# XXXVII Fonetiikan päivät
# Turku, 24.-25.4.2025

**Sisällys**

# Ohjelma

**TORSTAI 24.4.2025**

**9:00 Ilmoittautuminen ja posterien esillelaitto**

**10:00 Avaussanat** (vararehtori Maija S. Peltola)

**10:15 Sessio 1**, pj. Minnaleena Toivola
**10:15-10:40** Millaista on helposti ymmärrettävä puhe? Didaktisen kurssin vaikutuksia tulevien kieltenopettajien näkemyksiin (Heinonen et al.)
**10:40-11:05** Vieraskielisen suullisen vuorovaikutuksen automaattinen arviointi – AASIS-hankkeen lähtökohtia ja aineiston esittelyä (Ullakonoja et al.)
**11:05-11:30** Lukio-opiskelijoiden tunnekokemukset ja luokkahuoneessa saatu ääntämispalaute (Virkkunen et al.)
**11:30-11:55** Kielitaidon tukeminen tienä oppimisen yhdenvertaisuuteen ja osallisuuteen (Haapanen et al.)

**11:55 Lounas Galileissa**

**13:00 Sessio 2**, pj. Michael O'Dell
**13:00-13:25** The influence of the duration of voiced consonants on their recognition in sung vowel-consonant and consonant-vowel junctions (Vurma et al.)
**13:25-13:50** Shape-sound symbolism in Finnish and a word learning paradigm (Wikström et al.)
**13:50-14:15** The interlocutor factor and intonational variation in collaborative dialogues (Kachkovskaia)

**14:15 Posterit ja kahvitarjoilu**

**15:30 Sessio 3**, pj. Einar Meister
**15:30-15:55** Suomenruotsin prosodian tuottaminen erikielisillä alkeisoppijoilla (Kallio et al.)
**15:55-16:20** Exploring Speech Prosody of Finnish-Speaking Preadolescents with Autism Spectrum Disorder (ASD) – an Investigation on Silent Pauses and Turn-Maintenance (Myllylä et al.)
**16:20-16:45** Alignment of prosodic prominence and gesture in marking of negation in Estonian: further insights from a multimodal study (Lippus et al.)

**16:45 Vapaa-aikaa**

**19:00 Illallinen Ravintola Gräddassa**
Turun kaupungin tervehdys ja alkumalja
Illallinen
Ohjelmanumero
Jatkot olutravintola 5piste5

**PERJANTAI 25.4.2025**

**9:30 Sessio 4**, pj. Martti Vainio
**9:30-9:55** Emergence of Communication through Predictive Processing and Bayesian Inference (Šimko)
**9:55-10:20** Investigation of speech timing using excitable networks (O'Dell)
**10:20-10:45** Text-to-speech synthesis in infant-directed speaking style using noisy speech data from naturalistic infant-caregiver interactions (Grönlund et al.)
**10:45-11:10** Leveraging LLM in Speaking Assessment and Application for Learning Finnish (Phan et al.)

**11:10 Lounas Galileissa**

**12:15 Sessio 5**, pj. Okko Räsänen
**12:15-12:40** PFML: Self-Supervised Learning of Time-Series Data Without Representation Collapse (Vaaras et al.)
**12:40-13:05** Does multilingual and multi-speaker modeling improve low-resource TTS? Experiments on Sámi languages (Hiovain-Asikainen et al.)
**13:05-13:30** Suomen vokaalikvantiteetin oppiminen arabiankielisillä peliavusteisesti (Vakkilainen et al.)

**13:30 Kahvitauko**

**14:00 Sessio 6**, pj. Pärtel Lippus
**14:00-14:25** Suomenkielisten aikuispuhujien nasalanssiarvoja (Tivonen et al.)
**14:25-14:50** Introducing the FinnAffect Corpus (Lahtinen et al.)
**14:50-15:15** Large-Scale Self-Supervised Speech Representation Learning with Finnish National Archives (Getman et al.)

**15:15 Loppusanat**

# Suulliset esitykset
# Oral presentations

# Millaista on helposti ymmärrettävä puhe? Didaktisen kurssin vaikutuksia tulevien kieltenopettajien näkemyksiin

Henna Heinonen[1], Elina Vasu[1] & Maria Kautonen[2]

1) Tampereen yliopisto 2) Jyväskylän yliopisto

Toisen ja vieraan kielen ääntämisen ja ääntämisen opetuksen keskiössä ajatellaan usein olevan yksittäisten sanojen ja äänteiden hallinta (esim. Virkkunen & Toivola 2020). Äänteiden onnistuminen ei kuitenkaan ole niin ratkaisevassa roolissa puheen ymmärrettävyyden kuin esimerkiksi puheen rytmin ja painotusten onnistuminen (esim. Trofimovich & Isaacs 2012, Field 2005). Tässä esitelmässä kerromme tutkimuksestamme, jossa selvitämme, kuinka tulevien kieltenopettajien käsitykset ääntämisen ymmärrettävyydestä ja siihen vaikuttavista tekijöistä kehittyvät suullisen kielitaidon opetukseen liittyvän kurssin aikana.

Tutkimukseen osallistui maisterivaiheen kielten opettajaopiskelijoita, jotka osallistuivat kurssille, jolla käsiteltiin kattavasti suullisen kielitaidon teemoja niin teoreettisista kuin käytännöllisistä näkökulmista. Myös ymmärrettävyyteen vaikuttavia piirteitä sekä niiden opettamista opiskeltiin ja harjoiteltiin. Tutkimuksessamme selvitimme opiskelijoiden ymmärrettävyyteen liittyviä käsityksiä kyselylomakkeella, johon opiskelijat vastasivat kurssin alussa ja lopussa. Ymmärrettävyyteen liittyviä kysymyksiä oli kolme, joista ensimmäisessä opiskelijat saivat luonnehtia avoimesti, mitä helposti tai vaikeasti ymmärrettävä puhe on ja kahdessa seuraavassa arvioida annettujen suullisen kielitaidon osa-alueiden tärkeyttä sekä erinäisten tekijöiden vaikutusta ymmärrettävyyteen.

Alustavien tulosten perusteella opiskelijoiden arviot yksittäisten äänteiden vaikutuksesta ymmärrettävyyteen olivat selkeästi korkeammat kurssin alussa kuin kurssin loputtua. Alkukyselyssä noin puolet opiskelijoista vastasi, että äänteillä on paljon merkitystä ymmärrettävyyteen, kun osuus loppukyselyssä oli noin neljännes. Rytmin, sanapainon ja lausepainon, samoin kuin ulkoisten tekijöiden, vaikutus koettiin merkityksellisempänä kurssin loputtua kuin kurssin alussa.

## Lähteet

Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly, 39*(3), 399–423. https://doi.org/10.2307/3588487.

Trofimovich, P. & Isaacs, T. (2012). Disentangling accent from comprehensibility. *Bilingualism: Language & Cognition, 15(*4), 905–916. https://doi. org/10.1017/S1366728912000168.

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

## Vieraskielisen suullisen vuorovaikutuksen automaattinen arviointi – AASIS-hankkeen lähtökohtia ja aineiston esittelyä

Riikka Ullakonoja[1], Ilona Lähteenmäki[2], Nora Raud[3], Nhan Phan[3], Tamás Grósz[3], Raili Hilden[2], Mikko Kuronen[1] & Mikko Kurimo[3]

1) Jyväskylän yliopisto 2) Helsingin yliopisto 3) Aalto-yliopisto

Esitelmämme tarkoituksena on kuvailla monitieteisen AASIS-hankkeen (Automatic assessment of spoken interaction in a second language, Suomen Akatemia 2023-2027) teoreettisia ja metodologisia lähtökohtia. Hankkeessa keskitytään vieraskielisen suullisen vuorovaikutuksen automaattiseen arviointiin suomen kielessä. Projektin tavoitteet ovat uraauurtavia. Ensinnäkin tarkoituksenamme on kehittää automaattista puheen arviointi (ASA) systeemiä arvioimaan suullista kielitaitoa dialogipuheessa. Aiemmissa systeemeissä automaattinen arviointi on keskittynyt etupäässä monologipuheeseen. Toiseksi projektissa on tarkoituksena arvioida automaattisesti myös puheen nonverbaaleja piirteitä (esim. katsetta, ilmeitä, käden ja pään liikkeitä) videoaineistosta. Kolmanneksi tutkimuksen kohteena on suomen kieli, jonka automaattisesta arvioinnista on hyvin vähän aiempaa tutkimusta. Aiempi hankkeemme DigiTala on esimerkki aiemmasta tutkimuksesta, jossa automaattista puheentunnistinta on käytetty suomea ja ruotsia vieraana kielenään puhuvien monologipuheen automaattiseen arviointiin.

Automaattisen puheen arvioinnin lisäksi käytämme tutkimuksessamme myös ihmisarvioijia. Heitä varten on kehitetty Eurooppalaisen viitekehyksen pohjalta dialogipuheen arviointiin soveltuvat asteikot, jotka sisältävät myös nonverbaalien piirteiden arviointia. Tämä on uutta, sillä suullisenvuorovaikutuksen käsite ei perinteisesti ole sisältänyt nonverbaalisten piirteiden arviointia, vaikka niiden rooli ihmisten välisessä vuorovaikutuksessa tiedetään olevan suuri. Näin ollen Aasis-hankkeen tavoitteena on myös kehittää dialogipuheen arviointia niin, että suullisen vuorovaikutuksen käsite sisältäisi myös nonverbaalit piirteet ja niiden automaattisen arvioinnin. Tämä on tärkeää, koska dialogeja käytetään paljon kielten opetuksessa ja myös kielitaidon arvioinnissa. Esitelmässämme kuvailemme teoreettisten ja menetelmällisten lähtökohtien lisäksi myös keräämiämme aineistoja, joista osa on tarkoitus saattaa hankkeen päätyttyä muidenkin tutkijoiden käyttöön.

## Lukio-opiskelijoiden tunnekokemukset ja luokkahuoneessa saatu ääntämispalaute

Päivi Virkkunen, Minnaleena Toivola & Martti Vainio

Helsingin yliopisto

Suullinen kielitaito on keskeinen osa kielitaitoa. Vieraan kielen ääntämisen oppiminen edellyttää runsasta motorista harjoittelua, mutta myös palaute on tärkeää. Virkkusen ja Toivolan (2020) tutkimuksessa havaittiin, että monet kieltenopettajat pitivät palautteen antamista vaikeana ja välttelivät etenkin korjaavan palautteen antamista. Suomalaiset lukiolaiset, joista monilla on toiveena natiivinkaltainen ääntämistaito, toivoivat kuitenkin saavansa erityisesti korjaavaa palautetta omasta ääntämisestään ja tiedostivat sen merkityksen ääntämisen harjoittelussa (Virkkunen & Toivola, 2023).

Selvitämme tutkimuksessamme, minkälaisia tunteita suomalaiset lukiolaiset kokevat vieraan kielen opetuksessa luokkahuonetilanteessa. Tutkimuksen aineistona on laaja kyselytutkimus (n=1953). Kerromme tarkemmin opiskelijoiden ääntämispalautteen saamiseen liittyvistä tunteista ja pohdimme mm. opettajan antaman positiivisen ja korjaavan palautteen vaikutusta oppijan halukkuuteen puhua tunnilla.

Tulosten perusteella voidaan kehittää opetuskäytänteitä, jotka huomioivat paremmin oppijan toiveet ja tavoitteet ja edistävät siten oppimista. Tuloksia voidaan käyttää apuna myös kieltenopettajien koulutuksen ja täydennyskoulutuksen suunnittelussa.

### Lähteet

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.

## Kielitaidon tukeminen tienä oppimisen yhdenvertaisuuteen ja osallisuuteen

Katja Haapanen, Minna Rihko, Henna Tamminen, Kimmo U. Peltola & Maija S. Peltola

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

Kielitaidon tukeminen tienä oppimisen yhdenvertaisuuteen ja osallisuuteen (KiTu) on Turun kaupunkitutkimusohjelman rahoittama hanke, jonka päämääränä on löytää keinoja tukea monikielisten lasten suomen oppimista. Tutkimme erilaisten foneettisten puheen ymmärtämis- ja tuottamisharjoitteiden toimivuutta eri kielitaustoista tulevilla oppijoilla. Etsimme joukosta tehokkaita ja kouluympäristöön soveltuvia harjoitteita, ja kehitämme tutkimuksen tulosten pohjalta työvälineen, joka tarjoaa konkreettista apua kielitaidon tukemiseen monikielisissä luokkahuoneissa. Hankeen edetessä laajennamme tutkimusta muihin kieliin, kuten englantiin.

