TALLINN UNIVERSITY OF TECHNOLOGY

The 36th Finnic Phonetics Symposium (Fonetiikan Päivät)

April 25-26, 2024
Tallinn, Mektory Innovation and Business Centre

Hosted by the Laboratory of Language Technology

Supported by

REPUBLIC OF ESTONIA
MINISTRY OF EDUCATION AND RESEARCH

FP2024 programme

April 25, 2024

Registration starts 09:30

Opening 09:50

Session 1 ORAL Chair: Maija S. Peltola	10:00	Okko Räsänen, María Andrea Cruz Blandón, Khazar Khorrami, Daniil Kocharov	Modeling Child Language Development using Naturalistic Data at a Scale
	10:30	Juraj Šimko, Benjamin Elie, Alice Turk	A model of speech articulation based on optimal control theory
	11:00	Michael O'Dell	Heterokliinisen vekroston merkitys fonetiikassa

Coffee break 11:20-11:40

Session 2 ORAL Chair: Nele Ots	11:40	Martti Vainio, Antti Suni, Juraj Šimko, Sofoklis Kakouros	The Power of Prosody and Prosody of Power: An Acoustic Analysis of Finnish Parliamentary Speech
	12:00	Outi Tuomainen	Speakers in interaction: effortful speaking and listening in background noise across the lifespan
	12:20	Pärtel Lippus, Eva Liina Asu, Maarja-Liisa Pilvik, Liina Lindström	Alignment of prosodic prominence and gesture in marking of negation in Estonian: first insights from a multimodal study
	12:40	Eva Liina Asu, Pärtel Lippus, Heete Sahkai, Katrin Leppik	Which acoustic features contribute to the perception of surprise?

Buffet lunch 13:00-14:00

Session 3 ORAL Chair: Paavo Alku	14:00	Maija S. Peltola	Kielitaidon tukeminen tienä oppimisen yhdenvertaisuuteen ja osallisuuteen (KiTu) – uusi tutkimushanke pureutuu segregaatioilmiöihin kielen oppimisen kautta
	14:15	Peltola, K. U., Haapanen, K., Aerila, J-A., Gyekye, M., Kekki, N., Ruokonen, I., Thurin, N., Tyrer, M., Peltola, M. S.	KieliVertailu-työväline suomen kielen oppimisen tukena
	14:30	Heini Kallio	Englanninoppijoiden puheen prosodiset piirteet: lähtökielen vaikutusta selvittämässä
	14:45	Heini Kallio, Kamil Kaźmierski	Reduction of unstressed English vowels by L2 speakers with different language backgrounds
	15:00	Riikka Ullakonoja	Recognition of Russian accent in Finnish oral proficiency test
	15:15	Päivi Virkkunen, Minnaleena Toivola, Martti Vainio	Lukio-opiskelijoiden tunnekokemuksia ääntämisen opetuksessa saadusta palautteesta

Coffee break 15:30-16:00

Session 4 POSTER	16:00-17:00	Katja Haapanen, Antti Saloranta, Kimmo U. Peltola, Henna Tamminen, Maija S. Peltola	Ääntäminen ja transkriptio-kurssin vaikutus khoekhoegowabinkielisen puheen piirteiden havaitsemiseen logopedian ja fonetiikan yliopisto-opiskelijoilla
		Henna Tamminen, Katja Haapanen, Antti Saloranta, Kimmo U. Peltola, Lannie Uwu-khaeb, Maija S. Peltola	Sananalkuisten klusiilien sointi khoekhoegowabin puhujien Namibian englannissa
		Antti Saloranta, Katja Haapanen, Kimmo U. Peltola, Henna Tamminen, Meameno Shiweda, Napandulwe Shiweda, Maija S. Peltola	Namibianenglannin vokaalit auditiivisessa ja visuaalisessa tuottokokeessa oshiwambonpuhujilla
		Kalle Lahtinen, Liisa Mustanoja, Okko Räsänen	Building a Naturalistic and Representative Affective Speech Corpus for the Finnish Language
		Daniil Kocharov, Okko Räsänen	The effect of F0 measurements on prosody analysis in language development studies
		Khazar Khorrami, Okko Räsänen	Computational Investigation of the Feasibility of Statistical Learning for Early Word Comprehension using Realistic Input Statistics
		Liis Themas, Pärtel Lippus, Marika Padrik, Kairi Kreegipuu	Quantity perception among Estonian kindergarten children with developmental language disorder
		Pärtel Lippus, Liis Kask, Sofia Lutter, Nele Põldver, Kairi Kreegipuu	Further information on the perception of Estonian long-overlong quantity boundaries
		Pire Teras	The loss of word-internal laryngeal fricative in South Estonian Leivu dialect
		Tatiana Kachkovskaia, Michael O'Dell, Tommi Nieminen	Russian speakers' perception of stress in Finnish words
		Tatiana Kachkovskaia, Daniil Kocharov	Turn overlaps in collaborative dialogues and the factor of social distance
		Allan Vurma, Einar Meister, Lya Meister, Jaan Ross, Marju Raju, Veeda Kala, Tuuri Dede	Enhancing Plosive Recognition in Singing: The Impact of Elongated Plosive Closures in Varied Acoustics

Conference dinner	**19:00**	**KOHO restaurant**

April 26, 2024

Session 5 ORAL Chair: Martti Vainio	10:00	Paavo Alku, Manila Kodali, Sudarsana Reddy Kadiri	Machine learning-based prediction of SPL from healthy and pathological speech signals
	10:30	Mari Wiklund, Viljami Haakana, Ida-Lotta Myllylä, Martti Vainio	A Crosslinguistic Investigation of Prosodic Patterns Related to Autism Spectrum Disorder
	11:00	Alexandra Wikström, Lari Vainio, Martti Vainio	A cross-linguistic study of spatial sound symbolism
Coffee break	11:20 - 11:40		
Session 6 ORAL Chair: Mikko Kurimo	11:40	Tanel Alumäe	Language Identification is Difficult for Non-native Speech
	12:00	Katri Hiovain-Asikainen, Antti Suni, Sébastien Le Maguer	Neural Text-to-Speech for North Sámi: development and evaluation
	12:20	Tuukka Törö, Antti Suni, Juraj Šimko	Exploring the intersections of social dimensions and acoustics in Finnish text-to-speech synthesis
	12:40	Einari Vaaras, Manu Airaksinen, Okko Räsänen	IAR: Algoritmi epäkonsistenttien annotaatioiden siistimiseksi diskriminatiivisia luokittimia käyttäen
Buffet lunch	13:00-14:00		
Session 7 ORAL Chair: Tanel Alumäe	14:00	Yaroslav Getman, Nhan Phan, Ragheb Al-Ghezi, Ekaterina Voskoboinik, Tamás Grósz, Mikko Kurimo, Xinwei Cao, Giampiero Salvi, Torbjørn Svendsen, Sofia Strömbergsson, Anne Marte Haug Olstad, MInna Lehtonen, Anna Smolander, Sari Ylinen	Pronunciation Practicing App for Children Learning Nordic Languages
	14:15	Anton Malmi, Katrin Leppik	Training with the Estonian pronunciation app SayEst: vowel perception and user experience
	14:30	Sofoklis Kakouros	Enhancing Speech Emotion Recognition through Word Informativeness
	14:45	Xinyuan Wan, Lenka Kalvodová, Juraj Šimko	Perception and Production of Quantity in Finnish with Predictive Processing
	15:00	Marianne Kosin, Heini Kallio, Tiina Ihalainen, Nelly Penttilä	Prosodic Analysis of Speech Fluency in Finnish Speaking Elderly Women: Exploring Acoustic Correlations and Markers
Coffee break	15:15-15:45		
Session 8 ORAL Chair: Eva Liina Asu-Garcia	15:45	Natalia Kuznetsova, Oscar Cornelio Tiburcio, Hiroto Uchihara	Rare stress-induced lengthening in unstressed syllables in Finnic and Tlapanec: challenges for theory and typology
	16:00	Nele Ots	Utterance-initial intonation peaks in Estonian: A cognitive perspective
	16:15	Liis Ermus	Segmental an phrasal influences on the allophonic variation in short plosives in Estonian
	16:30	Kaidi Lõo, Pärtel Lippus, Benjamin V. Tucker	Part-of-speech and quantity interact in predicting acoustic durations of Estonian spontaneous speech

Closing **16:45**

Modeling Child Language Development using Naturalistic Data at a Scale

Okko Räsänen, María Andrea Cruz Blandón, Khazar Khorrami, Daniil Kocharov

Unit of Computing Sciences, Tampere University, Finland

In order to become proficient native language users, human children face several learning challenges. These include learning of the language's phonetic units, segmentation of words from running speech, association word forms with their meanings, and acquisition of the syntax. Typical empirical research on child language development (CLD) consists of well-controlled focused studies conducted in laboratory conditions. In contrast, only a few high-level theories, like NLM-e (Kuhl et al., 2008) and PRIMIR (Werker & Curtin, 2005), have aimed to integrate the present understanding of CLD into unified frameworks. However, these frameworks have not gained unanimous acceptance in the field, largely since they are relatively abstract, and only qualitatively describe the mechanisms and representations that could be involved in CLD. As a result, we still have limited understanding of the basic mechanisms and their interactions driving early language acquisition.

In principle, computational modelling of CLD is a potential solution to the so-called "integration problem" across empirical findings. This is because computational models can, and must, explicitly address all aspects of the information processing chain from input data to the resulting behavior. By formulating the theories as high-level computational goals and operations, implementing them as functional signal processing and machine learning algorithms, and finally exposing the models to realistic input data comparable to what real infants experience, ecological plausibility and validity of the underlying theories can be explicitly tested. However, the scientific impact of the existing modeling efforts has also been limited. Instead of addressing multiple aspects of language in one model, earlier models have usually focused on individual language phenomena (e.g., phonemic learning or word segmentation). Moreover, they have rarely investigated learning as a function of the developmental timeline of the learner.

We argue that two main factors currently hinder the development of more comprehensive, ecologically valid, and thus influential models of CLD: 1) lack of ecologically valid and openly available large-scale speech data to simulate child language experiences at a realistic scale, where various adult speech corpora are currently used instead, and 2) limited validation of the models with respect to empirical data on real infant language learning, where the current standard approach is to evaluate the models against linguistic theory of how speech is formally structured. Without fixing the data and evaluation problems, it is difficult to develop models that could truly help us to understand the big picture of CLD.

In this talk, we provide an overview of our ongoing "*Modeling Child Language Development using Naturalistic Data at a Scale*" (L-SCALE) project, where the aim is to enable development of more comprehensive and ecologically valid computational models of CLD. To achieve this, the L-SCALE project tackles the two challenges identified above: 1) solving the ecologically valid training data problem by creating a pipeline called *Generator of Infant Language ExperienceS* (GILES) for generation of ecologically relevant large-scale training data for computational models, and 2) solving the mismatch between human data and computational model evaluation by developing evaluation protocols that enable comparison of computational models against empirical data on CLD as a function of learner's age. We will also showcase some recent developments of the project, including a proof-of-concept demonstration of the GILES pipeline and a meta-analytic approach to compare models to empirical data on infant language learning.

The L-SCALE project is funded by Kone Foundation (2022–2026).

References

Kuhl, P.K., Conboy, B.T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded, (NLM-e). *Phil. Trans. Royal Society B*, 363, 979–1000.

Werker, J. F. & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, 1, 197–234.

A model of speech articulation based on optimal control theory

Juraj Šimko[1], Benjamin Elie[2], Alice Turk[2]

[1]University of Helsinki, Finland; *[2]The University of Edinburgh, UK*

Speech articulation is a complex process involving target-oriented temporal and spatial coordination of breathing patterns, laryngeal action and supralaryngeal articulation. Explaining the underlying principles of this coordination requires an account of harnessingthe vast number of degrees of freedom of the articulatory system. We present a development of such an account based on optimality control theory [1].

Our computational implementation of phonetic planning and articulatory action is built onthe assumption that speakers resolve the tradeoffs between the requirements of both the speaker and her listeners by adjusting their articulation to fit the situation and communication goals. The speaker and listener oriented aspects are conceptualized as mutually contradicting objectives of minimising articulatory effort and maximising intelligibility. The objectives are cast in a form of cost functions that are combined in the objective function driving the optimization process. In our model, articulatory effort is calculated as the sum of forces driving an adapted version of Maeda synthesizer using Tau guided movements as primitives [2]. The intelligibility of the produced output is approximated as a function of target phoneme recognition probability given a vector of articulatory synthesizer parameters.

We will present modelling results focusing on elicitation of hyper- and hypo-articulation patterns, namely vowel centralization and stop consonant lenition, by adjusting the weightsassigned to the component cost functions. In addition, we will introduce an extension of themodel incorporating the control of subglottal pressure used to simulate crucial aspects of Lombard speech production [3]. Finally, we will discuss the applicability of the model and itsplace in the comprehensive speech articulation planning platform, PlanArt [4].

References

[1] Elie, B., and Šimko, J., and Turk, A. (2023). Optimal control theory of speech productionusing probabilistic articulatory-acoustic models, *Proceedings of the 18th International Congress of Phonetic Sciences*, Prague, Czech Republic

[2] Lee, D. N. (1998) Guiding movement by coupling taus," *Ecological psychology*, vol. 10,no. 3-4, pp. 221–250.

[3] Elie, B., and Šimko, J., and Turk, A. (2024). Optimization-based modeling of Lombardspeech articulation: Supraglottal characteristics. *JASA Express Letters*, 4(1).

[4] Turk, A., Elie, B., and Šimko, J. (2023). PlanArt: A modular platform for computationalmodeling of articulatory planning, *Proceedings of the 18th International Congress of Phonetic Sciences*, Prague, Czech Republic

Heterokliinisen verkoston merkitys fonetiikassa

Michael O'Dell

Helsingin yliopisto

Dynaamisten systeemien teorialla on ollut merkittävä rooli foneettisten prosessien ymmärtämisen kannalta, esim. Task Dynamics, oskillaattorimallit, eksemplaarimallit. Parin vuosikymmenen ajan ns. heterokliinisia ver- kostoja ja 'voittajattoman kilpailemisen' (winnerless competition, WLC [1]) dynamiikkaa on tutkittu hyvin in- tensiivisesti dynaamisten systeemien teoriassa, mm. sen takia, että sellainen systeemi pystyy yhdistämään kaksi tärkeää, mutta näennäisesti ristiriitaista ominaisuutta, toistettavuus ja joustavuus (reproducibility and flexibility of transient behavior [7]). Toistaiseksi teoriaa on hyödynnetty fonetiikassa vain vähän (esim. [9]).

Päivillä esitän hyvin lyhyesti heterokliinisen verkoston teorian perusasioita, ja sitten joitakin teorian tuloksia, erityisesti sellaisia, jotka saattaisivat olla fonetiikassakin relevantteja [2]–[4], [8]. Kiinnitän erityistä huomiota hierarkkisiin heterokliinisiin verkostoihin. Hierarkkisesti organisoitu heterokliininen verkosto on ollut erityisen kiinnostuksen kohteena viime vuosina (esim. [5]), sillä monet käyttäytymisen lajit, puhe mukaan lukien, ovat luonnoltaan hierarkkisia. Hyvä tuore katsaus heterokliinisen verkoston teorian yleisistä käyttömahdollisuuk- sista löytyy Meyer-Ortmannsin artikkelista [6], jota käytän pohtiessani teorian yhteyksiä muihin dynaamisiin malleihin fonetiikassa sekä teorian mahdollista merkitystä fonetiikan kannalta yleisesti.

Viitteet

[1] V. S. Afraimovich, M. I. Rabinovich ja P. Varona, "Heteroclinic Contours in Neural Ensembles and the Winnerless Competition Principle," *International Journal of Bifurcation and Chaos*, vol. 14, nro 4, 2004. url: http://arxiv.org/abs/nlin/0304016.

[2] M. Aravind ja H. Meyer-Ortmanns, "On Relaxation Times of Heteroclinic Dynamics," *Chaos*, vol. 33, nro 10, 2023. DOI: 10.1063/5. 0166803.

[3] P. Ashwin ja C. Postlethwaite, "On Designing Heteroclinic Networks from Graphs," *Physica D: Nonlinear Phenomena*, vol. 265, s. 26–39, 2013. url: http://hdl.handle.net/10871/14534.

[4] Y. Bakhtin, "Noisy Heteroclinic Networks," *Probability Theory and Related Fields*, vol. 150, nro 1–2, s. 1–42, 2011. url: http://arxiv.org/pdf/0712.3952v3.pdf.

[5] J. Fonollosa, E. Neftci ja M. Rabinovich, "Learning of Chunking Sequences in Cognition and Behavior," *PLoS Computational Biology*, vol. 11, nro 11, 2015. DOI: 10.1371/journal.pcbi.1004592.

[6] H. Meyer-Ortmanns, "Heteroclinic Networks for Brain Dynamics," *Frontiers in Network Physiology*, vol. 3, nro 1276401, s. 1–16, 2023. DOI:10.3389/fnetp.2023.1276401.

[7] M. I. Rabinovich, R. Huerta, P. Varona ja V. S. Afraimovich, "Transient Cognitive Dynamics, Metastability, and Decision Making," *PLoS Computational Biology*, vol. 4, nro 5, e1000072, 2008.

[8] J. W. Reyn, "A Stability Criterion for Separatrix Polygons in the Phase Plane," teoksessa *International Conference on Nonlinear Oscillations, 8th, Prague, Czechoslovakia, September 11-15, 1978, Proceedings*, vol. 2, 1979, s. 595–600.

[9] I. B. Yildiz, K. von Kriegstein ja S. J. Kiebel, "From Birdsong to Human Speech Recognition: Bayesian Inference on a Hierarchy of Nonlinear Dynamical Systems," *PLoS Computational Biology*, vol. 9, nro 9, s. 1–16, 2013. DOI: 10.1371/journal.pcbi.1003219.

