Nordic Prosody XIV

Book of Abstracts

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

18–20 August 2026

An ISCA-Supported Event

Nordic Prosody XIV (NPXIV)
KTH Royal Institute of Technology
Department of Speech, Music and Hearing (TMH)
Stockholm, Sweden, 18–20 August 2026

Programme and practical information: https://npxiv.nordicprosody.org
Contact: npxiv@nordicprosody.org

Preface

This volume collects the abstracts of the work presented at Nordic Prosody XIV, held at KTH Royal Institute of Technology in Stockholm on 18–20 August 2026.

The texts gathered here are *meeting abstracts*: short accounts of work presented and discussed at the conference. They are not publications. They have not been through the kind of review a journal article or a proceedings paper goes through, they vary in length and format according to what their authors found useful, and they are not intended to be cited as finished, archival contributions.

This book exists so that participants can find their way through the programme while it is running, and so that, some weeks later, a title and half a page of text can bring back the talk itself and what was said about it. It is a mnemonic device more than anything else.

A number of the studies presented here will appear in full or extended form in the Nordic Prosody XIV book, to be published after the conference. For those contributions, the version to read — and the version to cite — is the one that will appear there.

The abstracts are reproduced exactly as submitted, and are arranged in the order in which they appear in the conference programme. Titles and author names in the contents and the author index follow the submitted abstracts.

The organising committee
Stockholm, August 2026

Contents

Session 5: Tools and demos

Session 6: Accents

Session 7: Prominence and boundaries

Session 8: Models

Session 9: Deviant speech

Author index

Effects of Body Posture on Voice Quality

Johan Sundberg[1,2], *Helena Engström*[1], *Christine Ericsdotter Nordgren*[1], *Susanne Fuchs*[4], *Mattias Heldner*[1], *Patrycja Strycharczuk*[3], *Marcin Włodarczak*[1]

[1]Stockholm University, Sweden
[2]KTH Royal Institute of Technology, Sweden
[3]University of Manchester, UK
[4]Leibniz-Centre General Linguistics (ZAS), Germany

jsu@kth.se

1. Introduction

Variation of voice quality (phonation type) is by now firmly established as a core dimension of prosodic expression [1, 2] with functions ranging from phonological to para- and extralinguistic [3]. Conveying such a wide range of meanings requires precise phonatory control, not least because voice production routinely accompanies other activities (such as walking or gesturing) which substantially alter the physiological settings of the entire speech apparatus.

In particular, all healthy speakers are able to produce speech while standing upright, lying down and hanging head-down in spite of the changing coupling between the respiratory and phonatory systems. Specifically, when standing up, the lungs are pulling caudally on the larynx, tending to lower it and exerting an abductive force on the vocal folds. This effect, known as tracheal pull, is stronger at high lung volumes, due to a lower position of the diaphragm, and tends to result in less hyperfunctional (pressed) phonation [4, 5]. By contrast, tracheal pull should be absent in both supine and, to an even greater extent, inverted body positions. In this condition, the abdominal content pressing against the diaphragm is expected to reduce the rib cage volume, generating increased subglottal pressure (P_s). It would also counteract the tracheal pull and, consequently, increase glottal adduction.

While there is some evidence that classically trained speakers are able to maintain unchanged P_s levels in the supine position [6], little is known about P_s regulation and posture-related phonatory effects in untrained subjects. In this study we analyse the relationship between P_s and glottal adduction across three body positions with the aim to provide some preliminary data on the relevance of gravity to phonation.

2. Method

Six speakers were recorded in upright, supine and partly inverted ($-30°$) positions. The recordings were part of a larger project studying speech adaptation to changes in body position and included co-registration of breathing kinematics (using Respiratory Inductance Plethysmography), vocal fold contact (using electroglottography), ultrasound tongue imaging and audio recordings. In this study, we use a subset of the data involving subglottal pressure and inverse filtered glottal source measurements in diminuendo sequences (i.e. with gradually decreasing loudness) of syllable [pæː] produced at stable f_0 by two male speakers.

Audio was recorded using an omnidirectional microphone (DPA 4061) attached to the ultrasound headset at a constant distance from the lip opening and calibrated in dB SPL using a sound level meter. The glottal airflow was derived by manually inverse-filtering the quasi-stationary portion of the vowel in Sopran [7]. The inverted formant frequencies and bandwidths were tuned applying as criteria ripple-free closed phase and a source spectrum envelope as smooth as possible near the formant frequencies. The resulting flow glottogram was subsequently analyzed with respect to the maximum flow declination rate (MFDR), the peak-to-peak amplitude of the glottal airflow pulse ($Flow_{ptp}$), the closed quotient (the duration ratio between the closed phase and the period, CQ), the skewing quotient (the duration ratio between the glottal opening and closing phases, SQ), the level difference between the first and the second voice source partials (L_1L_2) and amplitude quotient (the ratio between the $Flow_{ptp}$ and MFDR, AQ).

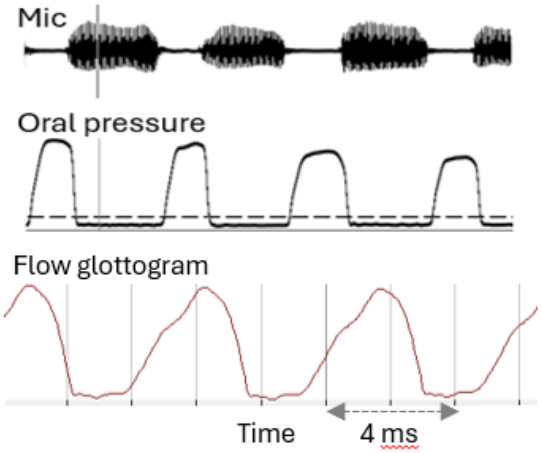


Figure 1: *The audio signal (top panel), P_s (middle panel) and a zoomed in section of the flow glottogram (bottom panel).*

Subglottal pressure was measured as the oral pressure during the occlusion phase of the consonant /p/. It was captured by means of a plastic tube, placed in the corner of the mouth and connected to a Subglottal Pressure Monitor PG-60 (Glottal Enterprises). The pressure recordings were calibrated by submerging the tube end in a glass of water at known depth.

An example of the three signals (audio, P_s, and flow glottogram) is presented in Figure 1.

3. Results

In both speakers, P_s produced higher MFDR values in Upright as compared with Inverted (Figure 2). Some other systematic effects of MFDR variation were found in Speaker 1. Pulse amplitude, which increased with MFDR, was highest in the upright and lowest in the inverted position (Figure 3). Dominance of the

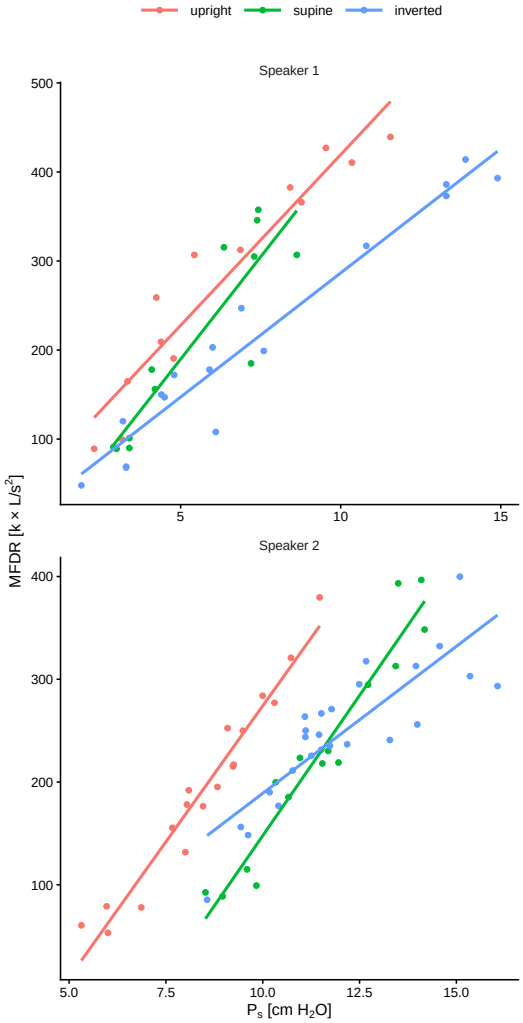


Figure 2: *MFDR as a function of P_s in the upright, supine and inverted body positions for Speaker 1 (top panel) and 2 (bottom panel).*

fundamental was strongest in the supine position but, surprisingly, weakest in upright position. In speaker 2, strong linear correlations were observed between f_0 and P_s, R^2 exceeding 0.93 in all three positions. For a given P_s, f_0 was highest in the upright and lowest in supine position.

4. Discussion

The results differed markedly between the two participants with regard to how systematic the glottal source parameters varied with P_s and with MFDR.

The systematic effects observed in Speaker 1 are in line with the earlier findings that tracheal pull tends to reduce glottal adduction during phonation. An increase of this adduction tends to reduce the flow pulse amplitude and the amplitude quotient. Generally, these effects are associated with a decreasing amplitude of the voice source fundamental. The reason why this was not observed in the present voices is left to future investigation. In Speaker 2, with the exception of f_0, the measured voice source properties exhibited an unsystematic variation with

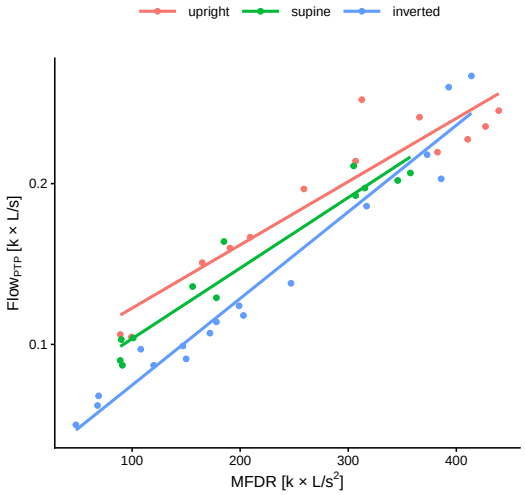


Figure 3: *$Flow_{ptp}$ as a function of MFDR in the upright, supine and inverted body positions for Speaker 1.*

MFDR for all body positions. Indeed, it is possible that the very strong relationship between MFDR and f_0 may have obscured the effects of body position.

Summarizing, changes between upright, supine and inverted body position affects the voice source in speaker-specific ways, some of which are compatible with the tracheal pull effects.

5. References

[1] N. Campbell and P. Mokhtari, "Voice quality: The 4th prosodic dimension," in *Proceedings of the 15 International Congress of Phonetic Sciences (ICPhS)*, 2003, pp. 2417–2420.

[2] A. N. Chasaide, I. Yanushevskaya, J. Kane, and C. Gobl, "The voice prominence hypothesis: The interplay of f0 and voice source features in accentuation." in *Proceedings of Interspeech 2013*, Lyon, France, 2013, pp. 3527–3531.

[3] C. Gobl and A. Ní Chasaide, "Voice source variation and its communicative functions," in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. Laver, and F. E. Gibbon, Eds. Oxford: Wiley-Blackwell, 2012, pp. 378–423.

[4] J. Iwarsson, M. Thomasson, and J. Sundberg, "Effects of lung volume on the glottal voice source," *Journal of Voice*, vol. 12, no. 4, pp. 424–433, 1998.

[5] J. Iwarsson and J. Sundberg, "Effects of lung volume on vertical larynx position during phonation," *Journal of Voice*, vol. 12, no. 2, pp. 159–165, 1998.

[6] J. Sundberg, R. Leanderson, C. von Euler, and E. Knutsson, "Influence of body posture and lung volume on subglottal pressure control during singing," *Journal of Voice*, vol. 5, no. 4, pp. 283–291, 1991.

