodic dominance respectively is apparent from research by Ohala (1984) and Liu and Xu (2014) on the frequency code and the expression of dominance-related emotions like anger through acoustically projected body size. We created a baseline utterance "Have a good day", which we then manipulated using PSOLA in Praat

(Boersma 2001) according to the two dominance dimensions (see Table 1).

Tab. 1: phy = physical dominance, mel = melodic dominance; H/L = high/low acoustic parameter setting

Variable	Manipulated dominance cues (4 cond.)			
	none	physical	melodic	phy&mel
F0 level	H	H	L	L
F0 range	H	H	L	L
F disp.	H	L	H	L
label	NN	NY	YN	YY

Questionnaire Composition

In a second step, we created a questionnaire as a between subjects experiment. In particular, in the first part of the survey, after the demographic questions, participants were asked to rate robots, which were presented to them in short videos and together with one of the four voices, according to the features engaging, enthusiastic, charming, persuasive, boring, passionate and convincing. In the second part of the experiment, audio files with the robot voices were presented one by one, together with images of the four robots the participants had previously seen in the videos. Their task was then to find the robot that best matches the voice heard. Potential carry-over effects are minimized by the fact that the video-voice pairings in part 1 of the experiment is different in each condition.

Robots

The robots used in the study are two small robots: Haru, an abstract desk-top robot built by the Honda Research Institute in Tokyo, and the EZ-bot built by JD-Robotics. The two large robots used are the robot developed in the project “Smooth” for laundry and garbage transportation, guiding and water serving in elderly care institutions, and Moxi, built by Diligent Robotics (see Figure 1).

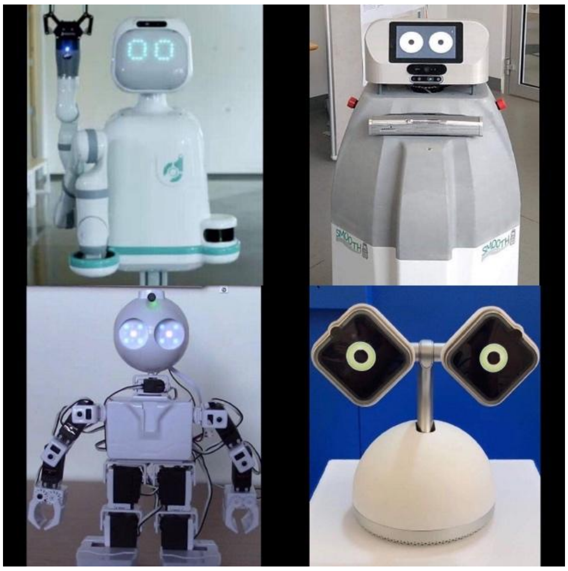


Fig. 1. The robots used

RESULTS

The questionnaire was sent to Prolific to gather data from 82 native speakers of English, 3 of whom did not finish the questionnaire, leaving 79 participants (47 female, 33male, 2 other).

The results reveal significant differences in the ratings of the voices concerning how convincing, charming, enthusiastic and passionate they are perceived. Ratings regarding charisma, boredom, persuasiveness and engagement were not significant. Interestingly, voices with low physical and melodic dominance received the highest ratings, with enthusiastic, charming, convincing and passionate reaching high levels of significance. Furthermore, the participants found the voices to fit particular robots to different degrees (see Table 2).

Table 2: Fitting voices (absolute frequencies): Relation robot x dominance condition.

voice	Smooth	Moxi	Haru	EZbot
NN	13	**33**	14	**19**
YN	**23**	16	**27**	13
NY	11	**29**	13	**26**
YY	**21**	9	**28**	21

How well a voice fits a robot seems to be strongly determined by acoustic features of melodic dominance than by acoustic features of physical dominance. The best fitting voices for Smooth and Haru are those with melodic dominance (YN,YY). The opposite applies to Moxi and, to some extent, also to the EZbot (NN, NY). These judgments are not correlated with robot size as Moxi and Smooth are bigger and Haru and EZbot smaller robots. However, what obviously coincides with the voice assignments is that, unlike Moxi and Ezbot, Smooth and Haru have no arms (which makes them look less approachable and inviting) and a somewhat strong, strict, and direct gaze. Future studies will show whether such visual attributes in fact determine vocal preferences.

DISCUSSION

Results regarding the perception of robot voices show that the features of physical and melodic dominance have significant effects on the ways the robots are rated in terms of how convincing, passionate, enthusiastic and charming robots are rated. Results regarding the relationship between robots and voices indicate that not all voices are suited for each robot. Particularly surprising is that, against expectation, a large robot should not necessarily have a voice that is high in physical dominance, whereas small robots may be associated with physical and melodic dominance. However, altogether, the voice without either physical or melodic dominance was perceived as the most charming, enthusiastic, passionate and convincing.

ACKNOWLEDGMENTS

The authors would like to thank the Danish Innovation Fund for their generous funding of the ‘Smooth’ project.

REFERENCES

[1] Boersma, P. (2001). Praat: A system for doing phonetics by computer. *Glot International 4*: 341-345.

[2] Crumpton, Joe & Bethel, Cindy (2016). A Survey of Using Vocal Prosody to Convey Emotion in Robot Speech. International *Journal of Social Robotics 8*: 271–285.

[3] Fischer, K. et al. (2020). Integrative Social Robotics Hands-On. *Interaction Studies* 21,1: 146-186.

[4] Liu, X & Xu, Y. (2014). Body size projection and its relation to emotional speech—Evidence from Mandarin Chinese. *Proc. 7th International Conference of Speech Prosody, Dublin, Ireland*, 974-977.

[5] McGinn, C. & Torre, I. (2019). Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. *Proceedings of HRI’19, Daegu*.

[6] Nass, C. & Brave, S. (2005). *Wired for Speech. How Voice Activates and Advances the Human-Computer Relationship*. The MIT Press, Cambridge, MA.

[7] Ohala, J. (1984). An ethological perspective on common cross-language utilization of F_0 of voice. *Phonetica 41*, 1-16.

[8] Riek, L. (2017). Healthcare robotics, *Communications of the ACM*, vol. 60, no. 11, pp. 68–78.

Speaker Transitions in English and Swedish Multiparty Casual Conversation

Emer Gilmartin[1*], Marcin Włodarczak[2], and Maria O'Reilly[1]

[1] *Trinity College, Dublin*
[2] *Stockholm University*
*gilmare@tcd.ie

Casual conversation proceeds as a series of contributions from participants, either speaking in the clear or in overlap. The pattern of who is speaking or not (the conversational floor state) changes constantly throughout a conversation. We examine the nature and frequency of these state changes or transitions in multiparty casual talk. We contrast between and within speaker transitions, analyzing the evolution of the conversational floor state from a stretch of single party speech in the clear to the next stretch of single party speech in the clear by the original or a different speaker. We investigate the effect of applying a minimum duration of single party speech in the clear to the incoming speaker's production, finding substantial differences in how transitions are categorized.

INTRODUCTION

Speech and silence timing data have been used to model several aspects of spoken interaction. Such research has generally followed two paths - large scale statistical modelling to build theory-neutral predictive models based on observance of the *presence* of speech and silence (chronemic analysis) [1], and minute examination of examples of interaction using Conversation Analysis. Chronemic analysis been used to generate predictions on several aspects of spoken interaction in a range of domains [2]–[5], while Conversation Analysis has given rise to observation and analysis of a wealth of conversational phenomena, generating insights into how speakers locally manage spoken interaction [6]–[9]. In this study, we thread a middle path between large scale prediction and explicative analysis in order to explore how multiparty casual conversation evolves around silence or overlap, and in particular to gain insight into patterns of speaker activity around these phenomena.

METHOD

We base our analysis on spoken interaction data manually segmented into interpausal units (IPUs), defined as a stretch of speech from a particular speaker bounded by silence from that speaker. We do this in order to avoid theory dependent concepts such as 'turns' or 'utterances'. We then define the `floor state' of a conversation at any time as the totality of participants speaking at the time, and represent interaction as a series of intervals of varying length where a particular floor state prevails. Relevant concepts discussed below are illustrated in Figure 1.

An n-party conversation, where each participant may be speaking or silent at any moment, has 2^n possible floor states, including global silence. In Figure 1, thirteen such states are shown. We use a shorthand of the sequence of the number of speakers in each floor state interval over longer stretches of conversation. The entire floor state sequence in Figure 1 would be represented as **A_AC_C_BC_ABC_AB_B_BC_C_GX_B_GX_A**, while the simpler numerical shorthand version is **1_2_1_2_3_2_1_2_1_0_1_0_1**. Note that this shorthand does not identify which speakers are involved in each interval, nor the length of the intervals.

We adopt the term **1Sp** to mean an interval of speech by a single speaker which is not overlapped by any other speaker. Note that this interval is a floor state interval and may not comprise a complete IPU. The type of transition between two **1Sp** intervals is also determined by the righthand context. To describe the transition types we adapt the terminology used in [10] for dyadic interaction, where a **1Sp** interval can transition back to the next **1Sp** interval with one intervening interval of silence or overlap, e.g. **1_0_1** or **1_2_1.** When the two **1Sp** intervals are produced by the same speaker, this results in within speaker silence or overlap, while production by different speakers result in between speaker silence or overlap.

Multi-party speech can transition back to 1Sp with one intervening interval of silence or overlap, as in the two-party condition, but the number of intervening intervals can increase once 3- or more party overlap occurs. Transitions from one substantial interval of single party speech to another, a situation loosely analagous to turn change or retention, can be operationalized by placing a lower bound on the first and final single speaker interval durations. We define **1Sp1** as such an interval of duration one second or more, and **1SpAny** as such an interval of any duration greater than zero.

For **1Sp1-1Sp1** transitions, the possibilities multiply, as the intervening intervals can include 1Sp segments shorter than the threshold duration, and thus sequences such as 1_2_1_0_1 are possible. For this work, we retain Heldner et al's notion of within and between speaker phenomena, but, as multiparty transitions can involve a combination of overlap and silence, we define only two transition types -- within speaker transitions (WST) where we examine transitions beginning and ending with the same speaker, and between speaker transitions (BST), which start with one single speaker and transition to another single speaker.

Below we investigate **Sp1-1Sp1** and **Sp1-1SpAny** transitions in two collections of spontaneous casual conversation – six 3-5 party conversations in English as described in [11], and eight 3-party conversations in Swedish described in [12].

The data were labelled for floor state using a custom Praat script [13], and **1Sp1** and **1SpAny** intervals were identified. We then extracted **Sp1-1Sp1** intervals. For each **1Sp1**, we searched forward to locate the next **1Sp1** and extracted the sequence of intervals (in terms of speaker numbers) from the initial **1Sp1** to the next **1Sp1**. As an example**, 1_2_3_2_1_0_1** contains 5 intervening intervals between the two stretches of **1Sp1**. A similar procedure was followed for **Sp1-1SpAny** intervals. The transitions were then labelled as Between (BST) or Within (WST) Speaker.

RESULTS

Distributions of **1Sp1--1Sp1** transitions for English and Swedish are shown in Figure , where it can be seen that the vast majority of intervening intervals are in stretches of odd numbers of intervals, with the number of cases dropping with increasing numbers of intervals. Overall, 95.1%) of all 8095 **1Sp1-** intervals are closed by a later **1Sp1** in fewer than 16 intervening intervals. Even-number cases accounted for only 2.1% of the 8095 **Sp1-1Sp1** transitions. This reflects the extreme scarcity of intervals where two or more speakers start or stop speaking at exactly the same time. The odd numbered cases and the 16+ interval bucket class were excluded from the **1Sp1-1Sp1** transition data, leaving 7542 transitions, comprising 73.7%.WST and 26.3% BST. The most frequent class of transitions are those with one intervening interval which account for 38% of cases overall. Figure 2 shows BST and WST distributions for English and Swedish, where it can be seen that one-interval transitions are most frequent overall, and particularly so in WST.

Distributions of **1Sp1-1SpAny** transitions for English and Swedish are shown in Figure 3, The vast bulk of **1Sp1-SpAny** transitions occur with one intervening interval (97.60% overall), while 3 and 5 intervals account for 1.37% and 0.16% overall respectively. Comparing **1Sp1-1SpAny** transitions sharing their left-hand **1Sp1**intervals with the 7542 **1Sp1-1Sp1** transitions results in the confusion matrix in Table 1. The transition type label changes for 24.5% of transitions overall depending on how the righthand interval is defined.

DISCUSSION & CONCLUSIONS

The results on **1Sp1--1Sp1** transitions show that WST are distributed more evenly over intervening intervals than BST, thus increasing the frequency of more complex transitions in BST. This could reflect more activity around turn change than around retention, or indeed more backchannels and acknowledgement tokens being contributed by more participants around speaker changes. One-interval transitions are the largest class, with a higher proportion of one-interval transitions in WST, perhaps reflecting breath pauses or single backchannels during monologic stretches. However, one-interval transitions only account for 38% of transitions overall, reflecting the need to consider more complex transitions around turn change and retention. It would be very interesting to separate within speaker breathing pauses from other transitions in order to better understand transitions around silence. Other future work involves further classification of transitions depending on the number of distinct speakers involved, and investigation of the duration of transitions. It is hoped that this study, and similar studies of other corpora, will allow us to inventory transition types in multiparty spoken interaction, and then analyse examples of the statistically more likely transitions in detail to better understand conversation dynamics.