The Power of Prosody and Prosody of Power: An Acoustic Analysis of Finnish Parliamentary Speech

Martti Vainio, Antti Suni, Juraj Šimko, and Sofoklis Kakouros

University of Helsinki, Finland

We present a study where we delved into the complex role of prosody in Finnish parliamentary speeches. Our study represents a comprehensive analysis of a corpus of parliamentary speeches spanning from 2008 to 2020, focusing particularly on how prosodic features, especially the fundamental frequency (f_0), varied in relation to the speakers' political positions – whether they belonged to the government or the opposition – and across different election terms.

Our methodological approach is designed to ensure the reliability of signal-based features despite the challenges posed by varying recording conditions. This includes the diversity of microphones, recording environments, and devices used in the parliament. To achieve a comprehensive analysis, we employ both traditional signal-based prosodic features and representations from state-of-the-art self-supervised speech models to dissect the acoustic nuances of the speeches.

Our findings reveal a systematic variation in prosody that correlates with the speaker's political alignment and the timing within their election term. We observed a tendency for parliament members to exhibit higher fundamental frequencies either at the onset of their term or when serving in the opposition, possibly reflecting heightened urgency or emotional engagement. This pattern is particularly evident in our statistical analysis, which shows significant differences in the mean f0 values between government and opposition speakers, with these discrepancies being more pronounced among male speakers. Our findings also corroborate earlier evidence about the characteristic changes in age related mean f0.

We also critically discuss the challenges associated with large scale analyses using self-supervised learning for prosodic feature analysis. We highlight the need to carefully consider speaker-dependent characteristics and the risk of inflated classification results due to data inconsistencies. Our research contributes valuable insights into how prosody and political dynamics intersect, offering a new perspective on how speech is employed by politicians to persuade and influence.

Our work not only illuminates the acoustic properties of political speech but also underscores the potential of prosodic analysis in decoding social phenomena such as affective polarization within parliamentary discourse. We propose that future studies could further investigate the links between prosodic features and various dimensions of emotional arousal, and examine the influence of long-term trends and electoral cycles on speech characteristics.

References

Vainio, M., Suni, A., Šimko, J., Kakouros, S. (2023). The Power of Prosody and Prosody of Power: An Acoustic Analysis of Finnish Parliamentary Speech. arXiv preprint arXiv:2305.16040.

Speakers in interaction: effortful speaking and listening in background noise across the lifespan

Outi Tuomainen

Department of Linguistics, University of Potsdam, Germany

A great majority of previous work on how we produce and understand speech has focused on laboratory experiments where participants are asked to read or listen to carefully controlled listsof words and sentences. Our everyday speech interactions are rarely like this: we normally converse with other people, while doing something else at the same time, and often in challenging listening conditions (e.g., in background noise). In these real-life scenarios, speakerstypically aim to convey the message to the listener in the most economical way possible, and to do this they need to find an appropriate balance between articulatory effort they invest and communicative success (how much was understood by the listener). For example, speakers continuously assess the level of understanding of their interlocutor via the appropriateness of their responses (e.g., the frequency of requests for clarification, pauses, and hesitations). If thereis a breakdown in communication, the speaker may speak more clearly to increase the intelligibility of their speech. If communication is progressing well, speakers might start to reducethe effort they are making, and switch to a more casual speaking style. Listeners, in turn, need tomodulate their listening effort to optimize the communicative success (e.g., invest more effort inchallenging conditions). Therefore, we can say that everyday interactive speech is highly dynamicconsisting of ongoing fluctuations between utterances understood/misunderstood and interlocutors' adaptations to these fluctuations.

In this talk, I will present some results from our recent study where we investigated how speakers clarify their speech and how listeners modulate their listening effort in order to compensate forthe impact of background noise in everyday settings. We recorded 114 individuals (8-80 years) while they carried out an interactive problem-solving task (*diapix*) in quiet, background speech and background non-speech noise. We analyzed vocal effort (correlation between f0 and energyin the mid-frequency range), listening effort (subjective ratings), cognitive load (performance ona secondary task) and task success (how quickly they completed the task). I will discuss our findings in the context of communication effort framework proposed by Beechey, Buchholz & Keidser (2020).

References

Beechey, T., Buchholz, J. M., & Keidser, G. (2020). Hearing Impairment Increases Communication Effort During Conversations in Noise. *Journal of Speech, Language, and HearingResearch*, 63(1), 305–320. DOI: https://doi.org/10.1044/2019_jslhr-19-10 00201

Alignment of prosodic prominence and gesture in marking of negation in Estonian: first insights from a multimodal study

Pärtel Lippus, Eva Liina Asu, Maarja-Liisa Pilvik, Liina Lindström

Institute of Estonian and General Linguistics, University of Tartu

Multimodal studies of speech prosody have shown that in addition to 'beat' gestures thatare used to mark stress and accent, the gestures that primarily have different communicativefunctions are often aligned with prosodically prominent units in speech [1]. The aim of the study is to find out which co-speech gestures (hand and head beats) accompany negation in Estonian.

In Estonian, standard (clausal) negation is asymmetric [2]: verb form in negation is differentfrom the affirmative form: negation particle *ei* + different connegative verb stem is used (seeexample 1). The position of the negation particle is fixed and is always just before the verb. We expect negation particle to be deaccented and the following verb to receive an accent.

(1)	Present:	*Ma jookse-n*	*Ma ei jookse*
		I run-1SG	I not run.CNG
	Simple past:	*Ma jooks-i-n*	*Ma ei jooks-nud*
		I run-PST-1SG	I not run-PST.PTCL

Among hand gestures there are few that have been associated with negation in other languages: namely the Palm-Down-Horizontal-Across and Open Hand Prone gestures [3], [4]. This study investigates whether similar gestures occur in Estonian and how these gestures are aligned with prosodically prominent speech material. Additionally the influence of word orderof Estonian negation on the timing of co-speech gestures will be analysed.

The analysis uses data from the Phonetic Corpus of Estonian Spontaneous Speech [5] which contains 23 dialogues with video recordings (à 30 min, total 13 h) where the speakers were sitting in opposite corners of the recording booth and recorded with GoPro cameras. The speech was annotated in Praat [6] (words, sounds, syllables, morphological categories). The video was analysed with OpenPose [7] and the gestures were manually annotated in Elan [8].

References

[1] P. Wagner, Z. Malisz, and S. Kopp, 'Gesture and speech in interaction: An overview', *SpeechCommunication*, vol. 57, pp. 209–232, Feb. 2014, doi: 10.1016/j.specom.2013.09.008.

[2] M. Miestamo, *Standard Negation: The Negation of Declarative Verbal Main Clauses in a Typological Perspective*. Mouton de Gruyter, 2005. doi: 10.1515/9783110197631.

[3] S. Harrison, 'The organisation of kinesic ensembles associated with negation', *GEST*, vol. 14, no. 2, pp. 117–140, Dec. 2014, doi: 10.1075/gest.14.2.01har.

[4] S. Harrison and P. Larrivée, 'Morphosyntactic Correlates of Gestures: A Gesture Associated with Negation in French and Its Organisation with Speech', in *Negation and Polarity: Experimental Perspectives*, vol. 1, P. Larrivée and C. Lee, Eds., in Language, Cognition, and Mind, vol. 1. , Cham: Springer International Publishing, 2016, pp. 75–94. doi: 10.1007/978-3-319-17464-8_4.

[5] P. Lippus, K. Aare, A. Malmi, T. Tuisk, and P. Teras, 'Phonetic Corpus of Estonian Spontaneous Speech v1.3'. Institute of Estonian and General Linguistics, University of Tartu, Oct. 20, 2023. doi: 10.23673/RE-438.

[6] P. Boersma and D. Weenink, 'Praat: doing phonetics by computer'. Feb. 27, 2021. [Online]. Available: http://www.praat.org

[7] Z. Cao, G. H. Martinez, T. Simon, S. Wei, and Y. A. Sheikh, 'OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 2019.

[8] 'ELAN [Computer software]'. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, 2023. [Online]. Available: https://archive.mpi.nl/tla/elan

Which acoustic features contribute to the perception of surprise?

Eva Liina Asu[a], Pärtel Lippus[a], Heete Sahkai[b], Katrin Leppik[a]

[a] *Institute of Estonian and General Linguistics, University of Tartu*
[b] *Institute of the Estonian Language, Tallinn*

This study, building on an earlier pilot study [1], reports the results of a large online perception experiment comparing the perception of surprise and information-seeking questions in Estonian. The stimuli of the perception experiment were recorded as part of a production study [2] and included 24 interrogative sentences: 12 with the question word *mis* 'what' (e.g. *Mis loom see on?* 'What animal is this?') and 12 with the question word *mida*, partitive case of *mis* 'what' (e.g. *Mida sa teed?* 'What are you doing?'). For each of these sentences an information-seeking reading and a surprise reading was elicited with the help of a context resulting in 48 utterances per speaker. All utterances from 21 different speakers were used as stimuli in the online perception experiment.

The task of the participants was to evaluate whether the speaker intended to ask for information or to express surprise. The results indicate a higher correct identification score for information-seeking questions as compared to surprise questions, which is in line with a similar study for French [3], where the correct perception rate of string-identical questions was also significantly higher than that of surprise questions. At the same time, the current perception study reveals a large variability depending on the test items and speaker.

Acoustic features that contribute to a higher perception rate of surprise questions included a longer duration, a wider pitch range, and a lower mean pitch of the utterance. The strongest acoustic cue correlating with the perception of surprise was the longer duration of the utterance, which has also been shown to be the main phonetic correlate of French surprise questions [3].

References

[1] P. Lippus, E. L. Asu, K. Leppik, and H. Sahkai, 'The perception of surprise questions in Estonian', in *Proceedings of the 20th International Congress of Phonetic Sciences*, R. Skarnitzl and J. Volín, Eds., Prague: Guarant International, 2023, pp. 1533–1537.

[2] E. L. Asu, H. Sahkai, and P. Lippus, 'The prosody of surprise questions in Estonian', *J. Ling.*, pp. 1–21, Mar. 2023, doi: 10.1017/S0022226723000014.

[3] A.Celle and M. Pélissier, 'Surprise questions in spoken French', *Linguistics Vanguard*, vol. 8, no. s2, pp. 287–302, Jan. 2022, doi: 10.1515/lingvan-2020-0109.

Kielitaidon tukeminen tienä oppimisen yhdenvertaisuuteen ja osallisuuteen (KiTu) – uusi tutkimushanke pureutuu segregaatioilmiöihin kielen oppimisen kautta

Maija S. Peltola

Learning, Age & Bilingualism -laboratorio, Turun yliopisto

Turun yliopiston fonetiikalle ja sen yhteydessä toimivalle Learning, Age & Bilingualism -laboratoriolle (LAB- lab) myönnettiin Turun Kaupunkitutkimuksen rahoitus (150000 €). KiTu -hankkeen päämääränä on löytää tutkitun tiedon avulla keinoja tukea maahanmuuttajataustaisten lasten suomen ja englannin oppimista. Tarve tälle nousee kärjistyvästä tilanteesta, jossa tarvitaan lisää osaavaa työvoimaa, korkeakoulutukseen kaivataan tulevaisuuden osaajia ja koulutuksen ja osallisuuden yhdenvertaisuus ovat koetuksella. Ratkaisuja etsitään kielellisestä osaamisesta, jolla varmistetaan koulutukseen sekä työelämään pääsy, kiinnittyminen yhteisöön ja siten yhteiskunnallinen osallisuus.

KiTu-hankkeessa tutkimme erilaisten harjoitteiden toimivuutta eritaustaisilla oppijoilla ja nojaamme tutkimuksessamme kansainväliseen teoreettiseen viitekehykseen vieraan kielen oppimisen ennakoitavuudesta suhteessa äidinkielen ja kohdekielen äänteellisiin suhteisiin (mm. Flege 1987, Best & Strange 1992). Testaamme miten suomen, ja myöhemmässä vaiheessa englannin, puheen ymmärtäminen (ID-kokeet, reaktioaikamittaukset) ja tuottaminen (yksittäisten äänteiden akustinen analyysi, pidempien puhunnosten raatiarviot) kehittyvät laboratorioharjoittein ja etsimme joukosta tehokkaita ja kouluympäristöön soveltuvia harjoitteita. Taustana toimivat LAB-labin aiemmat tutkimukset, joissa erilaistenharjoitteiden on todettu tukevat oppijoita moninaisin tavoin (Savo et al. 2019, Peltola KU et al. 2015, Tamminen et al. accepted, Saloranta et al. 2020, Immonen et al. 2022). Tieteellisten tutkimustulosten pohjalta tuotamme peruskoulujen opettajille ja oppilaille suunnatun nettisivuston, jossa on sekä taustat huomioivia harjoitteita, että tukea opettajille. Hankkeessa toimii monitieteinen ja -alainen tiimi: foneetikot tarjoavat akustiikan, artikulaation ja puheen neuraalis-behavioraalisen tutkimuksen asiantuntemusta, lisäksi mukana on kokenut kielten ja pedagogiikan asiantuntija, erityispedagogi sekä puheterapeutti ja nettisivustonlaatija. Hanke käynnistyy keväällä 2024 ja päättyy vuoden 2025 lopussa.

Esittelen Fonetiikan päivillä tutkimushankkeen tavoitteita, toteutustapoja sekä sovellusmahdollisuuksia, toiveena on myös löytää uusia yhteistyökumppaneita hankkeen tueksi.

Viitteet

Best, C. T. & Strange, W. (1992) Effects of phonological and phonetic factors on cross-language perception of approximants, *Journal of Phonetics*,20, 305–330.

Flege, J. E. (1987) The production of “new” and “similar” phones in a foreign language: evidence for the effect of equivalence classification, *Journal of Phonetics*, 15, 47–65.

Immonen, Katja, Alku, Paavo, Peltola, Maija S. (2022) Phonetic listen-and-repeat training alters 6–7-year-old children’s non-native vowel contrast production after one training session, *Journal of Second Language Pronunciation*, 95 - 115.

Peltola, Kimmo U., Tamminen, Henna, Alku, Paavo, Peltola, Maija S. (2015) Non-native production training with an acoustic model and orthographic or transcription cues. *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, UK, Paper number 0236.

Saloranta, Antti, Alku, Paavo, Peltola, Maija S. (2020) Listen-and-repeat training improves perception of second language vowel duration: Evidence from mismatch negativity (MMN) and N1 responses and behavioral discrimination, *International Journal of Psychophysiology*, 72-82.

Savo, Satu, Peltola, Maija S. (2019) Arabic-speakers Learning Finnish Vowels: Short-term phonetic Training Supports Second Language VowelProduction, *Journal of Language Teaching and Research*, 10(1), 45-50.

Tamminen, Henna, Kujala, Teija, Peltola, Maija S., Training non-native speech sounds results in permanent plastic changes – Hard-wiring new memory traces takes time, *Lingua*, (accepted).

KieliVertailu-työväline suomen kielen oppimisen tukena

Peltola, K. U., Haapanen, K., Aerila, J-A., Gyekye, M., Kekki, N., Ruokonen, I., Thurin, N., Tyrer, M. ja Peltola, M. S.

Äidinkielen äännejärjestelmällä on kiistatta keskeinen merkitys vieraan kielen äänteiden oppimisessa. Kohdekielen äänteiden suhde ensikielen kategorioihin vaikuttaa oleellisesti oppimisvaikeuksien vakavuuteen ja eroavaisuuksien pohjalta voidaan tehdä selkeitä ennusteita siitä, mitkä äänteet ovat vaikeimpia oppia (Flege 1987, Best & Strange 1992). Lisäksi eksplisiittinen tieto opittavan kielen fonetiikasta (Lintunen 2004) tai ääntämisen yksityiskohdista (Saloranta *et al*. 2015) voi nopeuttaa äänteiden tuottamisen oppimista merkittävästi. Monenlaiset harjoitteet toimivat myös oppimisen tukena siten, että kuuntele ja toista -harjoittelu muovaa oppimista tehokkaasti niin lapsilla (Immonen et al. 2022), aikuisilla (Peltola KU *et al*. 2017) kuin ikääntyneilläkin (Jähi *et al*. 2015). Näiden oppimiseen vaikuttavien tekijöiden yhdistelmänä suunniteltiin moniammatillisessa yhteistyössä harjoituspaketista (Aerila *et al*. 2022a) ja oppaasta (Aerila *et al*. 2022b) koostuva KieliVertailu (KiVe) –työväline, joita käytetään tällä hetkellä monikielisissä päiväkodeissa ympäri Suomea. Näiden pohjalta luotiin myös lasten kirja (Rönns 2022), jossa lapsille tarjotaan suomen kielen äänteiden oppimisen tukea sadun keinoin.

KiVe- harjoituspaketti koostuu lapsille kuvitetuista harjoitteista, peleistä ja leikeistä, joissa esiintyy suomen kielen äännejärjestelmän oppimista hankaloittavia ominaispiirteitä. Harjoitteissa korostuvat mm. Vokaalit /y/ ja /ø/ sekä vokaalien ja konsonanttien kestoerot. Leikit tuovat hankalat foneettiset piirteet korostetusti esille. Lisäksi mukana on päiväkodin henkilöstön tukena toimivat selkeät ohjeet leikkien ja pelien säännöistä sekä ohjeita soveltamiseen. KiVe-opas puolestaan sisältää lyhyen ja yleistajuisen fonetiikan perusteiden esittelyn, tietoa uuden kielen ja äidinkielen äännejärjestelmien välisten erojen vaikutuksista oppimiseen sekä parittaisia vertailuja suomen kielen ja päiväkodeissa eniten käytettyjen kielten välisistä eroista ja yhtäläisyyksistä. Työvälineen käytön vaikutuksia oppimiseen tulee seurata sekä oppijoiden että henkilöstön näkökulmista, sillä tavoitteena on sekä oppimisen helpottaminen, että päiväkotien henkilöstön tukeminen monimuotoisessa kieliympäristössä. Esitelmäni avaa työvälineen suunnittelun taustoja, toimintoja sekä käytöstä nousevia tutkimusmahdollisuuksia.