[7] Tolvan Software, "Sopran." [Online]. Available: https://www.tolvan.com/index.php?page=/sopran/sopran.php

Laryngeal and Supralaryngeal Contributions to Speaking Style

Donna Erickson[1], *Julián Villegas*[2], *Karen Perta*[3], *Zhaoyan Zhang*[4], *Jeff Moore*[5], *Zen Nakatsuka*[6]
Yuya Goto[7], *Seunghun Lee*[8]

[1]Haskins Labs, USA, [2]The University of Aizu, Japan,[3]Elmhurst University, USA,[4]University California Los Angeles, USA,[5]Tokyo University of Agriculture and Technology, [6]Zen Voice Factory, Japan, [7]the Voice Expert, LLC, Japan, [8]International Christian University, Japan

Ericksondonna2000@gmail.com, julian@u-aizu.ac.jp, karen.perta@elmhurst.edu, zyzhang@ucla.edu,jeffmoore.personal@gmail.com, zen.voice.factory@gmail.com, y.goto@thevoice-expert.com, seunghunjlee@gmail.com

Abstract

Voice quality is an essential component of spoken language (e.g., Laver 1980, Campbell and Mokhtari 2003), conveying information about the speaker's emotion (e.g., Scherer 2003, Perta et al. in press), attitude (e.g., Erickson et al. 2024), culture (e.g., Sadanobu et al. 2016, Erickson et al. 2018, Erickson et al. 2020), and language (Yanushevskaya et al. 2018). Given that both vocal fold and vocal tract adjustments contribute to voice quality differences (e.g. Zhang 2021, Obert et al. 2023), this pilot study investigates laryngeal and supralaryngeal interactions of voice production resulting in acoustic voice quality changes use for different communication purposes. The approach used here is the following.

At the International Christian University in Tokyo, Japan, two professional singers/instructors trained in the Estill Voice Model made simultaneous electroglottographic (EGG) and acoustic recordings (Estill et al. 2017). Across two F0s an octave apart (approximately 130 Hz and 260 Hz) and different vowel conditions (i.e., /a/, /e/, /i/), subjects performed the following tasks: (1) thin vocal folds, no pharyngeal narrowing, (2) thick vocal folds, no pharyngeal narrowing, (3) thin vocal folds, with pharyngeal narrowing, (4) thick vocal folds with pharyngeal narrowing. Each vowel was repeated three times, making a total of 216 productions for the two subjects. Vowels produced with narrowing in the mid-region of the pharynx are referred to as twang (Titze et al. 2003, Perta et al. 2021, Obert et al. 2023). Based on blinded perceptual assessment of all productions by an Estill-certified master trainer (third author), not all productions were perceived as intended. For this pilot study, we report first on a subset of the data: 33 vowels identified as having no-narrowing. For comparison of no-narrowing vowels with pharyngeal narrowing (twang) vowels, we report on a set of 11 /e/ vowels produced by both subjects on low F0 with thin folds, 6 vowels with no narrowing and 5 vowels with twang narrowing. EGG analysis was done to measure Close Quotient (CQ) and EGG wave shape.

Vowels with no pharyngeal narrowing

Wave shape analysis was done using Generalized Additive Models (GAMs), a non-parametric regression technique to generalize linear models, allowing flexible shapes for covariates rather than straight lines. For the GAMs analysis, five EGG periods of each vowel were analyzed. Normalized time was obtained by linear interpolating the EGG time of each cycle onto 101 equally spaced points between 0 and 1. Likewise, wave shape was first DC-filtered and then normalized to be ≤ 1. Closure was modeled by the fixed effects of frequency (LF0 vs. HF0) and thickness of vocal folds (thick vs. thin). In addition, frequency- and thickness-specific closure variations across normalized time were modeled via penalized regression splines. Speaker, manner, vowel, and repetition were included as random effects. The model fit was assessed using gam.check(), which confirmed an adequate residual distribution. The model explained 94.7% of the variance.

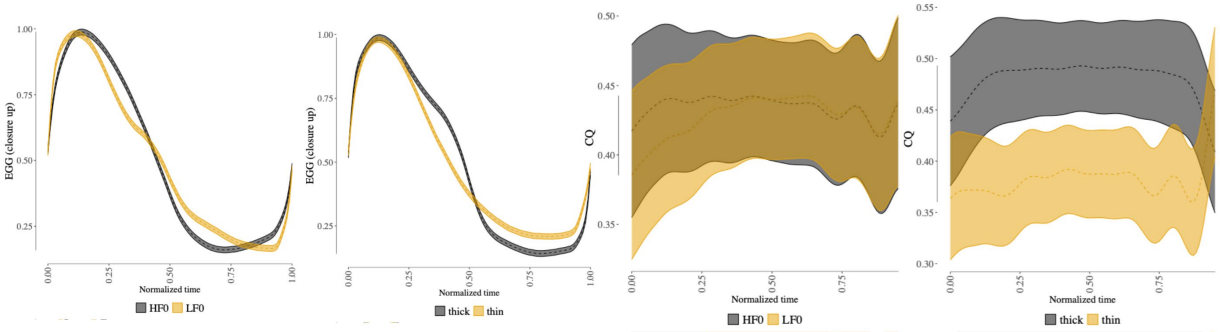


Figure 1: Predicted EGG signal for frequency (far left) and for vocal fold thickness (middle left). CQ as a function of F0 (middle right) and as a function of thickness of vocal folds (far right). X-axis indicates normalized time of vocal folds closing and opening.

The results suggest that overall, lower F0 produced slightly higher EGG amplitudes (about 2%, p = 0.046); thin manner, lower EGG amplitudes (about -2%, p=0.04). These differences change across the course of the EGG cycle. In terms of F0, (Figure 1 left side) , in

the opening phase of the EGG cycle, high F0 was about 20% more closed than the low F0 at the largest difference. In the second half of the EGG period, LF0 was about 20% more closed than HF0. In terms of thickness of vocal folds (Figure 1 middle left), thick was more than 25% more closed than thin during the opening phase, before the 50% of the EGG cycle. After that, thin was about 10% more closed.

With regard to CQ, a similar Generalized Additive Model was fit to the CQ of these recordings. The model explained 42.6% of the variance, reflecting the much noisier nature of the signals. In general, low F0 produced lower CQ than HF0 (~ -3%, p = .014). Thin manner also decreased CQ (~ -39%, p<.001), i.e., thin manner more effectively reduced CQ than LF0. The overall F0 differences did not translate to trajectory differences, as shown in the figure 1 (middle right). However, thick manner yielded higher CQs except at the extremes of the utterances, (Figure 1, far right).

Comparison of pharyngeal narrowing (twang) vowels with no narrowing vowels

In terms of CQ, there are no significant differences in the CQ trajectories across the utterance, as shown in Fig. 2, right side. In terms of EGG shape, relative to no narrowing, twang is more closed during the first 50% of the EGG cycle, and more open during the rest of the period; in both cases about 10% of the amplitude, as shown in Fig. 2 left side.

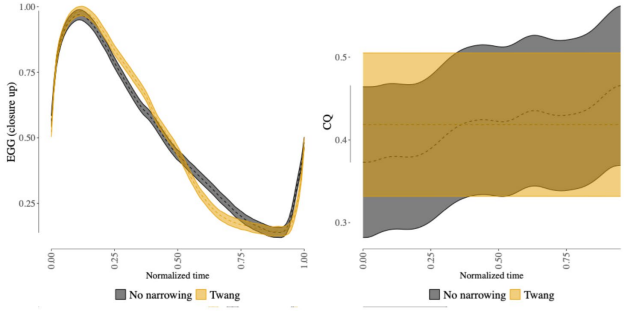


Figure 2: Predicted EGG signal for narrowing (left) and CQ as a function of thickness of vocal folds (right). X-axis indicates normalized time of vocal folds closing and opening.

Tentative results suggest that frequency, thickness of vocal folds, and pharyngeal narrowing affect the shape of the EGG signal. Analysis of EGG wave shapes is a novel contribution to the study of voice quality. Our observations support findings from vocal fold models of a complex interaction between vocal fold contact and supralaryngeal narrowing conditions that influence vocal fold dynamics. (Zhang, 2021). Future work will explore how speakers employ these mechanisms to change voice quality to impact communication styles, both within and across cultural/linguistic communities.

Index Terms: EGG, GAM, CQ, thick folds, thin folds, twang, F0, communication

Acknowledgements: Ian Wilson, Shigeto Kamano and Michinori Suzuki for help recording EGG and acoustics, and did first data organization,

References

Campbell, N. and Mokhtari, P. (2003) Voice quality: The 4th prosodic parameter, *Proc. 15th Int. Congr. Phonetic Sciences*, pp. 2417–2420.

Estill, J., Steinhauer, K., McDonald, M. (2017) *The Estill Voice Model: Theory and Translation*. Estill Voice International, LLC: Pittsburgh PA.

Erickson, D., Rilliard, A., Thurgood, E., De Moraes, J.A., Shochi, T. (2024) Acoustic and perceptual profiles of American English Social Affective Expressions: A case study. *Journal of Speech Sciences*, DOI: 10.20396/joss.v13i00.20015

Erickson, D., Kawahara, S., Rilliard, A., Hayashi, R., Sadanobu, T., Li, Yongwei., Daikuhara, H., de Moraes, J., Obert, K. (2020). Cross cultural differences in arousal and valence perceptions of voice quality, *Speech Prosody 2020.720-724*

Erickson, D., Sadanobu, T., Zhu, C., Obert, K., Daikuhara, H. (2018) Exploratory study in ethnophonetics: Comparison of cross-cultural perceptions of Japanese cake seller voices among Japanese, Chinese and American English listeners. *Speech Prosody 2018.*

Laver, J. (1980) *The Phonetic Description of Voice Quality*, Cambridge University Press. ISBN 0-521-23176-0, 1980.

Obert, K., Yun, J., Erickson, D., Reeve, M., Rowson, H., Møller, K. (2023) Voice quality: Interactions among F0, vowel quality, phonation mode and pharyngeal narrowing, *Studies in Laboratory Phonology. Language Science Press (langsci-press.org),* pp.190-199. *DOI: 10.2478/9788366675728-001.*

Perta, K., Bae, Y., & Obert, K. (2021). A pilot investigation of twang quality using magnetic resonance imaging. *Logopedics Phoniatrics Vocology*, 46(2), 77-85.

Perta, K., Zhang, Z., Erickson, D., Hayashi, R. Sadanobu, T. Submitted. MRI and acoustic study of angry screaming versus trained belting with vocal health implications. *JASA* Express Letters.

Boersma, P., & Weenink, D. (2026). Praat: doing phonetics by computer [Computer program]. Version 6.4.65, retrieved 24 April 2026 from https://www.praat.org/

Sadanobu, T., Zhu, C., Erickson, D., Obert, K. (2016) Japanese "street seller's voice". *Proc. Mtgs. Acoust*. **29**, 060003, doi: 10.1121/2.0000404.

Scherer, K. R. (2003)Vocal communication of emotion: A review ofvresearch paradigms, *Speech Commun*. **40(1)**, 227–256.

Titze, I.R., Bergan, C.C., Hunter, E.J., Story, B. (2003). Source and filter adjustments affecting the perception of the vocal qualities twang and yawn. *Logopedics Phoniatrics Vocology*, 28(4), 147-155.

Yanushevskaya, I., Gobl, C., and Ni Chasaide, A. (2018) Cross-language differences in how voice quality and *f*0 contours map to affect, *The Journal of the Acoustical Society of America*, vol. 144, pp. 27302750, https://doi.org/10.1121/1.5066448, 2018.

Zhang, Z. (**2021**). "Interaction between epilaryngeal and laryngeal adjustments in regulating vocal fold contact pressure," *JASA Express Lett.* **1**, 025201.

Realizations of sentence intonations in L1 and L2 Finland Swedish(es)

Heini Kallio[1], *Henna Heinonen*[1,2]

[1]Languages Unit, Tampere University, Finland
[2]2Department of Language and Communication Studies, University of Jyväskylä, Finland
heini.h.kallio@tuni.fi, henna.m.heinonen@jyu.fi

1. Introduction

Finland Swedish (hereafter referred to as FS), an official minority language of Finland and a non-dominant variety of Swedish in Scandinavia, is under-researched in terms of acoustic-phonetic investigations done with second language (L2) learners. The scarce previous research on L2 FS prosody has exclusively focused on one learner group: speakers with Finnish as their first language (L1) [1, 2, 3]. Today there is, however, an increasing number of FS learners with other language backgrounds, and the challenges these learners face in Swedish speech production may differ from those of Finnish-speaking FS learners. The objective of the current study is to compare intonation productions of L1 FS speakers with the ones of L2 learners with several L1 backgrounds: English, German, French, Russian, Vietnamese, and Sinhala. The study is part of the project "Pronunciation of Finland Swedish in beginner learners with different first languages" funded by Svenska Kulturfonden [4].