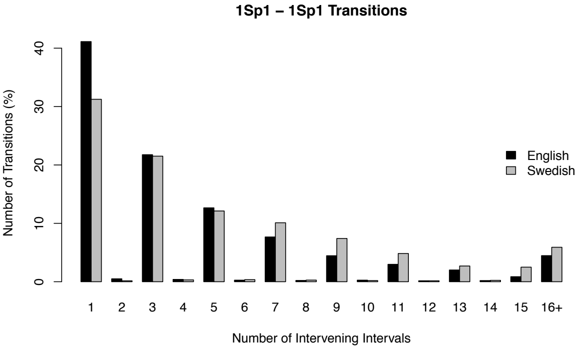


Fig. 1. 1Sp1-1Sp1 Transitions

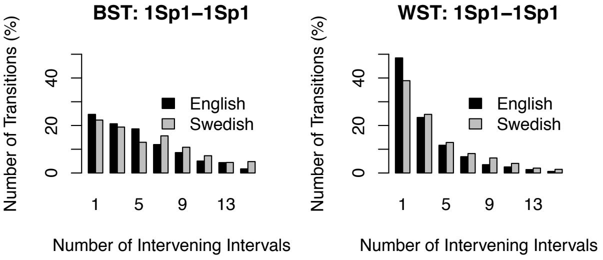


Fig. 2. Between and Within Speaker 1Sp1-1Sp1transitions

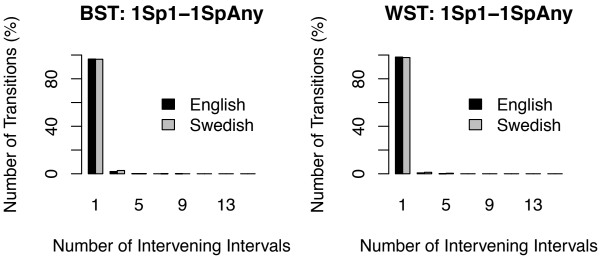


Fig. 3 Between and Within Speaker 1Sp1-1SpAny transitions

ACKNOWLEDGMENTS

This research was supported by Science Foundation Ireland (Grant 13/RC/2106) and the ADAPT Centre (www.adaptcentre.ie) at Trinity College, Dublin, and by a grant from the Swedish Research Council (2014-1072) to the project Breathing in Conversation at Stockholm University.

REFERENCES

[1] S. Feldstein and J. Welkowitz, “A chronography of conversation: In defense of an objective approach,” *Nonverbal Behav. Commun.*, pp. 329–378, 1978.

[2] B. Beebe, D. Alson, J. Jaffe, S. Feldstein, and C. Crown, “Vocal congruence in mother-infant play,” *J. Psycholinguist. Res.*, vol. 17, no. 3, pp. 245–259, 1988.

[3] J. Jaffe, S. Feldstein, and L. Cassotta, “Markovian models of dialogic time patterns.,” *Nature*, 1967.

[4] K. Laskowski, “Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation,” Carnegie Mellon University, 2011.

[5] J. D. Matarazzo and A. N. Wiens, “Interviewer Influence on Durations of Interviewee Silence.,” *J. Exp. Res. Personal.*, 1967.

[6] J. Heritage, “Conversation analysis at century’s end: practices of talk-in-interaction, their distributions, and their outcomes,” *Res. Lang. Soc. Interact.*, vol. 32, no. 1–2, pp. 69–76, 1999.

[7] G. Jefferson, “Preliminary notes on a possible metric which provides for a’standard maximum’silence of approximately one second in conversation.,” 1989.

[8] H. Sacks, E. A. Schegloff, and G. Jefferson, “A simplest systematics for the organization of turn-taking for conversation,” *Language*, pp. 696–735, 1974.

[9] E. A. Schegloff and H. Sacks, “Opening up closings,” *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

[10] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *J. Phon.*, vol. 38, no. 4, pp. 555–568, Oct. 2010.

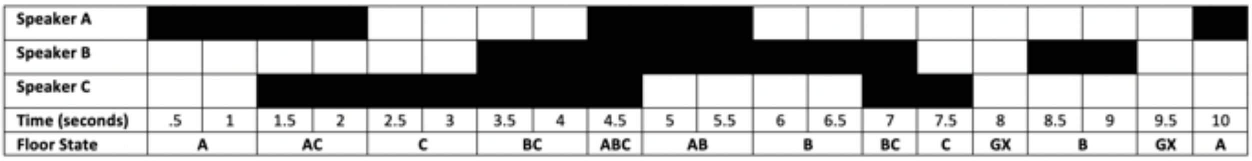
[11] E. Gilmartin and N. Campbell, “Capturing Chat: Annotation and Tools for Multiparty Casual Conversation.,” in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, 2016.

[12] M. Wlodarczak and M. Heldner, “Respiratory constraints in verbal and non-verbal communication,” *Front. Psychol.*, vol. 8, p. 708, 2017.

[13] P. Boersma and D. Weenink, *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*. 2010.

Tab 1: Confusion matrix for transitions depending on right hand interval duration threshold

		1Sp1-1Sp1		
		BST	WST	Total
1Sp1-1SpAny	BST	18.27	16.48	34.75
	WST	8.02	57.23	65.25
	Total	26.29	73.71	100



Speaker A	●	●	●	●					●	●	●									●
Speaker B							●	●	●	●	●	●	●	●			●	●		
Speaker C			●	●	●	●	●	●	●					●	●					
Time (seconds)	.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5	6	6.5	7	7.5	8	8.5	9	9.5	10
Floor State	A		AC		C		BC		ABC	AB		B		BC	C	GX	B		GX	A

Fig. 4 Three participants in a (fictitious) 10-second stretch of conversation, where black represents speech and white represents silence. Speaker A produces 3 inter-pausal units (IPUs) totalling 4 seconds (2+1.5+.5) of speech, Speaker B produces 2 IPUs totalling 5 seconds (4+1), and Speaker C speaks for a total of 4.5 seconds(3.5+1) across 2 IPUs. There is global silence for 1 second (.5+.5). There are a total of 13 floor state labels. The entire floor state sequence in Figure 1 would be represented as ***A_AC_C_BC_ABC_AB_B_BC_C_GX_B_GX_A****, while the simpler numerical shorthand version is* ***1_2_1_2_3_2_1_2_1_0_1_0_1****.*

Hesitations in the First (Japanese) and the Second (Russian as learned) Languages

Valeriya Prokaeva[1*]

[1]*Saint Petersburg State University, Dep.of Philology; Saint Petersburg, Russia*
*valeriya.prokaeva@yandex.ru

The following research aimed at the description of the characteristics of hesitation pauses in first and second languages produced by Japanese speakers who have been learning Russian. The study examined five adult native Japanese speakers learning Russian at levels B1-B2. The number of unfilled pauses in Japanese and Russian speech of the participants was counted and compared. Some extraordinary hesitation patterns in the second language speech production were examined.

INTRODUCTION

Even the most fluent speech contains frequent non-fluencies such as so-called non-juncture pauses, various repairs, repetitions, restarts, self-corrections and other suchlike elements that have been the subject of various psycholinguistic studies. The main functions of these elements, according to S. R. Rochester, are to «signal word choices» as well as to «reflect decisions at major constituent boundaries» [1]. Pauses representing the mentioned disfluencies, or hesitations, in the sustained spontaneous speech may either be silent (non-filled) and filled, implying a phonetic component: a sound, a sound lengthening or its qualitative sound change in a word.

Many previous studies have reported on the differences in hesitations between L1 and L2 learners, where the L2 speech being slower and more hesitant due to more complicated cognitive processes has more hesitation pauses. Moreover, such phenomena are reported to be phonetically and lexically different from hesitations in L1 speech as well [2, 3, 4].

The aim of the research was analyze by comparison the spoken language of the group of native Japanese students who had been learning Russian in order to find out in which ways the hesitation strategies differed in Japanese as L1 and Russian as L2. The correlation between the characteristics of pauses and the type of task was also tested. The hypotheses suggested that not only there would occur quantitative differences in hesitations, but also the features of filled pauses in L2 would be influenced by the characteristics of participants' native language, and the number of pauses during the static tasks performance in both languages would be higher than during dynamic tasks. The first suggestion proved to be true, whereas the second one seemed to prove itself only partly under the circumstances of participants performing tasks in their first language.

METHOD

The material chosen for the current research was 20 recordings of speech from five Japanese students aged 21-27 studying Russian (levels TORFL1/ TORFL 2) – 80 monologues and 20 dialogs with the interviewer, overall 259 minutes of recorded speech.

The study used the original experiment based on the method designed by N.V. Bogdanova-Beglaryan [5]. The experiment consisted of two interviews with the interval of 4 weeks between them. Each interview lasted for about 40 minutes. During the interview, I asked the participants to perform some tasks including:

- reading a text aloud

- answering interviewer's questions

- participation in a conversation with an interviewer

- describing a picture / making up a story based on the pictures (Fig. 1, 2).

Every task was presented in four versions in order to test the influence of different types of tasks on participants' speech performance: the interviewer asked the

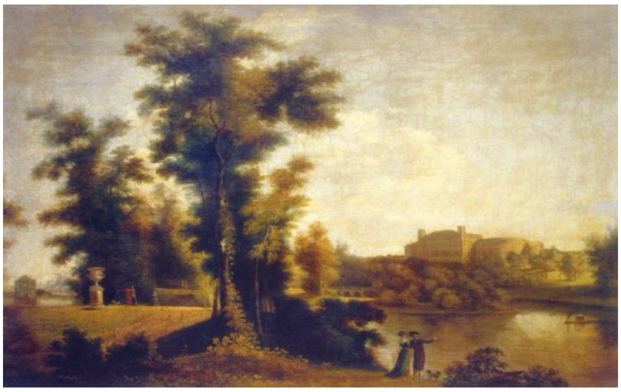


Fig. 1. Picture for a static story. S Shchedrin, View on the Gatchina Palace from Long island

participants to do the tasks sequentially in both Russian and Japanese and gave them either dynamic or static types of tasks on the different stages of the interview. The texts in Russian were adapted for Russian language learners using the text readability service with Flesh-Kincaid

Reading Ease readability score and some others.

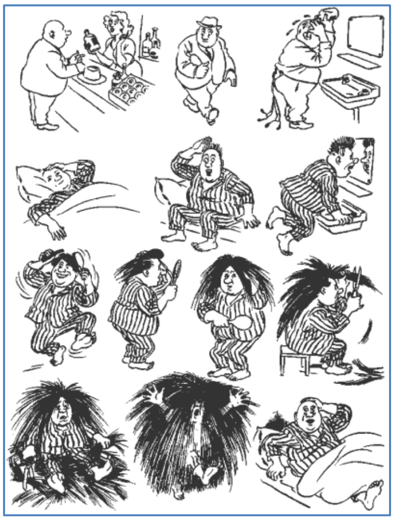


Fig. 2. Picture for a dynamic story. H. Bidstrup, Hair Remedy

The recordings were provided with orthographic annotations in PRAAT. The number of unfilled pauses was counted automatically, and the hesitation pauses were selected manually due to their high dependency on the context of the utterance. In this research I examined the two main types of pausal phenomena as following:

1) **filled pauses**

- pauses filled with speech-like sounds (*m-m; a-a*)
- vowel or consonant lengthening (*po-obezhal*)
- extralinguistic elements (laughter; sighs etc.)

2) **silent (unfilled) pauses**

- pauses disrupting the unity of a clause (200 ms or higher)
- pauses at clause boundaries (700 ms or higher)

The statistical significance of the silent pause difference in L1 and L2 was evaluated in JASP with paired T-student test.

RESULTS

Unfilled pauses

Table 1 shows the correlation of pauses with the language (df=9). The hypothesis about the surpassing number of unfilled pauses in L2 speech has proved to be true. The Participant 3 results might differ from others due to the participant’s Kansai dialect or individual strategies.

As can be seen from the table 2, the type of task does not affect the occurrence of pauses in L2 (df=9); the influence is present, however, in L1.

Tab. 1: the correlation of pauses with the language

P1	p=0.003
P2	p=0.042
P3	p=0.069
P4	p=0.009
P5	p<0.001

Tab. 2: the correlation of pauses with the type of task

L2		***L1***	
P1	*p=0.722*	*P1*	*p=0.346*
P2	*p=0.365*	*P2*	*p=0.655*
P3	*p=0.258*	*P3*	*p=0.087*
P4	*p=0.628*	*P4*	*p=0.025*
P5	*p=0.139*	*P5*	*p=0.058*

The number of pauses does not correlate with the type of text in L2 speech. This is assumed to be due to the task performed in L2 where supposedly both types of text are difficult to process.

The number of pauses in L2 is the closest to its amount in L1 among the participants who have been living in Russia for longest, and actively using L2 for everyday, educational and entertainment purposes.

Filled pauses

Some types of specific pauses in Russian language of the Japanese speakers (supposedly occurring as the result of the L1 influence):

1) **prolonged consonant vocalization**;

2) **dividing the words into syllables**;

3) **the break before the verb ending** (third person singular in the analyzed material).

Some specific **extralinguistic fillers**, such as *an interdental inhale, an inhale with a sound, lip smacking* and *tongue clicking* were described.

REFERENCES

[1] Rochester, S. R. (1973). The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research, 2(1)*, 64.

[2] Rose R. L. (1998). *The communicative value of filled pauses in spontaneous speech.* A Thesis submitted to the Faculty of Arts for the degree of Master of Arts in TEFL/TESL. The University of Birmingham.