Viitteet

Aerila, Juli-Anna, Gyekye, Marjaana, Haapanen, Katja, Kekki, Niina, Peltola, Kimmo U., Peltola, Maija S., Ruokonen, Inkeri, Thurin, Nina, Tyrer, Maria (2022a) KieliVertailu-työväline: harjoituspaketti. Kieltenvälinen vertailu varhaiskasvatuksen kielitietoisen pedagogiikan kehittämisessä, https://sites.utu.fi/kielivertailu/tyovaline/

Aerila, Juli-Anna, Gyekye, Marjaana, Haapanen, Katja, Kekki, Niina, Peltola, Kimmo U., Peltola, Maija S., Ruokonen, Inkeri, Thurin, Nina, Tyrer, Maria (2022b) KieliVertailu-työväline: opas. Kieltenvälinen vertailu varhaiskasvatuksen kielitietoisen pedagogiikan kehittämisessä, https://sites.utu.fi/kielivertailu/tyovaline/

Best, C. T. & Strange, W. (1992) Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics, 20, 305–330.*

Flege, J. E. (1987) The production of "new" and "similar" phones in a foreign language: evidence for the effect of equivalence classification. *Journal of Phonetics, 15, 47–65.*

Immonen, Katja, Alku, Paavo, Peltola, Maija S. (2022) Phonetic listen-and-repeat training alters 6–7-year-old children's non-native vowel contrast production after one training session, *Journal of Second Language Pronunciation*, 95 - 115.

Jähi, Katri, Alku, Paavo, Peltola, Maija S. (2015) Does interest in language learning affect the non-native phoneme production in elderly learners, *Proceeding of the XVIII International Congress of Phonetic Sciences*, Paper number 0234

Lintunen, Pekka (2004) Pronunciation and phonemic transcription: a study of advanced Finnish learners of English, University of Turku.

Peltola, Kimmo U., Alku, Paavo, Peltola, Maija S. (2017) Non-native speech sound production changes even with passive listening training, *Linguistica Lettica*, 25, 158-172.

Saloranta, Antti., Tamminen, Henna., Alku, Paavo., Peltola, Maija S. (2015) Learning of a non-native vowel through instructed production training, *Proceedings of the 18th International Congress of Phonetic Sciences,* Glasgow, UK, Paper number 235. Rönns, Christel (2022) Iso ja Pieni Talven arvoitus, Turun yliopisto, Turku.

Taimi, Laura, Jähi, Katri, Alku, Paavo, Peltola, Maija S. (2014) Children Learning a Non-native Vowel – The Effect of a Two-day Production Training, *Journal of Language Teaching and Research*, 5 (6): 1229-1235.

Englanninoppijoiden puheen prosodiset piirteet: lähtökielen vaikutusta selvittämässä

Heini Kallio

Tampereen yliopisto

Yksi englanninoppijoiden suurimmista haasteista on sana- ja lausepainojen tarkoituksenmukainen tuottaminen [1, 2]. Aiemmat tutkimukset ovat kuitenkin keskittyneet lähinnä sanapainon sijaintiin, ja painotusten akustisten toteutumien tarkastelu on jäänyt vähälle huomiolle. Lisäksi äidinkielen vaikutusta analysoivissa tutkimuksissa englanninoppijat tulevatusein typologisesti hyvin erilaisista kielistä. Tässä tutkimuksessa tarkastellaan englanninoppijoiden prosodiaa puhujilla, joiden äidinkieli on joko tšekki, slovakki, puola tai unkari. Kaikissa lähtökielissä on kiinteä sanapaino ja ne ovat typologisesti suhteellisen läheisiäkeskenään, mutta kielten prosodiassa on myös eroja [3, 4, 5]. Aiempi samalla aineistolla tehty tutkimus viittaa siihen, että puhujan äidinkieli vaikuttaa tapaan tuottaa englannin painotuksia, millä puolestaan on yhteys havaintoon puhujan taitotasosta [6]. Tämä tutkimus jatkaa englanninoppijoiden äidinkielen vaikutuksen tarkastelua keskittymällä erityisesti sellaisiin puheen prosodisiin piirteisiin, jotka voivat kummuta sana- ja lausepainon toteutumista.

Tšekin-, slovakin-, unkarin- ja puolankielisten englanninoppijoiden tuottamasta lukupuheesta mitattiin tavukestoihin perustuvia rytmiparametreja, perustaajuuden ja intensiteetin muutoksiasekä periodisuutta. Akustisten parametrien suhdetta erikielisten puhujien taitotasoarvioihin tarkasteltiin lineaarisilla regressiomalleilla. Tulokset paljastavat eroja ryhmien välillä niin puherytmissä, perustaajuuden muutoksissa kuin äänenlaadussakin. Osa ryhmien välisistä eroista voi johtua epätasaisesta taitotasojakaumasta, mutta jotkin piirteet todennäköisesti kumpuavat puhujien äidinkielten piirteistä.

Viitteet

[1] Hahn, L. D. (2004). Primary stress and intelligibility: Research to motivate the teachingof suprasegmentals. TESOL Quarterly, 38 (2), 201–223.

[2] Wennerström, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), Perspectives on fluency (pp. 102–127). University of Michigan Press.

[3] Cwiek, A., & Wagner, P. (2018). The acoustic realization of prosodic prominence in Polish: Word-level stress and phrase-level accent. Proceedings of Speech Prosody 2018.

[4] Duběda, T., & Mády, K. (2010). Nucleus position within the intonation phrase: a typological study of English, Czech and Hungarian. In Proceedings of INTERSPEECH 2010.

[5] Beňuš, Š., Reichel, U. D., & Mády, K. (2014). Modeling accentual phrase intonation in Slovak and Hungarian. In L. Veselovská & M. Janebová (Eds.), Complex Visibles Out There(Vol. 4, pp. 677–689). Olomouc, Czech Republic: Palacký University.

[6] Kallio, H., Suni, A., & Šimko, J. (2022). Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different languagebackgrounds. Language and Speech, 65(3), 571-597.

Reduction of unstressed English vowels by L2 speakers with different language backgrounds

Heini Kallio[1], Kamil Kaźmierski[2]

[1] Tampere University, [2]Adam Mickiewicz University

English has relatively strong stress contrasts, leading to the reduction of unstressed vowels into a schwa [1]. Language learners that don't have this feature in their L1 may find it difficult to achieve sufficient centralization of English unstressed vowels which, in turn, can hinder the intelligibility of their speech [2]. This study investigates the production of vowels in unstressed syllables by nonnative speakers of English with different language backgrounds.

The speech data consists of read utterances produced by native English speakers as well as L2 speakers of English with either Slovak, Hungarian, or Polish as their L1. In comparison to English, these languages have fixed word stress and little or no reduction is considered to manifest in unstressed vowels. However, some centralization effect has been found for all the three languages, but centralization seems to depend on the vowel category [3, 4, 5]. In this study we will compare the vowel space area (VSA) of unstressed vowels versus stressed vowels, produced by the L2 speakers and native speakers of English. The degree of shrinking of the vowel space for unstressed vowels compared to stressed vowels is expected to differ among the speaker groups. In particular, the difference between stressed and unstressed VSA is expected to be larger for natives than L2 speakers of English.

References

[1] Halle, M. and Keyser, S. J. *English stress.* Harper and Row New York, 1971.

[2] Lepage, A., & Busà, M. G. (2014). *Intelligibility of English L2: The effects of incorrect word stress placement and incorrect vowel reduction in the speech of French and Italian learners of English*. In Proceedings of the International Symposium on the Acquisition of Second Language Speech Concordia Working Papers in Applied Linguistics (Vol. 5, No. 2014, pp. 387-400).

[3] Rojczyk, A. (2019). *Quality and duration of unstressed vowels in Polish*. Lingua, 217, 80- 89.

[4] Markó, A., Bartók, M., Gráczi, T. E., Deme, A., & Csapó, T. G. (2018). *Prominence effects on Hungarian vowels: A pilot study.* In Proceedings of 9th International Conference on Speech Prosody 2018, Poznań, Poland.

[5] Beňuš, Štefan and Mády, Katalin (2010) *Effects of lexical stress and speech rate on the quantity and quality of Slovak vowels.* In: 5. Speech Prosody Conference, Chicago, Chicago, USA.

Recognition of Russian accent in Finnish oral proficiency test

Riikka Ullakonoja

Centre for Applied Language Studies, University of Jyväskylä

The presentation focuses on the Russian accent in the Finnish oral proficiency test in the National Certificates of Language Proficiency (NCLP). Previous studies on Russian accented Finnish exist, but a lot of them are on read-aloud speech and most lack a solid test-based assessment of the speaker's proficiency level. The data consist of spontaneous monologues on an intermediate proficiency level (B1-B2) of Finnish from 10 speakers (about 1,5 min each), whose first language is Russian. Trained raters (n=44) from NCLP were asked to rate the samples on the NCLP scale, identify the accent in the speech samples as well as to give justifications of their recognition in their own words. The analysis is based on a classification of the raters' descriptions as well as auditory and acoustic analysis by the author. The results show that the raters paid most attention to individual sounds (e.g. their palatalization), but also intonation was mentioned as the reason for recognizing the Russian accent. To conclude, further and more detailed acoustic analysis is needed to gain more understanding of the phonetic features of Russian accented Finnish. The work is funded by the project *'Broken Finnish': Accent perceptions in societal gatekeeping* (Research Council of Finland).

Lukio-opiskelijoiden tunnekokemuksia ääntämisen opetuksessa saadusta palautteesta

Päivi Virkkunen, Minnaleena Toivola & Martti Vainio
Helsingin yliopisto

Suullinen kielitaito on tärkeä kielitaidon osa-alue. Vieraan kielen ääntämisen oppiminen edellyttää runsasta motorista harjoittelua, mutta myös palautetta tarvitaan. Palaute tekee oppijan tietoiseksi oman ääntämisensä piirteistä (Tergujef, ym. 2019), mikä helpottaa harjoittelua, sillä ääntämisen vaikeudet johtuvat yleisemmin kognitiivisen kuin motorisen kielitaidon puutteesta (Fraser, 2000).

Virkkusen ja Toivolan (2020) tutkimuksessa nousi esille, että monet kieltenopettajat pitivät palautteen antamista vaikeana. Suomalaiset lukiolaiset kuitenkin toivovat saavansa erityisesti korjaavaa palautetta omasta ääntämisestään voidakseen kehittyä siinä paremmiksi (Virkkunen & Toivola, 2023).

Selvitämme nyt uudessa tutkimuksessamme, millaisia tunteita suomalaiset lukiolaiset (n=1953) kokevat saadessaan korjaavaa tai positiivista ääntämispalautetta ja miten nämä tunteet vertautuvat esimerkiksi opiskelijan oppimistavoitteisiin. Selvitämme myös onko eri palautetyyppien (palaute opettajalta, palaute toisilta opiskelijoilta, palaute tietokoneohjelmalta) herättämissä tunteissa eroja. Kerromme esityksessämme tutkimuksen tuloksista ja pohdimme niiden merkitystä suhteessa opetuskäytänteisiin.

Tutkimuksemme tuo uutta tietoa palautteen merkityksestä oppijalle. Tulosten perusteella voidaan kehittää opetuskäytänteitä, jotka huomioivat paremmin oppijan toiveet ja tavoitteet ja edistävät siten oppimista. Tavoitteemme on tarjota opettajille heidän kaipaamaansa tukea ääntämisen opetukseen ja palautteen antamiseen, joten tuloksia voidaan käyttää apuna myös kieltenopettajien koulutuksen ja täydennyskoulutuksen suunnittelussa.

Viitteet

Fraser, H. (2000). Coordinating improvements in pronunciation teaching for adult learners of English as a second language. ANTA Innovative Project. Canberra: DETYA.

Tergujef, E., Heinonen, H., Ilola, M., Salo, O.-P. & Kara, H. 2019. Suullisen kielitaidon opetus käytännössä. Teoksessa E. Tergujef & M. Kautonen (toim.) Suullinen kielitaito: opi, opeta, arvioi. Helsinki: Otava, 96–117.

Virkkunen, P., & Toivola, M. (2020). Foneettinen osaaminen helpottaa vieraan kielen ääntämisen opettamista – kyselytutkimus suomalaisten kieltenopettajien käyttämistä ääntämisen opetusmenetelmistä. Ainedidaktiikka, 4(1), 34-57.

Virkkunen, P., & Toivola, M. (2023). Lukio-opiskelijoiden käsityksiä vieraan kielen ääntämisen opetuksessa saadusta palautteesta. AFinLA-teema, 15, 40-59.

POSTER SESSION

Ääntäminen ja transkriptio -kurssin vaikutus khoekhoegowabinkielisen puheen piirteiden havaitsemiseen logopedian ja fonetiikan yliopisto-opiskelijoilla

Katja Haapanen, Antti Saloranta, Kimmo U. Peltola, Henna Tamminen & Maija S. Peltola

Fonetiikka ja Learning, Age & Bilingualism -laboratorio, Turun yliopisto

Tutkimuksessa selvitettiin, kuinka fonetiikan Ääntäminen ja transkriptio -kurssi vaikuttaa logopedian pääaineopiskelijoiden ja fonetiikan sivuaineopiskelijoiden kykyyn havaita heille ennestään tuntemattoman kielen foneettisia piirteitä. Logopedian opiskelijoille kurssi oli pakollinen osa pääaineopintoja. Kurssin aikana opiskelijat tutustuivat maailman äänteisiin IPA:n mukaisesti ja harjoittelivat transkriptioiden tekemistä. Kurssilla kiinnitettiin erityistä huomiota vokaalien ja pulmonisten konsonanttien lisäksi non-pulmonisiin klikkiäänteisiin. Tutkimuksessa opiskelijoille soitettiin katkelmia namibialaisten khoekhoegowabinpuhujien haastatteluista, koska khoekhoegowabissa esiintyy monia indoeurooppalaisiin ja uralilaisiin kieliin verrattuna harvinaisempia foneettisia piirteitä, kuten tooneja, nasaalivokaaleja ja klikkiäänteitä. Tutkimuksen tavoitteena on tarjota konkreettista tietoa siitä, kuinka yliopistossa tarjottava fonetiikan koulutus vaikuttaa tulevien puheterapeuttien ja kieliasiantuntijoiden kykyyn kuunnella ja havaita puhetta sen kielellisestä sisällöstä tai merkityksistä huolimatta.

Tutkimukseen osallistui kaksi kuuntelijaryhmää: logopedian pääaineopiskelijat (n=27, ikä 19–33 vuotta, keski-ikä 22 vuotta, naisia 27) ja fonetiikan sivuaineopiskelijat (n=10, ikä 24–46 vuotta, keski-ikä 25 vuotta, naisia 6). Fonetiikan sivuaineopiskelijoista 7 opiskeli pääaineenaan kieliä ja 3 tietojenkäsittelytieteitä. Kuuntelukoe toteutettiin verkkopohjaisena kyselynä kahdessa aikapisteessä 5 viikon välein: ennen Ääntäminen ja transkriptio -kurssia ja sen jälkeen. Kuuntelukokeessa opiskelijat kuuntelivat neljä Namibiassa nauhoitettua khoekhoegowabin puhenäytettä (2 nais- ja 2 miespuhujaa, kesto 30–50 s.) ja vastasivat kolmeen kysymykseen: mitkä piirteet tunnistit, mihin erikoisiin piirteisiin kiinnitit huomiota, ja miksi kiinnitit huomiota näihin piirteisiin? Vastakset luokiteltiin sisältölähtöisesti.

Tulokset osoittivat, että ennen kurssia logopedian opiskelijat kiinnittivät foneetikkoja vähemmän huomiota puheen segmentaalisiin piirteisiin, mutta erot tasoittuivat kurssin jälkeen. Molemmat ryhmät käyttivät vastauksissaan enemmän foneettisia termejä kurssin jälkeen, mikä näkyi erityisesti klikkiäänteiden kohdalla. Ennen kurssia molemmat ryhmät viittasivat khoekhoegowabin klikkiäänteisiin maiskauksina tai naksauksina, tai täysin puheen ulkopuolisina asioina, kuten taputuksena. Kurssin jälkeen molemmat ryhmät käyttivät vastauksissaan termejä klikki tai klikkiäänne.

Sananalkuisten klusiilien sointi khoekhoegowabin puhujien Namibian englannissa

Henna Tamminen[1], Katja Haapanen[1], Antti Saloranta[1], Kimmo U. Peltola[1], Lannie Uwu-khaeb[2],Maija S. Peltola[1]

[1]*Fonetiikka ja Learning, Age & Bilingualism -laboratorio, Turun yliopisto*
[2]*University of Turku in Windhoek, Namibia*

Englanti on ollut Namibian ainoa virallinen kieli maan itsenäistyttyä vuonna 1990. Tutkimukset ovat osoittaneet, että Namibiassa on kehittymässä Namibian englannin variantti, joka eroaa eteläafrikkalaisista englannin varianteista. Suurin osa namibialaisista on monikielisiä alkuperäiskielien puhujia, ja he oppivat englannin toisena tai kolmantena kielenä viimeistään koulussa. Neljännestä luokasta eteenpäin englanti on ainoa opetuskieli. On siis hyvin todennäköistä, että bantu- ja khoekielien fonologiset järjestelmät vaikuttavat namibianenglannin foneettisiin ominaisuuksiin. Tässä tutkimuksessa selvitettiin, onko L1 khoekhoegowabin puhujien namibianenglannissa sananalkuisten klusiilien /p, t, k, b, d, g/ sointioppositiota, ja miten se toteutuu vapaassa puheessa. Kuvaukset khoekhoegowabin klusiileista vaihtelevat suuresti. Joidenkin kuvausten mukaan kielessä on kolme klusiilia /b/, /d/ ja /g/, jotka ovat sananalkuisina soinnittomiasekä mahdollisesti hieman aspiroituneita, ja sanansisäisinä soinnillisia. Suurin osa khoekhoegowabinpuhujista puhuu myös afrikaansia. Afrikaansissa on /p–b/ ja /t–d/ kontrastit, joissa soinnillisten klusiilien sointi alkaa ennen eksploosiota (pre-voiced) ja soinnittomien klusiilien sointi alkaa hieman eksploosion jälkeen (short-lag).