In recent studies we have found that the production of sentence stress in terms of syllable duration differ between FS learners with respect to their L1 [5] and that speaker's L1 also predicts their f_0 slope values (rate of change in pitch over time) in read sentences [6]. As the global f_0 parameters provide very limited information on realizations of sentence intonation, we now investigate the speakers' f_0 contours in more detail, quantifying them with Discrete Cosine Transform (DCT) method.

The more the L1 of the L2 learner differs from the target language, the higher the possibility of L1 interference on learning L2 prosody is [7]. Therefore, we expect that f_0 contours of speakers with L1 prosody notably different from FS will deviate more from the f_0 contours of the natives than the contours of L2 speakers with L1 that is prosodically more similar to FS. We also expect that the level of complexity in sentence structures affects the variation in the f_0 contours of the L2 speakers.

2. Material and methods

2.1. Speech data

Speech data used here is part of a larger corpus [8] and consists of read speech samples from international students attending beginner-level Swedish courses in Finnish universities at the time of recording. The data was collected remotely using an online environment. The participants were instructed to use an external microphone, e.g., a headset, but the recording conditions were not controlled for, and the quality of the recordings varies.

For the current study, we chose a subset of recordings based on the following criteria: (1) at least six speakers with the same L1 were available, (2) content and acoustic quality enables the comparison of f_0 contours between speakers. These criteria leave us six speakers from English and French L1 groups and eight speakers from German, Russian, Vietnamese, and Sinhala L1 groups. Utterances of the following individual sentences, three declaratives and one interrogative, were selected for analysis:

- 1e) *Jag tycker om att sjunga.*
- 1f) *Han kommer från Kina.*
- 1g) *Vill du äta här?*
- 1h) *På kvällen är jag på gymmet.*

The first two sentences (1e and 1f) follow a simple SVO structure, while the fourth sentence 1h has inverted word order and therefore has a more complex structure. Sentence 1g is a typical yes/no question with no wh-word.

Reference data from six L1 FS speakers was collected for the purposes of the current project. The FS speakers were instructed to read the sentences using Standard Finland Swedish. In total, we have 176 utterance-sized L2 samples and 24 corresponding L1 samples.

2.2. Analysis

Manually corrected f_0 contours were extracted using Praat [9]. The corrected f_0s were subsequently interpolated and converted into semitones. Intonation was analyzed using Discrete Cosine Transform (DCT) coefficients that compactly represent complex contours as a weighted sum of cosine functions. The method enables the comparison of realizations of dynamic pitch movements between speakers. Each utterance-level f_0 contour was represented using four low-order DCT coefficients (DCT2–DCT5, respectively, as DCT1 depends primarily on the mean f_0 of the speaker), describing aspects of contour shape such as global slope and curvature at different levels.

Each DCT coefficient was used as a dependent variable in linear mixed effects (LME) models to study the effect of speaker L1 and sentence type on f_0 contours. The native speaker group was set as the reference speaker group. In addition, cosine distances of DCT coefficients were computed between all productions to assess how similar or dissimilar f_0 contours of a sentence are within and between speaker groups. The statistical analyses and visualizations were done using R [10].

3. Preliminary results and discussion

For brevity, we will report and discuss the full results of the LME models in our presentation and article. However, preliminary analysis suggests that there is a significant effect of sentence 1g (the interrogative "Vill du äta här?") for DTC2 and DTC4, indicating a contour shape different from the other sentences. A significant effect of sentence 1h ("På kvällen är jag på gymmet") was found for DTC3. The effects of speaker L1, in

turn, remained small and only some effects reached statistical significance. This may be due to within-group variation in L2 speakers and limited size of our data.

Figure 1 shows the relative similarity of f_0 contours between speaker groups for each sentence type. Out of the four sentence types, the contours for the declaratives 1e and 1f seem to be the most similar within and between speaker groups. This was expected, as these sentences are relatively simple in structure. For the sentence 1h with inverted word order, Figure 1 shows larger distances between the speaker groups. Interestingly, the German speaker group seems to differ more from the FS group than other L2 speakers, although German could be considered to be prosodically more similar to Swedish than the other L1s in this study. However, the cosine distance is also large within the German group, and the realizations of f_0 contours in this sentence should thus be scrutinized in more detail.

The contours of the interrogative sentence 1g ("Vill du äta här?"), in turn, seem to vary the most between the speaker groups. Figure 2 shows the mean reconstructed f_0 contours of the sentence per speaker group. The reconstructed contours suggest that native FS speakers produce a clear rise-fall f_0 contour, with a rise before the middle of the sentence (= the word "äta" carrying sentence stress) and a steep fall during the stressed syllable "ä". The contours for the French, Sinhala, and German groups seem to be the most dissimilar to the one of the FS speakers in terms of f_0 dynamics: the French speakers tend to produce a fall-rise pattern instead of rise-fall, Sinhala speakers a steadily descending, monotonous contour, and German speakers a less articulated late rise (possibly on the latter syllable in the stressed word "äta") with a high final rise. It is possible that some L2 learners tend to use a rising intonation in yes-no interrogatives when they are not aware of the intonation patterns of the target language. Interestingly, the average f_0 contour of the Vietnamese group is the most similar to the one of the FS group, with the f_0 peak aligned at the same location and the steep fall towards the end of the sentence. Vietnamese is a tone language and on one hand it could be considered prosodically quite different from Finland Swedish, but on the other hand the rich tonality of the language could also help the speakers to better perceive and produce diverse f_0 contours. The English and Russian f_0 contours also proximate the one of the FS group in terms of the first rise-fall pattern, although they still differ from it in timing and extent. However, since the cosine distances are large in this sentence, even within the speaker groups, and because our data is very small, we will investigate the individual reconstructed contours for a more comprehensible analysis. Variation is also evident in other sentences between but also within L1 groups. We will thus investigate the reconstructed contours of all sentences and discuss possible reasons for within and between group variation as well as implications of the study to language teaching.

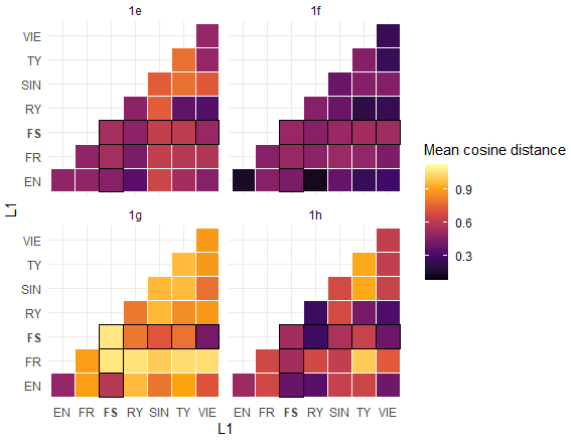


Figure 1: *Heatmap of mean cosine distances between speaker groups by L1. The darker the color, the more similar the f_0 contours between the speaker groups are.*

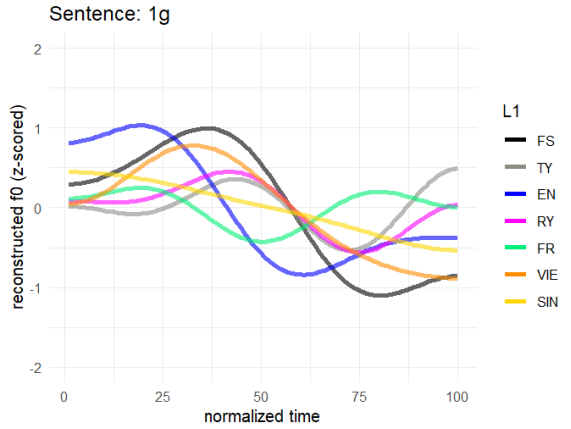


Figure 2: *Reconstructed mean f_0 contours of sentence 1g "Vill du äta här?". Speaker L1: FS = Finland Swedish, TY = German, EN = English, RY = Russian, FR = French, VIE = Vietnamese, and SIN = Sinhala.*

4. References

[1] M. Kautonen, "Finskspråkiga talares intonation av finlandssvenska i påståendeyttranden i fritt tal," *Folkmålsstudier*, vol. 55, 2017.

[2] H. Heinonen, "Durationsförhållandena i finskspråkiga gymnasisters uttal av L2-svenska: hur relaterar de till begripligheten?" in *Svenskans beskrivning*, no. 36. Uppsala universitet, 2019.

[3] H. Kallio, M. Kautonen, and M. Kuronen, "Prosody and fluency of Finland Swedish as a second language: Investigating global parameters for automated speaking assessment," *Speech Communication*, vol. 148, pp. 66–80, 2023.

[4] H. Kallio and H. Heinonen, "Hanke-esittely: Suomenruotsin ääntäminen ja ymmärrettävyys - puhujina eri kielitaustoista tulevat ruotsinoppijat," *Kieli, koulutus ja yhteiskunta*, vol. 16(3), 2025.

[5] H. Heinonen and H. Kallio, "Begriplighetsrelaterade uttalsdrag i finlandssvenska hos L2-talare med olika språkbakgrund," in *Svenskan i Finland 21*, 2026.

[6] H. Kallio and H. Heinonen, "Prosodic features in L2 Finland Swedish reveal differences in L1 effect," in *International Symposium of Applied Phonetics*, accepted.

[7] L. Rasier and P. Hiligsmann, "Prosodic transfer from L1 to L2. Theoretical and methodological issues," *Nouveaux cahiers de linguistique française*, vol. 28, no. 2007, pp. 41–66, 2007.

[8] H. Kallio, L. Tarvainen, A. von Zansen, Y. Getman, R. Al-Ghezi, E. Voskoboinik, A. Huhta, M. Kuronen, M. Kurimo, and R. Hildén, "Digitala: L2 swedish data from adult language learners, spring 2023. [speech corpus]." *The Language Bank of Finland*, 2023. [Online]. Available: http://urn.fi/urn:nbn:fi:lb-2023082921

[9] P. Boersma, "Praat: doing phonetics by computer," *http://www.praat.org/*, 2006.

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

Cross-Linguistic Influence at the Phonetic-Prosodic Interface: Asymmetric Vowel Reduction and Bidirectional Transfer in English-Icelandic Bilinguals

Christiane Ulbrich[1], Nicole Dehé[2]

[1]University of Cologne
[2]University of Constance
culbrich@uni-koeln.de, nicole.dehe@uni-konstanz.de

Abstract

1. Introduction

Lexical stress and phrasal accentuation constitute core organizational principles of spoken language, mapping phonological hierarchy onto acoustic salience. Cross-linguistically, the acoustic realization of stress typically involves a complex interplay of duration, intensity, and fundamental frequency (f0) [1], with f0 widely acknowledged as a primary perceptual cue in both English [2], [3] and Icelandic [4], [5]. However, the extent to which these phonetic-prosodic constraints interact within the bilingual mental lexicon remains an open empirical question. Specifically, this study investigates how two structurally divergent lexical accent and reduction systems influence each other in bilingual speech production in a heritage language context, and how these cross-linguistic patterns are systematically modulated under a structural prosodic mismatch condition. In English, lexical stress is variable, morpho-phonologically conditioned, and acts as a driver for rhythmic alternation [6]. Crucially, it is intrinsically coupled with vowel quality: English exhibits categorical qualitative vowel reduction, causing unstressed vowels to centralize dramatically toward schwa [ə] [7]. Conversely, Icelandic possesses a fixed, predictable stress pattern with primary stress strictly bound to the initial syllable [5], [8]. While unstressed syllables in Icelandic undergo quantitative compression in terms of duration shortening, the language inherently lacks systemic centralized reduction, meaning that – unlike in English – unstressed vowels do not reduce to schwa [ə] but maintain their distinct phonological categories, even if undergoing specific non-centralizing qualitative shifts [8].

Investigating these integrated prosodic structures in experimental production data can, however, be methodologically constrained by task-induced pragmatics. In the present study, a widespread production of target words with a rising terminal due to list intonation or uncertainty by our bilingual participants precluded a reliable pitch contour analysis. To circumvent this limitation and systematically assess the phonetic-prosodic interface, we shifted our empirical focus away from f0 contours to the strict domain of vocalic reduction in unstressed syllables – the phonetic interface where the two languages contrast sharply. This profound structural divergence provides an ideal testing ground for examining bidirectional cross-linguistic influence (CLI) in heritage, and second language (L2) acquisition contexts.