[3] Grosjean F. (1980). Spoken word recognition processes and the gating paradigm. *Perception and Psychophysics*. 28(4), 267-283.

[4] Watanabe M., Rose R. (2013). Pausology and hesitation phenomena in second language acquisition. *The Routledge Encyclopedia of Second Language Acquisition*/ Ed. Peter Robinson. NY.: Taylor & Francis, 480-483.

[5] Bogdanova-Beglarian, N. V. (ed.). (2013). *Zvukovoi korpus kak material dlia analiza russkoi rechi [Speech corpus as a source of Russian speech analysis]. Vol. 1: Чтение. Пересказ. Описание [Reading. Retelling. Description].* St. Petersburg, Philol. Faculty Publ. (in Russian).

On Acoustic Features of Inhalation Noises in Read and Spontaneous Speech

Jürgen Trouvain, Bernd Möbius and Raphael Werner

Language Science and Technology, Saarland University, Saarbrücken, Germany
trouvain@lst.uni-saarland.de

Breath noises are probably the most common non-verbal vocalisations in spoken communication. They can occur in a multitude of contexts and can serve as functional markers in various ways. Our goal in this paper is to offer some acoustic-phonetic descriptions of inhalation noises by considering read speech and spontaneous dialogues.

INTRODUCTION

Breath noises as acoustic and audible reflections of inhalation and exhalation are probably the most common non-verbal vocalisations in spoken communication. Breath noises can occur in a multitude of contexts and they can serve as functional markers in various ways. In contrast to acoustic and audible correlates of phonemes there are hardly any acoustic descriptions of inhalation noises and other respiratory signals from a phonetic perspective. Thus, the aim of this paper is to suggest some acoustic descriptors of inhalation noises to fill this particular research gap. This will be done for two different speech modes: on the one hand for read and highly controlled speech and on the other hand for dialogical and spontaneous speech. We start with a review of different functions of respiratory noises in spoken communication.

Inhalation noises frequently occur in speech pauses. Here, breath noises function as markers of prosodic-syntactic boundaries, which has motivated the use of the term breath-groups for intonation (or prosodic) phrases [13]. Phonetic studies have shown that duration and intensity of inhalation noises can be indicators of utterance planning in speech production and inform listeners about the length of the upcoming phrase [6,7]. A recent study also suggests that in read speech duration and intensity of inhalation noises are due to a 'recovery' from the effort of the prior utterance [10]. Interestingly, when speakers are under physical stress they show different forms of breath noises in speech pauses, e.g. with many exhalation noises [18].

A typical non-verbal vocalisation in spontaneous speech is laughter of which various forms can be described with characteristic noises of ex- and inhalation [1,20]. A strong inhalation noise can mark the offset of a long and complex laugh [4,20]. Also in (other) affect bursts, breath noises can play a crucial role, such as startle or in crying [16].

On the level of pragmatics, breath noises can be used as discourse markers, signalling an intent to take the turn, and in some cultures respiratory noises are markers of politeness, e.g. in Korean [22]. Breath noises also have a high potential of signalling individuality, either by idiosyncratic acoustics, e.g. by inhalation noises with [s↓], an ingressive alveolar fricative [15], or by different patterns of inhalation and exhalation [11,12]. The incomplete list above shows that breath noises are a rather rich source of information on the linguistic but also on the non-linguistic level.

Surprisingly, breath noises are often and maybe systematically ignored in speech analysis, speech synthesis and speech recognition. This is reflected for instance by the fact that in speech fluency research, pauses that contain breath noises are regarded as 'silent', although they are not silent from an acoustic point of view [3]. In some conversational corpora the annotation schemes do not have a category for breath noises [17]. Likewise, speech prosodists regularly ignore breath noises as important acoustic cues of prosodic phrase boundaries.

Pauses in synthesised speech are often not modelled naturalistically [19] and they virtually never contain breath noises. However, breath noises are likely to be beneficial for speech synthesis that is pleasant and memorisable [21], and they are necessary for expressive speech synthesis. Breath noise in automatic speech recognition is still an under-researched topic, although there are various approaches for explicit breath detection, e.g. [8].

While there are research groups working on the physiological, particularly the kinematic, bases of respiration in speech, e.g. [2,7,21], the link between kinematic and acoustic signals of inhalation and exhalation in speech is not yet fully understood. The disctintion between in- and exhalation in this paper is based on auditory assessment of acoustic data which were recorded under laboratory conditions. Adverse acoustic conditions might be challenging for this task.

GENERAL OBSERVATIONS

In this paper, which can be considered as a preliminary study, we offer some acoustic descriptions of inhalation noises. For read speech we selected some news produced by professional news casters [5]. Here, all breath noises investigated were inhalation noises that used a combination of oral and nasal airstream. Nearly all pauses were marked with these inbreath noises.

A typical acoustic feature of an inbreath noise is that it is sandwiched between short intervals of silence. The edges to the left and right of the breath noises show an average duration of 50 ms whereas the breath noises themselves have a duration between 200 and 500 ms (see Fig. 1).

Inhalation noises in the read speech samples reveal a relatively low intensity and the values for centre of gravity (COG) are below 2 kHz. The formant values seem to have rather stable values.

It might be of interest to compare inhalation noises with other 'breath' sounds, i.e. unvoiced fricative segments with inhalation or exhalation as their primary sound source. For German, two types of segments can play a role here: aspiration phases of the closure release of unvoiced stops, and unvoiced variants of the glottal fricative /h/. Regarding realisations of /h/, a preceding voiced context, for instance a vowel or a sonorant, usually leads to a voiced instantiation of /h/, which is more similar to a glide. Unvoiced productions obviously require a voiceless left context, for instance an unvoiced obstruent or a silence. This has been shown to be a regular pattern in German [14, 24] that probably functions in a similar way in other Germanic languages.

Figure 1 depicts an example where these three kinds of respiratory noises occur in close vicinity. In contrast to inhalation noises, aspiration phases of unvoiced stops are much shorter and rarely exceed 60 ms. Their intensity is much higher than those of breath noises. The COG values are above 2 kHz. The formant values do show more variable values than those for breath noises. Cases of unvoiced variants of /h/ were rather infrequent and therefore not considered for this preliminary investigation.

The spontaneous speech comes from the Lindenstraße dialogue corpus [9]. It contains dyads of friends (same sex) with the task of talking about video clips of an episode of a soap opera. The interlocutors could not see each other and were recorded by separate channels.

For these spontaneous dialogues the pattern regarding the inhalation noises is by far more variable. Breath noises are only observed in phases of vocal activity, for instance when having the turn, giving a comment, or providing a feedback (or backchannel) utterance. This means that for a given speaker a significant portion of the recorded dialogue is marked by the absence of vocalisations, which should not be confused with regular pauses in speech.

In contrast to the read speech of the professional news readers, the breath noises in the dialogues are sometimes only nasal inhalation. In some cases the dialogues also show some exhalations. In addition, some inbreath noises are apparently produced with an [s]-like tongue position giving this fricative an additional sound source.

A substantial difference to read speech is that laughter may occur in spontaneous speech. Often, the laughter episodes are marked by strong inhalation noises. In Figure 2 we show an example of a typical offset of a 'voiced' or 'song-like' laugh with a long duration.

It is no surprise that inhalation noises often occur at the turn-initial position. However, a form with a higher intensity can be assumed to function as a turn-claiming cue, as exemplified in Figure 3.

Finally, it should be mentioned that some inhalation noises are 'enriched' with tongue clicks [3], a discourse-related pause-internal particle that also occurs in languages that do not have clicks as phonemes.

DISCUSSION AND CONCLUSION

Although this study has only an exploratory character, it suggests that inhalation noises differ acoustically from other segments in spoken communication. Appropriate acoustic parameters that establish this difference include intensity, COG, duration, and formants of which often the first four are visible in the spectrogram. A special feature of inbreath noises in pauses is that inbreath noises are accompanied at the edges of short silent sections that separate the frication section from the prior and the upcoming speech sequences.

There are manifold functions in which inhalation noises are involved. A typical inbreath noise that occurs in a pause that marks a syntactic-prosodic break usually has a rather different acoustic form than an inbreath noise that marks the offset of a longer laugh. It is of general interest in phonetics to learn more about how a given phonetic form reflects certain functions, and vice versa. However, for the time being it is unclear how complex or simple the relationship between the acoustic shapes of inhalation noises and their (presumed) functions really are.

Thus, the next step is to perform a detailed and systematic study of the proposed acoustic parameters of inhalation noises. This should entail various speech styles as the above sketched differences between read and spontaneous speech samples have shown. Ideally, such a study would also compare speech data across languages.

ACKNOWLEDGMENTS

This research was funded in part by the German Research Foundation (DFG) under grants TR 468/3-1 and MO 597/10-1.

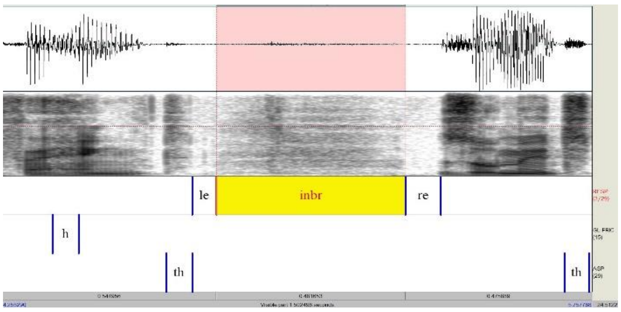


Fig. 1. Typical inbreath noise (spectrogram 0-8 kHz) with short silent edges on each side (top tier) in a 1.5-second section of read speech (... verhängt. Somit ...); realisation of /h/ as a glide (middle tier); aspiration of unvoiced alveolar stop (bottom tier).

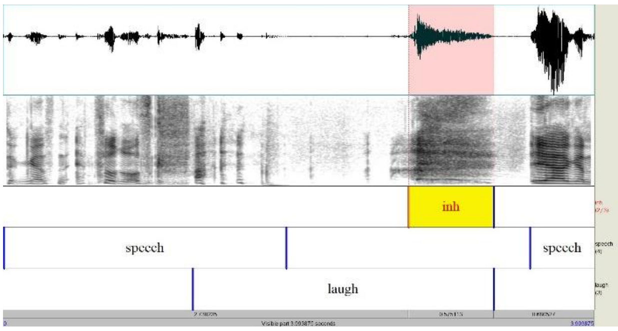


Fig. 2. Laugh with an inbreath noise (inh) as an offset (spectrogram 0-8 kHz).

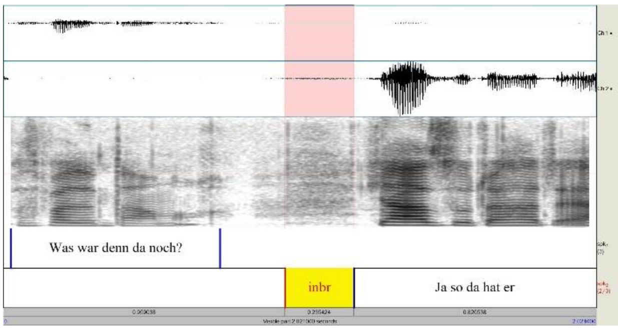


Fig. 3. 2-sec extract from a dialogue: speaker at top ends turn, speaker at bottom takes turn with turn-initial inhalation noise (spectrogram 0-8 kHz).

REFERENCES

[1] Bachorowski, J.A. & Owren, M.J. (2001). Not all laughs are alike: voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science* 12, pp. 252-257.

[2] Bailly, G., Rochet-Capellan, A. & Vilain, C. (2013). Adaptation of respiratory patterns in collaborative reading. Proc. *Interspeech*, Lyon. pp. 1653-1657.

[3] Belz, M. & Trouvain, J. (2019). Are 'silent' pauses always silent? Proc. *19th International Congress of Phonetic Sciences* (ICPhS), Melbourne, pp. 2744-2748.

[4] Chafe, W. (2007). *The Importance of Not Being Earnest*. Amsterdam: Benjamins.

[5] Eckart, K., Riester, A., Schweitzer, K. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In: Chiarcos, C., Nordhoff, S., Hellmann, S., (eds), *Linked Data in Linguistics*. Springer, pp. 65–75.

[6] Fuchs, S., Petrone, C., Krivokapic, J. & Hoole, Ph. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics* 41, pp. 9-47.

[7] Fuchs, S., Petrone, C., Rochet-Capellan, A., Reichel, U. & Koenig. L. (2015). Assessing respiratory contributions to f0 declination in German across varying speech tasks and respiratory demands. *Journal of Phonetics* 52, pp. 35-45.

[8] Fukuda, T., Ichikawa, O. & Nishimura, M. (2018). Detecting breathing sounds in realistic Japanese telephone conversations and its application to automatic speech recognition. *Speech Communication* 98, pp. 95-103.

[9] IPDS (2006). *Video Task Scenario: Lindenstraße – The Kiel Corpus of Spontaneous Speech*, Volume 4, DVD, Institut für Phonetik und Digitale Sprachsignalverarbeitung Universität Kiel.

[10] Kallay, J.E., Mayr, U. & Redford, M.A. (2019). Characterizing the coordination of speech production and breathing. Proc. *19th International Congress of Phonetic Sciences*, Melbourne, pp. 1412-1416.

[11] Kienast, M. & Glitza, F. (2003). Respiratory sounds as an idiosyncratic feature in speaker recognition. Proc. *15th International Congress of Phonetic Sciences*, Barcelona, pp. 1607-1610.