Tutkimukseen osallistui yhdeksän Namibian yliopiston opiskelijaa (ikä 20–23 vuotta, keski-ikä 21,4 vuotta, naisia 6), joiden äidinkieli on khoekhoegowab. Kaikki koehenkilöt raportoivat puhuvansa myös afrikaansia ja englantia. Muita raportoituja kieliä olivat saksa, ranska, otjiherero ja oshiwambo. Koehenkilöt olivat oppineet englannin 3–7-vuotiaina ja puhuivat sitä tutkimushetkellä päivittäin. Tutkimusaineisto koostui englanninkielisistä haastatteluista. Haastattelukysymykset koskivat koehenkilöiden jokapäiväistä elämää sekä heidän henkilökohtaisia kokemuksiaan ja mielipiteitään liittyen Suomen ja suomalaisten läsnäoloon Namibiassa. Analyysia varten haastatteluista poimittiin yhteensä 365 sanaa, jotka alkoivat klusiili-vokaaliyhtymällä, ja klusiileista analysoitiin soinnin alkamisajankohta (voice onset time, VOT). Sanat jaettiin soinnillisiin ja soinnittomiin muiden englannin varianttien fonologian mukaan.

Soinnittomien klusiilien soinnin alkamisajat olivat huomattavasti soinnillisia klusiileja pidempiä ja ne tuotettiin selvästi aspiroituneina. Soinnilliset klusiilit tuotettiin hyvin lyhyillä soinnin alkamisajoilla. Aspiroituneiden klusiilien soinnin alkamisajat olivat joihinkin aiempiin tutkimuksiin verrattuna hieman lyhyemmät. Tämä saattaa osittain johtua korkeasta puhenopeudesta. Tulokset osoittavat, että khoekhoegowabinkieliset englannin puhujat tekevät eron soinnittomien ja soinnillisten klusiilien välille ennemminkin soinnittomien klusiilien aspiraatiolla kuin soinnillisten klusiilien voimakkaalla soinnilla. Näyttää siis siltä, että sointiero tuotettiin kuten monissa muissakin englannin varianteissa eikä khoekhoegowabin tai afrikaansin vaikutusta löytynyt.

Namibianenglannin vokaalit auditiivisessa ja visuaalisessa tuottokokeessa oshiwambonpuhujilla

Antti Saloranta[1], Katja Haapanen[1], Kimmo U. Peltola[1], Henna Tamminen[1], Meameno Shiweda[2], NapandulweShiweda[2], Maija S. Peltola[1]

[1] *Fonetiikka ja Learning, Age & Bilingualism -laboratorio, Turun yliopisto*
[2] *University of Namibia (UNAM)*

Afrikan eteläosissa sijaitseva Namibia on kielellisesti monimuotoinen maa, jossa puhutaan n. 30 kieltä. Paikalliset kielet kuuluvat joko bantu- tai khoisankieliin, ja maassa puhutaan tämän lisäksi myös germaanisia kieliä, erityisesti afrikaansia, englantia ja saksaa. Paikallisista kielistä puhutuin on oshiwambo, bantukieliryhmä, jonka eri murteita puhutaan 45%:ssa kotitalouksista. Ryhmän kielissä on tyypillisesti viisi vokaalia /i, u, e, o a/, jotka esiintyvät sekä pitkinä että lyhyinä [1], [2]. Namibian virallinen kieli on englanti, mutta se on kuitenkin kotikielenä vain 3,4% kotitalouksista. Valtaosa namibialaisista puhuu arjessaan vähintään kahta kieltä, ja englanti onkin monille vasta kolmas tai neljäs opittava kieli. Namibianenglantia ei tyypillisesti pidetä omana englannin varianttinaan, mutta tuoreiden tutkimustulosten perusteella tietyt siinä esiintyvät foneettiset piirteet erottavat sen erityisesti Etelä-Afrikan englannista [3].

Tutkimuksen tarkoituksena oli osana Tanssi uhanalaisten kielten ja foneettisen maailman tulkkina -hanketta [4] tutkia englannin vokaalien tuottoa äidinkielisillä oshiwambon puhujilla sekä auditiivisessa että visuaalisessa koeasetelmassa. Osallistujia oli yhteensä 20 (ikä 19—68, KA 30,1, 12 naista), joista kaikki puhuivat englantia vähintään hyvällä tasolla (itse arvioitu taito KA 3,95, skaala 1 = alkeet, 5 natiivintasoinen). Käytetyt ärsykkeet olivat 20 englanninkielistä sanaa minimi- tai subminimipareina (heat-heed, hit-hid, bet-bed, hat-had, foot-hood, hoot, who'd, bought-board, hut-hud, tot-Todd, heart-hard), jotka sisälsivät 10 brittienglannin vokaalia soinnillisen ja soinnittoman klusiilin edellä.

Kummassakin koeasetelmassa kukin ärsykesana toistettiin kolme kertaa, eli toistoja oli yhteensä 60. Sanat esitettiin osallistujalle yksi kerrallaan kolmen sekunnin välein samassa pseudosatunnaisessa järjestyksessä, jossa sama sana ei esiintynyt peräkkäin. Osallistujia pyydettiin tuottamaan sanat normaalilla äänellä, ja kaikki tuotot nauhoitettiin. Osallistujista 11 aloitti kokeen visuaalisella ja 9 auditiivisella koeasetelmalla. Auditiivisessa koeasetelmassa kuulivat ärsykesanat miespuolisen brittienglannin puhujan tuottamana, ja visuaalisessa asetelmassa ärsykkeet esitettiin ortografisessa muodossa.

Kaikkien tuotettujen sanojen vokaalin keskikohdasta mitattiin F1 ja F2, joista laskettiin ryhmän keskiarvot. Alustavien tulosten perusteella formanttiarvot kasautuvat kummassakin koeasetelmassa noin viiteen ryhmään. Suurin yksittäinen ero asetelmien välillä on BED-BET ja HAD-HAT -sanaparien vokaalien läheiset formanttiarvot visuaalisessa koeasetelmassa. Ero johtuu HAD-HAT-parin vokaaleista, joiden formantit ovat siirtyneet kohti /e/-äännettä. Tämä saattaa johtua siitä, että visuaalisessa koeasetelmassa puhujat ovat tuottaneet sanat omalla englannillaan, kun taas auditiivisessa asetelmassa he ovat pyrkineet matkimaan brittienglannilla tuotettuja ärsykkeitä. Lopulliset tulokset esitellään Fonetiikan päivillä.

Viitteet

[1] D. Fivaz and S. Shikomba, *A Reference Grammar of Oshindonga (Wambo)*. Windhoek, Namibia: Star Binder &Printers, 1986.

[2] W. Zimmermann and P. Hasheela, *Oshikwanyama grammar*. Windhoek, Namibia: Gamsberg Macmillan, 1998.

[3] A. Schröder, F. Zähres, and A. Kautzsch, "The phonetics of Namibian English," in *The Dynamics of English in Namibia: Perspectives on an emerging variety*, A. Schröder, Ed., in Varieties of English Around the World. JohnBenjamins Publishing Company, 2021, pp. 111–134. doi: 10.1075/veaw.g65.06sch.

[4] K. Haapanen, A. Saloranta, K. U. Peltola, H. Tamminen, and M. S. Peltola, "Fonetiikan tutkijat etsivät keinoja englannin ääntämisen tukemiseen ja kieliperinnön säilyttämiseen Namibiassa," *Kieli, koulutus ja yhteiskunta*, vol.13, no. 6, 2022, Accessed: Jan. 23, 2024. [Online]. Available: https://www.kieliverkosto.fi/fi/journals/kieli- koulutus-ja-yhteiskunta-marraskuu-2022/fonetiikan-tutkijat-etsivat-keinoja-englannin-aantamisen-tukemiseen- ja-kieliperinnon-sailyttamiseen-namibiassa

Building a Naturalistic and Representative Affective Speech Corpus for the Finnish Language

Kalle Lahtinen[1], Liisa Mustanoja[2], Okko Räsänen[1]

[1]Unit of Computing Sciences Tampere University, Finland
[2]Language Studies, Tampere University, Finland

Spoken language contains affective (emotional) information, which is conveyed by suprasegmental variation (e.g prosody and fonation) in speech as well as by other situational variation such as word choices along with dialectal, syntactic and semantic variation. The information is ultimately perceived by the listener as subjective interpretations. While affect is part of everyday conversational communication, there is little existing research on expression and perception of affect in spoken Finnish [8][7], not to mention across different idiolectal subgroups such as speakers of different age or dialectal background. Since expression and interpretation of affect in language is known to depend on cultural and social conventions, better understanding of the expression of affect in Finnish would be desirable. The goal of our work is to research how affect is expressed in everyday spoken Finnish using large-scale data.

A prerequisite for our research on affective language is a speech corpus containing unscripted audio recordingsof speech paired with metadata (or annotations) containing information about the affective expression. However,we are aware of only two Finnish speech corpora related to affect, both consisting of acted emotional expressionswhile reading a pre-defined script and consisting only a small amount of speech in total [1][5]. In contrast, severallarge-scale datasets containing unscripted speech in Finnish exists [4][2][6], but they lack affect related metadata.Building an affective speech corpus can be done in several ways, typically by recording acted speech in a controlled setting or utilizing publicly available free speech audio sources from different medias such as podcasts,radio or television. The trade-off when building these types of datasets is typically between the richness and balanceof affective expression present in the data and the level of information the dataset contains about the expressionin the data [3]. In this presentation, we will describe our approach to compiling a spoken Finnish dataset for thestudy of affective expression by combining the LahjoitaPuhetta, HelPuhe and TamPuhe datasets. The datasetwill be built by aligning the audio recordings with their respective text transcriptions and split into individualutterance samples (consisting of audio and text). Each utterance sample in the dataset will be augmented with atext sentiment, speech-to-noise ratio and audio-based emotion estimates first by using automated tools and finallyannotating a subset of samples manually. The final dataset can be used to build better tools for automated affect related annotation providing more options for researching affect and idiolectical variation using large-scale data. The work is a part of the CONVERGENCE-project at Tampere University, funded by the Jane and Aatos Erkko Foundation.

References

[1] Matti Airas and Paavo Alku. “Emotions in Vowel Segments of Continuous Speech: Analysis of the Glottal Flow Using the Normalized Amplitude Quotient”. English. In: *PHONETICA* 63.1 (2006), pp. 26–46. ISSN: 0031-8388.

[2] *Longitudinal data of Tampere spoken language (Not yet public)* data set. URL: http://urn.fi/urn:nbn:fi:lb-2022090821.

[3] Reza Lotfian and Carlos Busso. “Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings”. In: *IEEE Transactions on Affective Computing* 10.4 (2019), pp. 471–483. DOI: 10.1109/TAFFC.2017.2736999.

[4] Anssi Moisio et al. “Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks”. English. In: *Language Resources and Evaluation* 57 (Aug. 2022). Publisher Copyright: © 2022, The Author(s)., pp. 1295–1327. ISSN: 1574-020X. DOI: 10.1007/s10579-022-09606-3.

[5] Tapio Seppänen, Juhani Toivanen, and Eero Väyrynen. “MediaTeam Speech Corpus : a first large Finnish emotional speech database”. In: 2003. URL: https://api.semanticscholar.org/CorpusID:15915991.

[6] *The Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)*. data set. URL: http://urn.fi/urn:nbn:fi:lb-2014073041.

[7] Laura Visapää. “Itsenäiset infinitiivit affektin ja empatian konstruktioina”. In: *Virittäjä* 117.4 (2013), pp. 524–550.

[8] Valma. Yli-Vakkuri. Suomen kieliopillisten muotojen toissijainen käyttö. fin. Turun yliopiston suomalaisen ja yleisen kielitieteen laitoksen julkaisuja ; 28. Turku: Turun yliopisto, 1986. isbn: 951-642-800-2.

The effect of F0 measurements on prosody analysis in language development studies

Daniil Kocharov and Okko Räsänen

Unit of Computing Sciences, Tampere University, Finland

Prosodic research on child-directed speech (CDS) usually focuses on measurements of F0 mean and standard deviation (SD). In general, there is an agreement that CDS has higher pitch mean and variability than ADS (adult-directed speech). However, earlier studies have reported conflicting findings on how F0 of CDS changes with the recipient child's age, as there is evidence for both presence and absence of age-related change of pitch variability. One possibility is that the disagreement might originate from the variability of speech behaviors of speakers in different languages and cultures. Alternatively, it could be explained by the F0 measurement methodology as well.

In this work, we investigate how the way we measure F0 influences the overall results on CDS change with child age. We investigated two different factors which could influence melodic analysis of speech: a) what sound segments are used to measure F0: vowels, sonorants, voiced obstruents, no account for sound segments; and b) whether the number of words within an utterance is taken into account in the analysis. We used a dataset of maternal utterances from the Providence Corpus (a collection of twice-monthly recordings of hour-long mother-child spontaneous interactions from six NA-English-speaking children) addressed at children in the age range of 1;0 to 3;0. The F0 values were calculated using OpenSMILE toolkit. The transcription-to-speech alignment was performed using WebMAUS online toolkit. The age- dependency of F0 feature was tested by means of the Spearman's rank correlation coefficient between quantized child age and the feature values associated with the age bin.

The results show that the estimation procedure can affect the findings, potentially affecting developmental interpretations. In case of calculating F0 on either voiced consonants and vowels, or vowels only, there is a significant age-dependent increase of F0 SD. This is in contrast to the case of using all F0 values calculated within an utterance by F0 detection algorithm, when no significant age-dependency for F0 SD is found. Thus, it matters whether we take into account segments that are known to have proper voicing or measure F0 from all speech that an automatic F0 estimator considers as voiced irrespectively of the underlying segments. Second, we found no age-related dependencies of F0 SD (whether we took sound segment identities into account or not) for the scenario where the analysis was controlled for the number of words within an utterance, i.e. comparing one-word utterances across all ages, two-word utterances across all ages, etc. The revealed age-dependency of F0 SD for the first scenario might be explained constantly increasing length of utterances in CDS in terms of number of pronounced words along with a child age. This is since the number of pronounced words in an utterance might influence melodic variability within the utterance, where more complex intonational structure may be required to prosodically structure the longer utterances.

Computational Investigation of the Feasibility of Statistical Learning for Early Word Comprehension using Realistic Input Statistics.

Khazar Khorrami & Okko Räsänen

Unit of Computing Sciences, Tampere University

Infants acquire their native language through interactions with their environment.Experimental studies have shown that by one year of age, infants develop some degree of sensitivity to the phonetic categories of their native language, can segment many words from continuous speech, and comprehend the meaning of several common words [1]. One possible mechanism behind early language comprehension, introduced and discussed in the literature, is that infants extract regularities and patterns in speech and language by being exposed to languageinput in repeating scenarios, known as statistical learning [2].

However, statistical learning is a data-hungry model; therefore, researchers have doubted its feasibility as a learning mechanism for early word comprehension due to the limited nature of children's input, especially the sparsity of audiovisual naming events that lead to a high level of referential ambiguity in the learner's mind [3]. Earlier computational studies have demonstrated the success of statistical learning as a mechanism for language learning, but these models were built upon unrealistic amounts of data, far from what is available for children [3]. We investigate if, in the presence of a realistic amount of audiovisual naming events, statistical learning can still lead to any word comprehension outcome.

We simulate a model of a 6–12-month-old infant statistical learner using a quantitatively similar amount of data accessible for an average child during different stages within their first year of age. We apply a model of auditory self-supervised learning in our speech processing pipeline, a modelof visual self-supervised learning, and a model of weakly-supervised learning as a cross-modal network that operates on the outputs of the modality-specific networks, connecting auditory and visual pipelines at the utterance and image levels. Our simulation data consists of photographs illustrating various everyday-life scenes and spoken utterances of read speech with varying lengths.In audiovisual learning, we assume that the utterances are contextually related to the contents of the visual scenes. The number of audiovisual pairs fed to the model is carefully selected to match the statistics of real-world naming events infants are exposed to, as reported in experimental works [3].

The results demonstrate that despite limited data, the model of self-supervised statistical learning can achieve some degree of phonemic, lexical, and semantic knowledge comparable to what is known about children's linguistic skills from experimental studies. We conclude that based on computational modeling, statistical learning without using any linguistic priors is itself sufficient to lead to some degree of phonemic, lexical, and word meaning perception knowledge observed in infants below 1 year of age.

References

[1] Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253-3258.

[2] Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current directions in psychological science*, *12*(4), 110-114.

[3] Clerkin, E. M., & Smith, L. B. (2022). Real-world statistics at two timescales and a mechanismfor infant learning of object names. *Proceedings of the National Academy of Sciences*, *119*(18), e2123239119.

Quantity perception among Estonian kindergarten children with developmental language disorder

Liis Themas[1], Pärtel Lippus[1], Marika Padrik[2], Kairi Kreegipuu[3]

[1]Institute of Estonian and General Linguistics, [2]Institute of Educational Science, [3]Institute of Psychology
University of Tartu

Estonian's word prosody features a unique three-way quantity distinction (short Q1, long Q2, overlong Q3) marked by syllable duration and pitch contour. The temporal pattern can be described by a reverse relation between the stressed and the unstressed syllables of a left-headed disyllabic foot, meaning thatthe stressed syllable is longer in the case of higher degrees of quantity while the unstressed syllable is compensatorily shortened [1].