Tutkimuksessa on kerätty erilaisia puheaineistoja monikielisiltä lapsilta (n=120) luokka-asteilta 2–6. Lisäksi olemme keränneet verrokkiaineiston suomea äidinkielenään puhuvilta lapsilta samoista ikäryhmistä (n=20). Kerätty aineisto koostuu yksittäisistä sanoista ja pidemmistä lukupuhunnoksista sekä tutkimustilanteiden videotallenteista. Ääntämistarkkuutta ja puheen ymmärrettävyyttä mitataan sekä akustisin mittauksin että raatiarvioin, minkä lisäksi erityispedagogiikan ja logopedian ammattilaiset läpikäyvät videoaineistot mahdollisten oppimis- ja puheentuottovaikeuksien kartoittamiseksi. Akustisten analyysien, raatiarvioiden sekä erityisopettajan ja puheterapeutin havaintojen perusteella pyrimme vastaamaan neljään kysymykseen: 1. Millaisia suomen ääntämiseen liittyviä haasteita eri kielten puhujien puheessa esiintyy? 2. Millaisista ääntämisharjoitteista eri kieli- ja ikäryhmät hyötyvät eniten? 3. Kuinka opettajat ja opettajaopiskelijat arvioivat monikielisten oppilaiden puheen ymmärrettävyyttä? 4. Liittyvätkö puheessa ja lukupuheessa esiintyvät ja ymmärrettävyyteen vaikuttavat piirteet kielenoppimiseen, yleisiin oppimisen vaikeuksiin vai puheen tuoton ongelmiin? Työtä tehdään tiiviissä yhteistyössä Turun kaupungin koulujen kanssa ja opettajien kokemuksia kartoitetaan haastatteluin koko hankkeen ajan.

Esitelmässä kerromme hankkeen taustoista, toteutuksesta sekä tulevaisuuden suunnista.

# The influence of the duration of voiced consonants on their recognition in sung vowel-consonant and consonant-vowel junctions

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

1) Estonian Academy of Music and Theatre 2) Tallinn University of Technology

In singing, poor intelligibility of sung text is a common issue, particularly in reverberant acoustics (Meyer, 2009) and at high pitches (Eberhart, 1962). Some voice teachers recommend lengthening consonants to enhance intelligibility (Nair, 2021), while others advocate against this approach as this may disrupt musical flow and prosody (Ware, 1998). However, little research has directly examined how consonant duration affects their recognition in singing.

This study investigates whether and how the recognition of voiced consonants */m/*, */n/*, */l/*, and */v/* in sung consonant-vowel (CV) and vowel-consonant (VC) junctions depends on consonant duration when sung in various acoustic environments, at different pitches, and in the presence of noise.

Recognition tests were conducted with 34 participants (12 male, 22 female, aged from 16 to 69 years). Stimuli consisted of sung VC and CV junctions (n = 2048) performed by a mezzo-soprano and a baritone at various pitches. Using PRAAT software, consonant durations were adjusted to various lengths ranging from 0 ms to 200 ms. Artificial reverberation, simulating environments such as a church or concert hall, or brown noise (to mimic the accompaniment) was added to some stimuli. Generalized Linear Models were used for the statistical analysis.

Recognition of more extended consonants always tended to be better. In the case of CV junctions, the influence of room acoustics on the recognition was small, and a consonant duration of 20 ms was sufficient to achieve 95% recognition at low pitches, except when noise was present. Recognition was significantly poorer at high pitches and in reverberant acoustics, especially for VC junctions. At low pitches, recognition was better than chance even when the stationary part of the consonant was absent and perception was based only on the transition of vowel formants.

The findings suggest that extending the duration of voiced consonants in singing improves their recognition, especially in the case of VC junctions in reverberant acoustics, and with the presence of accompaniment. These results provide evidence-based guidelines for optimizing consonant articulation in vocal pedagogy. Further research should explore how these adjustments impact perceived speech prosody.

**References**

Eberhart, C. (1962). Diction. *Journal of Singing, 15* (May), 8-34.
Meyer, J. (2009). *Acoustics and the performance of music*. Springer.
Nair, A. (2021). *The tongue as a gateway to voice, resonance, style, and intelligibility.* Plural Publishing Inc.
Ware, C. (1998). *Basics of vocal pedagogy*. McGraw-Hill.

## Shape-sound symbolism in Finnish and a word learning paradigm

Alexandra Wikstrom[1], Lari Vainio[1,2] & 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

Shape-sound symbolism is a well-known sound symbolistic phenomenon in which words that contain certain speech sounds are associated with either round or angular shapes ([1]). Except for some surface-level cross-linguistic explorations ([2]), research on the phenomenon in Finnish is still lacking. Due to shape-sound symbolism being so well-established, it is well suited for use in a word learning paradigm targeting the effects of sound symbolism in language learning. This experiment served two purposes: 1. to find the most fitting pseudowords per category to be used in a future word-learning paradigm, and 2. to validate known sound-symbolic associations of speech sounds in Finnish speakers, while observing possible effects that, for example, the modality of the stimulus could have.

By utilizing speech sounds that are known to be associated with the concepts of round and spiky ([3], [4]), round, spiky and neutral pseudowords were generated according to Finnish phonotactics. 96 pseudowords of each category were chosen and produced with speech synthesis for stimulus purposes. Using Gorilla Experiment Builder (www.gorilla.sc), 37 participants reacted to pseudowords presented as an auditory or visual stimulus at random. The words were rated by the participants on a scale of 1–6, indicating the perceived roundness or spikiness of the words. The preliminary results showed a preference for a word to be perceived as either rounder or spikier depending on the speech sounds it contained, which is in line with previous literature. In addition, some inter-group variation in the assessment of spiky-shape associated words could be seen. From the basis of this experiment, the words most associated with their corresponding concept will be used in the word learning paradigm. The goal of this paradigm is to observe the level at which participants are able to learn sound symbolic words when either disruptive articulatory movements are included or when no disruptions are present. Through this we aim to examine the neural mechanisms of sound symbolic associations which are known to aid in language learning ([5]).

### References

[1] Ramachandran, Vilayanur S. and Hubbard, Edward M. "Synaesthesia–a window into perception, thought and language". *Journal of Consciousness Studies 8* (2001), pp. 3–34.

[2] C´wiek, Aleksandra et al. "The bouba/kiki effect is robust across cultures and writing systems". *Philosophical transactions of the Royal Society of London. Series B. Biological sciences 377*.1841 (2022), pp. 20200390–20200390.

[3] Nielsen, Alan and Rendall, Drew. "The Sound of Round: Evaluating the Sound-Symbolic Role of Consonants in the Classic Takete-Maluma Phenomenon". *Canadian Journal of Experimental Psychology 65*.2 (2011), pp. 115–124.

[4] Cuskley, Christine, Simner, Julia, and Kirby, Simon. "Phonological and orthographic influences in the bouba–kiki effect". *Psychological Research 81* (2017), pp. 119–130.

[5] Lockwood, Gwilym, Dingemanse, Mark, and Hagoort, Peter. "Sound-Symbolism Boosts Novel Word Learning". *Journal of Experimental Psychology: Learning, Memory, and Cognition 42*.8 (2016), pp. 1274–1281. DOI:10.1037/xlm0000235.

# The interlocutor factor and intonational variation in collaborative dialogues

Tatiana Kachkovskaia

Helsinki Collegium for Advanced Studies, University of Helsinki

Speech behaviour is subject to variability due to social context. The observed differences are related to all linguistic features, including intonational. E.g. in formal settings, compared with everyday conversations, a different "intonation style" is usually used (Swerts et al. 1996; Hirschberg 2000). Other cases of variation in intonation style have been observed in research on attractiveness in dyadic conversations (Leongómez et al. 2014).

This research is a corpus-based study of intonational changes observed in the speech of the same person while involved in collaborative conversations with different interlocutors. The dataset consists of 90 annotated speech fragments from the SibLing speech corpus (Kachkovskaia et al. 2020), up to 1 minute each. Each fragment was chosen from a different collaborative dialogue where two interlocutors were searching for similarities in two decks of whimsical pictures. The key feature of this dataset is variation in the social distance between the interlocutors, ranging from siblings and close friends to strangers of the opposite gender and strangers with a significant age gap. The dataset also contains perceptual evaluation for each fragment, obtained from naive listeners via crowdsourcing. The evaluation included listeners' opinion on mutual attractiveness of the interlocutors, awkwardness of the conversation, the interlocutors feeling interested or bored, and other.

The main goal of this project is to look for connection between various parameters of the melodic contours and relationship between the interlocutors (both real and perceived). Intonational analysis includes such measurements as declination coefficient, wiggliness and spaciousness (Wehrle 2023), and shapes of the melodic contours for prosodically prominent words.

**References**

1. Leongómez, J. D., Binter, J., Kubicová, L., Stolařová, P., Klapilová, K., Havlíček, J., & Roberts, S. C. (2014). Vocal modulation during courtship increases proceptivity even in naive listeners. *Evolution and Human Behavior, 35*(6), pp. 489–496
2. Swerts, M. G. J., Strangert, E., & Heldner, M. (1996). F0 declination in read-aloud and spontaneous speech. In *Proceedings of the 4th International Conference on Spoken Language Processing,* ICSLP-96, Philadelphia, PA, USA, October 3-6, pp. 1501–1504
3. Wehrle, S. (2023). Conversation and intonation in autism: A multi-dimensional analysis. *(Studies in Laboratory Phonology 14)*. Berlin: Language Science Press
4. Hirschberg, J. (2000). A Corpus-Based Approach to the Study of Speaking Style. In: Horne, M. (eds) *Prosody: Theory and Experiment. Text, Speech and Language Technology, vol 14*. Springer, Dordrecht, pp. 335–350
5. Kachkovskaia T., Chukaeva T., Evdokimova V., Kholiavin P., Kriakina N., Kocharov D., Mamushina A., Menshikova A., & Zimina S. (2020). SibLing corpus of Russian dialogue speech designed for research on speech entrainment. *Proc. of the 12th LREC. Marseille, France: ELRA*, 2020, pp. 6556–6561

# Suomenruotsin prosodian tuottaminen erikielisillä alkeisoppijoilla

Heini Kallio & Henna Heinonen

Kielten yksikkö, Tampereen yliopisto

Tässä esitelmässä kerromme hankkeestamme *Suomenruotsin ääntäminen erikielisillä alkeisoppijoilla* [1] sekä hankkeen ensimmäisistä osatutkimuksista, joissa selvitämme ymmärrettävyyteen liittyvien prosodisten piirteiden onnistumista kuudesta eri lähtökielestä tulevilla puhujilla. Aiemmat suomenruotsin ja ruotsin L2-puheen tutkimukset Suomessa ovat keskittyneet pääasiassa suomenkielisiin oppijoihin. Tänä päivänä ruotsia opiskelee Suomessa kuitenkin monista muistakin kieliryhmistä tulevat oppijat, joiden ääntämisen haasteista ei vielä ole tutkimustietoa. Hanketta rahoittaa Svenska Kulturfonden (2024–2025).

Aiemmissa tutkimuksissa on etenkin ruotsin sana- ja lausepainon toteutumisella todettu olevan suuri merkitys puheen ymmärrettävyyteen [2, 3]. Tutkimme sana- ja lausepainojen toteutumista sekä puherytmiä suhteellisiin tavukestoihin perustuvilla parametreilla [3, 4]. Tavukestojen lisäksi vertailemme ruotsinoppijoiden (n=30) ja äidinkielisten puhujien (n=6) artikulaationopeuksia sekä perustaajuuden vaihtelua neljässä ääneen luetussa lauseessa. Ruotsinoppijoiden lähtökielet ovat saksa, englanti, ranska, venäjä, sinhali ja vietnam. Oletamme, että puhujat, joiden lähtökieli on typologisesti lähempänä ruotsia, hallitsevat ruotsin prosodian onnistuneemmin kuin typologisesti ruotsista poikkeavien kielten puhujat.