[12] Lauf, R. (2001). Aspekte der Sprechatmung: Zur Verteilung, Dauer und Struktur von Atemgeräuschen in abgelesenen Texten. In: Braun, A. (ed) *Beiträge zu Linguistik und Phonetik*. Stuttgart: Franz Steiner Verlag, pp. 406-420.

[13] Lieberman, Ph. (1967). *Intonation, Perception and Language*. Cambridge, Mass.: MIT Press.

[14] Möbius, B. (2004). Corpus-based investigations on the phonetics of consonant voicing. *Folia Linguistica* 38, pp. 5-26.

[15] Trouvain, J. 2010. Affektäußerungen in Sprachkorpora. Proc. *21st Konferenz Elektronische Sprachsignalverarbeitung*, Berlin, pp. 64-70.

[16] Trouvain, J. (2011). Zur Wahrnehmung von manipuliertem Weinen als Lachen. Proc. *22nd Konferenz Elektronische Sprachsignalverarbeitung*, Aachen, pp. 253-260.

[17] Trouvain, J. & Truong, K. (2012). Comparing non-verbal vocalisations in conversational speech corpora. Proc. *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals*, Istanbul, pp. 36-39.

[18] Trouvain, J. & Truong, K. (2015). Prosodic characteristics of read speech before and after treadmill running. Proc. *Interspeech*, Dresden, pp. 3700-3704.

[19] Trouvain, J. & Möbius, B. (2018). Zu Mustern der Pausengestaltung in natürlicher und synthetischer Lesesprache. Proc. *29th Konferenz Elektronische Sprachsignalverarbeitung*, Ulm, pp. 334-341.

[20] Truong, K.P., Trouvain, J. & Jansen, M.-P. (2019). Towards an annotation scheme for complex laughter in speech corpora. Proc. *Interspeech*, Graz, pp. 529-533.

[21] Whalen, D.H., Hoequist, Ch.E. & Sheffert, S. (1995). The effects of breath sounds on the perception of synthetic speech. *Journal of the Acoustical Society of America* 97, pp. 3147-315.

[22] Winter, B. & Grawunder, S. (2012). The phonetic profile of Korean formal and informal speech registers. *Journal of Phonetics* 40, pp. 808-815.

[23] Włodarczak, M. & Heldner, M. (2017). Respiratory constraints in verbal and non-verbal communication. *Frontiers in Psychology* 8, article id 708.

[24] Zimmerer, F. & Trouvain, J. 2015. Productions of /h/ in German: French vs. German speakers. Proc. *Interspeech*, Dresden, pp. 1922-1926.

Speech and breathing in different conditions of limb movements and over time

Hélène Serré[1], Marion Dohen[1], Susanne Fuchs[2], Silvain Gerber[1] and Amélie Rochet-Capellan[1]

[1] *Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France*
[2] *Leibniz-Centre General Linguistics (ZAS), 10117 Berlin, Germany*

Speech, respiration and limb movements seem interconnected in different ways. For example, speech production increases the duration of breathing cycles while physical effort shortens it. Physical activity may also improve creativity and semantic memory in relation with breathing. In this paper, we analyze changes in breath groups and breathing cycles during narration of short stories recalled with different motions of the limbs (hands free, hands blocked, hands biking, legs biking), and their progress over two days. The analyses suggest a significant increase of the breath group duration on day 2 as compared with day 1. This effect seems especially clear in the hand-blocked and hands-biking conditions. It is interpreted as an evidence of larger interferences between limbs movements and speech in these two conditions.

INTRODUCTION

Speech relies on specific adaptation of breathing control in close link with cognitive activity [1,2]. For instance, it is well known that during speech production, inhalation phases are shorter and exhalation phases longer than during quiet breathing [3]. The cognitive and physiological demands of the speech task affect the breathing profile [4,5]. At the same time, the respiratory constraint may shape speech planning [6,7]. But adaptation of breathing control is also involved in limb movements. For example, breathing frequency increases with physical effort [8]. It also seems that breathing-speech vs. breathing-limb coordination could be changed with training: theater actors [9] and athletes [10] might show different speech / breathing coordinative profiles. Previous works suggest that limb movement as well as learning could be relevant paradigms to further question the speech-breathing link. In line with this idea, we analyzed this link over time in different conditions of limb movements.

METHOD

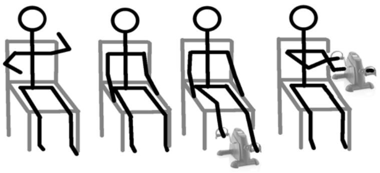


Figure 1: Experimental conditions, from left to right: Hands Free (HF), Hands Blocked (HBl), Legs Biking (LB), Hands Biking (HBi).

Eleven native speakers of German participated in this study. Their task was to watch short videos while sitting on a chair. They were then invited to retell the stories in different conditions: hands free (HF), hands blocked (HBl), hands biking on a mini-bike (HBi) vs. legs biking (LB) on the same mini-bike (Figure 1, left). Participants were recorded twice in the same conditions, on two different days. They also came back ten days later to retell the story in the HF condition only.

Speech and breathing were recorded synchronously using a microphone and the Inductance Plethysmograph system Respitrace®. Limb motion was also recorded using an Optitrack system but is not analyzed in the current paper. Interpausal units of speech (IPU) and breathing cycles were labeled automatically in Matlab and then checked and corrected when needed using Praat (Figure 2). The following measurements were considered to characterize the breathing cycles and the speech breath groups (in line with our previous work, cf. [11]):

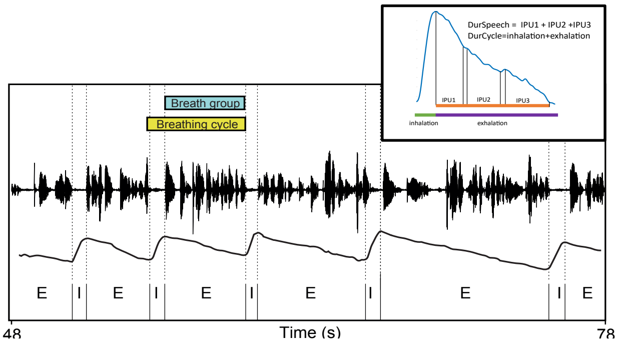


Figure 2: Identification of the breath group and the breathing cycle on the acoustic signal (top) and thorax trajectory (bottom)

- DurCycle: The duration of the breathing cycle (from the onset of the inhalation phase to the offset of the exhalation phase);

- SymCycle: The symmetry of the breathing cycle (duration of the inhalation phase divided by the total duration of the cycle);
- DurSpeech: The duration of the breath group (from the onset of the first interpausal unit to the offset of the last one produced on a same cycle);

We considered how these parameters changed according to the experimental condition and the day of recording.

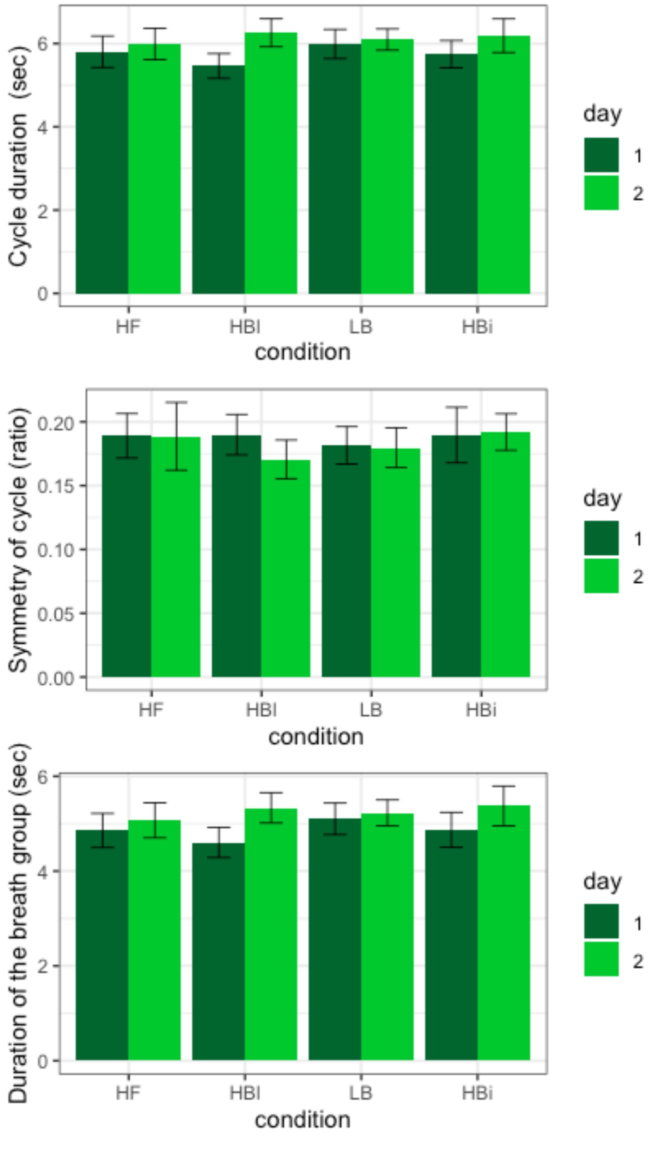


Figure 3: Measurements as a function of day (1 and 2) and experimental condition (HF, HBl, LB, HBi). Top: mean of the duration of the breathing cycles; Middle: mean of the *symmetry of the breathing cycles (inhalation duration/cycle duration); Bottom: mean of the duration of the breath groups*

RESULTS

The following linear mixed model (log(DurSpeech)~condition*day+(1|participant)) for DurSpeech measurements shows a significant effect of the day (z-value=2.271 p-value= 0.0232) but no effect of the condition. The amplitude of the effect was about +0.40 seconds from day 1 to day 2 in both cases. Neither the day, nor the condition significantly affected the asymmetry of the breathing cycle (Figure 3, middle).

Table 1: Means and standard deviations of DurCycle and DurSpeech as a function of day

Mean (± std)	**Cycle duration (sec)**	**Breath group duration (sec)**
day 1	5.75 (±1.10)	4.86 (±1.12)
day 2	6.13 (±1.12)	5.25 (±1.12)
Total	5.94 (±1.23)	5.05 (±1.13)

DISCUSSION

The increase in breathing cycle and breath group durations in these preliminary analyses might be related to a decrease of the cognitive load on day 2 as compared with day 1. Indeed, despite a non-significant effect of the condition, changes between day 1 and day 2 seem clearer for the unusual conditions of speech production (here, hands still and hands biking). These two conditions may induce a greater cognitive load than more usual conditions such as h ands free or biking with the legs. This could be evaluated by supplementary analyses of the linguistic content and the disfluencies. Further analyses of the correlation between the duration and amplitude of inhalation and the number of syllables per breath group should also help to understand to which extent these changes are related to changes in speech planning. Finally, analyses of the regularity of limb motion in the hand and leg biking conditions should help quantify the degree of familiarity to the task (e.g. in this case, smaller variability should be observed on day 2 with a different biking rate).

ACKNOWLEDGMENTS

This research is jointly supported by the French National Research Agency (ANR-17-FRAL- 0005-01) and the German Research Foundation (DFG) as part of the SALAMMBO project (Spoken Language in Motions: Learning and Adaptation of Speech Communication in the context of Body Motions)

REFERENCES

[1] Henderson, A., F. Goldman-Eisler, and A. Skarbek. "Temporal patterns of cognitive activity and breath control in speech." Language and Speech 8.4 (1965): 236-242.

[2] Grosjean, F., and M. Collins. "Breathing, pausing and reading." Phonetica 36.2 (1979): 98-114.

[3] Conrad, B., and P. Schönle. "Speech and respiration." Archiv für Psychiatrie und Nervenkrankheiten 226.4 (1979): 251-268.

[4] Wang, Y. T., Green, J. R., Nip, I. S., Kent, R. D., & Kent, J. F.. „Breath group analysis for reading and spontaneous speech in healthy adults." Folia Phoniatrica et Logopaedica, 62(6) (2010) : 297-302.

[5] Rochet-Capellan, A., and S. Fuchs. "Changes in breathing while listening to read speech: the effect of reader and speech mode." Frontiers in psychology 4 (2013): 906.

[6] Włodarczak, M., and M. Heldner. "Respiratory constraints in verbal and non-verbal communication." Frontiers in psychology 8 (2017): 708.

[7] Włodarczak, M., M. Heldner, and J. Edlund. "Communicative needs and respiratory constraints." Sixteenth Annual Conference of the International Speech Communication Association. 2015.

[8] Hoffmann, Charles P., and Benoît G. Bardy. "Dynamics of the locomotor–respiratory coupling at different frequencies." Experimental brain research 233.5 (2015): 1551-1561.

[9] Monteagudo, E., J. Sawyer, and A. Sivek-Eskra. "The effects of actors vocal exercises for relaxation on fluency: A preliminary study." Journal of fluency disorders 54 (2017): 50-57.

[10] McDermott, William J., R. EA Van Emmerik, and J. Hamill. "Running training and adaptive strategies of locomotor-respiratory coordination." European journal of applied physiology 89.5 (2003): 435-444.

[11] Rochet-Capellan, A., and S. Fuchs. "Take a breath and take the turn: how breathing meets turns in spontaneous dialogue." Philosophical Transactions of the Royal Society B: Biological Sciences 369.1658 (2014): 20130399.