Developmental language disorder is a heterogeneous category that encompasses a wide range of problems [2]. Impairment can occur on some or all levels of speech perception and/or production [3]. Previousstudies have shown delayed language processing in the DLD population [4], including difficulties in prosody perception that predict later impairments in language development [5], [6]. At school-age prosody perception remains compromised in children with DLD [7], [8]. There is some behavioural evidence that they have difficulties in distinguishing between the Estonian quantity degrees, pronouncing the quantities [9] and marking them correctly in orthography [10]. But to date there is no neurophysiological data about processing the quantity system by typically developing (TD) or children with DLD.

This study focuses on the neurophysiological and behavioural differences in perceiving Estonian's three-way quantity distinction between children with DLD and TD peers (ages: 4.6-6.5 years; DLD group N=25, TD group N=25) using psychometric testing, sleep-EEG, and auditory event-related potentials measured in a passive oddball paradigm with naturally produced stimuli: (1) *sada* (Q1 deviant; *hundred*, nom.sg), *saada* (Q2 standard; *send!* imp. sg) and *saada* (Q3 deviant; *to get*, imp. sg); (2) *vere* (Q1 deviant; *blood*, ptcp. sg), *veere* (Q2 standard; *slope*, ptcp. sg) and *veere* (Q3 deviant; *to roll*, imp. sg). Additionally, two computerized behavioural tasks were conducted: a quantity discrimination task and a lexical decision task.

Here we present the data of the first phase of our longitudinal study. The neural obligatory responses, whichreflect processes extracting auditory features, were present for all the stimuli in both groups. The findings of thecluster-based permutation tests reveal that in the vere-veere-veere condition, the typically developing (TD) group differentiated between Q2 and Q1/Q3. In the second condition the group displayed a response just to theQ2 vs Q1 contrast, and the reaction occurred in cortical areas atypical to the discrimination response. For both conditions no discrimination response was elicited in the DLD group. Behavioural task results showed better quantity discrimination in the TD group, with no significant differences in reaction times between groups.

References

[1] Lippus, P., Pajusalu, K., & Allik, J. (2009). The tonal component of Estonian quantity in native and non-native perception. J of Phonetics, 37(4), 388–396.

[2] Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & and the CATALISE-2 consortium. (2017). Phase 2 of CATALISE: A multinational and multidisciplinary Delphi consensus study of problems with language development: Terminology. J of Child Psychol and Psychiatry, 58(10), 1068–1080.

[3] Leonard, L. B. (2014). Children with Specific Language Impairment. (2. eds.). The MIT Press.

[4] Kujala, T., & Leminen, M. (2017). Low-level neural auditory discrimination dysfunctions in specific language impairment - A review on mismatch negativity findings. Developmental Cognitive Neuroscience, 28, 65–75.

[5] Weber, C., Hahne, A., Friedrich, M., & Friederici, A. D. (2005). Reduced stress pattern discrimination in 5-month-olds as a marker of risk for later language impairment: Neurophysiologial evidence. *Cogn Brain Res*, 25(1), 180–187.

[6] Cantiani, C., Riva, V., Piazza, C., Bettoni, R., Molteni, M., Choudhury, N., Marino, C. & Benasich, A. A. (2016). Auditory discrimination predicts linguistic outcome in Italian infants with and without familial risk for language learningimpairment. Dev Cog Neurosci, 20, 23-34.

[7] Datta, H., Shafer, V. L., Morr, M. L., Kurtzberg, D., & Schwartz, R. G. (2010). Electrophysiological Indices of Discrimination of Long-Duration, Phonetically Similar Vowels in Children With Typical and Atypical Language Development. J of Speech, Lang, and Hearing Res, 53(3), 757–777.

[8] Cheng, Y-Y, Wu, H-C., Shih, H-Y., Yeh, P-W., Yen, H-L., & Lee, C-Y. (2021). Deficits in Processing of LexicalTones in Mandarin-Speaking Children With Developmental Language Disorder: Electrophysiological Evidence. J of Speech, Lang, and Hearing Res, 64(4), 1176-1188.

[9] Padrik, M. & Hallap, M. (2016). Kommunikatsioonipuuded lastel ja täiaksvanutel: märkamine, hindamine ja teraapia. Tartu Ülikooli Kirjastus.

[10] Karlep, K. (1999). Emakeele abiõpe I. Tartu Ülikooli kirjastus.

Further information on the perception of Estonian long–overlong quantity boundaries

Pärtel Lippus [a], Liis Kask [b], Sofia Lutter [b], Nele Põldver [b], Kairi Kreegipuu [b]

[a] *Institute of Estonian and General Linguistics, University of Tartu*
[b] *Institute of Psychology, University of Tartu*

This paper presents the results of a large-scale web-based perception experiment that tested the distinction of long and overlong quantity in Estonian. The study has two main pointsof interest. Firstly, we observe segmental quality effect on the quantity perception. Most of the quantity experiments have used one or two triplets, e.g. [sɑtɑ] – [sɑːtɑ] – [sɑːːtɑ] [1], [2] or [jɑmɑ] – [jɑːmɑ] – [jɑːːmɑ] [3]. At the same time there is a significant effect of microprosodyon the perception of quantity category [4] and the temporal ratios can vary considerably dueto the differences of the segmental quality [5]. For this study the stimuli were created from words with 8 different segmental combinations: [sɑːtɑ] – [sɑːːtɑ], [sɑːke] – [sɑːːke], [sɑːki] – [sɑːːki], [lɑːtʲi] – [lɑːːtʲi], [liːkɑ] – [liːːkɑ], [viːte] – [viːːte], [viːki] – [viːːki], and a nonsense wordpair [tɑːtɑ] – [tɑːːtɑ]. The stimuli were created by resynthesis, manipulating the vowel durations or the pitch contour.

The second aim of this study is to map the possible dialectal variability in Estonian quantity perception. There has been some evidence that the listeners from East and South dialect areasare less sensitive to pitch cue than those from North and West [6]. The current study was carried out using the Kaemus online environment (https://kaemus.psych.ut.ee) with 290 native Estonian participants with various regional background.

The results showed that different segmental quality sets have slightly different Q2–Q3 category boundaries. Also, the precision of quantity identification was considerably lower in the case of nonsense word set. Surprisingly we were not able to find a clear dialectal background effect.

Acknowledgements

The title of the paper has been inspired by [3]. The work was supported by the Estonian Research Council grant number PRG1151.

References

[1] I. Lehiste, ‘Experiments with synthetic speech concerning quantity in Estonian’, in Congressus Tertius Internationalis Fenno-Ugristarum Tallinnae habitus 17.–23. VIII 1970. Pars I Acta linguistica, V. Hallap, Ed., Tallinn: Valgus, 1975, pp. 254–269.

[2] P. Lippus, K. Pajusalu, and J. Allik, ‘The tonal component of Estonian quantity in native and non-native perception’, Journal of Phonetics, vol. 37, no. 4, pp. 388–396, Oct. 2009, doi: 10.1016/j.wocn.2009.07.002.

[3] A. Eek, ‘Further information on the perception of Estonian quantity’, Estonian Papers in Phonetics, vol. 1979, pp. 31–57, 1980.

[4] E. Meister and S. Werner, ‘Duration affects vowel perception in Estonian and Finnish’, LU, vol. 45, no. 3, pp. 161–177, 2009, doi: 10.3176/lu.2009.3.01.

[5] P. Lippus and J. Šimko, ‘Segmental context effects on temporal realization of Estonian quantity’, in Proceedings of the 18th International Congress of Phonetic Sciences, M. Wolters, J. Livingstone, B. Beattie, R. Smith, M. MacMahon, J. Stuart-Smith, and J. Scobbie, Eds., Glasgow: University of Glasgow, 2015, pp. 1–5. [Online]. Available: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0318.pdf

[6] P. Lippus and K. Pajusalu, ‘Regional variation in the perception of Estonian quantity’, in Nordic prosody. Proceedings of the Xth Conference, Helsinki 2008, M. Vainio, R. Aulanko, and O. Aaltonen, Eds., Frankfurt, Berlin, New York: Peter Lang, 2009, pp. 151–157.

The loss of word-internal laryngeal fricative in South Estonian Leivu dialect

Pire Teras

University of Tartu

The occurrence of laryngeal fricative *h* is one of the phonological innovations in the Finnic languages (Pajusalu 2012). According to the *Uralic Areal Typology Online* (UraTyp) database, it also occurs in Saami languages, Hungarian, Kamas, and South Selkup (Norvik et al. 2022). However, the distribution of *h* in Southern Finnic languages (Estonian, including South Estonian, Livonian and Votic) is restricted (Pajusalu 2012).

In Livonian, like in its contact language Latvian, laryngeal fricative appears only marginally (in new loan words) (Pajusalu 2012). Recent studies about Estonian and Leivu South Estonian dialect have shown that an intervocalic *h* is more or less prone to complete loss also there. However, when in Estonian, the most frequent variant is a voiced variant, and the loss occurred in 21% of all analysed words (Teras 2018), in Leivu, an intervocalic *h* is almost always lost and occurred only in a few words as voiced fricative (Teras 2021).

This paper aims to find out what are the other conditions for the loss of word-medial *h*. What are other contexts where the loss of *h* occurs? How frequent is the loss and how does it influence the duration ratios? The spontaneous speech of three male speakers is analysed acoustically. Preliminary results show that in Leivu quantity 2 words *h* at the syllable boundary before a voiced consonant tends also to be lost. The loss of *h* usually lengthens the preceding short vowels. In quantity 3 words in the same context the syllable-final *h* is pronounced longer and is retained.

References

Norvik, Miina, Yingqi Jing, Michael Dunn, Robert Forkel, Terhi Honkola, Gerson Klumpp, Richard Kowalik, Helle Metslang, Karl Pajusalu, Minerva Piha, Eva Saar, Sirkka Saarinen, & Outi Vesakoski. (2022). *Uralic Typological database - UraTyp* (v1.1) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.6392555.

Pajusalu, Karl (2012). Phonological Innovations of the Southern Finnic Languages. *Linguistic Map of Prehistoric North Europe.* Eds. R. Grünthal, P. Kallio. Helsinki: Suomalais-Ugrilaisen Seura, 201−224.

Teras, Pire (2018). The phonetic variation of short intervocalic /h/ in Estonian. *Proceedings of the 9th international conference on Speech Prosody.* Eds. K. Klessa, J. Bachan, A. Wagner, M. Karpiński, D. Śledziński. Poznań: Adam Mickiewicz University, 883−887. https://doi.org/ 10.21437/SpeechProsody.2018-178.

Teras, Pire (2021). Broken tone in Leivu CV'V-words. *Journal of Estonian and Finno-Ugric Linguistics*, 12 (2), 169−190. https://doi.org/10.12697/jeful.2021.12.2.07.

Russian Speakers' Perception of Stress in Finnish Words

Tatiana Kachkovskaia, Michael O'Dell, Tommi Nieminen

Although Finnish is usually said to have fixed lexical stress on the first syllable of a word, speakers of other languages may occasionally perceive Finnish words as having stress on some other syllable, especially when the first syllable is short (cf. e.g. [1], [6]). There are several factors which may affect which syllables in a foreign language are perceived to be stressed, depending on the listener's native language [2].

In the case of Russian speakers listening to Finnish, the fact that duration is primarily associated with stress in Russian, which can occur on any syllable, not just the first ([3], [7]), means that short (CV) Finnish syllables may be less likely to be perceived as stressed. In addition, a possible half-long vowel in the second syllable after a short first syllable may further increase the probability of hearing second syllable stress.

Another possible influence is the identity of the vowel in various syllables of words in Finnish. Due to the fact that Russian has strong vowel reduction, only certain vowels normally occur in unstressed position, e.g. Russian /o/ almost never occurs in unstressed syllables. This being the case, it is possible that some vowels, for instance Finnish /o/, will attract more judgments of stress from Russian listeners than other vowels, regardless of their position in the word. Our research questions are therefore

- Are some Finnish words perceived by Russian speakers as having second syllable stress?
- Does perception of stress depend on the vowels of the first and second syllables?

We will report on the results of perception experiments designed to investigate these questions conducted with Russian speakers listening to Finnish words spoken by native speakers.

References

[1] E. Aho, M. Toivola, F. Karlsson, and M. Lennes, "Aikuisten maahanmuuttajien suomen ääntämisestä," *Puhe ja kieli*, vol. 36, no. 2, pp. 77–96, 2016.

[2] H. Altmann, "The perception and production of second language stress: A cross-linguistic experimental study," Ph.D. dissertation, University of Delaware, 2006.

[3] L. V. Bondarko, *Fonetika sovremennogo russkogo jazyka*. St. Petersburg: St. Petersburg University, 1998.

[4] M. Toivola and R. Ullakonoja, "Identification of Russian accented Finnish by native and non-native listeners with and without Finnish profi- ciency," in *Näkökulmia toisen kielen puheeseen — Insights into Second Language Speech*, ser. Soveltavan kielitieteen tutkimuksia 2017 10, M. Kuronen, P. Lintunen, and T. Nieminen, Eds., AFinLA-e, 2017, pp. 258–276.

[5] R. Ullakonoja, H. Dufva, M. Kuronen, and P. Hurme, "How to imitate an unknown language? Russians imitating Finnish," in *XXVIII Fonetiikan päivät — Turku 25.–26. lokakuuta 2013*, K. Jähi and L. Taimi, Eds., Turun yliopisto, 2014, pp. 10–18, ISBN: 978-951-29-5980-8. [Online]. Available: http://urn.fi/URN:ISBN:978-951-29-5980-8.

[6] V. V. Vihanta, "Suomi vieraana kielenä foneettiselta kannalta," in *Vieraan kielen ymmärtäminen ja tuottaminen. AFinLA:n vuosikirja 1990*, J.Tommola, Ed., AFinLA, 1990, pp. 199–225.

[7] I. Yanushevskaya and D. Bunc̆ic′, "Russian," *Journal of the International Phonetic Association*, vol. 45, no. 2, pp. 221–228, 2015. DOI: 10. 1017/S0025100314000395.

[8] R. Ylitalo, *The Realisation of Prominence in Three Varieties of Standard Spoken Finnish* (Acta Universitatis Ouluensis, Series B, Humaniora 88). University of Oulu, 2009. [Online]. Available: http://herkules.oulu.fi/isbn9789514291142/.

Turn overlaps in collaborative dialogues and the factor of social distance

Tatiana Kachkovskaia*, Daniil Kocharov**

**Independent Researcher, Finland*

*** Unit of Computing Sciences, Tampere University, Finland*

In dialogue speech, interlocutors' turns often overlap — i.e., there are fragments where both interlocutors speak simultaneously. This may occur at turn transitions (between-speaker overlaps, BSO) and without a change in speakers' roles (within-speaker overlaps, WSO), the latter group being mostly comprised of backchannels and unsuccessful attempts to interrupt the interlocutor.

In this research, we hypothesised that there might be variation in turn-taking strategies due to the factor of social distance, i.e. the relationship between the interlocutors. This idea was based on our previous work, where we found such variation for other speech phenomena, e.g. tempo and a range of paralinguistic events. This research uses the large speech corpus SibLing covering 5 degrees of social distance between the interlocutors: sibling-sibling (same gender), friend-friend (same gender), stranger- stranger (same gender), stranger-stranger (opposite gender), stranger-stranger with age difference (same gender; additionally, the elder interlocutor's job required leadership skills). The corpus contains 90 dialogues between 20 "core" speakers and their interlocutors; each of the 20 "core" speakers participated in 5 dialogues with different social distances. All dialogues included two collaborative tasks: a card-matching game and the classical map task. The results were obtained using repeated measures ANOVA. Apart from the factor of social distance, the factors of speaker's gender and role (for map task only) were analysed.

Between-speaker overlap (BSO) was measured as the percentage of turn transitions where a speaker started his/her turn before the interlocutor had finished speaking (with the threshold of 10 ms). In the card-matching game, we observed an interaction between the factors of gender and social distance ($p=0.002$). Our results revealed different strategies for male vs. female speakers in mixed- gender conversations. Within the female subgroup, the factor of social distance was significant ($p=0.026$); female speakers used less overlaps in dialogues with strangers of the opposite gender (mean 24% compared to 32-39% in other settings). Within the male subgroup, the factor of social distance was also significant ($p=0.027$), but less overlaps in dialogues with strangers of the same gender (mean 21% compared to 31-36% in other setting). An analysis of map task recordings showed similar results, but also enabled us to reveal the influence of conversational roles: Followers (those who followed the route and reported on mismatches between the maps) started their turns with more overlaps than Leaders (those who described the routes): $p=0.007$, mean frequency 33% vs 28%. Higher frequency of overlaps in Followers' speech in map task could be explained by the slight differences in the interlocutors' maps, which needed to be reported immediately in order to find the correct path.

Within-speaker overlap (WSO) was calculated as the frequency of a speaker's turns occurring fully within the interlocutor's turn, measured per second of the interlocutor's speech. Such turns are usually very short and could be false-started turns or backchannels. In card-matching games, we observed a statistically significant influence of the speaker's gender ($p=0.028$): WSO rate was higher in females' speech (0.035 items per second vs. 0.027). In the map task, social distance was a significant factor ($p<0.0001$), and so was role ($p<0.0001$) and their interaction ($p=0.003$). Leaders had higher WSO rate (mean 0.14 vs. 0.06). When acting as Leaders, speakers revealed significant differences in dialogues with strangers of greater age, in comparison with other dialogues. In such settings, WSO rate was higher: mean 0.25 vs. 0.08-0.14. That is, in settings with interlocutors of greater age, speakers used WSO about twice as frequently.

Enhancing Plosive Recognition in Singing: The Impact of Elongated Plosive Closures in Varied Acoustics

Allan Vurma,[1] Einar Meister,[2] Lya Meister,[2] Jaan Ross,[1] Marju Raju[1], Veeda Kala[1], and Tuuri Dede[1]

[1] *Estonian Academy of Music and Theatre*
[2] *Tallinn University of Technology*

Poor intelligibility of sung text is a common challenge for classically trained singers, particularly in reverberant rooms with orchestral accompaniment, which can mask the sung text [1, 2]. This study investigates whether elongating the duration of voiceless plosive closures improves the recognition of plosives during vowel-plosive-vowel junctions in reverberant acoustics.