Tulosten perusteella vaikuttaa siltä, että jo alkeistasolla saksankieliset ruotsinoppijat hyötyvät äidinkielensä ja kohdekielen prosodisesta samankaltaisuudesta, kun taas erityisesti ranskan- ja sinhalinkielisten puheessa näkyy äidinkielen siirtovaikutus. Vietnamin- ja englanninkielisten tuottamien prosodisten piirteiden onnistumisessa on sen sijaan vaihtelua. Venäjänkieliset puhujat onnistuvat ääntämisessä aiempiin tutkimustuloksiin nähden paremmin kuin odotimme [5]. Merkittävimmät erot ruotsin alkeisoppijoiden ja äidinkielisten puhujien välillä ovat lausepainon tuottamisessa ja artikulaationopeudessa.

**Viitteet**

[1] Suomenruotsin ääntäminen erikielisillä alkeisoppijoilla (29.1.2025) https://www.tuni.fi/fi/tutkimus/suomenruotsin-aantaminen-erikielisilla-alkeisoppijoilla

[2] Abelin, Å. & Thorén, B. (2015). What affects recognition most – wrong word stress or wrong word accent? Working Papers 55. *Proceedings from Fonetik 2015 Lund, June 8–10 2015*. Lund: Lund University. S. 7–10.

[3] Heinonen, H. (2019). Durationsförhållandena i finskspråkiga gymnasisters uttal av L2-svenska: hur relaterar de till begripligheten?. In *Svenskans beskrivning (No. 36)*. Uppsala universitet.

[4] Kallio, H., Kautonen, M., & Kuronen, M. (2023). Prosody and fluency of Finland Swedish as a second language: Investigating global parameters for automated speaking assessment. *Speech Communication, 148*, 66-80.

[5] Bannert, R. (2004). *På väg mot svenskt uttal*. Lund: Studentlitteratur.

# Exploring Speech Prosody of Finnish-Speaking Preadolescents with Autism Spectrum Disorder (ASD) – an Investigation on Silent Pauses and Turn-Maintenance

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

1) Department of Languages, University of Helsinki, Helsinki, Finland 2) Department of Digital Humanities, University of Helsinki, Helsinki, Finland

Speakers with ASD (autism spectrum disorder) may exhibit various atypical prosodic features in their speech (Asghari et al., 2021; Fusaroli et al., 2017). This includes atypical pausing – atypically long and unusually positioned silent pauses have been noted in prior literature (Fusaroli et al., 2021). In conversation, speakers with ASD may exhibit challenges in conversational pragmatics and delays in turn-management (Choi & Lee, 2013; Wehrle et al., 2023). These notions may negatively affect the conversational fluency and social acceptance of individuals with ASD (Paul et al., 2005).

In this study, we investigated silent pauses and turn-maintenance in Finnish-speaking preadolescents with ASD. All speakers were males aged 11 to 13. The findings were compared to those from a gender- and age-matched control group without ASD. Spontaneous speech from group conversations was used, and complete speaker turns were analyzed. Both grammatical and non-grammatical pauses were taken into account. The semantic and syntactic positioning of the pauses were also investigated. Using acoustic methods, the silent pauses and cues of turn-maintenance were analyzed. In addition, other prosodic features such as speech rate and the modulation of fundamental frequency (*f*0) were analyzed.

The ASD speaker group exhibited more silent pauses within turns and greater variation in pause durations when compared to controls. Speakers with ASD also produced longer pauses than controls. The speaker groups also varied in terms of their speech rates and their tendencies of modulating *f*0. In the more exploratory section of the study investigating turn-maintenance, it was found that within the ASD group, silent pauses appeared alongside various acoustic and articulatory cues that successfully indicated turn-maintenance despite frequent pausing. These included non-verbal vocalizations such as clicks as well as producing pauses as glottal stops. Such cues may compensate for frequent pausing to retain the turn in conversation.

These results of the analysis will be discussed in this presentation to investigate the interplay between pausing, other prosodic dimensions and conversational pragmatics. This study explores the complex phenomenon of frequent silent pauses and successful turn-maintenance cues in ASD conversational behavior utilizing multiple analytical frameworks. Pausing tendencies and conversational fluency in ASD remain under-researched in Finnish, and this preliminary study explores the topic from a multidisciplinary perspective.

**References**

Asghari, S. Z., Farashi, S., Bashirian, S., & Jenabi, E. (2021). Distinctive prosodic features of people with autism spectrum disorder: a systematic review and meta-analysis study. *Scientific Reports* , 11, 23093. https://doi.org/10.1038/s41598-021-02487-6

Choi, J., & Lee, Y. (2013). Conversational Turn-Taking and Topic Manipulation Skills of Children with High-Functioning Autism Spectrum Disorders. *Communication Sciences & Disorders*, 18(1), 12–23. https://doi.org/10.12963/csd.13002

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Richardt, J., Jepsen, M., & Weed, E. (2021). Toward a cumulative science of vocal markers of autism: A cross-linguistic meta-analysis-based investigation of acoustic markers in American and Danish autistic children. *Autism Research*, 15(4), 653–664 https://doi.org/10.1002/aur.2661

Fusaroli, R., Lambrechts, A., Bang, D., Bowler, D. M., & Gaigg, S. B. (2017). Is voice a marker for Autism spectrum disorder? A systematic review and meta-analysis. *Autism Research*, 10(3), 384–407. https://doi.org/10.1002/aur.1678

Paul, R., Shriberg, L. D., McSweeny, J., Cicchetti, D., Klin, A., & Volkmar, F. (2005). Brief Report: Relations between Prosodic Performance and Communication and Socialization Ratings in High Functioning Speakers with Autism Spectrum Disorders. *Journal of Autism and Developmental Disorders*, 35(6), 861–869. https://doi.org/10.1007/s10803-005-0031-8

Wehrle, S., Cangemi, F., Janz, A., Vogeley, K., & Grice, M. (2023). Turn-timing in conversations between autistic adults: Typical short-gap transitions are preferred, but not achieved instantly. *PLoS ONE*, 18(4 April). https://doi.org/10.1371/journal.pone.0284029

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

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

Institute of Estonian and General Linguistics, University of Tartu

This study is an elaboration of the preliminary study on negation gestures in Estonian presented at the Fonetiikan päivät 2024 in Tallinn. Earlier multimodal studies have shown that there are gestures that can be associated with negation in several languages [1], [2], [3]. While these gestures primarily have different communicative functions than the beat gestures, they are still often aligned with prosodically prominent units in speech [4]. The aim of this work is to establish the inventory of negation gestures in Estonian and to investigate the alignment of these gestures with the prosodically prominent syllables in the negation phrase.

In Estonian, standard (clausal) negation is asymmetric [5]: the negative verb form is different from the affirmative form (negation particle *ei* + different connegative verb stem, e.g. *Ma ei jookse* 'I not run.CNG'). The negation particle always immediately precedes the verb. In speech, the verb receives a pitch accent and the negation particle is therefore deaccented.

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

In the 20 dialogues analysed, 40 speakers produced a total of 2056 negation phrases. In 339 cases the negation phrase was accompanied with a head shake and in 136 cases with a hand gesture related to negation (including 52 open palm-up gestures, 25 shoulder shrugs, 17 throwing away gestures, 15 sweeping away gestures, 10 finger wags, 9 holding away gestures and 8 other gestures). The alignment of these gestures with the prosodic realisation of the negation phrase in speech will be discussed in more detail in the paper.

**References**

[1] 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.

[2] 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.

[3] P. Siahaan and G. P. Wijaya Rajeg, "Multimodal language use in Indonesian: Recurrent gestures associated with negation," p. 769328, 2023, doi: 10.17617/2.3527196.

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

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

[6] 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.

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

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

## Emergence of Communication through Predictive Processing and Bayesian Inference

Juraj Šimko

Department of Digital Humanities, University of Helsinki

We explore the emergence of speech communication as a probabilistic and adaptive process driven by predictive processing, Bayesian inference and Bayesian updating. Traditional models, such as Signaling Games (Lewis, 1969) and Rational Speech Act Theory (Frank & Goodman, 2012), often conceptualize communication as a discrete process in which predefined discrete set of signals map onto a finite set of meanings. This study presents an alternative perspective, framing speech communication as an iterative process where agents adjust their probabilistic beliefs about speechlike, continuously varied signals and responses through interaction, treating language as an evolving dynamic process created and shaped through its use (Vygotsky, 1978; Varela, Maturana, & Uribe, 1974).

Building on the Bayesian Brain Hypothesis and Predictive Processing paradigm (Friston, 2005; Clark, 2013; Hohwy, 2013), which posit that cognitive processes operate through hierarchical prediction and Bayesian inference, we propose a model in which speakers generate phonetic-like signals to elicit specific responses, and listeners interpret these signals using probabilistic inference. Through iterative learning, both parties refine their internal likelihood distributions, gradually developing a shared communicative system.

Computational simulations of this model demonstrate that groups of agents can develop stable signalresponse mappings over time thorugh repeated communication. Moreover, extending this framework by adding a hierarchical predictive processing level shows how structured "linguistic" patterns—such as word-like sequences—can emerge from low-level interactions.

These findings provide insights into the evolution of speech and language, suggesting that structured communication can arise naturally through iterative Bayesian updating. By treating communication as an emergent, self-organizing system rather than a static encoding-decoding mechanism, this work contributes to a broader understanding of language evolution within cognitive science, linguistics and phonetics.

### References

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*(6084), 998–998.

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Varela, F. J., Maturana, H. R., & Uribe, R. (1974). Autopoiesis: The organization of living systems, its characterization and a model. *BioSystems, 5*(4), 187–196.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456), 815–836.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Hohwy, J. (2013). *The Predictive Mind*. Oxford University Press.

# Investigation of speech timing using excitable networks

Michael O'Dell

University of Helsinki

For many years dynamical systems theory has played an important role in the investigation of phonetic processes in general and speech timing in particular. In recent years there has been a great deal of research in dynamical systems theory concerning excitable network attractors, including (noisy) heteroclinic networks and their close relatives (see [1] for an excellent short review). These models have many advantages ([8], [9]), including the ability to combine reproducibility and flexibility of transient behavior ([10]), and they bridge the gap between between classical discrete computation models of sequential categories and continuous-time dynamical systems ([4]). As dynamical systems, they are also well suited for investigating timing of observed behavior (cf. e.g. [6]).

An interesting question concerns what correlations, if any, can be found between phonotactics (syntagmatic and paradigmatic structure) and timing of (roughly) sequential categories (e.g. as in Articulatory Phonology and (Embodied) Task Dynamics). There now exist several general algorithms for constructing dynamical systems corresponding to finite state diagrams (or directed graphs; [2]-[4]). We are exploring the impact of various possible excitable network architectures applied to modeling speech timing and duration in particular. Hidden Markov Models (HMM), used extensively to model speech, provide a useful point of comparison, especially since it is known that excitable networks can exhibit non-Markovian behavior [5], [7]. Comparisons will be made with durational data from Finnish speech on the one hand, and HMM-based modeling on the other. Some results and discussion of these preliminary investigations will be presented at the conference.

**References**

[1] P. Ashwin, M. Fadera, and C. Postlethwaite, "Network attractors and nonlinear dynamics of neural computation," *Current Opinion in Neurobiology, vol. 84*, no. 102818, pp. 1–7, 2024.

[2] P. Ashwin and C. Postlethwaite, "On designing heteroclinic networks from graphs," *Physica D: Nonlinear Phenomena, vol. 265*, pp. 26–39, 2013. [Online]. Available: http://hdl.handle.net/10871/14534.

[3] P. Ashwin and C. Postlethwaite, "Designing heteroclinic and excitable networks in phase space using two populations of coupled cells," *Journal of Nonlinear Science, vol. 26*, no. 2, pp. 345–364, 2016. [Online]. Available: https://arxiv.org/pdf/1506.03212.pdf.