Revisiting rhetorical claims of breathing for persuasive speech

Plinio Barbosa[1*], Oliver Niebuhr*[2], Jana Neitsch[2]

[1]Instituto de Estudos da Linguagem, Unicamp, Brazil
[2]Centre for Industrial Electronics, Mads Clausen Institute, University of Southern Denmark, Sonderborg, Denmark
**olni@sdu.dk*

Evidence of a production experiment is presented suggesting that, unlike often claimed in rhetorical manuals and coachings, it is actually the chest breathing rather than the abdominal "belly" breathing that supports the acoustic-prosodic parameter settings of a (more) charismatic/persuasive tone of voice.

INTRODUCTION

Persuasive speech is often associated with charisma and, due to that, it is also called charismatic speech. We will use these terms interchangeably in the following, however, focusing on the more effect-oriented term of persuasion. Linguistic research on charismatic speech revealed specific changes of acoustic-phonetic parameters relative to matter-of-fact (neutral) speech, both at the segmental level [1] and at the prosodic level (see [2] for a summary). The changes at the prosodic level include higher values of f0 mean, f0 range, and f0 maximum, as well as a higher intensity level, and a greater spectral emphasis [3,4,5,6,7].

The starting point of the present study is that virtually all rhetoric manuals and coachings dealing with a speaker's persuasive 'delivery' on stage stress directly or indirectly associate the acoustic changes above with appropriate breathing patterns. More specifically, these manuals and coachings recommend a specific type of breathing, i.e. abdominal breathing that is dominated by muscular activity of the diaphragm. This is therefore also called diaphragmatic breathing, or, in more popular rhetorical terms, "belly breathing". For example, [8:192] reminds her readers: "make sure you're breathing deeply into your belly". Similarly, [9:223] claim that "deep breathing - breathing from the diaphragm - give[s] the voice a better support [and] a stronger resonance" both of which are implicitly stated to be key features of the art of (persuasive) public speaking; [10:132] draws a direct connection between belly breathing and persuasive (charismatic) speech by stating that "the deepest kind of breathing, which works from the stomach rather than the upper part of the lungs [...] works wonders for the voice: it gives it depth and power, and makes for a more convincing delivery". The latter quote illustrates that some rhetorical manuals and coachings not only recommend belly breathing, they also explicitly discourage speakers from using "chest breathing" on stage, i.e. breathing dominated by the intercostal muscles.

There is, in fact, empirical evidence that abdominal breathing is beneficial for singers [11] and successfully used to treat voice and breathing disorders [12]. But, to the best of our knowledge, it has never been tested so far whether there is also a link between abdominal "belly" breathing and persuasive speech. At least one fact casts doubt on the existence of this link: Singing as well as many voice/breathing disorder treatments rely on maintaining a long phase of powerful exhalation. In contrast, for persuasive speech, prosodic phrases should be fairly short [3,4], with many pauses in between. Thus, if persuasive speakers split up their messages into small bites of a few seconds, why should they employ and benefit from abdominal breathing?

The present study scrutinizes the prevailing recommendation of rhetoric on breathing. If the rhetorical claim about the superiority of abdominal breathing over chest breathing for a speaker's charismatic delivery is true, then we expect a positive correlation between measured variables of abdominal breathing and prosodic changes towards more persuasive prosodic parameter settings.

METHOD

Participants were asked to present a text about 200 words. The text is a successful English investor pitch taken from the e-learning course on "How to write a killer elevator pitch" by Mike Simpson[1]. It was selected firstly for its well-designed verbal charisma-inducing strategies [13] and, secondly, because the pitched business idea, a mobile app for employee work-time tracking, is relatively neutral with respect to gender stereotypes.

The elevator pitch was given in two conditions by our speakers: (i) an emotionally neutral, matter-of-fact presentation with no special audience in mind, here called the neutral presentation; and (ii) an expressive, committed investor-pitch presentation that was supposed to be emotionally "contagious" and persuade an imagined jury of potential investors to invest money into the new app. Condition (ii) is therefore referred to as the persuasive presentation condition henceforth.

The two presentations were performed in L2 English by 18 native speakers of German, 9 men and 9 women. The speakers' mean age 25.5 years (min 22, max 37 years). All 18 speakers were fluent speakers of English at level

[1] https://theinterviewguys.com/write-elevator-pitch/

B2 or higher according to SDU-internal study entry tests. All 18 speakers had basic experience with entrepreneurial activities, including giving elevator pitches[2]. In this context, they had also received formal training in charismatic public speaking at the SDU over several hours.

Recordings were conducted in individual sessions of about 20 minutes. The speakers' investor-pitch presentations were recorded simultaneously with a microphone and the Resp-Track device [14], measuring time-aligned volume changes of abdomen and chest. We refrained from a cross-speaker order balancing of the neutral and persuasive presentation conditions as, according to our experience, a persuasive presentation has a stronger influence (e.g., in the form of a prosodic "afterimage") on a subsequent neutral presentation than vice-versa. Therefore, all speakers started with the neutral presentation condition and then moved on to the persuasive presentation condition.

RESULTS

The overall results pattern of acoustic parameters is simple: Besides obvious significant difference in parameter level due to speaker sex, the mean values of all acoustic parameters are significantly higher for persuasive presentations than for neutral presentations. The f0 maximum (in semitones, st, re 100 Hz) is on average at 100 st in persuasive presentations and at 96 st in neutral presentations; the mean f0 range covers 15 st in persuasive presentations and only 11 st in neutral presentations. The mean spectral emphasis is 2.8 dB in persuasive presentations and only 1.6 dB in neutral presentations (with all differences at least ($p < 0.001$). Note that the higher spectral emphasis level produced by speakers in the persuasive presentation condition coincides with a 2 dB higher global breathing amplitude that we found for the chest. The f0-peak rate, i.e. the number of pitch accents per time unit, also showed an increase from the neutral to the persuasive presentation condition. However, this increase was only a significant trend ($p<0.1$).

The breath-cycle variables explain, in all linear regression models, a significant amount of the variance in the acoustic variables. The highest explained variance associated with a single acoustic parameter is 11 %: The higher the inhalation amplitude of the chest, the lower is the subsequently produced f0 minimum. This holds true for both persuasive and neutral presentations and is slightly more pronounced for female than for male speakers. Other correlations between individual variables of acoustics and breathing are significant, but very weak in terms of explained acoustic variance, particularly those related to abdominal breathing. Regarding correlations between breath-cycle variables, we found the expiration duration to be significantly correlated with the amplitude and the duration of inhalation.

DISCUSSION

Besides sex-related differences in breathing and f0 that are all explainable in physiological terms [15], our results show with respect to the acoustic measurements an intra-individually consistent increase of all parameters from the neutral to the persuasive presentation. Thus, in view of the known correlations between acoustic parameter settings and perceived speaker persuasion, the acoustic data suggest that all speakers performed better (i.e. were more charismatic) in the persuasive than in the neutral presentation condition.

Given that, the major new contribution of the present study is that our male and female speakers enhanced their chest breathing rather than their abdominal breathing when holding the persuasive investor pitch presentations, men even more so than women. Women switched more strongly from abdominal to chest breathing in the persuasive condition, but men breathed longer and far deeper on the chest than the women did. So, at least on the basis of the patterns of acoustics and speech breathing, there are no supporting empirical indications that belly breathing -- the training of which often fills a considerable amount of pages and personal coaching time in rhetoric -- has a positive effect on a speaker's persuasion and charisma. In fact, rather the opposite seems to be true. The better acoustic performances of the speakers coincided with stronger chest-breathing activities. Moreover, the significant correlation between a higher chest-inhalation amplitude and a lower f0 minimum (a key change in persuasive speech, both in its own right and in connection with an extended f0 range [3]) may be seen as direct evidence for the positive effect of chest breathing on acoustic persuasion. Thus, we have to reject our hypothesis based on the present data.

However, to date we have measured only a small selection of acoustic parameters. Relevant f0 parameters such as kurtosis [16] were excluded here, as were intensity (i.e. loudness) measures and voice parameters based on the long-term average spectrum (LTAS) of a speaker [17]. It is therefore important for future studies that we measure other prosodic parameters and correlate them with the findings on chest and abdominal breathing. Furthermore, we need to relate the breathing data from the present study to perceptual ratings of listeners. If chest breathing has a persuasion-enhancing effect and abdominal breathing has no or a less persuasion-enhancing effect, then there will be a clear correlation between perceived speaker persuasion and the amplitude and/or standard deviation of chest breathing, but not (to the same extent) of abdominal breathing. We are conducting this perception experiment at the moment, and initial results point exactly in the direction outlined above. Abdominal breathing creates a more pleasant and sonorant, but chest breathing a more persuasive and charismatic voice.

[2] An elevator pitch "is a concise, carefully planned, and well-practiced description of your company that your mother should be able to understand in the time it would take to ride up an elevator"[2], Robert Pagliarini, MIT Blossoms: https://blossoms.mit.edu/sites/default/files/video/download/The-Art-of-the-Elevator-Pitch.pdf

ACKNOWLEDGMENTS

The research presented here was founded by a grant of the 2nd author from the Danish Agency for Science and Higher Education entitled "Exploring the interplay between speech and breathing across speaking styles and languages: tools, data and experiments" (7057-00101A)

REFERENCES

[1] Niebuhr, O. & Ochoa, S. G. (2019). Do sound segments contribute to sounding charismatic? evidence from a case study of Steve Jobs' and Mark Zuckerberg's vowel spaces. International Journal of Acoustics and Vibration, 2019.

[2] Niebuhr, O., Tegtmeier, S., & Schweisfurth, T. (2019). Female Speakers Benefit More Than Male Speakers From Prosodic Charisma Training—A Before-After Analysis of 12-Weeks and 4-h Courses. Front. Commun. 4:12. doi: 10.3389/fcomm.2019.00012, 2019.

[3] Niebuhr, O. Brem, A., &. Tegtmeier, S. (2017) Advancing research and practice in entrepreneurship through speech analysis – From descriptive rhetorical terms to phonetically informed acoustic charisma profiles. Journal of Speech Sciences 6, 3–26.

[4] Rosenberg, A. & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic speech. Proc. 9th European Conference on Speech Communication and Technology, Lisbon, 2005, 513–516.

[5] Touati, P. (1994). Prosodic aspects of political rhetoric. Working Papers 41, Dept. of Linguistics and Phonetics, Lund, Sweden, 168–171.

[6] Berger, S. Niebuhr, O. & Peters, B. (2017). Winning over an audience – A perception-based analysis of prosodic features of charismatic speech. Proc. 43rd Annual Conference of the German Acoustical Society, Kiel, Germany, 1454–1457.

[7] D'Errico,F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from voice. a cross-cultural study. Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction, Geneva, Switzerland, 552–557.

[8] Cabane, O. F. (2012). The Charisma Myth: How Anyone Can Master the Art and Science of Personal Magnetism. New York: Penguin.

[9] Carnegie D & Eisenwein, J. B. (2011). *The art of public speaking.* West Valley City: Waking Lion.

[10] Barker, A. (2016). *Improve your communication skills*. London: Replika Press, 2011.

[11] Salomoni, S. van den Hoorn, W. & Hodges, P. (2016). *Breathing and singing: objective characterization of breathing patterns in classical singers*," PLoS ONE, vol. 11, p. e0155084, 2016.

[12] Xu, J., Ikeda, Y. & Komiyama, S. *Bio-feedback and the yawning breath pattern in voice therapy: a clinical trial.* Auris Nasus Larynx, vol. 18, pp. 67–77, 1991.

[13] Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. Academy of Management Learning & Education, 10(3), 374-396.

[14] Cwiek, A., Neueder, S., & Wagner, P. (2016). *Investigating the communicative function of breathing and non-breathing 'silent' pauses*. PundP 12 – Phonetik und Phonologie im deutschsprachigen Raum. München.

[15] Titze, I. R. (1989). *Physiological and acoustic differences between male and female voices*. J. Acoust. Soc. Am., vol. 85, no. 4, pp. 1699–1707.

[16] Niebuhr, O. & Skarnitzl, R. (2019). Measuring a speaker's acoustic correlates of pitch-but which? A contrastive analysis for perceived speaker charisma. In *19th International Congress of Phonetic Sciences: Endangered Languages, and Major Language Varieties*. Australasian Speech Science and Technology Association Inc. 1774-1778.

[17] Niebuhr, O., Skarnitzl, R. Tylečková, L. (2018). The acoustic fingerprint of a charismatic voice - Initial evidence from correlations between long-term spectral features and listener ratings, Proc. 9th *International Conference on Speech Prosody* 2018, 359-363.

The Influence of Alcohol Intoxication on Silent Pause Duration in Spontaneous Speech

Hong Zhang[1*]

[1]*Department of Linguistics, University of Pennsylvania, Philadelphia, PA, USA*

*zhangho@sas.upenn.edu

This study looks at the effect of alcohol intoxication on the distribution of silent pause durations in spontaneous monologue and dialogues using speech from the German Alcohol Language Corpus. Distribution of silent pause durations for each speaker is represented as the joint density function of silent pause duration against the following speech duration. Density estimations are then projected onto lower dimensional space through Singular Value Decomposition (SVD). Group difference between sober and intoxicated speakers can be effectively visualized using the first 3 dimensions in the derived space. Good classification results are achieved using Support Vector Machines (SVM) with gaussian kernel in the binary intoxication classification task. These results suggest that alcohol intoxication has global effect on the distribution of silent pause duration relative to the following speech utterance durations. The derived features can provide a representation for tasks such as alcohol intoxication detection.