First, we analyzed plosive closure durations in Romantic and Classical opera arias performed by 11 classically trained singers across different vocal ranges. Additionally, vocalists were requested to read the arias aloud in both conversational and oratorical styles. Notably, closure durations varied among singing (mean 74.7 ms), conversational speech (mean 84.7 ms), and oratorical reading (mean 93.8 ms). The outliers in all three cases extended to about 300 ms. Based on these data, the closure durations of 60 ms, 150 ms, and 260 ms were chosen for the VCV stimuli of the perception tests.

The recordings of /a-k-a/, /a-p-a/, and /a-t-a/ sequences sung by two classically trained professional opera singers – a mezzo-soprano and a tenor served as the basis for stimuli. The closure durations of plosives in VCV stimuli were manipulated in Praat [3], while burst durations and transitions to the following vowel were left intact for the sake of naturalness. The Praat Vocal Toolkit [4] was used to create stimulus sets with simulated reverberation (Church or Big Room) and with or without brown noise to mimic the masking effect of accompaniment. The final stimulus set included three series (I: based on tenor recordings at pitch G3, II and III: based on mezzo-soprano recordings at pitch G4 and F5, respectively). Consequently, the paradigm of the test series I and II included 90 stimuli (3 plosives x 3 different closure durations x 2 burst intensities x 5 acoustic conditions). In series III, the contrast of burst intensities was omitted, resulting in 54 stimuli.

34 listeners (11 males, 23 females) participated one-by-one Praat-administered perception tests in a soundproof booth using the same setup (a laptop, calibrated external audio card, Sennheiser HD 560s headphones). The results revealed a significant improvement (approximately 20 percentage points) in plosive recognition in simulated concert hall acoustics when the plosive closure was extended. However, this effect was weaker in church acoustics and absent in rooms without reverberation. Recognition did not improve when stimuli were presented with brown noise imitating accompaniment.

This study suggests that elongating plosive closures may enhance plosive recognition during singing in typical concert hall acoustics. Further investigation is needed to determine whether vocalists naturally adopt this technique in their performances. These findings contribute to our understanding of vocal techniques in varied acoustic environments and offer insights for improving sung text intelligibility.

References

[1] Meyer, J. (2009). Acoustics and the Performance of Music. 5th ed. Springer.
[2] Miller, R. (1996). On the Art of Singing. Oxford University Press.
[3] P. Boersma and D. Weenink, 2024. *Praat: doing phonetics by computer* [Computer program]. Version 6.4.03, retrieved 4 January 2024, from http://www.praat.org/
[4] R. Correte, *Praat vocal toolkit*, [computer program], 2021–2022 https://www.praatvocaltoolkit.com/index.html (Last viewed Dec 08, 2022).

April 26

Machine learning-based prediction of SPL from healthy and pathological speech signals

Paavo Alku, Manila Kodali, Sudarsana Reddy Kadiri

Aalto University, Finland

In everyday life, speakers regulate vocal intensity on many occasions to emphasise something or to be heard over a long distance. Vocal intensity is typically quantified using sound pressure level (SPL). In fundamental research of voice production, speech recordings are mostly conducted using a constant mouth-to-mic distance and by recording a standard calibration tone prior to voice recordings. This procedure enables computing the SPL of the recorded speech by comparing the RMS (root mean square) of the speech sample with that of the calibration tone. Unfortunately, speech recordings are today mostly conducted in speech technology research without recording the SPL calibration tone. Therefore, the original vocal intensity information, including SPL, is not available in most current speech databases and therefore speech signals of these databases are presented on arbitrary amplitude scales (e.g. by scaling the maximum amplitude value of the sound waveform to be 1.0). Despite being presented on an arbitrary amplitude scale, the speech waveform, however, contains acoustic cues about vocal intensity. This enables using machine learning (ML) in the automatic classification of speech intensity mode (i.e. a multi-class classification task) and in the prediction of SPL (i.e. a regression task) from speech signals which have been recorded without calibration information.

In the current study, we investigate the automatic ML-based prediction of SPL from speech signals whose original level information has been removed by presenting the waveform on an arbitrary amplitude scale. In other words, we simulate a scenario in which SPL is estimated from speech that has been recorded without an SPL calibration tone. We study this regression problem by comparing many acoustic feature representations as well as ML-based regression models. The speech data comprises sentences produced by healthy speakers and by speakers suffering from heart failure. The results are promising in showing that the best ML system was able to yield a smallish mean absolute error of 2 dB in the prediction of SPL when speech was presented on an arbitrary amplitude scale.

A Crosslinguistic Investigation of Prosodic Patterns Related to Autism Spectrum Disorder

Mari Wiklund[1], Viljami Haakana[1], Ida-Lotta Myllylä[1], Martti Vainio[2]

[1]*Department of Languages, University of Helsinki*
[2]*Department of Digital Humanities, University of Helsinki*

The research project focuses on the prosodic characteristics, such as intonation and stress patterns, of persons with autism spectrum disorder (ASD) whose native language is Finnish or French. The main objective of the project is to define 'prosodic patterns' based on features identified in the speech of persons with ASD independently of their native language.

It is known that the speech of individuals with ASD often has atypical prosodic features. Speakers with ASD exhibit prosodic patterns that stand out as atypical but tend to fall into phonetically definable categories (Wiklund & Vainio 2019). Previous studies suggest that neurotypical (i.e. non-autistic) individuals perceive autistic individuals' speech as atypical and often sounding as if it were produced by a non-native speaker (Wiklund *et al*., 2022). Thus, the prosodic features of speech in individuals with ASD stand out and are often confused with foreign language speakers' speech. A variety of atypical prosodic features related to ASD has been found in prior studies. For example, the speech can be melodically remarkably monotonous (Paul et al. 2005a) or, on the other hand, highly variable (Diehl & Paul 2013), or exceptionally fast, or very quiet. The speech might also have inconsistent pause structure, abnormal stressing, or a nasal voice quality (Shriberg *et al*. 2001). Such features constitute a significant obstacle to the social acceptance of the individual (Paul *et al*. 2005a: 205). Deviant prosodic features may create an immediate impression of "oddness" (Van Bourgondien & Woods 1992) and affect how speakers with ASD are rated in terms of social and communicative competence (Paul *et al*. 2005b).

The speech data of the project consist of spontaneous speech recorded from group therapy sessions in which 11-14-year-old Finnish- and French-speaking boys diagnosed with ASD talk about their lives, both among themselves and with their therapists. Finnish control data was collected from age- and gender-matched participants. French control group data, non-native Finnish data, and data from Finnish-speaking adults will be collected. The methods used in this project include perception tests where subjects evaluate the (a)typicality of speech samples from autistic
and neurotypical speakers. Various acoustic analyses are performed, such as Wavelet spectrum analysis (see *e.g*. Kallio *et al*. 2020) and principal component analysis of various characteristics measured from speech, including but not limited to tonal movement and amount of vocal fry (Lohi 2020).

Our results suggest that the perceptions of atypicality may be due to sing-song-like or bouncing pitch, disconnected speech rhythm, large pitch excursions or flatness of pitch, and atypical voice quality. The speech of the speakers with ASD has frequently been thought to have been produced by a non-native speaker of Finnish, although all the speakers were native speakers. (Wiklund *et al*. 2022) Principal component analysis has been able to separate the autistic group and the control group perfectly (Lohi 2020: 29). Further research focuses on how autistic listeners perceive the speech of autistic speakers in comparison to neurotypical listeners, how autistic speakers use gestures alongside prosody and how the atypical prosodic features of could be classified into more detailed categories.

References

Diehl, J.J. & Paul, R. (2013). Acoustic and perceptual measurements of prosody production on the profiling elements of prosodic systems in children by children with autism spectrum disorders. *Applied Psycholinguistics* 34, 135–161.

Kallio, H., Suni, A., Šimko, J., & Vainio, M. (2020). Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics*, 80, 100966.

Lohi, N. (2020). *Prosody of adolescent Finnish boys diagnosed with ASD: variation within the group and comparison with neurotypically developed adolescents*. MA thesis, University of Helsinki

Paul, R., Augustyn, A., Klin, A. & Volkmar, F. R. (2005a): Perception and Production of Prosody by Speakers with Autism Spectrum Disorders. *J. of Autism and Dev. Disorders*, 35(2), 205-220.

Paul, R., Shriberg, L., Mcsweeney, J., […], Volkmar, F. R. (2005b): Brief report: relations between prosodic performance and communication and socialization ratings in high functioning speakers with autism spectrum disorders. *J. of Autism and Dev. Disorders*, 35(6), 861-869.

Shriberg, L., Paul, R., Mcsweeny, J., Klin, A., Cohen, D., & Volkmar, F. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. Journal of speech, language, and hearing research: *JSLHR*, 44 5, 1097-115.

Van Bourgondien, M. E. & Woods, A. V. (1992): Vocational possibilities for high functioning adults with autism. In: Schopler, E. & Mesibov, G. (Eds.), *High functioning individuals with autism.* New York: Plenum Press, pp. 227-242.

Wiklund, M. & Vainio, M. (2019): Pitch-related features in the speech of Finnish- and French-speaking boys with autism in data coming from group therapy sessions. In: Lenk, H. E. H., Härmä, J., Sanromán Vilas, B. & Suomela-Härmä, E. (eds.), *Studies in Comparative Pragmatics*. Cambridge Scholars Publishing, Newcastle upon Tyne, 45-63.

Wiklund, M.; Vainio, L.; Saalasti, S. & Vainio, M. (2022): Puheen prosodian havaittu epätyypillisyys suomenkielisillä autismikirjon varhaisnuorilla [Perceived Atypicality of Speech Prosody of Finnish-Speaking Preadolescents with Autism Spectrum Disorder]. *Puhe ja kieli* 42(4), 309-330. (https://doi.org/10.23997/pk.127455)

A cross-linguistic study of spatial sound symbolism

Alexandra Wikström[1], Lari Vainio[1,2], and Martti Vainio[1]

[1]*Phonetics and Speech Synthesis Research Group, Department of Digital Humanities, University of Helsinki*
[2]*Perception, Action & Cognition Research Group, Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki*

Varying kinds of sound symbolic effects have been uncovered and researched in the past century, showing a non-arbitrary link between phonetic features and the meanings that they convey. This study investigated the association between spatial concepts and articulatory pat- terns across a wide array of languages. Recent behavioral evidence suggests that front vowels are associated with *front* concepts and that back vowels are associated with *back* concepts [1], while close vowels are associated with *high* concepts and open vowels are associated with *low* concepts [2]. Hence, this study aimed to examine if similar patterns were to be found from the vocabularies of the world's languages via the Database of Cross-Linguistic Colexifications (CLICS) [3].

Experiment 1 focused on inspecting the frequency counts of vowels and consonants in words pertaining to the concepts of *front* and *back* from 266 languages. It was found that front vowels and labiodental consonants appear more frequently with *front* concepts, whereas back vowels and uvular and velar consonants appear more in tandem with *back* concepts. In Exper- iment 2, the frequency counts of vowels and consonants in words pertaining to the concepts of *high* and *low* were analyzed from 517 languages. The vowel data mainly indicated an associa- tion of central to back, mid to open vowels with *low* concepts, while the consonant data showed prevalence of consonants produced with the body of the tongue active in *high* concepts and a prevalence of consonants produced with the body of the tongue inactive in *low* concepts.

The findings of this study are mainly in line with the data of behavioral experiments, and bring about evidence on how sound-space symbolic effects appear in consonants. As a whole, the patterns found in this study suggest a non-arbitrary relationship between sound and mean- ing that transcends linguistic boundaries, pointing towards a universal inclination for certain articulatory gestures to be associated with specific spatial meanings. The implications and causes of these effects are discussed both in regards to articulation and acoustics. Based on the current and previous research, further exploration into the involvement of different phonetic features in spatial sound symbolism is warranted, especially behavioral research into the role of consonants, which has not yet been conducted. A better understanding of sound symbolism altogether provides insight into its involvement in language evolution and human cognitive development, which in turn has implications on, for example, language learning and natural language modeling.

References

[1] Vainio, L., Kilpeläinen, M., Wikström, A., and Vainio, M. "Sound-space symbolism: Associating articulatory front and back positions of the tongue with the spatial concepts of forward/front and backward/back". *Journal of Memory and Language* 130 (2023), p. 104414. DOI: https://doi.org/10.1016/j.jml.2023.104414.

[2] Vainio, L., Wikström, A., Repetto, C., and Vainio, M. "Sound-symbolic association between speech sound and spatial meaning in relation to the concepts of up/down and above/below". *Language and Cognition* (2023), pp. 1–20. DOI: https://doi.org/10.1017/langcog.2023.31.

[3] Rzymski, C., Tresoldi, T., et al. *The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies*. 2019. URL: clics.clld.org.

Language Identification is Difficult for Non-native Speech

Tanel Alumäe

Laboratory of Language Technology, Tallinn University of Technology

Identifying languages from speech is a crucial initial step in many systems for processing spoken language. In recent years, the accuracy of state-of-the-art language identification systems has improved rapidly. This improvement is mainly due to the use of self-supervised models that are pretrained on multilingual data and the use of large training datasets, such as VoxLingua107 (Valk & Alumäe, 2021). This presentation demonstrates that when dealing with speech containing non-native or regional accents, or speech from highly conversational scenarios, the accuracy of spoken language identification systems significantly decreases. It also reveals that the accuracy of identifying the language is inversely correlated with the strength of the accent.

References

Valk, Jörgen; Alumäe, Tanel (2021). VoxLingua107: A dataset for spoken language recognition. 2021 IEEE Spoken Language Technology Workshop (SLT), SLT 2021: Proceedings, January 19-22, 2021, Online Conference. Piscataway, NJ: IEEE, 652−658. DOI: 10.1109/SLT48900.2021.9383459.

Neural Text-to-Speech for North Sámi: development and evaluation

Katri Hiovain-Asikainen[1], Antti Suni[2] and Sébastien Le Maguer[2]

UiT The Arctic University of Norway[1], The University of Helsinki[2]

Since the publication of the first North Sámi Text-to-Speech (TTS) as a closed-source project by the Divvun group and Acapela in 2015, the field of speech technology has been developing at an accelerating speed reaching higher quality. However, modeling low resource languages, such as the endangered North Sámi language, remains a challenge [6, 7]. The current presentation describes the steps we have taken in implementing a modern neural Text-to-Speech model for North Sámi language.

To achieve this, we reused and augmented the previously used North Sámi speech corpus and trained a new model using a transformer-based speech synthesis framework: FastPitch [1]. In comparison to other models such as Tacotron 2 [2, 6], FastPitch has several advantages: it is more controllable and requires less training data, both training and synthesis is fast, yet it still produces high-quality synthetic speech. As the original Acapela corpus contained only 4,3 hours of speech, we recorded 3,4 more hours of speech from the same speaker but in different recording conditions, resulting in an augmented speech corpus of altogether 7.7 hours.

To explore the capabilities of modern neural TTS models with the North Sámi language, we experimented with two North Sámi FastPitch variants, trained with different subsets of the speech corpus: 1) model trained with the original Divvun/Acapela speech corpus only and 2) model trained with the augmented speech data. We then conducted a subjective evaluation relying on standard protocols used to evaluate synthetic speech [3, 4] and analyzed the result using a Bonferroni corrected Wilcoxon ranking test as recommended in [5]. We have focused our evaluation and analysis to answer two questions: do our neural TTS systems produce a more comprehensible and more pleasant synthetic voice than the historical systems? To which extent the variations between the training datasets impact the quality of the synthesized speech?

The results of the Mean Opinion Score (MOS) test show that the new neural TTS is found to be better than the previous North Sámi TTS by Acapela and Divvun. The results of this test also suggest that recording and adding more speech data (despite the different recording conditions) for training instead of only training with the original Acapela North Sámi TTS corpus significantly increased the comprehensibility and pleasantness of the synthesized speech. Our work shows that even very small training corpora can be used to achieve end-user suitable TTS for low-resource languages.

References

[1] Łańcucki, A. (2021, June). Fastpitch: Parallel text-to-speech with pitch prediction. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6588-6592). IEEE.

[2] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., ... & Wu, Y. (2018, April). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 4779-4783). IEEE.

[3] ITU, Methods for subjective determination of transmission quality, ITU-T Recommendation P.800, International Telecommunication Union (ITU-P), Geneva (1996).

[4] ITU-T, A method for subjective performance assessment of the quality of speech voice output devices, Tech. Rep. P.85, International Telecommunication Union (ITU-R) (1994).

[5] R. A. Clark, M. Podsiadlo, M. Fraser, C. Mayo, S. King, Statistical analysis of the blizzard challenge 2007 listening test results, in: The Blizzard Challenge Workshop, 2007, http://festvox.org/blizzard/bc2007/blizzard 2007/full papers/blz3 003.pdf.

[6] Makashova, L. (2021). SPEECH SYNTHESIS AND RECOGNITION FOR A LOW-RESOURCE LANGUAGE Connecting TTS and ASR for mutual benefit.

[7] Rätsep, L., & Fishel, M. (2023, May). Neural Text-to-Speech Synthesis for Võro. In Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa) (pp. 723-727).

Exploring the intersections of social dimensions and acoustics in Finnish text-to- speech synthesis

Tuukka Törö, Antti Suni, Juraj Šimko

University of Helsinki

State-of-the-art text-to-speech (TTS) synthesizers are able to elicit expressive speech in various speaking styles, such as angry, sad and happy. This can be achieved by style labels, acoustic prompting or by direct manipulation of their latent embedding spaces (Li *et al*., 2022, Skerry-Ryan *et al*., 2018). Latent space manipulation can also be used for explicit control of acoustic features of the output such as f_0, speaking rate and voice quality (Šimko *et al*., 2023). While these models offer controllability and expressiveness, addressing language variation becomes essential to cater to diverse user needs.