[4] P. Ashwin and C. Postlethwaite, "Excitable networks for finite state computation with continuous time recurrent neural networks," *Biological Cybernetics, vol. 115*, no. 5, pp. 519–538, 2021.

[5] Y. Bakhtin, "Small noise limit for diffusions near heteroclinic networks," *Dynamical Systems, vol. 25*, pp. 413–431, 2010.

[6] J. Creaser, P. Ashwin, C. Postlethwaite, and J. Britz, "Noisy network attractor models for transitions between EEG microstates," *Journal of Mathematical Neuroscience, vol. 11*, no. 1, pp. 1–25, 2021.

[7] G. Manicom, V. Kirk, and C. Postlethwaite, "Non-Markovian processes on heteroclinic networks," *Chaos, vol. 34*, no. 033120, pp. 1–15, 2024. DOI: 10.1063/5.0176205.

[8] H. Meyer-Ortmanns, "Heteroclinic networks for brain dynamics," *Frontiers in Network Physiology, vol. 3*, no. 1276401, pp. 1–16, 2023. DOI: 10.3389/fnetp.2023.1276401.

[9] C. M. Postlethwaite, P. Ashwin, and M. Egbert, "A continuous time dynamical Turing machine," *IEEE Transactions on Neural Networks and Learning Systems*, in press.

[10] M. I. Rabinovich, R. Huerta, P. Varona, and V. S. Afraimovich, "Transient cognitive dynamics, metastability, and decision making," *PLoS Computational Biology, vol. 4*, no. 5, e1000072, 2008.

# Text-to-speech synthesis in infant-directed speaking style using noisy speech data from naturalistic infant-caregiver interactions

Elmeri Grönlund, Daniil Kocharov & Okko Räsänen

Signal Processing Research Centre, Tampere University, Finland

Speech synthesis systems have undergone substantial development during the last few years, best systems reaching human-like sound quality and capability for different speaking styles. However, there is no existing system for producing natural-sounding infant-directed speech (IDS). One problem in developing a high-quality IDS synthesizer is the lack of suitable speech data, where realistic IDS from real-world communicative contexts is often of poor or varying audio quality.

In this study, we present results from a series of experiments to generate IDS using a zeroshot TTS system, XTTS (Casanova et al., 2024), capable of copying the speaking style of a provided reference speech sample. Since naturalistic IDS speech samples are often noisy, resulting in poorquality synthesis outputs, we employed special techniques for selecting and enhancing the samples before using them as a reference in XTTS. More specifically, we generated speech using IDS reference samples from the Providence corpus (Demuth et al., 2006) consisting of real-world infantcaregiver interactions and compared different techniques for the sample selection: (1) UTMOS (Saeki et al., 2022), a neural-network based automatic predictor for human quality ratings, and (2) intelligibility measured by word error rate (WER), computed using the small model of Whisper by OpenAI (Radford et al., 2023), (3) Manual selection of high-quality samples based on listening. We also compared enhanced reference samples against the original noisy samples, where the speech was enhanced using the Resemble Enhance tool by ResembleAI (https://github.com/resembleai/ resemble-enhance).

To test the system, 600 IDS speech samples were synthesized using reference audios with low, medium, and high F0 in order to study capability of the synthesizer to mimic IDS in different pitch registers. Each synthesized speech transcript was automatically generated using a recently developed generator of infant-language experiences, as described in (Räsänen and Kocharov, 2024) and included up to 15 sentences. To evaluate the quality of the generated speech outputs, UTMOS, WER, and character-error rates (CER) were measured for each condition.

Our results indicate that the best performing approach was to use enhanced IDS reference speech samples that were automatically selected using both the ASR- and UTMOS-based metrics, resulting in average UTMOS score of 3.63 and WER of 14%. Manual listening and prosodic feature analysis further indicated that the system was able to produce high-quality IDS speech, and that several key prosodic features of the reference audio files were transferred to the synthesized speech.

**References**

Casanova E. et al., XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model. *Proc. of Interspeech 2024*, 4978–4982.

Demuth J. et al. Word-minimality, epenthesis and coda licensing in the early acquisition of English, *Language & Speech, 49*(2), 137–173, 2006

Radford et al. Robust speech recognition via large-scale weak supervision. *Proc. of International Conference on Machine Learning, 2023*, 28492–28518

Räsänen O. & Kocharov D., Age-dependent analysis and stochastic generation of child-directed speech. *Proc. of CogSci 2024.*

Saeki T. et al. UTMOS: UTokyo-SaruLab System for VoiceMOS Challenge 2022. *Proc. of Interspeech 2022*, 4521–4525

## Leveraging LLM in Speaking Assessment and Application for Learning Finnish

Nhan Phan[1], Anna von Zansen[2], Maria Kautonen[3], Ekaterina Voskoboinik[1], Tamás Grósz[1], Raili Hildén[2] & Mikko Kurimo[1]

1) Aalto University, Finland 2) University of Helsinki, Finland 3) University of Jyväskylä, Finland

We propose a framework to address several unsolved challenges in second language (L2) automatic speaking assessment (ASA) and feedback. The challenges include: 1. ASA of visual task completion, 2. automated content grading and explanation of spontaneous L2 speech, 3. corrective feedback generation for L2 learners, and 4. all the above for languages with minimal L2 speech data. Our system combines visual natural language generation, automatic speech recognition and prompting a large language model for low-resource L2 learners. We describe the framework and the outcomes of our a case study on a picture description task in Finnish. Our results indicate substantial agreement with human experts in grading, explanation and feedback [1].

This framework has the potential for a significant impact in constructing next-generation computer-assisted language learning systems to provide automatic scoring with targeted feedback for learners of low-resource languages. We also present our initial findings from our development of the language learning mobile app that leverages this framework [2].

[1] N. Phan, A. von Zansen, M. Kautonen, E. Voskoboinik, T. Grosz, R. Hilden, and M. Kurimo, "Automated content assessment and feedback for Finnish L2 learners in a picture description speaking task," in *Interspeech 2024*, ISCA, Sep. 2024, pp. 317–321. doi: 10.21437/Interspeech.2024-1166.

[2] N. Phan, A. von Zansen, M. Kautonen, T. Grósz, and M. Kurimo, "CaptainA self-study mobile app for practising speaking: task completion assessment and feedback with generative AI," in *Interspeech 2024*, ISCA, Sep. 2024, pp. 5212–5213.

# PFML: Self-Supervised Learning of Time-Series Data Without Representation Collapse

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

1) Signal Processing Research Centre, Tampere University, Finland 2) BABA Center, University of Helsinki, Finland

Self-supervised learning (SSL) is a data-driven approach that leverages the innate structure of the data to guide the learning process [1]. SSL allows the model to learn rich feature representations from vast amounts of unlabeled data, which can serve as a foundation for downstream tasks, either directly or by fine-tuning the feature extractor to better address specific classification tasks [2]. Given the typical abundance of unlabeled data and the scarcity of labeled data, SSL has been demonstrated to reduce the reliance on large, manually annotated datasets [3, 4, 5]. Unlike supervised learning, which relies on external labels, SSL uses the intrinsic properties of the data to generate its own supervisory signal. However, a common problem with many SSL methods is representation collapse, where the model produces a constant, input-invariant feature representation [1, 6, 7, 8]. This problem limits the potential application of SSL methods to new data modalities, as efforts to prevent representation collapse wastes researchers' time and effort.

In this study, we introduce a novel SSL algorithm for time-series data, such as speech or EEG data, named Prediction of Functionals from Masked Latents (PFML). PFML aims to predict statistical functionals of the input signal associated with masked embeddings, based on a sequence of unmasked embeddings. The primary methodological goal of our approach is to develop an SSL algorithm that is easy to apply across various time-series data domains with minimal hyperparameter tuning, and without the risk of representation collapse. We demonstrate the effectiveness of PFML using three different data modalities with complex, real-life classification tasks. Here, we focus on one of these three tasks, namely automatic speech emotion recognition from child-centered audio recordings recorded at the neonatal intensive care unit of Turku University Hospital [9].

As a result, our experiments show that PFML obtains superior results against both a conceptually similar SSL method [10] and a contrastive learning-based SSL method [11]. Additionally, PFML is on par with the current state-of-the-art data modality agnostic SSL method [7], while also being conceptually simpler and without suffering from representation collapse. Compared to other SSL methods for time-series data, this renders PFML more straightforward to apply to new time-series data domains, such as in the case of clinical time-series data.

**References**

[1] Balestriero et al., "A Cookbook of Self-Supervised Learning," arXiv preprint arXiv: 2304.12210, 2023.
[2] Erhan et al., "Why Does Unsupervised Pre-training Help Deep Learning?", *Journal of Machine Learning Research, vol. 11,* no. 19, pp. 625–660, 2010.
[3] van den Oord, Y. Li, and O. Vinyals, "Representation Learning with Contrastive Predictive Coding," arXiv preprint arXiv: 1807.03748, 2018.
[4] Baevski et al., "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," in *Proc. NeurIPS, 2020*, pp. 12 449–12 460.
[5] Chen et al., "A simple framework for contrastive learning of visual representations," in *Proc. ICML, 2020*, pp. 1597–1607.
[6] Akbari et al., "VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text," in *Proc. NeurIPS, 2021*, pp. 24 206–24 221.
[7] Baevski et al., "data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language," in *Proc. ICML, 2022*, pp. 1298–1312.
[8] Wang et al., "Image as a Foreign Language: BEIT Pretraining for Vision and Vision-Language Tasks," in *Proc. IEEE CVPR, 2023,* pp. 19 175–19 186.
[9] Vaaras et al., "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.
[10] He et al., "Masked Autoencoders Are Scalable Vision Learners," in *Proc. IEEE CVPR, 2022*, pp. 15 979–15 988.
[11] Yue et al., "TS2Vec: Towards Universal Representation of Time Series," in *Proc. AAAI, 2022*, pp. 8980–8987.

## Does multilingual and multi-speaker modeling improve low-resource TTS? Experiments on Sámi languages

Katri Hiovain-Asikainen[1,2] & Antti Suni[2]

1) UiT The Arctic University of Norway 2) University of Helsinki

In 2024, the Divvun group at UiT The Arctic University of Norway [1] published TTS applications for North, Lule and South Sámi, all of which are official languages in Norway. The development process included also creating entirely new speech corpora from scratch [2,3]. While these applications improve the situation, there is room for further improvement. When working with these low-resource languages, the collected training corpora are inevitably smaller than ideal, resulting in somewhat brittle synthesis; pronunciation errors and unsuitable prosody can occur, and the number of voices per language is limited to one or very few. Especially in very small language communities such as Lule (~2000 speakers) and South Sámi (~500-600 speakers), many people would immediately recognize the voice talent in the TTS, which is sometimes undesirable. In addition, the natural multilinguality of these speech communities can not be modelled, thus, for example, the pronunciation of proper names of majority languages does not reflect the speakers' competence properly.

Here, we are examining if training multiple speakers and languages jointly in a single neural TTS model could alleviate these issues. Specifically, we are using material from closely related languages. Firstly, we are training a model with data from all three Sámi languages, and secondly, we are adding material from well-resourced Finnic languages, Finnish and Estonian. Ideally, this should result in improving the quality of the low-resource language synthesis and allow for language transfer of the voices, but interference or 'foreign accent' could also result. We will report our preliminary experiences.

[1] Pirinen, F., Moshagen, S., & Hiovain-Asikainen, K. (2023, May). GiellaLT—a stable infrastructure for Nordic minority languages and beyond. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)* (pp. 643-649).

[2] Hiovain-Asikainen, K., & De la Rosa, J. (2023). Developing TTS and ASR for Lule and North Sámi languages. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)* (pp. 48-52).

[3] Hiovain-Asikainen, K., & Moshagen, S. (2022, June). Building open-source speech technology for low-resource minority languages with Sámi as an example–tools, methods and experiments. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages* (pp. 169-175).