INTRODUCTION

Alcohol intoxication can cause deterioration in various aspects of cognitive processing, which may not only lead to problems in the motor control of speech production, but also result in deficits in speech planning [1]. Previous research has shown that speakers under the influence of alcohol intoxication tend to produce higher overall fundamental frequency (F0) [2], increased rate of disfluencies [3] and changed short-term energy function and F0 contour [2,4]. Practically speaking, successful detection based on altered speech signal caused by alcohol intoxication can be helpful in the prevention of alcohol related health issues, such as drunk and drive. To facilitate the development of systems that improve the efficiency of alcohol intoxication detection, Alcohol Language Corpus (ALC) [5] has been developed and used for a speaker state detection challenge [6]. In the challenge, a common set of acoustic features were used to train systems on utterance level classification with a baseline test accuracy (Unweighted Average Recall, UAR) of 65.9%. The best system [7] following the paradigm of this challenge achieved a UAR score of 71.4%. Here we ask the question of how the distribution of silent pause durations changes when the speaker is alcohol intoxicated.

In this study, we take the same ALC and ask if the distribution of pause durations changes at individual level when the speaker is intoxicated. As suggested in [2,4], the effect of alcohol intoxication on speech is highly speaker dependent, meaning that the same effect may surface in the opposite direction on the same acoustic measures for different individuals. This property of intoxicated speech may partly explain the relatively poor performance of utterance level classifiers, even if trained using state-of-the-art neural network architecture with rich acoustic representation [8]. Therefore we take a global perspective, with the goal of exploring the feature space that can efficiently represent the change induced by alcohol intoxication.

METHOD

The speech data

ALC is a collection of speech from a total of 162 German speakers (85 males, 77 females) produced in two conditions: sober and alcohol intoxication at a self-chosen intoxication level. The actual blood alcohol concentration (BAC) level was measured immediately before recording. Speech tasks used in the corpus include read speech, monologue (such as picture description, commands and instructions) and short conversations. Speech from the picture description task and short dialogues with the interviewer is chosen for the current study. The speech is recorded with a sample rate of 44.1 kHz with 16 bit rate. Verbatim transcriptions at phoneme level are available and the recordings are aligned.

Feature generation

The distribution of pause durations is represented as the joint density function of silent pause duration against the speech segment duration immediately following the silence. For each individual in each intoxication state, the joint distribution function is derived from all the selected speech in that condition. All silent pauses longer than 50ms are considered in the calculation. A 100x100 grid is used to sample from the 2D density function. Therefore the joint distribution is represented by a 100x100 matrix per speaker condition.

To reduce the sparsity of this representation and achieve a compact representation of the distributions, each 100x100 matrix is flattened as a 1x10,000 vector. SVD is then performed on the full 162x10,000 matrix stacked from all the individual feature vectors in each intoxication condition. The left singular matrix (dimension 162x162) is used as the final feature representation of all the speakers in each state, where each row corresponds to an individual in the given condition.

Feature evaluation

The derived feature vector is first evaluated by visualizing individual speakers in two conditions using the first three dimensions in the derived 162-dimensional feature space. A binary classification task is then performed using a simple SVM with the full feature vector to classify each individual as sober or intoxicated.

RESULTS

Fig 1 illustrates the difference in the joint distribution of silent pause duration and the following speech segment duration for a single speaker in intoxicated (left) and sober (right) conditions. A clear distinction between the two joint distributions can be observed. Silent pauses produced in intoxicated condition appear to be shorter, and the overall distribution is multi-modal compared to the sober condition.

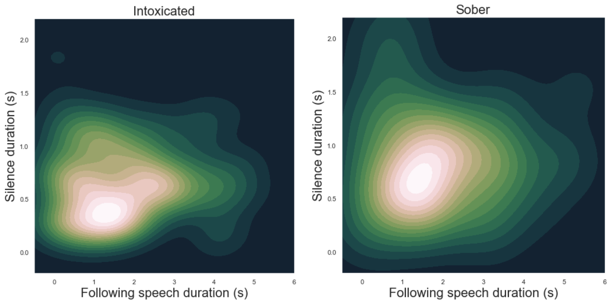


Fig. 1. 2D density plot of the joint distribution of silent pause duration (y-axis) against following speech segment duration (x-axis) for a single speaker in intoxicated (left) and sober (right) conditions.

The scattered plots for all speakers in sober and intoxicated conditions in the first three dimensions in the derived space are plotted in Fig 2. In the coordinate defined by the first and second dimension, intoxicated speakers are distributed mainly in the lower right corner, while in the coordinate defined by the second and third dimension, intoxicated speakers are mainly distributed to the left of the vertical line as shown in the figure. Thus the derived feature space is able to represent the group difference in the distribution of silent pause durations as measured by its relation with the following speech duration.

To test the performance of this derived feature space in distinguishing speakers in intoxicated from sober condition, speakers are randomly divided into training and testing set with a 3-to-1 ratio. The training set contains 122 speakers in both intoxicated and sober states, whereas the testing set includes the paired intoxicated and sober states for the rest of the speakers. Therefore the task can be understood as distinguishing between sober and intoxicated states when the speaker is given.

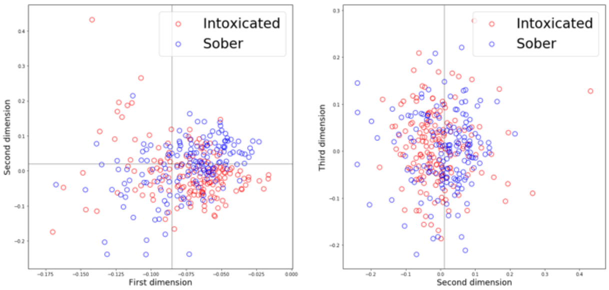


Fig. 2. Scattered plots of individual speakers in the derived space in intoxicated and sober conditions. The left plot shows the distribution along the first (x-axis) and second (y-axis) dimensions, and the right plot shows the distribution along the second (x-axis) and third (y-axis) dimensions.

A simple SVM classifier with gaussian kernel is used for this classification task. Tuning parameters are selected through a grid search with 10-fold cross-validation on the training set. Table 1 reports the simple testing accuracy for models trained on different percentages of the training data.

Results from this simple classification experiment suggest that the derived features are efficient in classifying speaker intoxication states between sober and intoxicated, as 20% of the training data can already achieve above-chance performance. Training on the full training set is able to yield a pretty high testing accuracy (93.75%) on the unseen test examples.

Tab. 1: Testing accuracy for SVMs training on different percentages of the training data

Training data	20%	40%	60%	80%	100%
Accuracy	0.65	0.7125	0.8875	0.7375	0.9375

DISCUSSION

In this study, we addressed the question of how alcohol intoxication affects the distribution of pause durations in spontaneous monologue and conversations at individual speaker level. Using speech produced from the picture description and short conversation tasks in ALC, we demonstrated that the effect of alcohol intoxication on silent pause duration can be effectively represented through the relation between silent pause duration and the following speech segment duration. Dimensionality reduction techniques, such as SVD, are able to offer compact parameterization of the differences observed from the joint distribution. The derived features appear to be highly efficient in intoxication identification.

The good performance of our feature space in the simple SVM setting also shows that although individual variation in particular acoustic dimensions can be problematic in deriving good representations of alcohol intoxicated speech, features derived from rich characterization of joint distributions of related variables can generate robust parameterizations for speaker state detection.

REFERENCES

[1] Peterson, J. B., Rothfleisch, J., Zelazo, P. D., & Pihl, R. O. (1990). Acute alcohol intoxication and cognitive functioning. *Journal of studies on alcohol*, *51*(2), 114-122.

[2] Baumeister, B., Heinrich, C., & Schiel, F. (2012). The influence of alcoholic intoxication on the fundamental frequency of female and male speakers. *The Journal of the Acoustical Society of America*, *132*(1), 442-451.

[3] Schiel, F., & Heinrich, C. (2015). Disfluencies in the speech of intoxicated speakers. *International Journal of Speech, Language & the Law*, *22*(1).

[4] Heinrich, C., & Schiel, F. (2014). The influence of alcoholic intoxication on the short-time energy function of speech. *The Journal of the Acoustical Society of America*, *135*(5), 2942-2951.

[5] Schiel, F., Heinrich, C., & Barfüsser, S. (2012). Alcohol language corpus: the first public corpus of alcoholized German speech. *Language resources and evaluation*, *46*(3), 503-521.

[6] Schuller, B., Steidl, S., Batliner, A., Schiel, F., & Krajewski, J. (2011). The INTERSPEECH 2011 speaker state challenge. In *12th Annual Conference of the International Speech Communication Association*. Florence, Italy, 2011, pp. 3201–3204.

[7] Bone, D., Li, M., Black, M. P., & Narayanan, S. S. (2014). Intoxicated speech detection: A fusion framework with speaker-normalized hierarchical functionals and GMM supervectors. *Computer speech & language*, *28*(2), 375-391.

[8] Berninger, K., Hoppe, J., & Milde, B., (2016) Classification of Speaker Intoxication Using a Bidirectional Recurrent Neural Network, in: P. Sojka, A. Horák, I. Kopeček, K. Pala (Eds.), *Text, Speech, and Dialogue,* Springer International Publishing, Brno, Czech Republic, pp. 435–442.

Functions of Pause in Italian Television News Broadcasts

Jessica Di Napoli[1*]

[1]*Institute of English, American and Romance Studies, RWTH Aachen University, Germany*
*jessica.dinapoli@ifaar.rwth-aachen.de

This study explores three types of pause occurring within and between news items in Italian regional news broadcasts: breath pauses, silent pauses, and filled pauses. Pause frequency, duration, and syntactic position are investigated. Findings show that, within news items, breath pauses are the longest and most frequently occurring type of pause. Breath pauses occur in clause-final, clause-medial, and phrase-medial position, and no effects of syntactic position on pause length are observed. This suggests that pause placement relates more to a physiological need to breathe than to the marking of syntactic structure. Between news items, pauses are much longer than within news items, and have both a structural and physiological role – to signal a new topic in the news broadcast (structural), and to give the news anchor time to breathe before resuming speech (physiological).

INTRODUCTION

The style of Italian news reading in both radio and television has changed dramatically over the last 50 years – namely, it is becoming faster [1,2,3]. However, this increase in pace is not necessarily attributable to news readers speaking more quickly. Rather, it results from a change in the form and function of *pauses* [3].

Recent studies have shown that, when comparing news reports from the 1950's and 1960's to news reports in the modern era (1990's and later), speech rates have increased, but articulation rates (which exclude pauses) remain relatively stable [1,2,3]. In fact, silent pauses in modern news reporting are shorter, less frequent, and are almost exclusively breath pauses (see also [4,5]). These breath pauses do not necessarily occur at major syntactic boundaries, such as those which would be marked with punctuation in a written text [1,3]. This suggests that pause placement, in particular for breath pauses, may be driven more by a physiological need to breathe [6], regardless of syntactic structure [3].

Though previous work has demonstrated a change in pace for Italian news over time, a comprehensive analysis of the types of pauses occurring in modern news reporting, and an account of their syntactic position in the clause, is lacking. Additionally, previous studies have focused on isolated news items, without considering pauses occurring between news items in a broadcast.

The present study fills this gap by undertaking a detailed analysis of pauses in the speech of four news anchors, with the goal of understanding the functions of pause in modern news reporting. Pauses are analyzed in terms of their form, frequency, and syntactic position.

METHOD

A corpus of news reporting consisting of speech from four news anchors (two male, two female, aged 37-47) in eight TGR Lazio regional television news broadcasts (two per speaker) was collected from the TGR online news archive. Each broadcast was segmented first into single speaker stretches, that is, contiguous stretches of speech from a single news anchor. This excluded speech from pre-recorded news reports from other contributors. The single speaker stretches were then broken down into news items, that is, stretches of speech from a news anchor on a single news item.

For each news item, we then labeled intervals of speech and silence and marked syllable nuclei in Praat [7], using a version of the script by de Jong & Wempe [8] modified for our purposes to allow for manual correction of annotation as it proceeds. Pauses (greater than 100 ms, in line with [4] and [5]) were manually labeled during this process and categorized first as filled or unfilled pauses [9]; unfilled pauses were then further specified as (1) containing inhalation noise, or (2) silent, possibly with a glottal closure (see [9] and [10]). This yielded three types of pause, respectively: filled, breath, and silent.

The number of pauses in each news item, as well as the duration of each pause was subsequently determined by script. Pause rate in syllables per pause was also calculated for each news item, by dividing the total number of syllables in a news item by the total number of pauses plus one, to account for the pause that always occurs at the start of a news item. The duration of pauses *between* news items was measured manually.

In addition to the type, number, and duration of pauses occurring within news items, we also investigated the syntactic context of each pause. Pauses were classified as occurring in one of four syntactic positions: (1) *clause final*, for a pause which occurs at the end of an independent clause; (2) *clause medial*, for a pause which occurs within a clause but at the end of a syntactic phrase corresponding to a major functional unit at the clause level (e.g. subject or object NP; verb VP); (3) *phrase medial*, for a pause which occurs inside a syntactic phrase, such as NP or PP; and (4) *word medial*, for a pause which occurs within a word.