We posit that while social dimensions affect speech in complex ways – with their fuzzy and overlapping limits – the latent embeddings of a neural TTS system grasp acoustic underpinnings behind them. To explore this, we trained a multi-speaker TTS model with the Donate Speech Corpus (Moisio *et al*., 2023), a large dataset of colloquial Finnish accompanied with metadata such as dialect region, age, gender and education level.

Leveraging linear regression models with latent space embeddings as explanatory variables, we identified directions between social categories, allowing for interpolation to control the output of the synthesizer.

Our presentation will delve into the challenges encountered when training TTS models with 'wild' data, emphasizing the importance of addressing variability in real-world speech. We will explore whether differences between social categories can be discerned purely based on acoustic qualities and discuss the potential research questions that TTS systems can help answer. By shedding light on these aspects, our work contributes to the broader understanding of how social dimensions and their interplay affect speech.

References

Li, T., Wang, X., Xie, Q., Wang, Z., & Xie, L. (2022). Cross-speaker emotion disentangling and transfer for end-to-end speech synthesis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 1448-1460.

Moisio, A., Porjazovski, D., Rouhe, A., Getman, Y., Virkkunen, A., AlGhezi, R., ... & Kurimo, M. (2023). Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks. *Language Resources and Evaluation*, *57*(3), 1295-1327.

Skerry-Ryan, R. J., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., ... & Saurous, R.A. (2018, July). Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning* (pp. 4693-4702). PMLR.

Šimko, J., Törö, T., Vainio, M., & Suni, A. (2023). Prosody under control: Controlling prosody in text-to-speech synthesis by adjustments in latent reference space. In *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 3086-3090).

IAR: Algoritmi epäkonsistenttien annotaatioiden siistimiseksi diskriminatiivisia luokittimia käyttäen

Einari Vaaras[1], Manu Airaksinen[2] & Okko Räsänen[1]

[1]Tietotekniikan yksikkö, Tampereen yliopisto
[2]BABA Center, Lastentautien tutkimuskeskus, Uusi lastensairaala, HUS

Diskriminatiiviset koneoppimisluokittimet kuten neuroverkot pyrkivät luomaan epälineaarisia luokitusrajapintoja piirreavaruuteen siten, että halutut kohdeluokat voidaan erottaa mahdollisimman hyvin toisistaan [1]. Tällaiset luokittimet hyötyvät konsistenteista opetusannotaatioista [2, 3], ja korkeat erimielisyydet annotaattoreiden välillä johtavat heikkoon luokittelutarkkuuteen opetetuilla luokittimilla [3, 4]. Esimerkiksi lääketieteen eri sovelluskohteissa annotoitavat ilmiöt ovat usein monitulkintaisia, ja täten edes asiantuntijatkaan eivät aina ole täysin yksimielisiä tulkinnoissaan. Tämä puolestaan voi johtaa korkeisiin erimielisyyksiin annotaattoreiden välillä, mikä toisaalta myös johtaa heikompaan koneoppimismallien luokittelutarkkuuteen kuin mitä olisi mahdollista saavuttaa konsistentimmeilla annotaatioilla [5, 6].

Tässä työssä esittelemme uuden algoritmin nimeltään Iterative Annotation Refinement (IAR), jonka avulla luodaan epäkonsistenteista annotaatioista konsistentimpia diskriminatiivisia koneoppimisluokittimia varten. IAR:ssä perusajatus on yhdistää ihmisannotoijien luomat alkuperäiset annotaatiot opetetun koneoppimismallin tuottamien posterior-todennäköisyyksien kanssa uusiksi annotaatioiksi. Tämä on iteratiivinen prosessi, jossa ensin opetetaan koneoppimisluokitin alkuperäisillä ihmisannotaatioilla, jonka jälkeen opetetun luokittimen avulla luodaan uudet annotaatiot summaamalla alkuperäiset annotaatiot ja luokittimen ennusteet yhteen. Sitten sama prosessi toistetaan uudestaan ja uudestaan käyttäen aina edellisen iteraation tuottamia annotaatioita opetusaineistona luokittimelle.

Aiemmassa tutkimuksessa [7] tarkastelimme eri tapoja kehittää automaattista puheen emootiontunnistinta vastasyntyneiden teho-osastonauhoitteiden emotionaalisen sisällön analysointiin. Tässä työssä kehittämäämme IAR-algoritmia sovellettiin sekä Turun yliopistollisen keskussairaalan että Tallinnan lastensairaalan vastasyntyneiden teho-osastojen nauhoitteiden analysointia varten kehitettäviin koneoppimisluokittimiin. Näitä emootioluokittimia oli yhteensä kahdeksan kappaletta: suomenkielinen ja vironkielinen luokitin emotionaalisen valenssin (engl. valence) ja virittyneisyyden (engl. arousal) automaattiseen tunnistamiseen sekä mies- että naispuhujille. Lopputuloksena saimme parannettua luokitustarkkuutta kaikilla luokittimilla IAR:n avulla, mikä osoittaa sekä algoritmin toimivuuden käytännön sovelluskohteessa että konsistenttien annotaatioiden hyödyllisyyden koneoppimisluokittimien opetuksessa. Lisäksi suorittamissamme kokeissa simuloidulla aikasarjadatalla havaitsimme, että on ainoastaan kourallinen tapauksia joissa IAR-algoritmi epäonnistuu, sekä myös että näitä tapauksia esiintyy melko harvoin realistisissa datoissa.

Viitteet

[1] T. Mitchell, “Machine Learning”, 1st edition, USA: McGraw-Hill, Inc., 1997.
[2] H. Song, M. Kim, D. Park, Y. Shin, ja J. Lee, “Learning from Noisy Labels with Deep Neural Networks: A Survey”, IEEE Trans. on Neural Networks and Learning Systems, vol. 34, no. 11, pp. 8135–8153, 2023.
[3] B. Han, Q. Yao, X. Yu et al., “Co-teaching: robust training of deep neural networks with extremely noisy labels”, Proc. NeurIPS, pp. 8536–8546, 2018.
[4] C. Northcutt, L. Jiang, ja I. Chuang, “Confident Learning: Estimating Uncertainty in Dataset Labels”, Journal of Artificial Intelligence Research, vol. 70, pp. 1373–1411, 2021.
[5] B. Farzin, R. Fahed, F. Guilbert et al., “Early CT changes in patients admitted for thrombectomy: Intrarater and interrater agreement”, vol. 87, no. 3, pp. 249–256, 2016.
[6] N. Stevenson, R. Clancy, S. Vanhatalo, I. Rosén, J. Rennie, ja G. Boylan, “Interobserver agreement for neonatal seizure detection using multichannel EEG”, vol. 2, no. 11, pp. 1002–1011, 2015.
[7] E. Vaaras, S. Ahlqvist-Björkroth, K. Drossos, L. Lehtonen, ja O. Räsänen, “Development of a speech emotion recognizer for large-scale child-centered audio recordings from a hospital environment”, Speech Communication, vol. 148, pp. 9–22, 2023.

Pronunciation Practicing App for Children Learning Nordic Languages

Yaroslav Getman[1], Nhan Phan[1], Ragheb Al-Ghezi[1], Ekaterina Voskoboinik[1], Tamás Grósz[1], Mikko Kurimo[1], Xinwei Cao[2], Giampiero Salvi[2], Torbjørn Svendsen[2], Sofia Strömbergsson[3], Anne Marte Haug Olstad[4], MInna Lehtonen[4], Anna Smolander[5], and Sari Ylinen[5]

[1]Aalto University, Finland; [2]Norwegian University of Science and Technology, Norway; [3]Karolinska Institutet, Sweden; [4]University of Oslo, Norway; [5]Tampere University, Finland

Computer-assisted pronunciation training (CAPT) is a rapidly developing area accelerated by advancements in the field of AI. With a well-designed and reliable mobile CAPT application students can practice pronunciation outside of the classroom at any time and place. Furthermore, using captivating games in mobile applications has shown encouraging results on learning outcomes by motivating young users to practice more and perceive pronunciation learning as a positive experience (*Junttila et al*., 2022; Uther *et al.*, 2018). In the TEFLON project, we have collected data and developed speech technology to create a Pop2Talk-Nordic mobile app which is an online pronunciation practice system for young children learning Finnish, Swedish, and Norwegian. In addition to foreign and second language pronunciation, children with speech sound disorders can use it as part of their speech therapy.

Our experimental results show that by fine-tuning multilingual speech models (Baevski *et al.*, 2020) we can develop speech recognizers and pronunciation rating systems without collecting excessive amounts of new speech data (Getman *et al*., 2023b). Additionally, the results reveal that by fine-tuning the models in a multi-task manner, we can integrate large speech models into an online game as they simultaneously perform speech recognition and pronunciation ratings (Getman *et al*., 2023a). In addition to evaluations of our speech recognizers and pronunciation raters, we describe our experience in collecting a number of corpora of Finnish, Swedish, and Norwegian spoken by children. The challenges encountered include not only recording and annotating the data but also making it publicly available. We hope that sharing our experience will help others to collect and publish similar corpora for other languages. So far, we have been able to make our models and the Norwegian corpus publicly available to serve as an example and benchmark for developing children's speech recognition and pronunciation rating systems for low-resource tasks.

Viitteet

Katja Junttila, Anna-Riikka Smolander, Reima Karhila, Anastasia Giannakopoulou, Maria Uther, Mikko Kurimo, and Sari Ylinen. Gaming enhances learning-induced plastic changes in the brain. Brain and Language, 230:105124, 2022.

Maria Uther, Anna-Riikka Smolander, Katja Junttila, Mikko Kurimo, Reima Karhila, Seppo Enarvi, and Sari Ylinen. User Experiences from L2 Children Using a Speech Learning Application. Advances in Human Computer Interaction, 7345397:6, 2018.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec2.0: A framework for self-supervised learning of speech representations. In Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Yaroslav Getman, Ragheb Al-Ghezi, Tamas Grosz, and Mikko Kurimo. 2023a. Multi-task wav2vec2 Serving as a Pronunciation Training System for Children. In Proc. 9th Workshop on Speech and Language Technology in Education (SLaTE), pages 36–40.

Yaroslav Getman, Nhan Phan, Ragheb Al-Ghezi, Ekaterina Voskoboinik, Mittul Singh, Tamás Grósz, Mikko Kurimo, Giampiero Salvi, Torbjørn Svendsen, Sofia Strömbergsson, Anna Smolander, and Sari Ylinen. 2023b. Developing an AI-assisted low-resource spoken language learning app for children. IEEE Access, 11:86025–86037.

Training with the Estonian pronunciation app SayEst: vowel perception and user experience

Anton Malmi, Katrin Leppik

University of Tartu

SayEst is an Android mobile app designed to help users improve their Estonian pronunciation. Previous studies have shown that such tools and phonetic speech training can help to improve pronunciation (Savo & Peltola 2019; Taimi *et al*. 2014; Leppik & Tejedor-García 2019; Hacking, Smith & Johnson 2017). The app is available in English and Russian, and it focuses on vowels and consonants. It has three different exercises: exposure, discrimination, and pronunciation, and theoretical videos explaining the pronunciation of Estonian vowels and consonants.

To assess the efficacy of the app, we asked 30 Russian L1 learners of Estonian (mean age 31, SD = 9.3) with different proficiency levels (A - beginner, B - intermediate, C - advanced) to participate in a pre- and post-test design study involving an unsupervised training period. First, the participants came to the University of Tartu to complete two perception tasks (vowel identification and minimal pairs identification) and a reading task. The participants' reaction time was recorded during the perception tasks. The reading task consisted of a list of 180 short words that were similar to the words used in the app. The same procedure was repeated after the participant finished the training period. The training period ended when the participant finished all the exercises in the app.

The participants filled out a self-assessment questionnaire before and after using the app and were interviewed after the training. The results showed that the self-rating of the participants' pronunciation improved after using the app; they became more confident in speaking Estonian and paid more attention to how they spoke. The participants liked the exposure and discrimination task but struggled with the production exercise.

The results of the vowel identification task show that the learners improved slightly in the vowel identification task during the training period. The learners with lower proficiency levels (Group A) of Estonian made more mistakes in vowel identification. They had a lower percentage of correct answers compared to learners with higher proficiency levels (Groups B and C). The learners misidentified the vowels /ɤ/ and /y/ most often, these vowels were confused with each other, /u/ and /i/. The longer the reaction time, the higher the probability of misidentification, regardless of the learners' proficiency level.

The answers for the perception test of minimal pairs showed that there were no differences between the pre- and post-tests. The answers were pooled together, and the results showed that the overall percentage of misidentification decreased with proficiency level. Group A misidentified the highest percentage of pairs. Groups B and C were similar and made only a few mistakes. Incorrect answers correlated with slower reaction times.

References

Hacking, Jane F., Bruce L. Smith & Eric M. Johnson. 2017. Utilizing electropalatography to train palatalized versus unpalatalized consonant productions by native speakers of American English learning Russian. *Journal of Second Language Pronunciation*. John Benjamins Publishing Company 3(1). 9–33. https://doi.org/10.1075/jslp.3.1.01hac.

Leppik, Katrin & Cristian Tejedor-García. 2019. Estoñol, a computer-assisted pronunciation training tool for Spanish L1 speakers to improve the pronunciation and perception of Estonian vowels. *Eesti ja soome-ugri keeleteaduse ajakiri. Journal of Estonian and Finno-Ugric Linguistics* 10(1 SE-Articles). 89–104. https://doi.org/10.12697/jeful.2019.10.1.05.

Savo, Satu & Maija S. Peltola. 2019. Arabic speakers learning Finnish vowels: short-term phonetic training supports second language vowel production. *Journal of Language Teaching and Research*. Academy Publication 10(1).45. https://doi.org/10.17507/jltr.1001.05.

Taimi, Laura, Katri Jähi, Paavo Alku & Maija S Peltola. 2014. Children learning a non-native vowel - The effect of a two-day production training. *Journal of Language Teaching and Research* 5(6). 1229–1235. https://doi.org/10.4304/jltr.5.6.1229-1235.

Enhancing Speech Emotion Recognition through Word Informativeness

Sofoklis Kakouros

University of Helsinki, Finland

The study of emotion recognition from speech encompasses a crucial challenge: the iden- tification of speech signal segments carrying acoustic variations that are most relevant for distinguishing specific emotions. Traditional methodologies in this field have focused on computing metrics, such as feature functionals, for features like energy and fundamental frequency (f_0) over entire sentences or even longer speech segments. These functionals, for example, encompass the first and second-order statistics of the features. However, applying these statistics across larger speech segments risks diminishing fine-grained variation, as they tend to average out the details, leading to a smoothing effect that could obscure subtle but important characteristics. Thus, this approach, risks overlooking the nuanced variations present within speech that are critical for accurately identifying emotions.

This research presents a novel approach to speech emotion recognition (SER) by incor- porating the concept of word informativeness, as derived from a pre-trained language model, to accurately identify semantically important segments within speech. By focusing on these segments, the method computes acoustic features specifically for these parts of the speech signal aiming at capturing the most relevant information for conveying emotional states. This technique diverges from traditional methods that uniformly analyze entire sentences or larger speech portions. The utilization of word informativeness allows for a more targeted analysis, ensuring that the acoustic features such as f_0 and energy, along with their functionals, are calculated where they are most likely to contribute to the accurate recognition of emotions. This refined focus on semantically important segments leverages the inherent linguistic context to guide the emotional analysis.

The results indicate an improvement in recognition performance when features are com- puted on segments selected based on word informativeness. This highlights the potential of leveraging state-of-the-art large language models to refine the process of emotion recognition in speech. This method not only improves the accuracy of emotion recognition but also sets a new standard for how linguistic and acoustic information can be synergistically used to enhance the understanding of emotional expressions in speech.

Perception and Production of Quantity in Finnish with Predictive Processing

Xinyuan Wan, Lenka Kalvodová, Juraj Šimko

University of Helsinki

This pilot study examined whether training solely in speech perception has any effect on the production of Finnish quantity in Mandarin speakers (L2 group).

Various current cognitive models [1, 5, 6, 7] have investigated the relationship between speech perception and production. Some (e.g., the predictive processing framework [1]) suggest that the processes of speech production (action) and perception share critical stages at the higher cognitive level of speech. This means that phonologisation only takes place once. Therefore, we hypothesise that once phonological categories of Finnish quantity are formed in perception, they should also manifest in production, causing an improvement in both perception and production.

Finnish has a two-way quantity system of long and short for both vowels and consonants [8]. These are distinguished using phonological (categorical) knowledge. We generated continua based on the duration of the target words produced by native Finnish (L1) speakers and trained the L2 group in distinguishing Finnish quantity [4]. For this, we used a repeated two-alternative forced choice identification task with the target words embedded in a carrier sentence [3, 4]. Immediate feedback [2] based on L1 speakers' responses in the same task was given after each trial. As for production data, we recorded the L2 group reading aloud the same sentences before and after perception training. We then compared the L2s' quantity ratios in production before and after training with each other, as well as with the L1 speakers' ratios. The L2's production improved significantly, although there were differences between target phonemes. The results for L2 perception indicated improvement, too, but were not significant for all target phonemes. Based on this pilot, a formal experiment is being conducted. We also plan to run a 'vice versa' experiment in the future by training production only.

References

[1] Clark, Andy. "Whatever next? Predictive brains, situated agents, and the future of cognitive science." Behavioral and brain sciences 36.3 (2013): 181-204.

[2] Lee, Andrew H., and Roy Lyster. 'Can Corrective Feedback on Second Language Speech Perception Errors Affect Production Accuracy?' *Applied Psycholinguistics* 38, no. 2 (March 2017): 371–93. https://doi.org/10.1017/S0142716416000254.

[3] Nagle, C. L. (2021). Revisiting perception-production relationships: Exploring a new approach to investigate perception as a time- varying predictor. Language Learning, 71(1), 243-279.