## Suomen vokaalikvantiteetin oppiminen arabiankielisillä peliavusteisesti

Joonas Vakkilainen[1] & Michael O'Dell[2]

1) Tampereen yliopisto 2) Helsingin yliopisto

Tämä tutkimus tarkastelee arabiankielisten suomen oppijoiden harjoittelua Tell Me - pelillä. Tell Me perustuu aiempaan Say it again, kid -peliin (Karhila ym. 2017; Uther ym. 2018). Pelaajat toistavat kuulemansa sanan tai fraasin. Tutkimuksessamme analysoimme pelaajien edistymistä (lähentymistä mallipuhunnokseen) sekä vertailemme heitä toisiinsa. Analyysi keskittyy eri kvantiteettihahmoihin. My¨os arabiassa on kvantiteettijärjestelmä, ja tutkimuksemme tarkastelee, esiintyykö arabiankielisillä oppijoilla vaikeuksia suomen kvantiteetin oppimisessa.

### Viitteet

Karhila, R., Ylinen, S., Enarvi, S., Palom¨aki, K., Nikulin, A., Rantula, O., . . . Kurimo, M. (2017). Siak—a game for foreign language pronunciation learning. Teoksessa *Interspeech* (s. 3429–3430).

Uther, M., Smolander, A.-R., Junttila, K., Kurimo, M., Karhila, R., Enarvi, S., & Ylinen, S. (2018). User experiences from L2 children using a speech learning application: implications for developing speech training applications for children. *Advances in Human-Computer Interaction*, 2018.

## Suomenkielisten aikuispuhujien nasalanssiarvoja

Kaisa Tivonen[1], Aino Ervasti[1], Meri Ollikainen[1], Marja Laasonen[1] & Nelly Penttilä[2]

1) Logopedia, filosofinen tiedekunta, Itä-Suomen yliopisto 2) Logopedia, yhteiskuntatieteiden tiedekunta, Tampereen yliopisto

**Johdanto:** Nasalanssi eli nenäsointisuus tarkoittaa objektiivisella mittarilla, esimerkiksi nasometrillä, mitattavaa nenä- ja suuontelosta vapautuvan energian suhdetta, joka tyypillisimmin ilmaistaan prosenttiosuutena. Liiallinen ilmavirran pakeneminen nenäontelosta alentaa suuontelon ilmanpainetta ja aiheuttaa hypernasaalisuutta. Tämä puolestaan heikentää puheen ymmärrettävyyttä vaikuttaen vokaalien erottuvuuteen ja konsonanttien tuottoon. Hypernasaalisuus voi johtua useista eri häiriöistä ja syistä. Rakenteellisia syitä ovat esimerkiksi huuli- tai suulakihalkiot, sensorisia esimerkiksi kuulovamma ja neurologisia esimerkiksi aivoverenkiertohäiriöt tai neurologiset sairaudet, kuten amyotrofinen lateraaliskleroosi (ALS). Puheterapeutin kliinisessä työssä hypernasaalisuutta arvioidaan useimmiten perkeptuaalisesti, vaikka menetelmä subjektiivisuutensa vuoksi onkin herkkä inhimillisille virheille. Perkeptuaalisten menetelmien rinnalle kliiniseen työhön tarvitaankin objektiivisia, erityisesti akustisia ja aerodynaamisia menetelmiä esimerkiksi subkliinisten oraalimotoristen häiriöiden tunnistamiseen ja sairauksien etenemisen seurantaan. Tavanomaisen nasaalisuuden on havaittu tutkimuksissa kielikohtaisuuden lisäksi olevan useimmiten sukupuoli- ja joissakin kielissä myös ikäsidonnaista. Suomen kielen nasalanssiarvoista on tehty tietääksemme vain yksi aiempi tutkimus, jossa tutkimusaineisto painottui lapsiin ja nuoriin aikuisiin. Suomen kielellä ei ole tutkittu mahdollisia sukupuoli-, ikä-, tai puhetehtäväkohtaisia eroja nasalanssiarvoissa. Jotta nasalanssiarvoja voidaan suomenkielisessä ympäristössä hyödyntää diagnostiikan ja arvioinnin tukena, ovat käyttöön soveltuvat viitearvot välttämättömät.

**Menetelmät:** Tutkimuksessa tarkastellaan suomenkielisten puhujen puheessa esiintyvän nasalanssin määrää eri puhetehtävissä icSpeech-nasometriä hyödyntäen. Aineisto koostuu neurologisesti terveistä yli 40-vuotiaista miehistä ja naisista (n = 60). Tutkimuksessa tarkastellaan nasalanssin määrää ja mahdollisia eroja sukupuoli ja ikä huomioiden erityyppisissä puhetehtävissä.

**Tulokset:** Tutkimuksen alustavia tuloksia tarkastellaan Fonetiikan päivillä keväällä 2025.

Allison, K. M., Yunusova, Y., Campbell, T. F., Wang, J., Berry, J. D., & Green, J. R. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to ALS. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 18* (5–6), 358–366. https://doi.org/10.1080/21678421.2017.1303515.

Borrie, S. A., McAuliffe, M. J. & Liss, J. M. (2012). Perceptual learning of dysarthric speech: A review of experimental studies. *Journal of Speech, Language, and Hearing Research, 55* (1), 290– 305. https://doi.org/10.1044/1092-4388(2011/10-0349).

Eshghi, M., Richburg, B., Yunusova, Y. & Green, J. R. (2019). Instrumental Evaluation of Velopharyngeal Dysfunction in Amyotrophic Lateral Sclerosis. *Conference Paper of Proceedings of ICPhS in 2019*.

Green, J. R., Yunusova, Y., Kuruvilla, M. S., Wang, J., Pattee, G. L., Synhorst, L., Zinman, L. & Berry, J. D. (2013). Bulbar and speech motor assessment in ALS: Challenges and future directions. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration, 14* (7–8), 494–500. https://doi.org/10.3109/21678421. 2013.817585.

Haapanen, M. L. (1991). Nasalance scores in normal Finnish speech. *Folia Phoniatrica, 43*(4), 197–203. https://doi.org/10.1159/000266124

Karakoc, O., Akcam, T., Birkent, H., Arslan, H. H., & Gerek, M. (2013). Nasalance scores for normal-speaking Turkish population. *The Journal of Craniofacial Surgery, 24*(2), 520–522. https://doi.org/10.1097/SCS.0b013e3182802361.

Kim, H.-K., Yu, X., Cao, Y., Liu, X., & Huang, Z.-M. (2016). Dialectal and gender differences in nasalance for a Mandarin population. *Clinical Linguistics & Phonetics, 30*(2), 119-130. https://doi.org/10.3109/02699206.2015.1116111.

Stipancic, K. L., Yunusova, Y., Berry, J. D. & Green, J. R. (2018). Minimally detectable change and minimal clinically important difference of a decline in sentence intelligibility and speaking rate for individuals with amyotrophic lateral sclerosis. *Journal of Speech, Language, and Hearing Research, 61* (11), 2757–2771. https://doi.org/10.1044/2018_JSLHR-S-17- 0366.

Van Lierde, K. M., Wuyts, F. L., De Bodt, M., & Van Cauwenberge, P. (2001). Nasometric Values for Normal Nasal Resonance in the Speech of Young Flemish Adults. *The Cleft Palate Craniofacial Journal, 38*(2), 112–118. https://doi.org/10.1597/1545-1569_2001_038_0112_nvfnnr_2.0.co_2

Watterson, T. (2020). The Use of the Nasometer and Interpretation of Nasalance Scores. *Perspectives of the ASHA Special Interest Groups 5* (1), 155–163. https://doi.org/10.1044/2019_PERSP-19-00029.

# Introducing the FinnAffect Corpus

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

1) Signal Processing Research Centre, Tampere University, Finland 2) Languages Unit, Tampere University, Finland

Spoken language contains affective (emotional or emotion inducing) information, which is conveyed in speech as both segmental and suprasegmental variation, as well as in the lexical, morphological, and syntactic properties of the expressed linguistic content. The affective cues are ultimately perceived by the listener through subjective interpretations. Although affective expression is part of everyday conversational communication, there is little existing research on expression and perception of affect in spontaneous spoken Finnish, not to mention across different idiolect subgroups such as speakers of different ages or dialectal backgrounds. Although human emotional experiences appear to have cross-cultural similarities (Prinz, 2004, Russell et al., 1989, Sianipar et al., 2016), the expression and perception of affective states through language are also dependent on cultural social conventions (Wierzbicka, 1986, Evans and Levinson, 2009).

In order to study the specific nature of affective expression in everyday spoken Finnish, a dataset of spontaneous Finnish speech with affect-related metadata is required. While large corpora of spontaneous Finnish speech have been available (Moisio et al., 2022, University of Helsinki, 2014, Tampere University, n.d.) they do not contain affect-related annotations with which affective language could be examined. Hence, datasets containing rich and representative spontaneous Finnish paired with affect related labels are needed to advance both the linguistic research as well as speech technology related to speech emotion recognition (SER)

This presentation describes the creation of the first affective speech corpus of spontaneous Finnish language. We first introduce the corpus compiled from samples of three already existing large-scale, non-scripted Finnish speech corpora, for which affect related labels were obtained through manual annotation efforts. We briefly explain the sample selection process and the annotation process, then describing the resulting dataset characteristics. We then report results from SER classification experiments making use of the corpus, highlighting the feasibility of the dataset for training machine learning algorithms for SER in spontaneous Finnish. As an outcome, this corpus enables future research of affective expression in spontaneous Finnish.

## References

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *The Behavioral and Brain Sciences*, 32 (5), 429–448.

Moisio, A., Porjazovski, D., Rouhe, A., Getman, Y., Virkkunen, A., AlGhezi, R., Lennes, M., Gr´osz, T., Lind´en, K., & Kurimo, M. (2022). Lahjoita puhetta: A large-scale corpus of spoken finnish with some benchmarks [Publisher Copyright:©2022, The Author(s).]. *Language Resources and Evaluation, 57*, 1295–1327. https://doi.org/10.1007/s10579-022-09606-3

Prinz, J. (2004). Which emotions are basic. *Emotion, Evolution, and Rationality, 69*, 88.

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology, 57* (5), 848–856.

Sianipar, A., van Groenestijn, P., & Dijkstra, T. (2016). Affective meaning, concreteness, and subjective frequency norms for indonesian words. *Frontiers in Psychology, 7*. https://doi.org/10.3389/fpsyg.2016.01907

Tampere University. (n.d.). Longitudinal data of Tampere spoken language [Tampere University, The Unit of Languages and Institute for the Languages of Finland and Liisa Mustanoja]. http://urn.fi/urn:nbn:fi:lb-2022090821

University of Helsinki. (2014). *The Downloadable Version of the Longitudinal Corpus of Finnish Spoken in Helsinki (1970s, 1990s and 2010s)* [University of Helsinki, The Department of Finnish, Finno-Ugrian and Scandinavian Studies and Institute for the Languages of Finland and Heikki Paunonen]. http://urn.fi/urn:nbn: fi:lb-201609142

Wierzbicka, A. (1986). Human emotions: Universal or culture-specific? *American Anthropologist, 88* (3), 584–594

.

## Acknowledgement

The work is a part of the CONVERGENCE-project at Tampere University, funded by the Jane and Aatos Erkko Foundation.

## Large-Scale Self-Supervised Speech Representation Learning with Finnish National Archives

Yaroslav Getman, Tamás Grósz & Mikko Kurimo

Dept. lnformation and Communications Engineering, Aalto University, Finland

Self-supervised representation learning has become a standard approach across many fields, including Automatic Speech Recognition (ASR). However, most available speech foundation models are pre-trained on English or multilingual datasets dominated by high-resource languages, offering limited support for languages with fewer speakers. To bridge this gap, we introduce monolingual speech foundation models for Finnish, Finland-Swedish, and Northern Sámi, pre-trained on large-scale collections of TV and radio broadcasts sourced from the Radio and Television Archive (RTVA), provided by the Finnish National Audiovisual Institute (KAVI). The training was performed on the LUMI supercomputer. For Northern Sámi – spoken by only about 25,000 people – we utilized 22,000 hours of speech, while Finnish and Finland-Swedish models were trained on up to 150,000 hours per language, which is, to the best of our knowledge, the largest monolingual data used for self-supervised non-English speech representation learning. Our models demonstrate superior downstream ASR performance in low-resource settings and improved generalization compared to prior work.