2. Methodology

Production data were elicited via a controlled picture-naming task across various speaker groups representing different configurations of bilingualism and language dominance: North American English monolinguals (NAmEng), Modern Icelandic monolingual controls (ModIce), L2 learners, and three generations of bilinguals (NAmEng, North American Icelandic (NAmIce), including speakers of moribund North American Icelandic (NAmIce-mor) and sequential generations (NAmIce-Gen1, NAmIce-Gen2). Target items consisted of English and Icelandic cognates (e.g., banana / banani) presented in two prosodic conditions: a Match condition (retaining baseline initial stress, e.g., hamburger / hamborgari) and a Mismatch condition (triggering a prosodic conflict, e.g., professor / prófessor). Acoustic analyses targeted the first (F1) and second (F2) formants of the unstressed syllables. To isolate the target vowels and effectively avoid potential confounds from secondary word stress, measurements for the Match condition were strictly taken from the second syllable. To prevent structural distortion of the acoustic vowel space caused by the highly asymmetric lexical distribution of underlying target vowels in the stimulus material (predominantly open vowels like /a/ alongside a lower token frequency of closed corner vowels like /u/ or /i/), a category-based normalization was implemented. Speaker-specific parameters for a Lobanov transformation [9] were calculated using unweighted category means extracted exclusively from primary stressed baseline tokens. Unshifted target tokens were then z-transformed relative to this stable, balanced baseline. To collapse multidimensional quality shifts into a single metric, the Euclidean Distance (ED) from the individual speaker's system center was computed. Linear mixed-effects models (LMMs) were fitted in R using the lme4 package [10], with ED entered as the dependent variable. In line with modern psycholinguistic and phonetic standards [11], the models included crossed random effects for Speaker and Item, allowing us to simultaneously account for variation across individual participants and lexical stimuli. Group, Condition, and their interaction were specified as fixed effects, while Vowel Type was included as a fixed covariate to filter out the inherent topography of the vowel space.

3. Results

In the Match condition (initial stress), a clear linguistic separation emerged regarding the acoustic reduction of the second, unstressed syllable. Crucially, no significant differences were found among the Modern Icelandic (ModIce)

controls and the various bilingual groups (NAmIce-mor, NAmIce- Gen1, NAmIce-Gen2, L2); under these baseline conditions, their vowel realizations clustered within a stable, intermediate-to-high ED range, maintaining peripheral vowel qualities. In sharp contrast, the North American English (NAmEng) monolingual controls exhibited a distinct baseline pattern, consistently producing schwa-like vowel qualities with ED values ranging below 0.8.

In the Mismatch condition, however, divergent patterns suggesting structural attrition and phonetic transfer emerged across languages:

- English targets: Vowels in the initial unstressed syllable produced by NAmEng monolinguals exhibited a collapse in distance toward the system center, indicating significantly stronger vowel reduction compared to all bilingual groups (NAmIce-mor, NAmIce- Gen1, NAmIce-Gen2, L2) and ModIce controls ($p < .05$). Crucially, the bilinguals' vowel realizations remained consistently and significantly more peripheral, resisting native-like reduction.
- Icelandic targets: In the baseline unstressed positions (second syllable), a systematic stratification emerged: Mor B and L2 speakers patterned together, realizing unstressed syllables with significantly less peripheral vowel qualities ($p < .05$) than ModIce, NAmIce- Gen1, -Gen2 speakers. A parallel pattern was observed when comparing the ED in the initial syllable of the Icelandic mismatch condition: while the prosodic conflict triggered an overall phonetic enhancement of peripherality across all bilingual groups ($p < .05$), the L2 and NAmIce-mor groups again realized the initial syllable with significantly less enhanced peripheral vowel qualities than NAmIce-Gen1, -Gen2, and monolingual controls.

4. Discussion and Conclusion

Taken together, these findings provide compelling evidence for complex, bidirectional cross-linguistic influence at the phonetic-prosodic interface. On the one hand, the phonetic subsystem of the L1 (Icelandic) acts as an acoustic shield in L2 English production: bilingual speakers consistently resist native-like vowel reduction to schwa in unstressed positions, retaining a significantly more peripheral, Icelandic-like vowel space. On the other hand, the systemic pressure of English-type reduction permeates upon the heritage and L2 Icelandic systems, as evidenced by the significantly diminished peripherality in the NAmIce-mor and L2 groups across both syllabic positions. Crucially, the amplification of these stratification patterns under prosodic mismatch conditions underscores that cross-linguistic influence is a fundamentally bidirectional, mutual process that cannot be analyzed as a purely segmental or isolated phenomenon. Instead, phonetic degradation and phonetic resilience are dynamically modulated by structural prosodic constraints, reflecting an integrated, multi-dimensional phonetic-prosodic interface within the bilingual mental lexicon. The data clearly demonstrate that cross-linguistic transfer does not operate as a unidirectional intrusion from a dominant into a heritage language, but rather as a continuous, systemic interplay where both linguistic subsystems concurrently shape phonetic outcomes.

5. Acknowledgements

This research was funded by the Deutsche Forschungsgemeinschaft (DFG) – Project numbers UL 458/2-1 and DE 876/4-1. We sincerely thank our local collaborators in Manitoba for their invaluable support with participant recruitment: Elva Simundsson and Julianna Roberts (Gimli), Star Johnson (Arborg), Keith Eliasson (Riverton), and Katrín Nielsdóttir (Winnipeg). We also extend our gratitude to all the bilingual speakers who generously shared their time and language for this study.

6. References

[1] Gordon, Matthew & Roettger, Timo. (2017). Acoustic correlates of word stress: A cross-linguistic survey. Linguistics Vanguard. 3. 10.1515/lingvan-2017-0007.

[2] Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1(2), 126-152.

[3] Lieberman, P. (1960). Some effects of fundamental frequency and morphologic accent. *JASA*, 32(4), 451-454.

[4] Árnason, K. (1980). *Quantity in Historical Phonology: Icelandic and Related Dialects*. Cambridge University Press.

[5] Thráinsson, H. (1994). Icelandic. In E. König & J. van der Auwera (Eds.), *The Germanic Languages* (142–189). Routledge.

[6] Chomsky, N., & Halle, M. (1968). *The Sound Pattern of English*. Harper & Row.

[7] Flemming, Edward. 2009. The phonetics of schwa vowels. Phonological weakness in English: From Old to Present-day English, ed. by Minkova, Donka, 78–95.

[8] Árnason, K. (2011). *The Phonology of Icelandic*. Oxford University Press.

[9] Lobanov, B.M. (1971) Classification of Russian Vowels Spoken by Different Speakers. JASA, 49, 606-608.

[10] Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.

[11] Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412.

Swedish prosody transferred into L2 Spanish: Pragmatic effects

Berit Aronsson, Lars Fant

Umeå University and Stockholm University

The present study investigates possible pragmatic effects of non-native intonation, spoken by Swedish learners of Spanish, on the personal traits that native speakers attribute to the L2 speakers. Native Spanish speakers' judgments are compared with those of Swedish L1 speakers. The evaluated task consists of a reservation at a restaurant.

The data of consist of four recorded telephone spontaneous dialogues in a simulated situation where four Swedish participants speaking L2 Spanish were instructed to make a table reservation at a restaurant, interacting with a 'head waitress' played by a native speaker of Spanish. The duration of each dialogue varies between 32-61 turn-constructional units (TCUs). The 'callers' were language students at a Swedish University with a command of Spanish that corresponded to the B1 level of the Common European Framework of Reference (CEFR) scale, whereas the part of the 'waitress' answering the four phone calls was performed by one and the same native speaker, whose identity was unknown to the subjects. The evaluations of the four dialogues were made by 32 Spanish L1 speakers (22 female and 10 male) and by 20 Swedish L1 speakers (16 female and 4 male). Results show that the judgments made by the Spanish L1 evaluators when describing the speakers are more negative when non-native speech is evaluated.

A combination of acoustic analysis in terms of prosodic units and pragmatic analysis in terms of speech acts revealed contradictory signals between intended speech acts and acoustic cues in the L2 but not the L1 data. Triangulating these data using linear mixed effects models (LMM) further shows that a higher proportion of tonal cues misaligned with the corresponding speech acts is associated with an increase in negative traits, such as anger, unpleasantness, and arrogance, but only as perceived by the Spanish evaluators. This association is not observed among Swedish evaluators, who evaluate their L2-Spanish-speaking countrymen more favourably. Different language specific cues associated with rising endtones: information seeking at the transactional level in Spanish ("I request information") versus intersubjectivity seeking at the interpersonal level in Swedish ("I'm trying to be nice, do you like me"), are suggested as additional explanations to the results.

Uncovering Interlanguage systematicity: Micro-analysis of L2 Japanese interrogative intonation by L1 Swedish speakers

Natsumi Goto[1], Mitsuhiko Ota[2], Shinichiro Ishihara[1]

Lund University[1], University of Edinburgh[2]

Intro: This study investigates the inter-speaker variation in the production of interrogative intonation in Japanese *Wh*-questions and Yes/No questions by L1 Japanese controls and L1 Swedish/L2 Japanese speakers. It explores whether individual intonational patterns stem from L1 phonetic transfer, distinct learner-specific strategies, or an interaction of both. While traditional L2 speech research relies heavily on macro-level group comparisons with native baselines, a comparative approach that often obscures systematic patterns of individual speakers (Colantoni et al., 2015). In the broader field of L2 acquisition, it primarily confines individual variation in the domain of "Individual Differences (IDs)", analysing how individuals' external, cognitive, or socio-cultural factors impact learning outcomes (Li et al., 2022). This study introduces a paradigm shift by utilising unsupervised, bottom-up cluster analysis not to correlate learners' production with external factors, but instead as a tool of unveiling systematic intonational patterns both in native controls and L2 speakers, contributing to the studies documenting variation in both L1/L2 prosody (Niebuhr et al., 2011). Our results reveal four distinct prosodic patterns based on sentence-final intonation. Crucially, the key finding emerged not from the specific phonetic contours themselves, but from the highly systematic distribution of these strategies across individual speakers.

Background: In Swedish, question intonation typically features a higher pitch register and a higher final focal accent compared to statements: a sentence-final rise is optional, conditioned by syntax and dialect (Gårding,1979; Hadding-Koch and Studdert-Kennedy, 1964; House, 2002, 2004, 2005). In Central Swedish, questions rely on a global F0 rise and a delayed utterance-final focal peak (House, 2003). In South Swedish, a wider F0 range rise on first prosodic word to the following prominent word is found in Yes/No questions compared to *Wh*-questions (Horne and Roll, 2021). Conversely, Japanese lacks syntactic inverted question word order, relying on prosodic cues. Japanese *Wh*-questions exhibit "focal prosody" (Ishihara, 2003, 2007), which entails focal F0-rise on the *Wh*-phrase and is followed by post-focal reduction. In Yes/No questions, this phrasal prominence is observed on the sentence-final verb (Ishihara, 2017). Crucially, both question types typically realise with a scooped sentence-final rise (LH%) (Maekawa et al. 2002; Venditti et al., 2008) as Japanese lacks the syntactic cue of inverted word order in question sentences. In sum, the cross-linguistic features are summarised as follows: (1) an optional, dialect-dependent terminal rise vs. a compulsory Japanese LH% rise; (2) an expanded initial F0 range in Swedish Yes/No questions vs. a downstepped Japanese contour leading to final verb prominence; (3) the absence of structural prominence on Swedish Wh-words vs. an obligatory, highly salient Japanese focus peak; and (4) gradual phrasal decay in Swedish vs. systematic, structural post-focal compression in Japanese. **Method**: Speech data were elicited from 26 L1 Swedish L2 Japanese speakers and 15 native Japanese speakers as controls. One Swedish participant (S20) was excluded due to a persistent creaky voice. L1 Swedish regional variations were assessed via their Swedish production data (Table 2).

Stimuli: Both languages included three sentence types: declaratives, Yes/No questions, and *Wh*-questions. For Swedish, 18 stimulus sentences balancing Accent 1 and Accent 2 across three lexical sets (3 sentence types ×2 accents ×3 sets) are prepared and repeated 5 times for the recoding. For Japanese, 9 stimulus sentences for three accented lexical sets (3 sentence types ×3 sets) are drafted, and each is repeated 5 times. While Swedish data was segmented in a standard manual word segmentation, Japanese data were phrase-level segmented up to the final verb. To analyse final-rise dynamics, the last verb is annotated with the rest of the verb and the last mora of the verb. With Prosody pro script (Xu, 2013), the semitone value at the time-normalised points was calculated and later converted to a z-scored semitone for plotting. The annotation rule and the corresponding time-normalised points for example stimuli are summarised in Table 1.