[4] Nagle, Charles L., and Melissa M. Baese-Berk. 'ADVANCING THE STATE OF THE ART IN L2 SPEECH PERCEPTION- PRODUCTION RESEARCH: REVISITING THEORETICAL ASSUMPTIONS AND METHODOLOGICAL PRACTICES'. *Studies in Second Language Acquisition* 44, no. 2 (May 2022): 580–605. https://doi.org/10.1017/S0272263121000371.

[5] Pardo, Jennifer S., Lynne C. Nygaard, Robert E. Remez, and David B. Pisoni, eds. 'Front Matter'. In *The Handbook of Speech Perception*, 1st ed. Wiley, 2021. https://doi.org/10.1002/9781119184096.fmatter.

[6] Pickering, Martin J., and Simon Garrod. 'An Integrated Theory of Language Production and Comprehension'. *Behavioral and Brain Sciences* 36, no. 4 (August 2013): 329–47. https://doi.org/10.1017/S0140525X12001495.

[7] Schwartz, Jean-Luc, et al. "The Perception-for-Action-Control Theory (PACT): A perceptuo-motor theory of speech perception." Journal of Neurolinguistics 25.5 (2012): 336-354.

[8] Vainio, M., Järvikivi, J., Aalto, D., & Suni, A. (2010). Phonetic tone signals phonological quantity and word structure. *The Journal of the Acoustical Society of America*, *128*(3), 1313-1321.

Prosodic Analysis of Speech Fluency in Finnish-Speaking Elderly Women: Exploring Acoustic Correlations and Markers

Marianne Kosin[1], Heini Kallio[2], Tiina Ihalainen[1], Nelly Penttilä[1]

[1]*Logopedics, Faculty of Social Sciences, Tampere University*
[2]*Phonetics, Faculty of Information Technology and Communication Sciences, Tampere University*

Motivated by recent advancements in acoustic and disfluency-based analyses (Deng et al., 2020; Mehrotra et al., 2021; Vincze et al., 2021; Yuan et al., 2020), this study investigates the prosodic features of speech in Finnish-speaking elderly women to discern potential acoustic markers indicating disfluencies or correlating with disfluency frequency and specific types of disfluencies. Building on recent progress in identifying speech irregularities and their links to cognitive decline, our research aims to contribute valuable insights into the acoustic parameters associated with disfluent speech.

The research focuses on the prosodic characteristics of speech in a cohort of 17 healthy, Finnish-speaking women aged over 65. Participants engaged in a semi-spontaneous speech task, narrating a comic strip. Speech samples were annotated to the syllable level using Praat software. Disfluent speech segments were categorized, including filled pauses (e.g., "uh"), interjections (e.g., "well"), interruptions, reformulations, and repetitions. Additionally, silent pauses (>200ms) were marked. Subsequently, an acoustic-phonetic analysis was conducted, encompassing various measurements of f0 change, rhythm, speech fluency, and prominence across the entire speech signal. Acoustic parameters were then examined for potential correlations with disfluency frequency and specific disfluency types.

The study aims to identify acoustic markers and correlations that may indicate disfluencies, providingvaluable insights into the acoustic parameters associated with various types and frequencies of speechdisruptions. This research could contribute to a deeper understanding of impaired speech by uncovering potential acoustic cues linked to linguistic challenges, ultimately aiding in the identification of language impairments and cognitive decline solely through acoustic signals.

References

Deng, H., Lin, Y., Utsuro, T., Kobayashi, A., Nishizaki, H., & Hoshino, J. (2020). Integrating Disfluency-based and Prosodic Features with Acoustics in Automatic Fluency Evaluation of Spontaneous Speech. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pages 6429–6437, Marseille, France. European Language Resources Association.

Mehrotra, U., Garg, S., Krishna, G., & Vuppala, A. K. (2021). Detecting Multiple Disfluencies from Speech using Pre- linguistic Automatic Syllabification with Acoustic and Prosody Features. 2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Tokyo, Japan, 2021, pp. 761-768.

Vincze, V., Szatlóczki, G., Tóth, L., Gosztolya, G., Pákáski, M., Hoffmann, I., & Kálmán, J. (2021). Telltale silence: temporal speech parameters discriminate between prodromal dementia and mild Alzheimer's disease, Clinical Linguistics& Phonetics, 35:8, 727-742.

Yuan, J., Bian, Y., Cai, X., Huang, J., Ye, Z., Church, K. (2020). Disfluencies and Fine-Tuning Pre-Trained Language Models for Detection of Alzheimer's Disease. Proc. Interspeech 2020, 2162-2166, doi: 10.21437/Interspeech.2020-2516.

Rare stress-induced lengthening in unstressed syllables in Finnic and Tlapanec: challenges for theory and typology

Natalia Kuznetsova[1], Oscar Cornelio Tiburcio[2], Hiroto Uchihara[3]

[1] Università Cattolica del Sacro Cuore, Milan, Italy; [2] Centro de Investigaciones y Estudios Superiores en Antropología Social, Mexico; [3] Tokyo University of Foreign Studies, Japan

Stress implies metrical inequality of syllables, namely that one syllable is prosodically highlighted over others in a given prosodic domain. Recent studies (Himmelmann, 2023; Ladd & Arvaniti, 2023) list various dimensions and meanings of stress and note that this complex notion is mostly based on Germanic languages, while few other language groups provide evidence for all the dimensions.

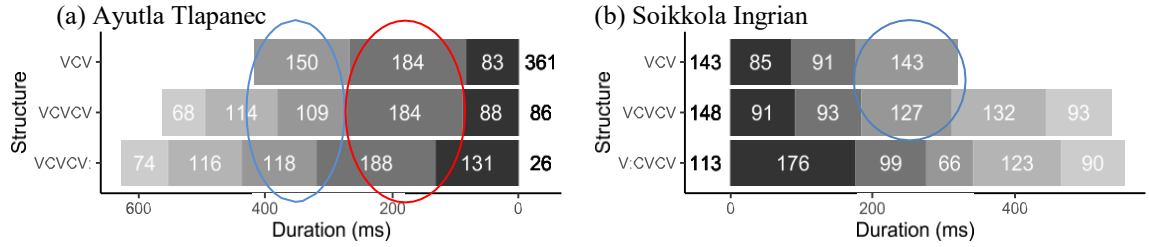
Of the acoustic correlates of lexical stress, duration has been statistically found to be the most cross-linguistically reliable measure (Gordon & Roettger, 2017, p. 8), and Himmelmann (2023) cites it as the most obvious candidate for a robust typological study on lexical stress. The common implicit assumption about stress (shared also by Articulatory Phonology, see e.g. Tilsen, 2019, p. 50) is that stress-induced durational lengthening should appear within the stressed syllable. There are, however, exceptions to this. Our talk discusses two such exceptions attested in Finnic (Uralic) and Tlapanec (Otomanguean) language groups (Kuznetsova et al., under subm.). We present our experimental acoustic field data on Soikkola Ingrian and Ayutla Tlapanec, studied with mixed linear regression statistical modelling. Stress-related lengthening here is attested post-tonically (in the trochaic foot of Soikkola Ingrian) or pre-tonically (in the iambic foot of Ayutla Tlapanec), rather than in the stressed syllable vowel.

In Finnic, lengthening mostly concerns the post-tonic syllable vowels of the trochaic light foot (C)VCV(S). This effect is highlighted in blue on the post-tonic vowels of di-syllabic and trisyllabic light feet (VCV and VCVCV structures) in Figure 1b. The trochaic heavy foot (any other type of foot; exemplified by the V:CVCV structure in 1b), on the contrary, manifests post-tonic vowel reduction.

In Tlapanec, pre-tonic lengthening is found in all kinds of iambic feet and concerns both pre-tonic vowels (blue circle in Figure 1a) and pre-tonic consonants (pink circle in 1a). Phonological quantity contrast exists here only in the vowels of the tonic (word-final) syllable, as illustrated with the VCVCV and VCVCV: structures in 1a. Pre-tonic lengthening effects are all phonetic and stress-induced.

The talk discusses why stress in a structural sense is unanimously considered by the specialists in respective language groups as anchored on a non-lengthened syllable (word-initial in Finnic and word-final in Tlapanec) and why this lengthening can indeed be seen as namely stress-induced. We also draw cross-linguistic parallels to these cases and discuss challenges posed by such placement mismatch between phonetic cues and phonological anchoring to the general theory and typology of metrical stress.

Figure 1. Mean durations of segments in three comparable structures of Tlapanec and Ingrian.



References

Gordon, M. K., & Roettger, T. (2017). Acoustic correlates of word stress: A cross-linguistic survey. *Linguistics Vanguard*, *3*(1), 1–11. https://doi.org/10.1515/lingvan-2017-0007

Himmelmann, N. P. (2023). On the comparability of prosodic categories: Why 'stress' is difficult. *Linguistic Typology*, *27*(2), 341–361. https://doi.org/10.1515/lingty-2022-0041

Kuznetsova, N., Cornelio Tiburcio, O., & Uchihara, H. (under subm.). Stress-induced lengthening in unstressed syllables (Finnic and Tlapanec): Challenge for the theory and typology of metrical prominence. *Language and Speech (Special Issue): Interdisciplinary Approaches to Speech Prosody*.

Ladd, D. R., & Arvaniti, A. (2023). Prosodic prominence across languages. *Annual Review of Linguistics*, *9*(1), 171–193. https://doi.org/10.1146/annurev-linguistics-031120-101954

Tilsen, S. (2019). Space and time in models of speech rhythm. *Annals of the New York Academy of Sciences*, *1453*(1), 47–66. https://doi.org/10.1111/nyas.14102

Utterance-initial intonation peaks in Estonian: A cognitive perspective

Nele Ots

Goethe University of Frankfurt am Main, Germany

The study examined pitch peaks in longer vs. shorter utterances [1, 3, 6, 7, 8], investigating the link between utterance-initial pitch raising and cognitive demands of speech production. Drawing on evidence that voice pitch raises in response to heightened cognitive load [2, 4, 5],the experiment tested whether the observed pitch rise in longer utterances is linked to increasedcognitive efforts in generating messages for sentence production.

To modulate cognitive resources available for message generation in a visual world speech production task, the study implemented a dual-task paradigm. Eighty-four native Estonianspeakers described 64 pictures depicting events with multiple actors. In half of the descrip- tions, participants memorized three nouns, later recalling them and answering related questions.Under the high cognitive load, the language-related memory span was expected to decrease, leading to failure to pre-plan the height of utterance-initial intonation peaks. Thus, the pitch difference between short and long utterances was predicted to diminish or disappear under the high cognitive load.

Unexpectedly, longer utterances consistently exhibited higher pitch peaks, regardless of cog- nitive load. However, under high cognitive load, pitch peaks in both short and long utterances were lower than under low load. In essence, the study identified a lowering effect of cogni-tive load on sentence intonation. This effect tentatively implies interference between the verbal memory task and linguistic planning processes. Otherwise, the pitch peaks should have been higher under cognitive load, as observed in previous studies [2, 4, 5]. The lowering effect of high cognitive load is proposed to indicate restricted activation of lexical-semantic representa- tions due to a smaller working memory span. Consequently, the study suggests that sentence intonation might be sensitive to lexical activation, emphasizing the need for future investigationsinto the relationship between language-related cognitive processes and sentence intonation.

References

[1] Asu, E. L. et al. (2016). "F0 declination in spontaneous Estonian: implications for pitch-related preplanning". In: *Proceedings of Speech Prosody, Boston 31 May – 3 June 2016*.

[2] Huttunen, K. et al. (2011). "Effect of cognitive load on speech prosody in aviation: Evidencefrom military simulator flights". In: *Applied Ergonomics* 42.2.

[3] Liberman, M. & J. Pierrehumbert (1984). "Intonational Invariance under Changes in Pitch Range and Length". In: *Language Sound Structure*. Ed. by M. Aronoff & R. T. Oehrle. Studies inPhonology. Presented to Morris Halle by his Teacher and Students. The Massachusetts Insti- tute of Technology.

[4] Lively, S. E. et al. (1993). "Effects of cognitive workload on speech production: Acoustic analyses and perceptual consequences". In: *The Journal of the Acoustical Society of America* 93.5.

[5] Mersbergen, M. van & A. E. Payne (2020). "Cognitive, Emotional, and Social Influences on Voice Production Elicited by Three Different Stroop Tasks". In: *Folia Phoniatr Logop*.

[6] Prieto, P. et al. (2006). "Evidence for 'soft' preplanning in tonal production: Initial scaling in Romance." In: *Speech Prosody, May 2006, Dresden, Germany*.

[7] Thorsen, N. G. (1985). "Intonation and text in Standard Danish". In: *The Journal of the Acoustical Society of America* 77.3. eprint: https://doi.org/10.1121/1.392187.

[8] Yuan, J. & M. Liberman (2014). "F0 declination in English and Mandarin Broadcast News Speech". In: *Speech Communication* 65.

Segmental and phrasal influences on the allophonic variation in short plosives in Estonian

Liis Ermus

Institute of the Estonian Language, University of Tartu

The allophonic variation of short plosives in Estonian has so far been observed mainly word-medially and in vocal context (Raasik 2010; Ermus 2017), the consonant context has only been touched briefly (Ermus, Mihkla 2019). This study examines how the allophonic variation of short plosives is influenced by the segmental context and the position of the plosive in a word and sentence.

The analysis of natural speech (Ermus 2017) distinguished five main types of allophones: **voiceless**, **partly voiced**, **voiced with burst**, **voiced without burst**, **total loss**. Ermus and Mihkla (2019) searched for segmental and phrase traits that could be used to predict the occurrence of allophones. They found the biggest influence by the segmental context, but the position of the plosive in the word and the position of the word containing the plosive in the phrase also had some importance.

The speech material for this study comes from a male speaker in the Speech Synthesis Corpus of the Institute of the Estonian Language (Piits 2016). I analysed short word-initial, word-medial and word-final short plosives (/p/ N=1372, /t/ N=340, /k/ N=2873). In the main part, the same factors are studied as were used in the study of Ermus and Mihkla (2019), but in this study I looked at all sounds separately to find more fine-grained traits.

In the analysis, I used regression trees (ctree) to identify the most important factors influencing the allophonic variation. Similar to the findings of Ermus and Mihkla (2019), the allophonic division was mainly influenced by the segmental context, but the influence of the location of the plosive also played a role. While the proximity of the vowels mainly affected the voicing (most of the sentence-medial tokens were at least partly voiced), the proximity of the consonants also had an impact on the burst phase. The effect of nasals was particularly strong: the plosive after the nasal was almost always voiced, the plosive before the nasal was mostly without the burst phase, with the effect also occurring over the word boundary.

References

Ermus, Liis. 2017. Eesti keele lühikeste klusiilide häälduse variatsioon ja seda mõjutavad tegurid. *Mäetagused* 68. 27–52. https://doi.org/10.7592/MT2017.68.ermus.

Ermus, Liis & Meelis Mihkla. 2019. Predictability of plosive reduction from written text in Estonian. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*, 2635–2639. Canberra, Australasia: Australasian Speech Science and Technology Association Inc.

Piits, Liisi. 2016. Eesti Keele Instituudi kõnesünteesikorpus. Center of Estonian Language Resources. https://doi.org/10.15155/3-00-0000-0000-0000-05BDAL.

Raasik, Liis. 2010. Intervokaalsete lühikeste klusiilide laad eesti keele spontaankõnes. Master thesis. University of Tartu.

Part-of-speech and quantity interact in predicting acoustic durations of Estonian spontaneous speech

Kaidi Lõo[1], Pärtel Lippus[1], Benjamin V. Tucker[2]

[1]University of Tartu, [2]Northern Arizona University

Spontaneous speech is highly variable (Tucker and Mukai 2023). When controlling for the number of phones, frequency and other predictors, content words are produced with longer durations than function words (Dilts 2013, Gahl et al 2012, Seifart 2018). However, these effects have mainly been investigated in English. The current study focuses on Estonian which has a different word class and phonemic quantity system.

The materials (7158 disyllabic function words and 16566 content words in three quantities) for the study were extracted from the Phonetic Corpus of Estonian Spontaneous Speech (Lippus et al 2021). We excluded words containing preceding and following disfluencies, pauses, and proper nouns as well as words at the beginning and end of the phrase.

For the analysis, we used generalized additive mixed effects models (Wood 2017). Our analyses indicate that content words (nouns and adjectives) are longer than function words (pronouns and conjunctions) and that verbs and adverbs fall between the two classes in Estonian. Whereas words are longer in the second and third quantities than in the first quantity, the effect is not the same for all parts-of-speech. Semantically rich part-of-speech, nouns and adjectives are more affected by the quantity distinction than other parts-of-speech. Further, the duration of both content and function words decreases with increasing predictability and frequency. In summary, we show that although the effects of part-of-speech in Estonian are similar to English, their exact influence is modulated by characteristic properties of the language in use, such as part-of-speech and quantity.

References

P. Dilts (2013). *Modelling phonetic reduction in a corpus of spoken English using random forests and mixed-effects regression*. University of Alberta (Canada).

S. Gahl, Y. Yao, and K. Johnson (2012). Why reduce? phonological neighborhood density and phonetic reduction in spontaneous speech, *Journal of Memory and Language*, vol. 66, no. 4, pp. 789–806.

P. Lippus, K. Aare, A. Malmi, T. Tuisk, and P. Teras (2021). Phonetic Corpus of Estonian Spontaneous Speech V1.2 [Online]. Available: https://www.doi.org/10.23673/RE-293.

F. Seifart, J. Strunk, S. Danielsen … B. Bickel (2018). Nouns slow down speech across structurally and culturally diverse languages, *Proceedings of the National Academy of Sciences*, vol. 115, no. 22, pp. 5720–5725.

B. V. Tucker and Y. Mukai (2023). Spontaneous Speech in *Elements in Phonetics*. Cambridge: Cambidge University Press.

S. N. Wood (2017). *Generalized additive models: an introduction with R*. CRC press.