# Posteriesitykset
# Poster presentations

# The durational characteristics of devoicing in Inari Saami words in utterance-final position

Helen Wilbur

University of Tartu

The goal of this study is to present a preliminary acoustic-phonetic investigation of the utterance-final devoicing phenomenon in Inari Saami which is a Finno- Ugric language spoken in Northern Finland by about 450 speakers. Similar to Finnish, Inari Saami exhibits devoicing of segments at the end of utterances. For Finnish, in the study of quantity Lehtonen [1] marks that the sentence endings in the informants' speech are often weakened to the extent of being completely voiceless whisper. In addition, Suomi et al. [3] note that voicelessness is also preceded by breathy voice in Finnish. There are no previous studies considering utterance-final devoicing in Inari Saami. Türk et al. [4] mention devoicing in Inari Saami with regard to the analysis of fundamental frequency values as some of their data was discarded due to creaky voice quality or devoicing.

This study aims at giving insights into the utterance-final devoicing feature in Inari Saami from the durational point of view along with the changes in voice quality within the word. The data of this study comes from the Inari Saami Prosody corpus [2] that contains sentences with di-, tri- and tetrasyllabic test words located at the end of a phrase and at the end of a sentence read by four male speakers. For the analysis of this study, the sentences with test words at the very end of the utterances were used. The durations of the whole word and the devoiced parts of the word were measured along with the durations of segments pronounced with creaky voice. The results showed that devoicing clearly occurred at the sentence-final position and was dependent on the speaker. It spanned from one segment to up to two syllables of a word. Devoiced word endings were often preceded by creaky segments. Similar patterns were found in different word structures, however, in tetrasyllabic words the average duration of the devoiced part was the longest. The results of this study indicate that devoicing of segments at the end of utterances in Inari Saami denote the end of a speech unit for some speakers but can be optional for others.

## References

[1] Jaakko Lehtonen. Aspects of quantity in standard Finnish. *Studia philologica Jyväskyläensia* 6. Jyväskylä: Jyväskylä, 1970.

[2] Pärtel Lippus et al. Inari Sami prosody corpus. Jan. 2018. url: https: //doi.org/10.15155/1-00-0000-0000-0000-00150L.

[3] Kari Suomi, Juhani Toivanen, and Riikka Ylitalo. "Finnish sound structure". In: *Studia Humaniora Ouluensia 9* (2008).

[4] Helen Türk et al. "The acoustic correlates of quantity in Inari Saami". In: *Journal of Phonetics 72* (2019), pp. 35–51. issn: 0095-4470. doi: doi.org/ 10.1016/j.wocn.2018.11.001.

# Can political science help us improve speech synthesis evaluation?

Sébastien Le Maguer & Juraj Šimko
University of Helsinki, Helsinki, Finland

**1. A bit of context.** Speech synthesis evaluation is now at a crossroad. As speech synthesis models have improved during the last decade, the direct consequence is the obsolescence of the current standard protocols. Indeed, the Absolute Category Rating (ACR) [1] (or MOS-test) and its corresponding score MOS have now shown problematic limits [2].

Alternative protocols such as the MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA) [3] or the preference tests exists. Yet, MUSHRA is improperly used as its recommendation explicitly requires the introduction of anchors (i.e., artificially constructed, degraded speech signals aiming to stabilise the rating scale). Such anchors are yet to be defined in the context of a speech synthesis evaluation. Researchers and engineers simply overlook them when conducting a MUSHRA in the context of speech synthesis evaluation; violating de-facto the recommendation. It is also well kown that using "out-of context" (generic) evaluation protocols does not capture well the differences between modern systems [4]. Yet, the metric to be used while considering the context A proposed alternative which have been advocated for [5] is the use of preference tests (or AB(X) test) to conduct the evaluation as it leads to a relative analysis. However, conducting such evaluation remains cumbersome as comparing multiple systems implies multiplying the number of steps.

Relying on the same premise - the analysis needs to be conducted in a relative manner - we propose to rely on a political science staple: the Ranked-Choice Voting (RCV)

**2. (Adapted) RCV at the rescue.** RCV is not new1 and has been well studied in computational social sciences [6] for a long time. It relies on following premise: during an election, instead of selecting a preferred candidate, the voter will rank the candidate by decreasing order of preference. The appeal of the RCV is its ability to provide a more diverse and representative outcome as the second choices are also taken into account. The difficulty mainly lies on the selection of the "winner" and multiple strategies to address it have been proposed [6].

Our goal is not to determine the best system but how a given system compares to given anchors. To do so, we first propose to consider standard systems trained using standard datasets and a given configuration as anchors. This is widely available and we should take opportunity of it. Then, as we are not interested in a score but as relative comparison of systems, we do not faces the constraint of MUSHRA. Then, we propose to adapt the RCV paradigm to fit our 1it can be traced back at least to the 13th century. needs: 1) we only consider one round; 2) listeners can associate the same rank to multiple systems (equivalent to "no-preference"); 3) listeners can choose to not rank samples which they deem of a too poor quality; 4) the samples are randomized and their corresponding system are hidden to the listener. For the present submission, no control over the samples (except play and repeat) is given to the listener. The investigation of a more refined control (as proposed for MUSHRA) is left for future work.

**3. Experimental protocol**. In order to have a comparison point, we propose to conduct the RCV evaluation based on the material of Experiment 4 presented in [2]. This experiment consisted on evaluation the naturalness of four speech synthesis systems, obtained by combining two acoustic models (FastPitch/Tacotron) and two neural vocoders (WaveNet/WaveGAN), and the natural voice as the hidden reference. The results of this experiments show that almost all the synthesis systems were deemed equivalent with the only exception being FastPitch/WaveNet being considered significantly better. By using RCV, we want to determine 1) if a clear ranking emerges 2) how different strategies can lead to a different "winner".

Experiment 4 consisted of each listener evaluating all the synthesis combinations for all utterances. This allows us to convert the scores to a system ranking for each utterance and each listener. By doing so, we can now compare the rankings obtained by the RCV procedure to the ones obtained using the ACR. The results and discussion will be presented during the conference.

## References

[1] "Methods for subjective determination of transmission quality," ITU-T, Geneva, Tech. Rep. P.800, 1996.
[2] S. Le Maguer, S. King, and N. Harte, "The limits of the mean opinion score for speech synthesis evaluation," *Computer Speech & Language, vol. 84*, p. 101577, 2024.
[3] ITU-T, "Method for the subjective assessment of intermediate sound quality (MUSHRA)," *International Telecommunication Union (ITU-R), Tech. Rep. BS*.1534-1, 2001.
[4] J. O'Mahony, P. Oplustil-Gallegos, C. Lai, and S. King, "Factors Affecting the Evaluation of Synthetic Speech in Context," in *ISCA Speech Synthesis Workshop, 2021*, pp. 148–153.
[5] S. Shirali-Shahreza and G. Penn, "Mos naturalness and the quest for human-like speech," *IEEE Spoken Language Technology Workshop (SLT)*, Dec 2018.
[6] H. Nurmi and R. P. Palha, "A theoretical examination of the ranked choice voting procedure," in *Transactions on Computational Collective Intelligence XXXVI*. Springer, 2021, pp. 1–16.

## *D*-äänne ja sen vastineet 2020-luvun puhutussa suomessa

Emma Heikkinen & Michael O'Dell

Helsingin yliopisto

Yksi suomen kielen uusimmista äänteistä, *d*, on ollut erityisesti 1970–90-luvuilla tehdyissä paikkakuntakohtaisissa murretutkimuksissa yksi tarkastelluimpia ilmiöitä. *D*-äänteen ja sen vastineiden käytön on huomattu riippuvan mm. maantieteellisestä alueesta sekä puhujan iästä ja sukupuolesta. Tämän lisäksi *d*:n on tuolloin nähty olevan vahvassa murroksessa koko maassa ja tämä muutos on vaikuttanut olevan kaksisuuntaista: samalla, kun yleiskielinen *d* on vallannut alaa murteellisten varianttien kustannuksella, osa *d*-äänteen vastineista on pyrkinyt leviämään perinteisten alueidensa ulkopuolelle. Me halusimme selvittää, mikä *d*-äänteen ajankohtainen tila 2020-luvun puhutussa suomessa on. Suurta datamassaa apunamme käyttäen pyrimme osoittamaan, missä päin Suomea *d* on yleinen ja missä harvinainen ottaen huomioon sekä iän että sukupuolen tuoman vaikutuksen. Tutkimuksessa tarkasteltiin myös, mitä äänteitä *d*:n paikalla käytetään, kun *d*:tä ei käytetä ja miten äänneympäristö vaikuttaa *d*:n ja sen vastineiden esiintymiseen. Samalla pyrimme arvioimaan, mihin kehitys saattaa olla menossa ja pohdimme, mikä on muuttunut 1900-luvun tutkimusten jälkeen.

Tutkimuksen aineistona käytettiin viime vuosina Lahjoita puhetta -kampanjassa digitaalisesti kerättyjen puhetiedostojen transkriptioita yhteensä noin 15 000 puhujalta [1]. Aineistoa oli kaikista Suomen maakunnista (tässä tutkimuksessa pl. Ahvenanmaa) ja yhdistettynä metatietoihin, kuten ikään ja sukupuoleen, pystyttiin tekemään vertailevia tilastollisia analyysejä *d*:n yleisyydestä eri ryhmissä. Tulokset osoittavat, että *d* on yleisimmillään eteläisessä ja lounaisessa Suomessa ja sen käyttö vähenee asteittain pohjoista ja itää kohti siirryttäessä. *D*:n yleisin vastine koko maassa on itämurteille tyypillinen kato, jonka edustus on vahvimmillaan pohjoisessa ja itäisessä Suomessa, eli käänteisesti *d*:n yleisyyteen nähden. Länsimurteisten vastineiden (*r*, *l*) määrät ovat suhteellisen pieniä ja tutkimus antaakin viitteitä *l*:n lähes täydellisestä häviämisestä *d*:n vastineena, kuten myös *r*-vastineen melko kriittisestä tilasta etenkin Etelä-Pohjanmaan ulkopuolella. Äänneympäristön vaikutus *d*:n esiintymiseen on selvä ja mukailee aiemmissa tutkimuksissa tehtyjä löydöksiä: *d* ääntyy kaikilla alueilla varmimmin lyhyen vokaalin jäljessä, harvemmin pitkän vokaaliaineksen jäljessä ja harvimmin *hd*-yhtymässä.

Murretaustalla vaikuttaa olevan tutkituista muuttujista selvin vaikutus *d*:n esiintyvyyteen, mikä oli aiemman tutkimuksen perusteella odotettavissa. Eroja löytyy myös sukupuolten ja ikäluokkien väliltä. Aineiston miehillä *d* on jonkin verran naisia yleisempi, mikä eroaa monista aiemmista tutkimuksista. Iän vaikutus *d*:n esiintyvyyteen on valtakunnallisesti vähäistä, mutta alueellisesti merkittävää. Tulokset osoittavat, että nuoremmat puhujat usein käyttävät *d*:tä vanhempia puhujia enemmän alueilla, joilla vastineiden osuus on korkea ja vanhempia vähemmän niillä, joilla *d* on yleisempi. Vaikuttaa siis siltä, että *d* on vahvistamassa asemaansa vahvempien murre-edustusten alueilla, kun taas katoedustus on valtaamassa alaa yleiskielisimmillä alueilla. Täten erot eri alueiden välillä näyttävät olevan kapenemassa molemmista suunnista, mitä myös aikaisempi tutkimus on ennustanut yksittäisten murteiden osalta.