Table 1: Example of stimulus

Time-normalised point	0-10	10-20	20-30	30-40	40-50
Syntactic structure	Subject phrase	Object phrase	Location phrase	rest of the verb	last mora of the verb
Statement Yes/No question	*Onéesan wa*	*wáin o*	*ie de*	*nó*	*mu*
	Sister-TOP	wine-ACC	home-LOC	drink	
Wh-Question	*Onéesan wa*	*náin o*	*ie de*	*nó*	*mu*
	Sister-TOP	What-ACC	home-LOC	drink	

Table 2: L1 Swedish profile

Dialect variation	Speaker IDs
Two-peak (n=17)	S01-13, S16, S23–25
One-peak (n =2)	S15, S22
Uncategorized (n=6)	S14, S17, S18, S19, S21,S26

Result/discussion: For the focal prosody analysis, a new Swedish subset was created, excluding speakers S02, S04, and S12, due to sharp rises in particles and S20 due to a creaky voice. To compare distinct intonational strategies, the pitch difference (Question – Statement) was calculated across question types (Figure 1). In the left panel (*Wh*-questions), Japanese native speakers display a substantial positive difference driven by the focal rise on the *Wh*-phrase, whereas the Swedish group shows an attenuated focal rise. After the *Wh*-phrase, the L1 native group exhibits a sharp negative difference reflecting post-focal reduction; the Swedish learners mirror this negative trend around the same points, though to a lesser degree. A similar pattern emerges for Yes/No question prosody:

Swedish learners show smaller F0 prominence on the verb than Japanese natives. To break down another core part of the question, intonation, namely, sentence-final intonation, independent Principal Component Analyses (PCA) and Hierarchical Clustering (HCPC) were performed for each question type using FactoMineR package (Lê et al., 2008) in R programme, focusing on the final verb window (Points 30-50). Figure 2 plots the mean of z-scored semitone contours for each clustered type. Figure 3 shows the percentage of each cluster's prosody by individual speakers. In detail, *Wh*-Cluster 1 (Target-like feature integration) is observed predominantly among native controls and a subset of Swedish speakers, reflecting robust post-focal compression at the verb onset, followed by a scooped final rise (LH%). *Wh*-Cluster 2 (Syntax-Prosody Boundary Shifting) yielded a target-like final dip-rising, yet it preserves a pitch accent (H*+L) at the first part of the verb, indicating each phrase forms a separate prosodic phrase because the speakers produced each phrase separately with a short pause, which is confirmed by the acoustic analysis of individual production. *Wh*-cluster 3 (Early terminal fall from L1) is mainly produced by Swedish speakers (except speaker J07), having a higher plateau on the verb compared to cluster 1, followed by a shallower dip-rising, resembling an early fall of their L1 Accent 2 focal accent (H*LH) with the final rising. Cluster 4 (Interlanguage) is almost exclusively among Swedish speakers, showing the intermediate state; target-like scooped dip-rise, yet not producing *Wh*-derived post-focal compression on the verb. The difference between *Wh*-clusters and Yes/No clusters is subtle yet can be seen at the beginning of the verb, where *Wh*-clusters have a lower pitch range than the Yes/No clusters. **Conclusion**: Taken together, the combined macro-micro level analysis reveals that both native and Swedish speakers do produce similar cluster prosody, yet the distribution of intonation types shapes the group difference and the complex, intermediate status of L2 intonation; the mixture of influence from L1 in L2 intonation and a partial acquisition of target intonation.

Figure 1: Comparison of the difference of z-scored semitone

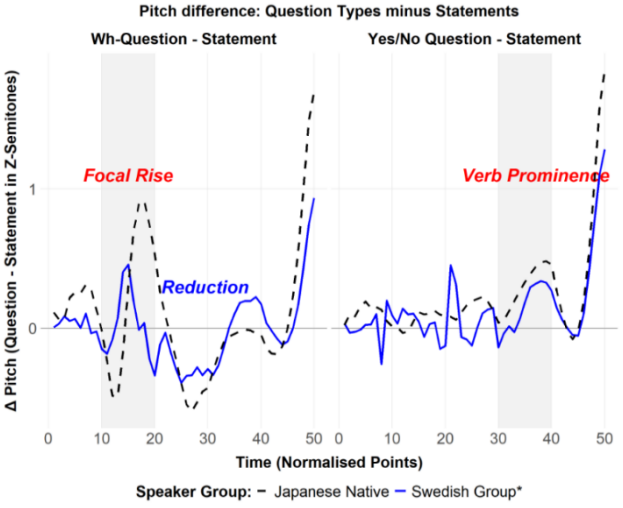


Figure 2: Mean z-scored semitone contour by the clusters

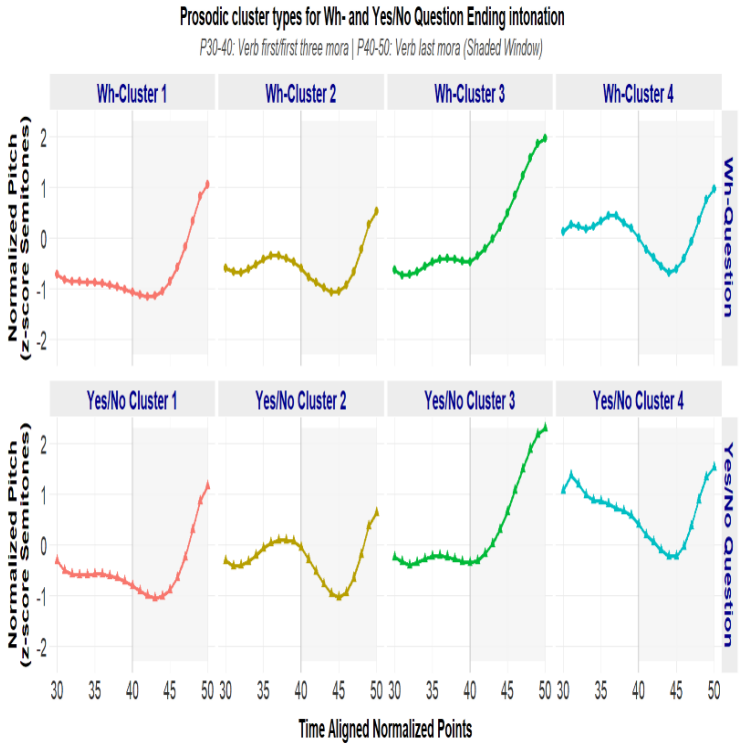
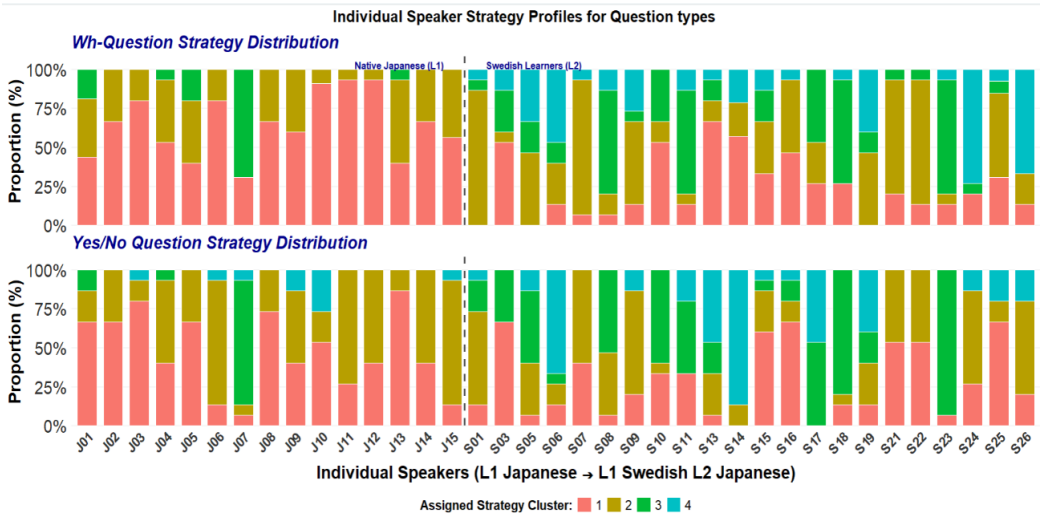


Figure 3: Proportion of each clustered production by individuals



Reference

Colantoni, L., Steele, J., & Escudero, P. (2015). *Second language speech: theory and practice*. Cambridge University Press.

Gårding, E. (1979). Sentence intonation in Swedish. *Phonetica*, *36*, 207–215.

Hadding-Koch, K and Studdert-Kennedy, M. (1964). An experimental study of some intonation contours. *Phonetica 11*, 175-185.

Horne, M and Roll, M. (2021). Question intonation in South Swedish. *Proceedings of Fonetik 2021*, 54–57.

House, D. (2002). Intonational and visual cues in the perception of interrogative mode in Swedish. *Proceedings of the 7th International Conference on Spoken Language Processing* 1957–1960.

House, D. (2003). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. *Proceedings of the 15th ICPhS*: 755-758.

House, D. (2004). Final rises and Swedish question intonation. *Proceedings of FONETIK.* 56-59.

House, D. (2005). Phrasal- final rises as a prosodic feature in Wh-questions in Swedish human-machine dialogue. *Speech communication* 46, 268-283.

Ishihara, S. (2003). *Intonation and interface conditions*. PhD thesis, Massachusetts Institute of Technology. http://hdl.handle.net/1721.1/17020.

Ishihara, S. (2007). Major phrase, focus intonation, and multiple spell-out (MaP, FI, MSO). *The Linguistic Review 24:* 137–167. doi:10.1515/TLR.2007.006.

Ishihara, S. (2017). The Intonation of Wh- and Yes/No-Questions in Tokyo Japanese. In C. Lee, F. Kiefer, & M. Krifka (Eds.), *Contrastiveness in Information Structure, Alternatives and Scalar Implicatures* (pp. 399–415). Springer International Publishing. https://doi.org/10.1007/978-3-319-10106-4_19

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R package for multivariate analysis. *Journal of Statistical Software*, 25(1), 1–18.

Li, S., Hiver, P., & Papi, M. (2022). The routledge handbook of second language acquisition and individual differences. Routledge.

Maekawa, K., Kikuchi, H., Igarashi, Y and Venditti, J. (2002). X-JToBI: An extended J_ToBI for spontaneous speech. *Proceedings of the 7th International Conference on Spoken Lan guage Processing* (ICSLP 2002), Denver, Colorado, 1545–1548.

Niebuhr, O., D'Imperio, M., Gili FIvela, B., & Cangemi, F. (2011). Are there "shapers" and "aligners"? Individual differences in signalling pitch accent category. *In Proceedings of the 17th ICPhS, Hong Kong, China: Special session on tones and shapes* (pp. 120-123)

Venditti, J., Maekawa, K and Beckman, M.E. (2008). Prominence marking in the Japanese intonation system. In: Miyagawa, S and Saito, M(eds.), *The Oxford handbook of Japanese linguistics,* 456–512. New York: Oxford University Press.

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), Aix-en-Provence, France. 7-10. https://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/

The alignment of beat gestures with accented syllables in Estonian

Pärtel Lippus, Andra Annuka-Loik, Eva Liina Asu, Liisa Kai Siimut, Len Toots
University of Tartu, Estonia

Gesturing has been shown to form an integral aspect of speech production, e.g. [1]. In gesture research, several types of gestures have been outlined, see [2], [3], among others, for example, iconic gestures (that are meant to mimic an object or action that corresponds with a referent within the speaker's utterance) or deictic gestures (which reference the physical or abstract space or objects within them, e.g. pointing at something).

In this paper, the gesture type of interest is the 'beat gesture' – a rhythmic movement "serving to mark out segments of the discourse or the rhythmic structure of the speech" [2, p. 100]. As the name implies, beat gestures can be physically executed in a more forceful or sudden manner, for instance, by a punch on the knee [2]. Beat gestures are without referential meaning, and primarily have a discursive function. They serve to add emphasis to something in the utterance, and tend to be associated with prosodic prominence through prominence-lending intonational pitch accents [3], [4], [5], [6].

This paper studies the alignment of beat gestures (produced by hands, shoulders, elbows or fingers) with prosodic prominences in Estonian spontaneous dialogue. The focus is on singular beat gestures as opposed to recurrent gestures or more complex gesture sequences.