RESULTS

A total of approximately 38 minutes of speech, across a total of 163 news items, underwent analysis for the type, duration, and syntactic position of pauses. We first present the results for pauses occurring within news items. In total, 285 pauses were identified, the majority of which were *unfilled* pauses, and in particular, breath pauses (see Tab. 1). Filled pauses were more infrequent in the corpus, but they did occur. In addition, a small number of complex pauses occurred, characterized by a combination of a filled and an unfilled pause. These pauses always consisted of an unfilled pause followed by a filled pause. Both breath and silent pauses occurred in combination with filled pauses in this way.

The mean duration of simple (i.e., non-complex) pauses in the corpus is 267 ms, while for complex pauses, the mean duration is 485 ms. There is a significant effect of pause type on duration for simple pauses (see Fig. 1): breath pauses are significantly longer than both filled and silent pauses ($\chi^2(2)=350.21$, $p<2e\text{-}16$).

On the whole, pauses were not particularly frequent, and many news items (40 percent) were produced without any pauses at all (see Fig. 2). This is likely linked to the fact that the news items tended to be relatively short (mean duration = 13.4 s). Across speakers, the mean pause rate is 38.22 syllables per pause, equivalent to approximately 19 words per pause.

The results for the syntactic position of pauses within news items are summarized in Table 1 according to type of pause. Breath pauses occur most frequently in clause-medial (but phrase-final) position (see Fig. 3), often between optional adverbial elements in a clause, or between adverbial elements and core clause elements (e.g., subject, verb, object). They also occur frequently between subjects and a following verb and between verbs and objects. Breath pauses also occur in phrase-medial position, most frequently within noun phrases, between a noun and pre- or post-modifiers in the form of adjectives, prepositional phrases, or restrictive relative clauses. The duration of breath pauses remains constant across the different syntactic contexts (see Fig. 4).

Silent pauses can also occur in a number of different syntactic contexts, in particular in clause-final, clause-medial, and phrase-medial position. They occur most frequently in phrase-medial position, in noun phrases between nouns and their pre- or post-modifiers. In one case, a silent pause also marks a word-internal disfluency.

Filled pauses pattern rather differently than the unfilled pauses discussed above. They occur almost exclusively in phrase-medial position, most frequently within noun phrases between a determiner and a noun. The filled pauses present in this corpus were all schwa-like vowel articulations indicative of hesitations, and they occurred primarily after consonant-final grammatical items, such as articles, prepositions, and conjunctions.

Tab. 1: Number of pauses in news items according to type and syntactic position.

Pause position	Pause type			
	Breath	Silent	Filled	Complex
clause final	44	8	0	0
clause medial	103	13	1	5
phrase medial	62	17	26	4
word medial	0	1	1	0
TOTAL	209	39	28	9

Between news items, there were a total of 32 breath pauses. Their average duration (1581 ms) was much longer than the breath pauses occurring within news items. Their composition also differed – while breath pauses within news items consist almost entirely of inhalation noise, breath pauses between news items consist of a long portion of silence, followed by inhalation noise in the final portion of the pause. Frequently, tongue clicks also occur, at the juncture between silence and the start of inhalation.

DISCUSSION

In this study, we see that, within news items, three main types of pauses occur (breath, silent, filled), and the majority of pauses (73 percent) are breath pauses. These pauses are also the longest (simple) pauses in the corpus. Breath pauses can occur in a number of different syntactic positions, ranging from phrase-medial to clause-final position. While some breath pauses coincide with major syntactic boundaries such as those which would have punctuation marks in written texts (such as clause-final position and clause-medial position between optional adverbial elements or in lists), many do not (such as clause-medially between a subject and a verb, or phrase-medially). The results suggest that breath pauses occur where they do primarily due to the news anchor's physiological need to breathe (see [6]). There is no evidence to suggest that news anchors plan breath pauses for major syntactic boundaries, in line with [3].

Between news items, breath pauses serve both a structural and physiological role. These long pauses signal the start of a new topic (often together with tongue clicks), and give the news anchor time for a deep breath before beginning the next news item.

The present study has shown that unfilled pauses play a structural role in news reporting, but not at the clause level. It is only at higher discourse levels that pauses serve to signal larger units. This study also provides further evidence that unfilled pauses are not always acoustically silent (see also [11]), and that a great deal of insight can be gained from considering the phonetic characteristics of these pauses.

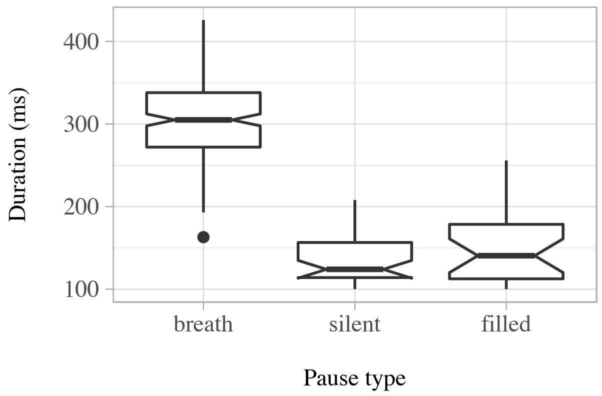


Fig. 1. Pause duration by type for pauses within news items. Breath pauses are significantly longer than both silent and filled pauses.

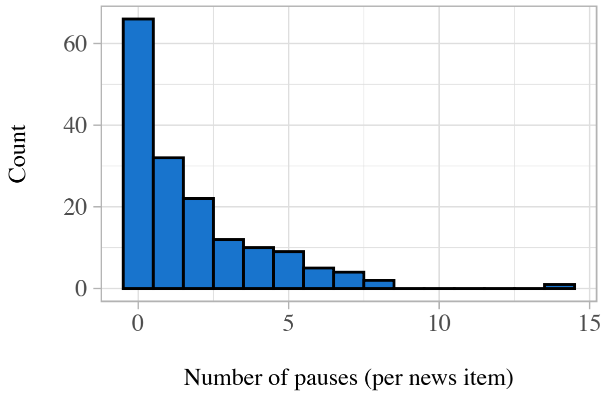


Fig. 2. Distribution of the number of pauses occurring within individual news items.

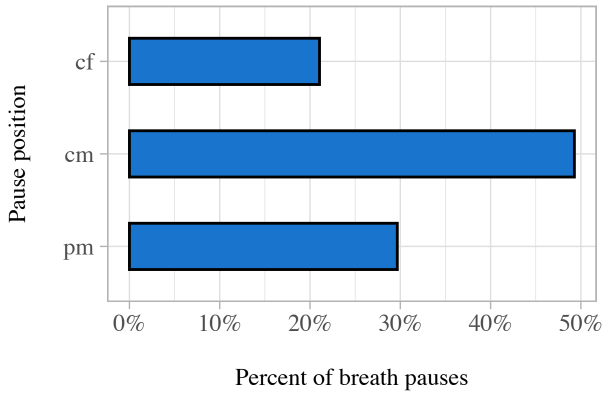


Fig. 3. Percentage of breath pauses occurring in clause-final (cf), clause-medial (cm), and phrase-medial (pm) position. Percentages calculated out of 209 total breath pauses.

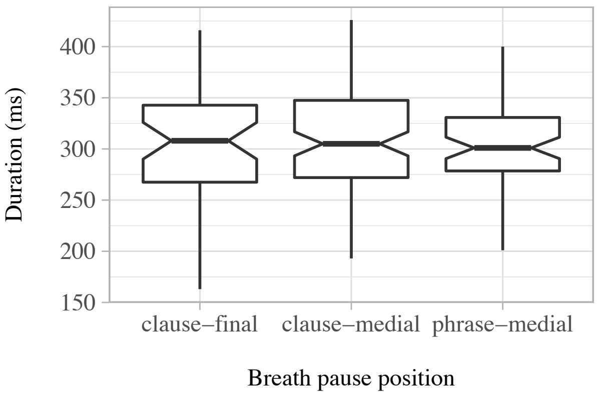


Fig. 4. Duration of breath pauses within news items according to their position in the clause.

REFERENCES

[1] Giannini, A. & Pettorino, M. (1999). I cambiamenti dell'italiano radiofonico negli ultimi 50 anni: aspetti ritmico-prosodici e segmentali. In: Delmonte, R. & Bristot, A. (Eds.), *Aspetti computazionali in fonetica, linguistica e didattica delle lingue: modelli e algoritmi* (pp. 65-81). Padova: Unipress.

[2] Giannini, A. (2005). Analisi acustica dell'italiano televisivo. In: Cosi, P. (Ed.), *Misura dei Parametri, aspetti tecnologici ed implicazioni nei modelli linguistici* (pp. 49-61). Padova: EDK.

[3] Pettorino, M. & Giannini, A. (2010). Il parlato dei mass media. In: Pettorino, M., Giannini, A., & Dovetto, F. M. (Eds.), *La Communicazione Parlata 3*, vol. 2 (pp. 71-83). Napoli: OPAR.

[4] Heinz, M. (2006). *Textsortenprosodie: Eine korpusgestützte Studie zu textsortenspezifischen prosodischen Mustern im Italianischen mit Ausblick auf das Franzözische.* Tübingen: Max Niemeyer Verlag.

[5] Rodero, E. (2012). A comparative analysis of speech rate and perception in radio bulletins. *Text & Talk*, 32(3), 391-411.

[6] Grosjean, F. & Collins, M. (1979). Breathing, pausing and reading. *Phonetica*, 36(2), 98-114.

[7] Boersma, P. & Weenink, D. (2011). *Praat: doing phonetics by computer*. Computer program.

[8] de Jong, N. H. & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2), 385-390.

[9] Magno Caldognetto, E., Zmarich, C., & Ferrero, F. (1997). A comparative acoustic study of spontaneous and read Italian speech. In: Kokkinakis, G., Fakotakis, N., & Dermatas, E. (Eds.), *Proc. of EUROSPEECH* (pp. 779-782). Rhodes: ISCA.

[10] Di Napoli, J. (2018). *The phonetics and phonology of glottalization in Italian*. Berlin: Peter Lang

[11] Trouvain, J., Fauth, C., & Möbius, B. (2016). Breath and non-breath pauses in fluent and disfluent phases of German and French L1 and L2 read speech. In: Barnes, J., Brugos, A., Shattuck-Hufnagel, S., & Veilleux, N. (Eds.), *Proc. of Speech Prosody* (SP8), Boston, 31-35.

Pragmatics of Pauses

Liudmila Savinitch

Institute for Information Transmission Problems
Russian Academy of Sciences
savinitch@iitp.ru

This article analyzes a pause in communicative aspect. Combined with other prosodic traits and some acoustic features, the pause may be used to convey the implicit meanings of the utterance. We consider pauses in the function of understanding the ironic meaning and pauses which contribute to convey implication. The examples are examined with the program Speech Analyzer and illustrated with graphs displaying pitch fluctuations, sound intensity, pauses, and other prosodic features.

INTRODUCTION

There are many works of linguistic literature devoted to the study of forms, types, and functions of pauses. There are logical and psychological pauses [1], hesitation pauses [2], intonational or syntactic pauses [3], filled and unfilled pauses [4], and breathing pauses [5] among others. As we see, even from a far incomplete list, it is obvious that the types of pauses are very diverse and carry out various functions. In this article we describe another, not previously noted, strategic use of pauses when conveying implicit meanings.

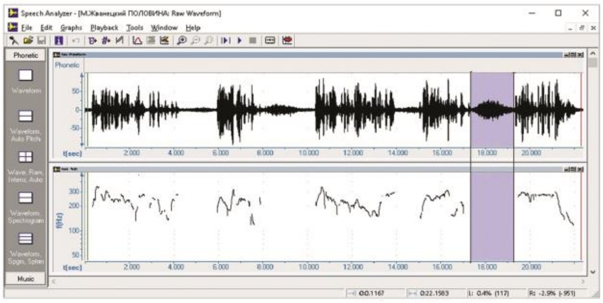
INTERPRETATION PAUSES

Our first example translated into English, was recorded from a performance of a famous satirist Mikhail Zhvanetsky, who ridicules particular unpleasant phenomena of Soviet life in the mid-twentieth century.

(1) *Half of the passengers make sure the other half takes their tickets*↱ *on the tram, and the first half does not take*↗ *any – this half is in the service*↘. *Thus*↱, *only half*↘ *of the population travels by ticket.*

In the example, arrows located after the accented wordforms indicate changes in fundamental frequency of the speech. In the first sentence, the irony is expressed by the collision of simultaneously activated meanings in the word '*service*' – lexical and contextual. The main lexical meaning of the word 'service' is "work, execution of any duties". In the situation described by the satirist, the word acquires a new, contextual, meaning: to voluntarily assume the duties of a public 'controller' (as if 'in the service') and therefore not to buy a ticket for the fare. Both meanings are in obvious contradiction with each other. Further, after the emphatic highlighting of the word *in the service* ↘, the author maintains a long pause, which lasts 22 seconds. On Tonogram 1, the pause is framed by two vertical lines. On the bottom panel of the segment that contains the pause, the tonogram graph is empty, indicating an empty pause (i.e. silence). Meanwhile, increasing and decreasing fluctuations are recorded for the pause section on the top panel's oscillogram, caused by background noise of audience laughter.