### Viitteet

[1] Moisio, Anssi – Porjazovski, Dejan – Rouhe, Aku – Getman, Yaroslav – Virkkunen, Anja – AlGhezi, Ragheb – Lennes, Mietta – Grósz, Tamás – Lindén, Krister –Kurimo, Mikko 2023: Lahjoita puhetta: a large-scale corpus of spoken Finnish with some benchmarks. *Lang Resources & Evaluation* 57, 1295–1327. Saatavissa: https://doi.org/10.1007/s10579-022-09606-3

## Does orthography affect the perception of Estonian vowels?

Katrin Leppik, Kaidi Lõo, Eva Liina Asu & Pärtel Lippus

Institute of Estonian and General Linguistics, University of Tartu

Previous studies (Leppik et al., 2023) have shown that Spanish L1 learners of Estonian struggle differentiating Estonian mid vowels. In production the learners merge /ø/ and /ɤ/ into an ambiguous mid vowel and often confuse it with /o/. In perception the learners identified /ø/ most often as /ɤ/ (29%) and /e/ (21%), and only in 5% of the cases was /ø/ identified as /o/. This indicates a discrepancy between perception and production which suggests that the orthography might have an influence on the perception of L2 learners. Most of the Estonian vowels that are difficult for Spanish L1 learners are presented in orthography with modified Latin letters *ä*, *ö*, *õ*, *ü*. Thus, it might be the case that the learners simply ignored the diacritics above the letters and merged the unfamiliar vowel categories with the vowels that are represented with the unmodified character (*a*, *u*, *o*).

There are several studies that have noted the interference of orthography in L2 production (e.g., Escudero and Wanrooij, 2010; Young-Scholten et al., 2002). Regarding Estonian, it has been suggested that orthography also interferes with quantity production (Meister et al., 2015).

A Visual World Paradigm eye tracking study is planned to investigate the effect of orthography on vowel perception. The study combines pictures and printed words which are presented in quadruplets, where every quadruplet has a target (e.g., *söök*), orthographic competitor (e.g., *sool*), phonetic competitor (e.g., *seen*) and unrelated distractor (e.g., *pall*). The target and competitors always start with the same consonant. During the experiment the participant hears a sentence, e.g., '*Vali pilt, kus on köök'* and has to click on the named picture (or printed word). Half of the stimuli are presented as pictures and the rest as printed words. It is expected that the printed words with similar letters receive more looks than the same stimuli presented with pictures. The paper will present the preliminary results from the study.

### References

Escudero, P., & Wanrooij, K. (2010). The effect of L1 orthography on non-native vowel perception. *Language and speech, 53*(3), 343-365.

Leppik, K., Lippus, P., & Asu, E. L. (2023). The perception and production of Estonian vowels and vocalic quantity contrasts by Spanish L1 learners. *Ampersand, 11*, 100147. DOI: 10.1016/j.amper.2023.100147.

Meister, E., Nemoto, R., & Meister, L. (2015). Production of Estonian quantity contrasts by Japanese speakers. Eesti ja soome-ugri keeleteaduse ajakiri. *Journal of Estonian and Finno-Ugric Linguistics, 6*(3), 79-96.

Young-Scholten, M. (2002). Orthographic input in L2 phonological development. An integrated view of language development: *Papers in honor of Henning Wode.*

## Kveenin kielen vokaalilaadut ja -kestot ensikielisillä ja ei-ensikielisillä puhujilla: tutkimushankkeen esittely

Saloranta, Antti[1,2] & Heikkola, Leena Maria[1,3]

1) Kielten ja kulttuurin laitos, UiT – Norjan arktinen yliopisto 2) Fonetiikka ja Learning, Age & Bilingualism -laboratorio, Turun yliopisto 3) Suomen kieli, Åbo Akademi

Kveeni on suomen peräpohjalaisiin murteisiin perustuva noin kahden tuhannen puhujan kieli, joka ollut virallinen vähemmistökieli Norjassa vuodesta 2005. Kveenin puhujia on historiallisesti yritetty norjalaistaa, mutta nykyisin kieltä elvytetään aktiivisesti. Kveenin kieltä on kuitenkin tutkittu varsin vähän, erityisesti (sosio)foneettisesta näkökulmasta. Nyt esiteltävän vuosina 2025–2026 toteutettavan tutkimushankkeen tavoitteena on kartoittaa nykyisin puhuttavan kveenin äännejärjestelmää sekä vertailla äännejärjestelmiä ensikielisten ja ei-ensikielisten kveeninpuhujien välillä.

Kveenin kielen tarkastelu äännetasolla on tärkeää, sillä norjalaistamispolitiikka ja ympäröivän yhteiskunnan norjankielisyys on voinut vaikuttaa kveenin äänteisiin useilla eri tavoilla. Muistuttavatko vokaalien laadut suomea, vai onko niissä tapahtunut muutoksia norjan kielen vaikutuksesta? Tuotetaanko kveenissä esiintyvä vokaalien ja konsonanttien kestoero puhtaasti kestoerona kuten suomessa, vai onko kestoilla myös laadullisia eroja? Onko näissä piirteissä eroja lapsuudessaan kveeniä puhuneiden ja kveenin aikuisella iällä oppineiden välillä?

Sekä kveenissä, suomessa että norjassa äänteiden kestolla on sanojen merkitystä erotteleva vaikutus. Kveenissä on suomen tapaan kahdeksan vokaalia, /i/, /y/, /u/, /e/, /ø/, /o/, /æ/ ja /ɑ/, joista kaikki voivat esiintyä sekä lyhyinä että pitkinä. Norjassa kuvataan tyypillisesti olevan yhdeksän pitkää (/iː, yː, ʉː, uː, eː, øː, oː, æː, ɑː/) ja yhdeksän lyhyttä vokaalia (/i, y, ʉ, u, ɛ, oe, ɔ, æ, ɑ/). Suomessa erikestoisten vokaalien laadussa on pientä eroa, mutta sitä pidetään tyypillisesti havaitsemisen kannalta epäolennaisena. Norjassa pitkien ja lyhyiden vokaalien laadullinen ero vaihtelee vokaalista riippuen, mutta on merkittävämpää kuin suomessa. Kveenin osalta vastaavaa tutkimusta ei ole vielä tehty.

Hankkeen päätutkimuskysymykset ovat:

1) Onko kveenin vokaaleissa laatueroja ensikielisten ja ei-ensikielisten puhujien välillä?

2) Toteutuuko kveenin vokaalien kestoero vain ajallisena erona vai kuuluuko siihen myös laadullista vaihtelua?

3) Miten norjan mahdollinen vaikutus ilmenee näissä piirteissä?

Kysymyksiin vastaamiseksi hankkeessa sekä kerätään uutta puheaineistoa että hyödynnetään olemassa olevaa korpusaineistoa. Uutena aineistona kerätään vapaata puhetta sekä foneettisesti kontrolloituja sanalistoja. Aineistoista analysoidaan akustisesti kohdeäänteiden kestot sekä vokaalien F1- ja F2-formantit. Uusissa nauhoituksissa puhujilta kerätään sekä kveenin- että norjankielistä aineistoa. Tutkimuksesta saatavat tulokset tarjoavat täysin uutta tutkimustietoa suomen lähikielistä ja niiden muutoksesta ajan myötä. Hankkeen tarkempaa toteutusta ja mahdollisia alustavia tuloksia esitellään Fonetiikan päivillä.

# Presenting an Alternative Creak Detection Algorithm

Viljami Haakana & Reetta Jokinen

University of Helsinki

Creaky voice detection is notoriously difficult and has been attempted in the past. A notable example of a creaky voice detection algorithm is the Covarep algorithm (Drugman et al. submitted, Kane et al. 2013, Drugman et al. 2012), written in MATLAB. However, when using this algorithm to detect creaky voice in the French language child speech data collected in 2023 and 2024 for the project A Crosslinguistic Investigation of Prosodic Patterns Related to Autism Spectrum Disorder: Acoustic, Experimental and Multimodal Analysis (PROSO-ASD), led by Mari Wiklund, a very large number of false positives was found.

It has been previously noted (e.g. Watkins et al. 2024) that "octave jumps" that are normally regarded as errors in automatic pitch detection, are often the result of creaky voice. Therefore, we thought that if we analyzed the fundamental frequency (f0) with two different algorithms, one that is prone to octave jumps and one that is not, by comparing those two algorithms' outputs we could detect creaky voice. If the algorithm less prone to octave jumps, in our case Antti Suni's PitchSqueezer (Suni, 2023), and the one more prone to octave jumps (Praat's and Parselmouth's Cross-Correlation "cc"), disagree by a large enough margin on the f0, that segment is identified as creaky voice. We also developed a variant of this algorithm ("shs variant") where regions that are completely voiceless according to a third pitch detection algorithm, namely "shs" in Praat/Parselmouth, are excluded.

Preliminary results show that for the few samples of the neurotypical Swiss French child speech data where we have tested, our algorithms work significantly better than Covarep. 20 % of the non-creaky audio signal was falsely identified as creaky by Covarep, but only 6 % by our algorithms. The "true accuracy" (correctly identified creaks / (correctly identified creaks + unidentified obvious creaks + misidentified obvious non-creaks)) of our algorithms for the French neurotypical data was 57 %, but only 37 % for Covarep. Our algorithm has worked better than Covarep also for the autistic French data we have tested it on, but to a lesser extent. For both neurotypical and autistic Finnish data so far tested, Covarep has worked better, and the same is true for the neurotypical data in the Slovak Autistic and Non-Autistic Child Speech Corpus (SANACS). For autistic Slovak data, our algorithms have so far worked better than Covarep.

Preliminary results also show that [s] is one of the most common sounds during which our algorithms output false positives, accounting for 9 % of them in the non-shs variant and 8 % in the shs variant. Thus, reducing the volume of [s], as is often done in music production (see Jared H. 2021), could also improve these algorithms.

## References

Drugman, T., Kane, J., & Gobl, C., (2012). *Resonator-based creaky voice detection*, Interspeech 2012, Portland, Oregon, USA.

Drugman, T., Kane, J., & Gobl, C., *Automatic Analysis of Creaky Excitation Patterns*, Submitted to Computer Speech and Language.

Jadoul, Y., Thompson, B., & de Boer, B. (2018). Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71, 1-15. https://doi.org/10.1016/j.wocn.2018.07.001

Jared H. (2021). *De-esser: The Guide for Sibilant-Free Vocal Recordings*. https://ledgernote.com/columns/mixing-mastering/de-esser/

Kane, J., Drugman, T., & Gobl, C., (2013). *Improved automatic detection of creak*, Computer Speech and Language 27(4), pp. 1028-1047.

Kruyt, J., Sabo, R., Polónyiová, K., Ostatníková, D. & Beňuš, Š. (2024). The Slovak Autistic and Non-Autistic Child Speech Corpus:Task-Oriented Child-Adult Interactions. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 16094–16099, Torino, Italia. ELRA and ICCL.

Watkins, M., Boersma, P., & Hamann, S. (2024). *Revisiting pitch jumps: F0 ratio in Seoul Korean*. Interspeech 2024, 3135–3139. https://doi.org/10.21437/interspeech.2024-1184

Suni, A. (2023). *PitchSqueezer*. https://github.com/asuni/PitchSqueezer

Covarep creaky voice detection algorithm source code: https://github.com/covarep/covarep/tree/master/glottalsource/creaky_voice_detection

PROSO-ASD project website: https://blogs.helsinki.fi/asd-prosody-research

# Modeling Cross-Linguistic Dialectal Variation with Self-Supervised Speech Representations

Tuukka Törö[1], Priyankoo Sarmah[2], Antti Suni[1] & Juraj Šimko[1,2]

1) University of Helsinki, Finland 2) Indian Institute of Technology, Guwahati, India

Quantifying and comparing dialectal variation across languages presents several challenges. As languages differ on multiple levels—from individual phonemes to prosodic patterns and from morphology to syntactic structure—it is difficult to establish cross-linguistic categories for comparison without introducing theoretical bias and to quantify them in a mutually compatible way. A recently developed approach using speech representations derived from self-supervised machine-learning models has been shown to overcome some of these challenges, capturing relationships among languages and dialects in terms of their sound patterns, phonetic and prosodic characteristics, as well as sociolinguistic influences (Suni et al., 2019; Hiovain et al., 2022; Törö et al., 2024a, 2024b).

Here, we present a comparative analysis of two unrelated languages: Finnish (a Uralic language spoken in Europe) and Assamese (an Indo-Aryan language spoken in Northeastern India). We focus on dialectal distributions in the two languages and how dialectal diversity interacts with factors such as age and gender. The analysis is based purely on the speech signal and uses self-supervised speech representations fine-tuned for the language identification task (Kukk & Alumäe, 2022).