Materials and Method

The data comes from the University of Tartu Phonetic Corpus of Spontaneous Estonian [7], and comprises a subset of 23 half-hour dialogues with a total duration of 13 hours. During the recording, the speakers were seated diagonally from each other in the opposite corners of the sound booth. The audio was recorded with DPA 4088 head mounted microphones and the video was captured with two GoPro Hero 5 cameras placed between the speakers (positioned to include the head, torso and hands).

With four files excluded due to technical reasons, the analysed dataset consists of 42 speakers (24 female and 18 male, age 20–74 years). The speakers were predominantly with an academic background who spoke the standard variety of Estonian. The partners in each dialogue knew each other beforehand. The participants were instructed to talk freely about any topic. They were aware of the purpose of the recording and gave a written informed consent for participating.

The corpus has been manually segmented and annotated on word and phoneme levels in Praat. The subset has also been annotated on the gesture level (head, right hand, left hand) in ELAN. Each gesture was segmented into phases and labelled with a gesture type (*beat*, *iconic*, *pragmatic*, etc.). After that, accented syllables within the utterance where the beat gestures occurred were annotated in Praat.

For the current analysis, the times of the strokes of the beat gestures were extracted from the ELAN files using the *readelan* package and the times of the accents from Praat TextGrids using the *rPraat* package in R.

Results

By the time of abstract submission, the preliminary analysis of 29 speakers includes a total of 875 beat gestures (554 full hand gestures, 158 only fingers, 108 shoulder shrugs and 28 elbow gestures): 453 of the gesture strokes co-occurred with one accent in speech, 201 of the gesture strokes occurred when there were no accented words in speech, 221 gesture strokes overlapped with more than one accent.

The preliminary analysis of alignment between the timing of the gesture's stroke phase with the accented word is presented in Figure 1. The results show that the onset of the stroke is best aligned with the onset of the word (Fig 1 top left panel) while in reference to the beginning of the syllable nucleus strokes start a bit earlier (Fig 1 bottom left panel). If there are no accented syllables overlapping with the gesture then the gesture slightly leads the nearest accent in speech rather than follows it. The strokes end quite closely before the end of the accented syllable (Fig 1 bottom right panel).

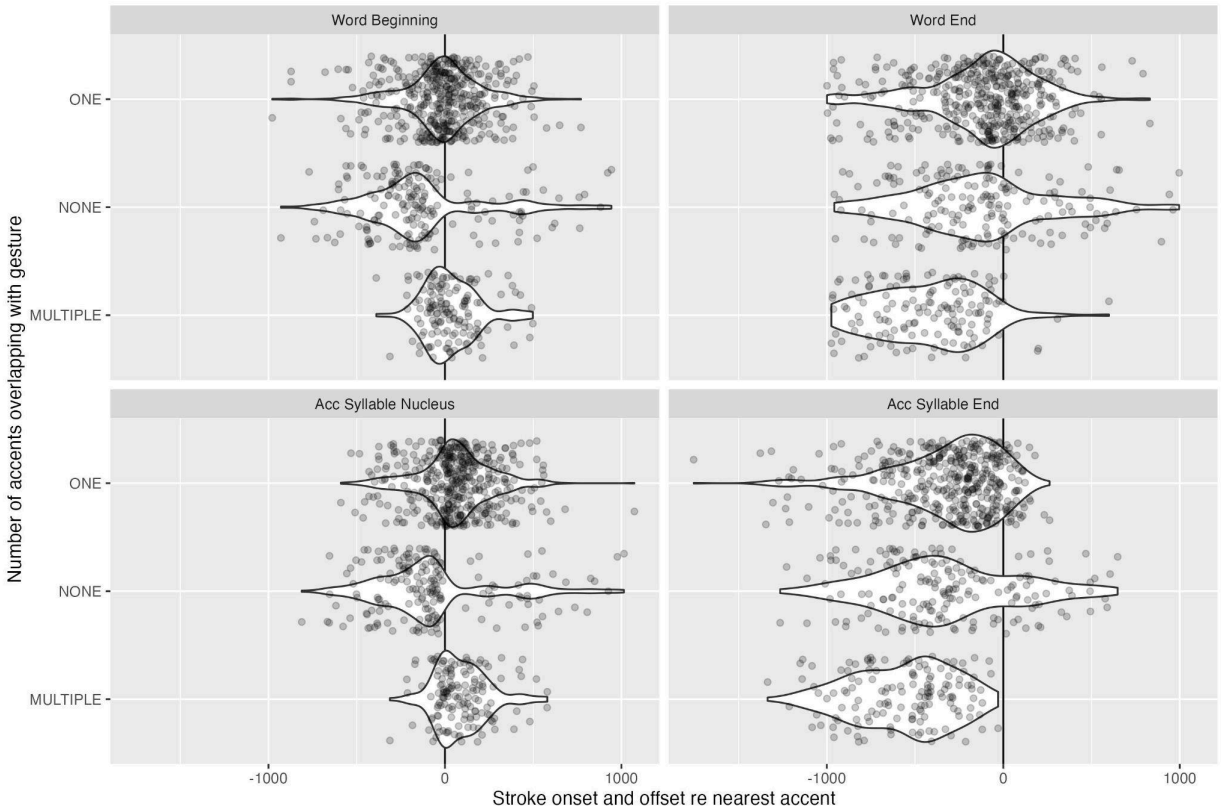


Figure 1. Alignment of stroke onset with the beginning of the accented word and syllable nucleus (left panels) and stroke offset with the end of the accented word and syllable nucleus (right panels).

Conclusion

The current results show that beat gestures in Estonian spontaneous dialogues are mostly aligned with the accented syllable boundaries or realised a bit before the lexical affiliate. The next steps of the analysis include investigating the alignment of F0 peaks and the peaks of the gestures according to the velocity curve. We will also take a closer look at the descriptive categorisation of the beat gestures.

References

[1] J. M. Iverson and S. Goldin-Meadow, "Why people gesture when they speak," *Nature*, vol. 396, no. 6708, pp. 228–228, Nov. 1998, doi: 10.1038/24300.

[2] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge: Cambridge University Press, 2004. doi: 10.1017/CBO9780511807572.

[3] D. Casasanto, "Gesture and language processing," in *Encyclopedia of the Mind*, Los Angeles, London, New Delhi, Singapore, Washington DC: Sage, 2013, pp. 372–374.

[4] G. Ambrazaitis and D. House, "Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings," *Speech Communication*, vol. 95, pp. 100–113, Dec. 2017, doi: 10.1016/j.specom.2017.08.008.

[5] C. Carignan, N. Esteve-Gibert, H. Lœvenbruck, M. Dohen, and M. D'Imperio, "Co-speech head nods are used to enhance prosodic prominence at different levels of narrow focus in French," *The Journal of the Acoustical Society of America*, vol. 156, no. 3, pp. 1720–1733, Sep. 2024, doi: 10.1121/10.0028585.

[6] D. P. Loehr, "Temporal, structural, and pragmatic synchrony between intonation and gesture," *Laboratory Phonology*, vol. 3, no. 1, Jan. 2012, doi: 10.1515/lp-2012-0006.

[7] 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.

Multimodal prosody and syntactic completion in overlapping speech in Central Swedish conversation

Margaret Zellers[1]

[1]Department of Linguistics, Stockholm University, Sweden

margaret.zellers@ling.su.se

1. Introduction

Speech in dialogue has been reported to be produced largely by one speaker at a time [1]. Despite this, speakers in conversations regularly overlap with one another; for example, [2] report that as many as 54.4% of talk spurts in two-party telephone conversations are produced in at least partial overlap with another speaker. Overlapping speech is not necessarily problematic for interactants, and must have a minimum duration of about 120ms to be detected [3]. Research in both qualitative and quantitative traditions, e.g. [4, 5, 6, 7], has shown that variability in prosody, i.e. speech melody, loudness, rhythm, or phonation quality, is important for signalling speakers' intentions as to whether overlapping speech is intended to be turn-competitive, that is, whether the incoming speaker wants to take the conversational floor. Prosodic information alongside potential syntactic completion is used more generally by speakers and listeners to signal and interpret intentions about floor-holding or floor-release in conversation [8, 9, 10, 11, 12, 13].

A growing body of research supports treating gesture as an integrated part of linguistic prosodic systems (see discussion in [14]). Visual and auditory information are integrated during speech perception [15], and gesture can be used to disambiguate meaning when the speech signal provides insufficient information [16]. Gesture appears to be implemented and interpreted early relative to spoken cues; for example, gestural behavior has been shown to differ systematically as early as 3 seconds before the offset of a speech turn depending on whether the speaker intends to continue speaking or to stop and let an interlocutor take up a turn [17], and gestures have been shown to facilitate semantic prediction and speed question response in conversation [18, 19]. Visual information could be especially useful in overlapping speech in dialogue, since gestures are produced in a different modality and thus do not contribute to problems of understanding that can otherwise arise in overlap.

2. Methods

This exploratory study investigates variability in multimodal turn-end marking in the context of overlapping speech in Central Swedish two-party conversations from the Spontal corpus [20]. The dataset used in this study comprises 10 5-minute excerpts from 9 dialogues. Hand annotation of syntactic completion of talk spurts and hand gestures was carried out as part of previous research [12, 17]. Talk spurts with duration of at least 500ms were included in the analysis, and overlapping speech was defined as an overlap of the end of one talk spurt with the beginning of a next speaker's talk spurt with a duration of at least 120ms. Hand gestures were divided into phases: *preparation, stroke, hold, retraction*, cf. [21]. Fundamental frequency (F0) and amplitude information were extracted for the whole talk spurt as well as for overlapping portions only using Praat [22]. In total, 1057 talk spurts were analyzed, 304 (28.8%) of which were produced with overlap at their offsets.

3. Preliminary results

A χ^2 test found no difference in the presence or absence of hand gesture at syntactically complete versus incomplete talk spurt ends, regardless of whether these were overlapped by the incoming speaker (χ^2 (1, 1057) = 0.90773, p = .34). However, the acoustic prosodic aspects of the overlapped speech varied depending on whether the overlapped speech was accompanied by gesture, but the differences only achieved statistical significance if the overlapped chunk was syntactically incomplete. Figure 1 shows that the duration of overlaps in the case of incomplete talk spurt ends is longer than if there is no gesture accompanying the incomplete talk spurt end; model results are given in Table 1. Figure 2 shows that intensity is reduced less in overlapped turn ends in syntactically incomplete talk spurts that are accompanied by gesture than in those which are not accompanied by gesture; model results are given in Table 2. No significant results were found for F0 in overlapped speech, which may be due to data loss from non-modal voice qualities near speech offset.

Table 1: *Linear mixed model results for duration of overlapping speech. Conditional R^2 = 0.062; marginal R^2 = 0.034.*

	Estimate	**SE**	**t**	**p**
(Intercept)	0.9412	0.1046	9.00	0.0000*
gestT	-0.2937	0.2180	-1.35	0.1791
compI	-0.1293	0.1695	-0.76	0.4463
gestT:compI	0.9927	0.3642	2.73	0.0069*

Table 2: *Linear mixed model results for difference of intensity between overlapped speech compared to its full turn. Conditional R^2 = 0.042; marginal R^2 = 0.023.*

	Estimate	**SE**	**t**	**p**
(Intercept)	-6.3787	0.6580	-9.69	0.0000*
gestT	0.7760	1.3712	0.57	0.5720
compI	1.8393	1.0661	1.73	0.0858
gestT:compI	-4.9379	2.2913	-2.16	0.0322*

4. Discussion

The preliminary results indicate that interlocutors adapt their prosodic and gestural behavior in the context of overlapping

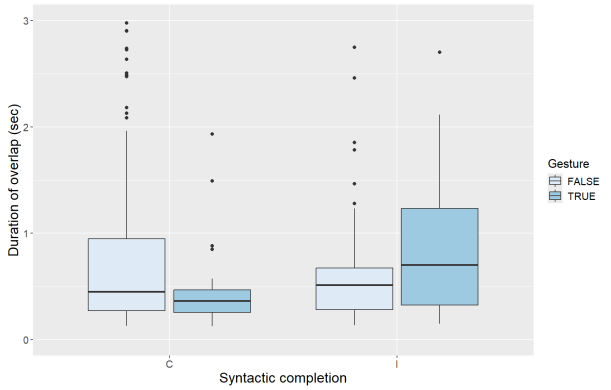


Figure 1: *Difference in duration of overlapped speech compared to full turn in complete and incomplete talk spurts, with and without hand gesture accompanying the overlap.*

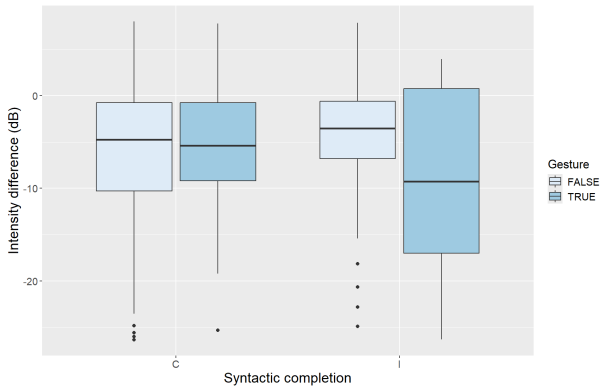


Figure 2: *Difference in intensity in overlapped speech compared to full turn in complete and incomplete talk spurts, with and without hand gesture accompanying the overlap.*

speech, and that these adaptations also show a sensitivity to the syntactic completeness of the overlapped talk spurt. Specifically in the context of overlapped syntactically incomplete talk spurt ends, gesturing is associated with a stronger reduction in amplitude compared to overlaps when the speaker is not gesturing. In these contexts, gesture from the overlappee is, rather surprisingly, also associated with a relatively longer overlap duration. At incomplete talk spurt ends with gesture, variability in the acoustic features is also greater than in other conditions, perhaps suggesting that gesture has more scope to add ambiguity to the syntactic incompletion compared with the lexical and acoustic information alone.

5. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organisation of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[2] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: findings and implications for automatic processing of multiparty conversation," in *Proceedings of 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 1359–1362.

[3] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *Journal of the Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[4] E. A. Schegloff, "Overlapping talk and the organization of turn-taking for conversation," *Language in Society*, vol. 29, no. 1, pp. 1–63, 2000.

[5] C. Oertel, M. Włodarczak, A. Tarasov, N. Campbell, and P. Wagner, "Context cues for classification of competitive and collaborative overlaps," in *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, China, 2012, pp. 22–25.

[6] K. P. Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in *Proceedings of Interspeech 2013*, 2013, pp. 1404–1408.

[7] E. Kurtić and J. Gorisch, "F0 accommodation and turn competition in overlapping talk," *Journal of Phonetics*, vol. 71, pp. 376–394, 2018.

[8] C. Ford and S. Thompson, "Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns," in *Interaction and grammar*, E. Ochs, E. Schegloff, and S. A. Thompson, Eds. Cambridge, UK: Cambridge University Press, 1996, pp. 134–184.

[9] J. Caspers, "Local speech melody as a limiting factor in the turn-taking system in Dutch," *Journal of Phonetics*, vol. 31, pp. 251–276, 2003.

[10] D. Barth-Weingarten, "Contrasting and turn transition: Prosodic projection with parallel-opposition constructions," *Journal of Pragmatics*, vol. 41, no. 11, pp. 2271–2294, 2009.

[11] S. Bögels and F. Torreira, "Listeners use intonational phrase boundaries to project turn ends in spoken interaction," *Journal of Phonetics*, pp. 46–57, 2015.

[12] M. Zellers, "Prosodic variation and segmental reduction and their roles in cuing turn transition in Swedish," *Language and Speech*, vol. 60, no. 3, pp. 454–478, 2017.

[13] M. Włodarczak, M. Heldner, A. Bruggeman, and P. Wagner, "Voice quality dynamics of turn-taking events in Swedish and German," in *International Congress of Phonetic Sciences (ICPhS)*, 2023, pp. 3477–3481.

[14] P. Wagner, Z. Malisz, and S. Kopp, "Gesture and speech in interaction: An overview," *Speech Communication*, vol. 57, pp. 209–232, 2014.

[15] S. D. Kelly, P. Creigh, and J. Bartolotti, "Integrating speech and iconic gestures in a Stroop-like task: evidence for automatic processing," *Journal of Cognitive Neuroscience*, vol. 22, no. 4, pp. 683–694, 2010.

[16] B. Guellaï, A. Langus, and M. Nespor, "Prosody in the hands of the speaker," *Frontiers in Psychology*, vol. 5, p. 700, 2014.

[17] M. Zellers, J. Gorisch, and D. House, "Temporal relationships between speech and hand gestures in the vicinity of potential turn boundaries in German and Swedish conversation," *Language and Cognition*, vol. 17, p. e57, 2025.

[18] K. H. Kendrick, J. Holler, and S. C. Levinson, "Turn-taking in human face-to-face interaction is multimodal: gaze direction and manual gestures aid the coordination of turn transitions," *Philosophical Transactions of the Royal Society B*, vol. 378, no. 1875, p. 20210473, 2023.

[19] M. Ter Bekke, L. Drijvers, and J. Holler, "Hand gestures have predictive potential during conversation: An investigation of the timing of gestures in relation to speech," *Cognitive Science*, vol. 48, no. 1, p. e13407, 2024.

[20] J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House, "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture," in *Proceedings of LREC 2010, Valetta, Malta*, 2010.

[21] A. Kendon, *Gesture: Visible action as utterance*. Cambridge, UK: Cambridge University Press, 2004.

[22] P. Boersma and D. Weenink, "Praat: doing phonetics by computer," http://www.praat.org/, 2025, version 6.4.42.

Embodied behavior during pausing in L2 Finnish at different fluency levels

Riikka Ullakonoja[1]*, Pauliina Peltonen*[2]*, Mikko Kuronen*[1]*, Asko Tolvanen*[3]*,*

[1]Department of Language and Communication Studies, University of Jyväskylä
[2]Department of English, University of Turku
[3]Faculty of Education and Psychology, University of Jyväskylä
riikka.ullakonoja@jyu.fi, paupelt@utu.fi, mikko.j.kuronen@jyu.fi, asko.j.tolvanen@jyu.fi

Abstract

There is abundant research on fluency as an indicator of L2 speaking proficiency, but previous studies have mainly focused on speech fluency without exploring embodied behavior. The current study addresses this gap by focusing on embodied behavior (i.e. hand, head and body movements) during two (dis)fluency indicators, silent and filled pauses. The sample consists of videoed dialogues form 38 speakers of L2 Finnish (assessed at A2, B1, and B2 CEFR levels for their fluency by 21 trained raters). The preliminary results show that embodied elements were used very frequently during pausing across all fluency levels (A2-B2). There were some differences between the fluency levels and speakers in the types of embodied elements used, but the differences remained statistically insignificant.

Index Terms: fluency, embodied behavior, language assessment, pausing, Finnish

1. Introduction

Previous research has identified fluency as a key factor of second language (L2) speaking proficiency. Features of fluent speech in terms of measuring them in spoken output are rather well established. Segalowitz [1] uses the term *utterance fluency*, which refers to the measurable temporal features, such as pauses and speech rate contributing to the *perceived fluency* of the speech sample.

However, research on the interaction of multimodal features (i.e. embodied behavior such as body movements) with speech fluency features is scarce [see however 2, 3, 4]. To address this gap, we studied embodied behaviour (that is hand, head and body movements) during silent pauses (SPs) and filled pauses (FPs) at different fluency levels (A2, B1 and B2) on the CERF-based [5, 6] scale using videoed dialogue data of L2 Finnish learners.

Previous studies have shown SPs to be shorter and less frequent at higher proficiency levels (see 7 for a meta-analysis). The link between FPs and L2 speaking proficiency is not as well established (e.g. 8, 9), as there is much individual variation in FP use (e.g. 10) and their functions (e.g. 11), and their effect on perceived fluency can therefore vary.

Our approach of studying fluency from a multimodal point of view is novel, but it has been called for in L2 assessment contexts [12, 13, 14, 15, 16, 17].

2. Material and methods

For the analysis, we used a subset of the data from a larger corpus of the AASIS project, which aims to develop automatic assessment of interactional speech in L2 Finnish [18]. The data are spontaneous, paired dialogues from a problem-solving task between L2 learners of Finnish (n=38). The dialogues were videoed and manually transcribed for fluency-related and embodied features using ELAN [19] and Praat [20]. The task is publicly available in a repository [21]. A pause threshold of 200 ms for annotating SPs was used following [22]. The annotated embodied features were hand, head and body movements, each feature containing three categories (hand: pointing, fidgeting and other; head: nod, shake and other; body: lean forward, lean backward and other). The duration of the dialogues was from 3 to 5 minutes, and the duration of the total speaking time per participant in our sample varied from 1.2 to 4.3 minutes.

Expert human raters (n=21) rated the samples on a CERF-based scale (overall speaking proficiency and the subskills: range, accuracy, fluency, pronunciation and interaction). In this study, only the fluency ratings are used. A more detailed description of the speakers, rating, recording setup and annotation is provided in [18].

For the analysis we measured the duration of SPs and FPs and checked if they contained any embodied behavior. The durations were standardized in terms of total time per speaker.

3. Results

The preliminary results showed that pause durations were rather similar at all the investigated fluency levels and the differences were not statistically significant across levels. SPs were on average longer than FPs.

Embodied behaviour was present in almost all (95%) pauses (n=2705). About half (54%) of the pauses co-occurred with two different types of embodied behaviour, whereas having all three types of embodied behaviour was less frequent (13%). The frequency of embodied behaviour was similar in SPs and FPs. Body movement was on average the most frequent type (90%), hand movement the second (73%) and head movement proportionally the least frequent (22%). However, individual variation in the occurrence of embodied behavior was substantial.

The highest proficiency level (B2) was characterized by more frequent body movement than the lower levels (A2 and B1), which, in turn, were characterized by more frequent hand movements. This difference was not statistically significant.

4. Discussion and conclusions

The high frequency of body movements in the data can be explained by the physiological need to change a position while sitting in the data collection whereas the head and hand movements cannot be explained by this. The higher frequency of hand movements in the lower fluency levels (A2-B1) can be related to the compensatory functions of some hand movements, which are likely to become less prevalent when the proficiency increases.

To conclude, we would like to emphasise that there is a need to further studies on embodied behaviour using larger datasets and potentially more refined annotation protocols to better understand its role in dialogic learner speech.

5. Acknowledgements

The study was conducted within the project AASIS (*Automatic assessment of spoken interaction in second language*), funded by the Research Council of Finland (grant numbers 355586, 355587, 355588). The research design and data collection were joint efforts, for which we are thankful to the project team.

6. References

[1] Segalowitz, N. (2010). The cognitive bases of second language fluency. Routledge.

[2] Kosmala, L. (2024). Beyond disfluency: The interplay of speech, gesture, and interaction. John Benjamins. https://doi.org/10.1075/ais.11

[3] Lopez-Ozieblo, R. (2022). Cut-offs and co-occurring gestures: Similarities between speakers' first and second languages. International Review of Applied Linguistics in Language Teaching, 60(3), 647–677. https://doi.org/10.1515/iral-2017-0117

[4] Peltonen, P., Kosmala, L., Götz, S., Lintunen, P. (2023). The interplay between speech fluency and gesture in L1 Finnish and L2 English task-based interactions. Proc. Disfluency in Spontaneous Speech (DiSS) Workshop 2023, 8-12, doi: 10.21437/DiSS.2023-2

[5] Council of Europe (2001). Common European Framework of Reference for Languages: learning, teaching, assessment. Council of Europe. Retrieved 20th March 2025, from https://rm.coe.int/16802fc1bf

[6] Council of Europe (2020). Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume. Council of Europe Publishing. Retrieved 20th March 2025, from www.coe.int/lang-cefr

[7] Yan, X., Chuang, P.-L., Pan, Y., Cai, H., Staples, S., & Bertho, M. C. (2025). Disfluency doesn't happen in isolation: Exploring how individual disfluency features co-occur in L2 speaking performances. Studies in Second Language Acquisition, 1–32. https://doi.org/10.1017/S0272263125000245

[8] Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. The Modern Language Journal, 104, 169–191. https://doi.org/10.1111/modl.12620

[9] Kautonen, M., Kuronen, M., Kallio, H., Risto, E., Kosunen, R., Lepistö, K., Ohtamaa, M. & Rossi, P. (2026). Automatisk bedömning av tal på L2-svenska med fokus på flyt. Svenskan i Finland. https://urn.fi/URN:NBN:fi:jyu-202605194345

[10] Götz, S. (2019). Do learning context variables have an effect on learners' (dis)fluency? Language-specific vs. universal patterns in advanced learners' use of filled pauses. In L. Degand, G. Gilquin, L. Meurant & A. C. Simon (Eds.), Fluency and disfluency across languages and language varieties (pp. 177–196). Presses Universitaires de Louvain.