Tonogram 1

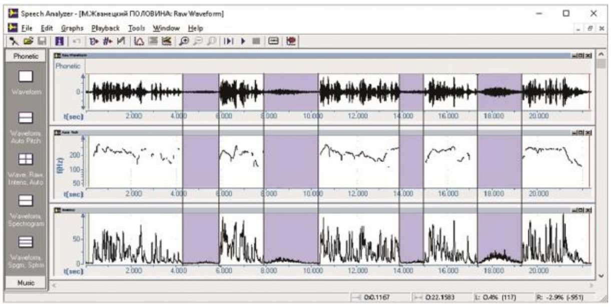


на службе
'in the servise'

That is, the planned pause is sustained by the satirist after the word that bears high ironic implicit content, which is often difficult to guess at once. The critical evaluation of the satirist, implicitly expressed in irony, is that the so-called 'generous moral impulse' of a half of the population actually reveals a striking truth: this half of the population itself travels without tickets and thus causes significant damage to the economy.

This pause strategy is consistently maintained by the satirist throughout the performance. On Tonogram 2, the framed segments mark the boundaries of the various pauses in the miniature recorded.

Tonogram 2



свистнули *'pinched'* *заняты* *'occupy'* *билеты* *'tickets'* *на службе* *'in the service'*

The acoustic correlate of a pause is drop in sound intensity to zero. In the segments of the bottom panel of Tonogram 2, the intensity graph falls close to zero. It is important to note that each pause appears after the words of salient lexical content explained in the article:

‘*pinched*’, ‘*occupy*’, ‘*tickets*’, ‘*in the service*’. These wordforms are also bearers of accents due to their status as communicatively significant units. Therefore, the satirist leaves long pauses after them to give the audience time to perceive the implicit meaning of the miniature. This type of pause can be called an interpretation (or understanding) pause.

‘EXPRESSIVENESS PAUSE’

Our next example is taken from a recorded work of art read by a professional actor. This passage describes events in Makuto, a state ruled by a dictator. The author ends the passage with a metaphorical conclusion, which ultimately serves to transfer an implicit meaning – a hint about the illegal measures used:

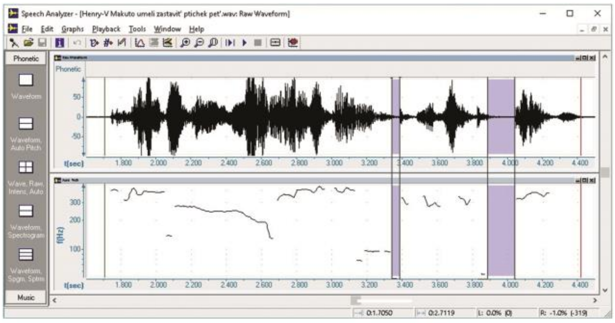
(2) *В Макуто*↦ *умели* ↘ *заставить птичек … петь.*
‘*In Makuto*↦*, <they> could* ↘ *force birdies … to sing.*’

While reading the sentence the actor increases expressiveness by maintaining an extended pause before the last wordform. The question arises: why does the professional actor intentionally sustain a long pause before the finishing word?

In linguistics, a pause is most often defined as a break in sound or cessation of phonation for a certain time. On Tonogram 3, the oscillogram of the upper panel and the tonogram of the lower panel show vertical lines that represent ‘empty’ segments corresponding to pauses.

In the second, larger segment, the pause before the final wordform *петь* ‘*to sing*’ lasts 140 milliseconds. This pause lasts almost three times longer the one in the previous segment between the wordforms *заставить птичек* ‘*force birdies*’, which lasts 50 milliseconds.

Tonogram 3



заставить птичек петь
‘force birdies to sing’

The pause before the last wordform does not act as an intonational (syntactic) pause, which divides speech into intonational–semantic units (phrases and syntagmata) and indicates semantic relationships between adjacent words by the presence or absence of breaks in certain areas of the speech flow. In our sentence, the verb group *force birdies to sing* is a linear sequence of verb phrase governed by the verb *to force*. A pause is undesirable within this word group, since the pause between words breaks or significantly weakens the connection between them. If such a pause does occur, it will appear not as a planned intonation pause, but as a pause of hesitation, reflecting the process of search and rearrangement during the generation of speech, i.e. an act of complex speech planning [2]. In our example, the type of pause is most closely related to the pragmatic aspect of language use. Following the preliminary long, unfilled pause, the waiting listener is struck by the sound of the final verb *петь* [p’et’] ‘*to sing*’, which incorporates intensive exhalation of air (aspiration) on the plosive consonants [p’] and [t’] and an increase in overall volume. Despite (or likely because of) its unconventionality, the actor’s speech strategy is effective because the pause in speech causes the next spoken word to become distinguished and strongly accentuated due to the impression of initiation or (re-)starting – *a beginning effect*.

Delivering this final sentence, the professional actor maintains an extended pause, articulates voiceless stops [p’] and [t’] with strong aspiration, intensity, and tension, as if imitating pronunciation or characterizing a certain peculiarity of voice (voices). If we turn to phonetics, the articulation characteristics described above can be attributed to two types of phonation, or methods of airflow modulation occurring in different configurations of vocal tract, known as “breathy voice” and “stiff voice” [6]. For the speaker, pronunciation becomes functionally significant and creates an impression of confidence, firmness, rigidity, and perseverance for the listener. This strengthens the meaning of the full passage and ultimately contributes to convey the implication by reflecting arrogance of authorities and their boundless power.

ACKNOWLEDGMENTS

The presented research was partially supported by Russian Foundation for Basic Research Grant № 18-07-00736.

REFERENCES

[1] Volkonsky, S. (1913). An expressive word. St. Petersburg Sirius, 166–180.

[2] Nikolaeva, T. M. (1970). A new direction in the study of spontaneous speech (on the so- called speech hesitations. *Questions of Linguistics*, 3, 117–123.

[3] Svetozarova, N. D. (1993). Accent-rhythmical innovations in Russian spontaneous speech. *Phonetics Problems*, 1, 189–199.

[4] Cruttenden, A. Intonation. (1997). New York: Cambridge University Press, 30.

[5] Krivnova, O. F. (2017). Phonetic characteristics of breathing pauses with different text localization // Computational linguistics and intellectual technologies / Papers of the annual international conference «Dialogue». Issue 16. Vol. 2. Moscow: Russian State University for the Humanities.

[6] Kodzasov, S. V. (2009). Researches in the field of Russian prosody. Moscow: Yazyki Slavianskih Kultur.

Author list

Program

Day 1 (SUN)		Day 2 (MON)		Day 3 (TUE)	
		8:30 – 9:00	**Opening Session**	8:30 – 9:30	**Keynote:** Jeremy Day-O'Connell *Voice, ear, and mind: The foundations of speech and song*
		9:00 – 10:00	**Keynote:** Jens Edlund *Breathing in interaction between humans and between humans, machinesand robots*	9:30 – 10:50	**Voice3**, Chair: Jeremy Day-O'Connell Mikkel Ploug & Oliver Niebuhr *There is music in speech melody*
		10:00 – 10:15	BREAK		Frederikke Dam Hansen, Lærke Hansen Siedentopp, Ågot Møller Grøntved & Trine Printz *Speaking Intensity Potential (SIP) – A new clinical measure for the voice disordered population*
		10:15 – 11:00	**Breathing1**, Chair: Jens Edlund		
			Marcin Włodarczak & Matthias Heldner *Breathing in conversation – What we've learned*		Frederico Cavalcante, Plinio A. Barbosa & Tommaso Raso *The prosodic features of the topic information unit in spontaneous speech from a crosslinguistic perspective*
			Matthias Heldner, Marcin Włodarczak, Peter Branderud & Johan Stark *The RespTrack system*		Jan Michalsky *Prosodic correlates of dominance and self-assurance. Acoustic cuesto testosterone related personality states of male speakers*
		11:00 – 11:20	BREAK	10:50 – 11:20	BREAK
		11:20 – 12:40	**Voice1**, Chair: Donna Erickson	11:20 – 12:40	**Voice4**, Chair: Oliver Niebuhr
			Kerstin Fischer & Nathalie Schümchen *Hesitation markers and audience design: Position matters*		Jana Neitsch, Oliver Niebuhr, Nicole Baumgartner & Andrea Kleene The perception of hate speech in German
			Charlotte Bellinghausen, Bernhard Schröder, Thomas Fangmeier, Andreas Riedel & Ludger Tebartz van Elst *Producing and perceiving prosody in Autism Spectrum Disorder*		Kerstin Fischer, Oliver Niebuhr, Rosalyn M. Langedijk & Selina Eisenberger *I shall know you by your voice – Melodic and physical dominance in the design of robot voices*
			Maria Di Maro, Jana Voße, Francesco Cutugno & Petra Wagner *Perception break down recoveryin computer-directed dialogues*		Emer Gilmartin, Marcin Włodarczak & Maria O'Reilly *Speaker transitions in English and Swedish multiparty casual conversation*
			Srikanth Nallanthighal *Deep learning methods for sensing breathing signal from from speech*		Valeriya Prokaeva *Hesitations in the first (Japanese)and the second (Russianas learned) languages*
		12:40 – 13:45	LUNCH	12:40 – 13:45	LUNCH
		13:45 – 14:45	**Keynote:** Donna Erickson *Voice: A multifaceted finely-tuned instrument for any occasionand culture*	13:45 – 14:45	**Keynote:** Plinio Barbosa *Stylistic and cross-linguistic differences in the prosodic organization of breathing, stressing, and pausing*
		14:45 – 15:00	BREAK	14:45 – 15:00	BREAK
15:00 – 17:30	**Registration**	15:00 – 15:40	**Voice2**, Chair: Kerstin Fischer	15:00 – 15:40	**Breathing2**, Chair: Oliver Niebuhr
			Cécile Fougeron, Angélina Bourbon & Véronique Delvaux *Age effects on voice and rate in French according to sex*		Jürgen Trouvain, Bernd Möbius & Raphael Werner *On acoustic features of inhalation noises in read and spontaneous speech*
			Alexsandro R. Meireles, Beatriz Raposo de Medeiros & João P. Cabral *Voice quality comparison between MPBsinging and speech*		Hélène Serré, Marion Dohen, Susanne Fuchs & Amélie Rochet-Capellan *Speech and breathing in different condition of limb movement and over time*
			Míša Hejná *Menstrual cycle effectson phonatory aspects of sustained vowels: It's about the onsetand the offset*		Oliver Niebuhr & Plinio Barbosa *Revisiting rhetorical claims of breathing for persuasive speech*
		15:40 – 15:50	BREAK	15:40 – 15:50	BREAK

Program

	Day 1 (SUN)		Day 2 (MON)		Day 3 (TUE)
	Registration	**15:50 – 17:10**	**Breathing & Pausing**, Chair: Plinio Barbosa	**15:50 – 17:10**	**Pausing**, Chair: Jens Edlund
			Christine Mooshammer, Oksana Rasskazova, Alina Zöllner & Susanne Fuchs *Effect of breathing on reaction time in a simple naming experiment: Evidence from a pilot experiment*		Hong Zhang *The influence of alcohol intoxication on silent pause duration in spontaneous speech*
			Heather Weston *A new type of analysis for assessing the temporal organization of breath, (non-breath) pause and speechintervals in spontaneous speech*		Jessica Di Napoli *Functions of pause in Itaian television news broadcasts*
			Charlotte Bellinghausen, Simon Betz, Katharina Zahner, Alina Sasdrich, Marin Schröer & Bernhard Schröder *Disfluenciesin German adult- and infant-directed speech*		Liudmila Savinitch *Pragmatics of pauses*
			Loulou Kosmala *Onthe multifunctionality and multimodality of silent pausesin native and non-native interactions*		
		17:10 – 17:30	BREAK	**17:10 – 17:30**	**Closing session**
17:30 – 18:30	**Lab tour**	**17:30 – 18:30**	**Poster session**, Chair: Jan Michalsky		
			Dorottya Gyarmathy & Valéria Krepsz *Temporal characteristics of silent pauses and breathing: The effects of speakers' age*		
			Míša Hejná *In support of the laryngeal articulator model? A casestudy of vowel height and glottalization in Czech*		
			Julie Kairet *Silent pause duration and distribution in Older-Women's Speech: A case study with 4 within-speaker comparison*		
			Maria Alm *Final lengthening in Danish: Conversational questions*		
			Jana Neitsch & Oliver Niebuhr *Questions as prosodic configurations: How prosody and context shape the multiparametric acoustic nature of rhetorical questions in German*		
			Nathalie Schümchen *Teaching strategic use of hesitation markers*		
			Stephanie Berger, Oliver Niebuhr & Margaret Zellers *Alternative phrase boundary symbolization and Its effectson pause duration*		
			Tirza Biron *Automatic segmentation of spontaneous speech and the "ideal pause"*		
			Oliver Niebuhr & Jana Neitsch *Raiseyour skills, then raise your voice: Comparing speech-melodyvisualization tools for acoustic leadershiptraining*		
			Kristina Diekjobst *Mellow the cello! –Determining correlations between human voicesand instrument voicesas a new sourcefor innovating cello strings*		
			Silke Tegtmeier, Tim Schweisfurth & Oliver Niebuhr *How voiceson stage affect investor funding. A pilot analysis of "TechCrunch"start-up presentations*		
			Barbara Heloha *Developmentof models for the automatic detection of prosodic boundaries in spontaneous speech*		
			Anna Grutnyk, Oliver Niebuhr & Wentao Gu *Public-speaking across languages and culture*		
19:00 – 21:00	**Welcome reception**	**19:00 – 21:00**	**Conference dinner**		

Legend

Opening & closing	Keynotes	Voice	Breathing	Breathing & Pausing
Breaks	Pausing	Poster session	Lab tour	Social events