This approach enables a holistic and relatively theory-independent way of studying how dialectal variation manifests in different languages and among various speaker groups based on gender and age.

## References

Kukk, K., & Alumäe, T. (2022). Improving language identification of accented speech. arXiv preprint arXiv:2203.16972.

Suni A, Włodarczak M, Vainio M, Šimko J. (2019). Comparative analysis of prosodic characteristics using WaveNet embeddings. In: *Annual Conference of the International Speech Communication Association: Interspeech 2019*. ISCA, p. 2538–2542.

Hiovain K, Suni A, Kakouros S, Šimko J. (2022). Comparative analysis of majority language influence on North Sami prosody using WaveNet-based modeling. *Language and Speech; 65*(4):859–888.

Törö T, Suni A, Šimko J. (2024). Quantifying sociolinguistic change: Effects of age and gender on dialectal variation in a large corpus of spontaneous Finnish speech. In: *Proceedings of ISAPh 2024*, p. 105–109.

Törö T, Suni A, Šimko J. (2024). Emergent Dialectal Patterns: Analysis of regional variants in a vast corpus of Finnish spontaneous speech using a large-scale self-supervised model. In: *Proceedings of Speech Prosody 2024*, p. 37–41.

Hsieh SK, Tseng YH, Lian DC, Wang CW. (2024). Self-supervised learning for Formosan speech representation and linguistic phylogeny. *Frontiers in Language Sciences; 3*:1338684.

# Exploring the nature of cross-language (mis)perception of lexical stress: a case of Finnish and Russian

Tatiana Kachkovskaia[1,2], Michael O'Dell[2] & Tommi Nieminen[3]

1) Helsinki Collegium for Advanced Studies, 2) University of Helsinki, 3) Omnia

Cross-linguistic perception of word stress seems to vary depending on the cues used in the listeners' native language [4]. Thus, in an experiment with manipulated preudo-word stimuli for Estonian and Russian, Estonian listeners were found to be more sensitive to pitch, and Russian listeners to duration [2]. The case of Finnish, a language often described as having fixed stress on the first syllable, is particularly interesting: in some Finnish words, speakers of other languages are reported to perceive stress elsewhere in the word, especially if the first syllable of the word is light (cf. e.g. [1], [5]).

Here, we present intermediate results of on-going research on the perception of stress in Finnish words by speakers of Russian as L1 who have no experience with Finnish [3]. (Importantly, this research is not directly concerned with L1-L2 interaction as it deals with naïve listeners.) We continue to explore the phonetic and phonological factors underlying stress perception. The main goal of this research is to answer the following question: What is the role of vowel category in the perception of stress in Finnish words by Russian speakers?

Our preliminary perception experiment showed that, indeed, Russian listeners sometimes perceive Finnish words as having 2nd syllable stress. However, despite all target words having a CVCVCV structure, there was significant variation between the stimuli; at the same time, our analysis indicated a tendency for the vowel /i/ to "attract" lexical stress.

At this stage, we are trying to trace the complex interaction of vowel quality and vowel duration in the perception of stress in Finnish by naïve Russian listeners. We will report on vowel formant measurements, more sophisticated duration analysis, and frequency of vowel categories in Russian stressed and unstressed syllables.

## References

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

[2] A. Eek, "*The Perception of Word Stress: A Comparison of Estonian and Russian*". In honor of Ilse Lehiste: Ilse Lehiste Pühendusteos, edited by Robert Channon and Linda Shockey, Berlin, New York: De Gruyter Mouton, 1987, pp. 19–32

[3] T. Kachkovskaia, M. O'Dell, T. Nieminen, "Russian speakers' perception of stress in Finnish words". Presented *at The 36th Finnic Phonetics Symposium (Fonetiikan Päivät),* Tallinn, April 25, 2024

[4] A. Tremblay, "The Past, Present, and Future of Lexical Stress in Second Language Speech Production and Perception". In: Wayland R, ed. *Second Language Speech Learning: Theoretical and Empirical Progress*. Cambridge University Press; 2021:175–192

[5] V. V. Vihanta, "Suomi vieraana kielena foneettiselta kannalta." In *Vieraan kielen ymmartaminen ja tuottaminen. AFinLA:n vuosikirja* 1990, J. Tommola, Ed., AFinLA, 1990, pp. 199–225

## Foneettisten harjoitteiden vaikutus monikielisten lasten suomen äänteiden tuottoon

Minna Rihko, Anna Nurmi, Katja Haapanen, Henna Tamminen, Kimmo U. Peltola & Maija S. Peltola

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

Turun kaupungin perusopetuksessa on n. 2600 (yli 20 %) maahanmuuttajataustaista oppilasta. Monikielisten lasten oppimisen pulmista nousi tarve tutkia asiaa tarkemmin ja kehittää opettajille oppilaan kielitaustat huomioivaa tukimateriaalia. KiTu - hankkeen päämääränä on löytää tutkitun tiedon avulla keinoja tukea monikielisten lasten suomen kielen oppimista. Hyvän suomen kielen taidon saavuttaminen auttaa lapsia integroitumaan koulumaailmaan, myöhemmin jatko-opintoihin ja lopulta työelämään, jolloin maahanmuuttajien kiinnittyminen yhteisöön ja yhteiskunnallinen osallisuus mahdollistuvat kaikille yhdenvertaisesti.

KiTu-hankkeessa tutkimme neljän erilaisen harjoitteen toimivuutta erikielisillä oppijoilla. Tutkimukseen valitut 2-6-luokkalaiset lapset edustavat Turun kaupungin koulujen kuutta suurinta S2-kieliryhmää: arabia, somali, kurdi, albania, ukraina ja venäjä. Harjoitteiden suunnittelussa hyödynnettiin vieraan kielen oppimisen kontrastiivisia teorioita (mm. Flege & Bohn 2021, Best & Tyler 2007). Taustana toimivat LAB-labin aiemmat tutkimukset, joissa erilaisten harjoitteiden on todettu tukevan oppijoita moninaisin tavoin (Savo et al. 2019, Peltola KU et al. 2015, Saloranta et al. 2020, Immonen et al. 2022, Tamminen et al. 2015). Tässä tutkimuksessa testasimme, miten suomen sanojen tuottaminen kehittyi neljän eri harjoitteen (audio, kuuntele ja toista, ortografia ja huulio) avulla. Harjoitusten vaikutusta äänteiden laatuun ja kestoon mitattiin akustisella analyysilla. Analyysin tuloksia hyödynnetään hankkeessa kehitettävän työvälineen suunnittelussa. Työvälineen tavoitteena on tarjota konkreettisia keinoja suomen suullisen kielitaidon oppimiseen alakouluissa.

Posterissamme esittelemme akustisen analyysin ensimmäisiä tuloksia harjoitteiden vaikutuksesta äänteiden laatuun eri kieliryhmillä.

### Viitteet

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. Teoksessa O.-S. Bohn & M. J. Munro (Toim.), *Language Learning & Language Teaching* (Vsk. 17, ss. 13–34). John Benjamins Publishing Company. https://benjamins.com/catalog/lllt.17.07bes

Flege, J. E., & Bohn, O.-S. (2021). The revised speech learning model (SLM-r). Teoksessa R. Wayland (Toim.*), Second language speech learning: Theoretical and empirical progress* (ss. 3–83). Cambridge University Press.

Immonen, K., Alku, P., & Peltola M. 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, K. U., Tamminen, H., Alku, P., & Peltola, M. 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, A., Alku, P., & Peltola, M. 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 147*, 72-82.

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

Tamminen, H., Peltola, M. S., Kujala, T., & Näätänen, R. (2015). Phonetic training and non-native speech perception—New memory traces evolve in just three days as indexed by the mismatch negativity (MMN) and behavioural measures. *International Journal of Psychophysiology, 97*(1), 23–29. https://doi.org/10.1016/j.ijpsycho.2015.04.020

# Continuation rise in contacting tone languages: A study of postlexical tone variation in Mano and Kpelle (Guinea)

Maria Konoshenko & Maria Khachaturyan
University of Helsinki

To date, research on contact-induced variation in intonation has mainly explored European languages (Elordieta & Romera 2021; McCarthy & De Leeuw 2022), with some rare exceptions (Lai and Gooden 2022). Lexical tone languages, which are the focus of this paper, typically feature a complex interplay of lexical and postlexical tone phenomena (Downing and Rialland 2017), and yet, little is known about contact-induced intonational variation in lexical tone languages, the existing studies mostly focussing on Southeast Asian languages (Jingwei et al. 2023).

This paper reports the results of a project exploring a critically understudied phenomenon, multilingual variation in prosody involving two contacting African tone languages from the Mande family, Guinean Mano [iso 639-3:mev], featuring a three-way L/M/H lexical tone contrast, and Guinean Kpelle [iso 639-3:gkp], having a binary L/H contrast. While Continuation Rise (CR) as a boundary phenomenon is rather rare in baseline (L1) Mano; in baseline Kpelle, CR is an optional boundary H% tone indicating non-finality, it most often appears on non-utterance-final syntactic clause boundaries, or after hanging topics. H% surfaces on the final tone bearing unit constituent-final non-monomoraic syllable, or on an additional mora due to lengthening of monomoraic syllables. It is usually realised at least twice as high as the highest H of the utterance.

This study is based on two datasets: (i) a semi-spontaneous picture questionnaire originally devised to study reflexives (Khachaturyan et al. 2024) comprising data from 22 adult speakers (686 utterances) and (ii) spontaneous folk tale narratives recorded by co-authors from 8 adult speakers (1432 chunks) with various Mano-Kpelle acquisition trajectories (simultaneous and sequential Kpelle-Mano bilinguals). Kpelle L1 speakers were recorded in Kpelle; all other speaker types recorded in Mano. Both datasets were annotated in Praat and processed in Excel, Power Pivot and R by the first author. The observations are of two types: manual categorical labels and semitone-transformed acoustic *f*0 features (mean *f*0 in preceding context, min *f*0, max *f*0 of the rise) automatically extracted with Praat scripts written by Juraj Šimko and the first author.

Our findings are two-fold. First, we show that CR is produced by Mano speakers (2) having a higher degree of Kpelle exposure, hence we attribute CR in Mano to the cross-linguistic influence of Kpelle. However, we further demonstrate that the phonetic implementation of the Continuation rise also differs across speakers depending on their Mano-Kpelle exposure trajectories: the maximal *f*0 in CR is generally higher in baseline Kpelle; and speakers with higher Kpelle exposure have more Kpelle-like CR production (Fig.3). Using a linear regression model, we also show that, overall, the highest point of Continuation rise (max *f*0) is predicted not only by the speaker type, but also by mean *f*0 of immediate left context chunk; suggesting a joint influence of phonetic and social factors on the realisation of Continuation rises in multilingual Mano-Kpelle ecology.

## References

Downing, Laura J., and Annie Rialland, eds. 2017. Intonation in African Tone Languages. De Gruyter Mouton.

Elordieta, Gorka, and Magdalena Romera. 2021. "The Influence of Social Factors on the Prosody of Spanish in Contact with Basque." *International Journal of Bilingualism 25*(1):286–317.

Gussenhoven, Carlos. 2004. The phonology of tone and intonation. Cambridge: Cambridge University Press.

Jingwei, Zhang, Tan Weijie, and Christopher Strelluf. 2023. "*Sociophonetics and Chinese.*" in *The Routledge Handbook of Sociophonetics*. Routledge.

Khachaturyan, M., Moroz, G., & Mamy, P. (2024). Do languages spoken in multilingual communities converge? A case study of reflexivity marking in Mano and Kpelle. *Linguistics*. https://doi.org/10.1515/ling-2022-0111

Lai, Li-Fang, and Shelome Gooden. 2022. "Language Contact, Language Ecology, and Intonational Variation in the Yami Community." *Language and Speech 65*(4):791–832. doi: 10.1177/00238309221115636.

McCarthy, Kathleen M., and Esther de Leeuw. 2022. "Prosodic Patterns in Sylheti-English Bilinguals." *Studies in Second Language Acquisition 44*(2):562–79.