[11] Kosmala, L., & Crible, L. (2021). The dual status of filled pauses: Evidence from genre, proficiency and co-occurrence. Language and Speech, 65(1), 216–239. https://doi.org/10.1177/00238309211010862

[12] Burton, J. D. (2024). Evaluating the impact of nonverbal behavior on language ability ratings. Language Testing, 41(4), 729–758. https://doi.org/10.1177/02655322241255709

[13] Galaczi, E. D., & Taylor, L. (2018). Interactional competence: Conceptualisations, operationalisations, and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236. https://doi.org/10.1080/15434303.2018.1453816

[14] Glasson, N., & Halley, K. (2025). Less talk, more action? Exploring proficiency scores and embodied resources in online L2 interactions. Classroom Discourse, 16(3), 318–339. https://doi.org/10.1080/19463014.2024.2408421

[15] Jenkins, S., & Parra, I. (2003). Multiple layers of meaning in an oral proficiency test: The complementary roles of nonverbal, paralinguistic, and verbal behaviors in assessment decisions. Modern Language Journal, 87(1), 90–107. https://doi.org/10.1111/1540-4781.00180

[16] May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. Language Assessment Quarterly, 8(2), 127–145. https://doi.org/10.1080/15434303.2011.56

[17] Plough, I., Banerjee, J., & Iwashita, N. (2018). Interactional competence: Genie out of the bottle. Language Testing, 35, 427–445. http://dx.doi.org/10.1177/0265532218772325

[18] Ullakonoja, R., Lähteenmäki, I., Raud, N., Phan, N., Grósz, T., Suuronen, H., Hilden, R., Kurimo, M., Kuronen, M., Von Zansen, A., & Kautonen, M. (2025). Building the foundations for automatic assessment of verbal and nonverbal aspects of spoken interaction in Finnish as a second language. Studies in Language Assessment, 14(2), 28–57. https://doi.org/10.58379/FFNO9141

[19] Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. Behavior Research Methods, 41, 841–849. https://doi.org/10.3758/BRM.41.3.841

[20] Boersma, P., & Weenink, D. (2025). Praat: doing phonetics by computer [Computer software]. Version 6.4.33, retrieved 18 May 2025 from https://www.praat.org/

[21] Von Zansen, A. (2024). Aasis speaking tasks for Finnish L2 learners (2024). Zenodo. https://doi.org/10.5281/zenodo.14892701

[22] J. Gao, P. P. Sun, and C. Li. (2025)." Exploring the optimal thresholds of silent pauses for measuring second language utterance fluency in monologic and dialogic speaking". *Language Testing,* vol. 42, no. 3, pp. 283–311. 2025 https://doi.org/10.1177/02655322251315792

STS-Net: A Phonological Property Prediction Model for Swedish Sign Language

Jonas Beskow[1]

[1]Department of Speech, Music and Hearing, KTH Royal Institute of Technology, Sweden

beskow@kth.se

1. Introduction

Sign languages have a rich sublexical phonological structure: each sign is characterised by handshape, hand orientation (attitude), place of articulation, contact type, movement, and number of hands. Computational modelling of these properties supports recognition, segmentation, and linguistic analysis. Tavella et al. [1] introduced WLASL-LEX, a dataset for recognising phonological properties in ASL from skeleton features, demonstrating the feasibility of automatic phonological annotation from pose. Labelled phonological data remain scarce for smaller sign languages, however, and continuous-signing corpora are typically unannotated at the phonological level.

Swedish Sign Language (STS) is one such language. The Swedish Sign Language Lexicon (SSLL)[1] provides roughly 20,000 isolated-sign video clips, each accompanied by a textual description from which phonological targets can be automatically parsed. The Swedish Sign Language Corpus (SSLC), by contrast, contains approximately 210,000 annotated gloss instances in continuous conversational signing from 37 signers, but carries no phonological labels [2].

This paper presents STS-Net, which learns six phonological feature categories from SSLL isolated clips and demonstrates three downstream applications: (1) mining SSLC for additional training data via embedding-space retrieval, (2) gloss recognition in SSLC using the learned representations, and (3) sign and phrase boundary segmentation of SSLC continuous video. The segmentation experiments additionally reveal a prosodic limitation of clip-level training: STS-Net features encode sign-level phonological properties but discard the supra-lexical temporal dynamics that underlie prosodic phrasing, and a TCN trained end-to-end on raw pose sequences recovers phrase boundaries substantially better.

2. Model

Each clip is processed with MediaPipe [3] at 25 fps to yield four parallel pose streams: (i) dominant hand — 21 wrist-centred keypoints; (ii) non-dominant hand — 21 keypoints; (iii) upper body — 12 shoulder-normalised keypoints; (iv) face — 25 nose-normalised keypoints. All keypoints are 2D (x, y); invalid frames (missing wrist detection) are zeroed.

Each stream is encoded by a dedicated **FrameEncoder**: a linear projection to $D{=}256$ dimensions, layer normalisation, ReLU activation, and three stacked temporal convolution blocks (Conv1d, kernel 5, residual connections, layer norm, ReLU, dropout). The four 256-dim outputs are concatenated and passed through a **Fusion MLP** (1024→256, layer norm, ReLU, dropout). A **masked attention-pooling** layer then aggregates frames within the annotated sign window $[\text{start}, \text{end})$ into a single 256-dim clip embedding. Six multi-label **classification heads** read from this embedding: handshape (42 classes, multi-hot BCE), attitude/orientation (34 classes, BCE), contact location (22 classes, BCE), contact type (4 classes, BCE), motion direction (8 classes including *none*, BCE), and hand type (2 classes, cross-entropy). Total parameters: ≈4.3M.

The model is trained on ≈17,700 SSLL clips (signer-based 85/15 train/val split) with AdamW (lr = 3e-4), batch size 32, cosine LR schedule, dropout 0.2, and Gaussian pose noise ($\sigma = 0.005$). Phonological targets are parsed automatically from the textual sign descriptions. Base validation accuracy: handshape 80.1%, attitude 77.4%, hand type 96.3%.

[1]https://teckenspraklexikon.su.se

3. Experiments

3.1. Input representations

Table 1 compares five input representations on SSLL. **MediaPipe (MP)** is the 2D skeleton baseline. **WiLoR** replaces hand streams with 3D MANO keypoints [4]. **WiLoR+MP** combines all six streams. **DINO+MP** adds a DINOv3 ViT-L/16 CLS token [5] to the skeleton. **Sapiens (1B model)** uses a whole-body pose foundation-model stream in place of MediaPipe.

WiLoR+MP scores best across all properties, confirming 3D hand geometry complements the 2D skeleton. Sapiens matches the MP baseline despite using a single stream. MediaPipe was chosen for the remainder of the experiments since feature extraction with Sapiens 1B and WiLoR is around one order of magnitude slower.

Table 1: *Top-1 accuracy (%) by input representation (SSLL only, val set).*

Property	MP	WiLoR	WiLoR+MP	DINO+MP	Sap.
Shape (dom.)	85.1	83.5	**85.5**	82.5	83.0
Attitude	78.8	79.3	**79.3**	76.7	78.4
Contact loc.	81.2	79.7	**81.8**	77.7	81.1
Contact type	75.8	75.7	**76.3**	73.3	76.1
Motion dir.	63.6	63.1	**65.1**	60.5	64.4
Hand type	97.0	96.7	**97.3**	96.5	96.7

3.2. Sign mining from continuous data

Phonological labels cannot be derived directly from SSLC gloss strings. Instead, we transfer labels via embedding-space matching. For each SSLL training clip we compute a 256-dim clip embedding; embeddings are grouped by gloss. For each SSLC gloss window we compute the same embedding and retrieve the

nearest SSLL variant per gloss. Instances with cosine distance below 0.4 are retained, and the SSLL clip's phonological targets are transferred. This yields **11,280 instances** spanning 1,284 glosses from 28 SSLC signers. Re-training on SSLL + mined SSLC with stronger regularisation (dropout 0.3, noise $\sigma = 0.02$, label smoothing 0.1, $\pm 15\%$ temporal stretch) gives the improvements in Table 2.

Table 2: *Top-1 accuracy on all predicted phonological properties before and after SSLC mining. Both models use 2D pose, 4 streams, and a non-dominant handshape head. Evaluated on the SSLL validation set (3,100 clips).*

Property	SSLL only	+mined SSLC
Handshape (dom.)	79.9%	**85.1%**
Attitude (dom.)	75.9%	**79.3%**
Contact location	78.3%	**80.6%**
Contact type	72.3%	**76.8%**
Motion direction	57.4%	**62.4%**
Hand type	96.6%	**97.1%**
Handshape (nondom.)	81.3%	**85.7%**

3.3. Gloss recognition on SSLC

Zero-shot retrieval. Each of 64,462 SSLC test clips is ranked against per-gloss mean SSLL embeddings over 2,010 glosses. Mining raises Rec@1 from 1.0% to **5.8%** and Rec@10 from 5.8% to **22.2%** without any direct SSLC supervision for this task.

Supervised linear probe. A linear classifier is trained on frozen 256-dim clip embeddings from 116k SSLC clips and tested on 13k clips (1,824-gloss closed vocabulary), as shown in Table 3. The mining-augmented model outperforms the SSLL-only baseline by $\approx$4 percentage points, with median rank improving from 12 to 8.

Table 3: *Supervised linear-probe gloss recognition on SSLC.*

Model	Rec@1	Rec@5	Med. rank
SSLL only	18.6%	38.3%	12 / 1824
+mined SSLC	**22.8%**	**44.5%**	**8 / 1824**

3.4. Sign segmentation on SSLC

We address BIO sequence labelling (B = begin, I = inside, O = outside) for sign and phrase boundaries on SSLC continuous video, evaluated on 44 held-out test videos (signer-based split). Sign boundaries are decoded via Moryossef-style peak detection on B-class probabilities [6]. We compare four SSLC-trained models: (i) a **windowed linear classifier** on frozen STS-Net features; (ii) a **TCN on STS-Net features** (6 dilation levels, receptive field $\approx$10 s); (iii) a **TCN on raw pose** that bypasses the STS-Net backbone and is trained directly on the concatenated pose streams; and (iv) the pretrained BiLSTM of Moryossef et al. applied zero-shot from German Sign Language (DGS).

Both TCN models achieve the best sign-boundary F1 (S-F1 = 0.547) with near-ideal sign count ratios. The key contrast is in phrase IoU: the TCN on STS-Net features reaches P-IoU = 0.572, while the TCN trained directly on raw pose improves this to 0.602 — despite identical architecture. This gap reflects a fundamental limitation of the clip-level training objective: STS-Net encodes sign-level phonological properties but the masked attention pooling discards the long-range temporal context that encodes prosodic phrasing. The raw-pose TCN has access to the full continuous signal and thus learns the dynamics of both sign boundaries and phrase rhythm jointly. The pretrained DGS model leads on phrase IoU in zero-shot transfer (0.612), further suggesting that training on continuous signing — regardless of language — is more important than language match for capturing prosodic structure.

Table 4: *Sign/phrase segmentation on SSLC (44 test videos). Sign % = predicted / gold segment count (ideal: 1.00). The DGS rows are not directly comparable (different language/corpus).*

Model	S-F1	S-IoU	P-F1	P-IoU
Windowed linear (STS-Net)	0.540	0.519	0.486	0.438
TCN on STS-Net	**0.547**	0.511	0.535	0.572
TCN on raw pose	**0.547**	**0.553**	**0.551**	0.602
Pretrained DGS→SSLC	0.498	0.442	0.569	**0.612**
Moryossef '23 (DGS)	*0.630*	*0.690*	*0.650*	*0.820*

4. Conclusions

STS-Net is a compact, pose-based model that predicts six phonological feature categories from short clips. A nearest-neighbour mining step bridges isolated and continuous signing corpora, yielding a 5 pp gain in handshape accuracy and a sixfold improvement in zero-shot gloss retrieval. The frozen backbone supports competitive sign boundary segmentation on SSLC continuous video. Critically, the segmentation experiments reveal a prosody-relevant asymmetry: clip-level training preserves sign-level phonological detail but discards the prosodic structure of continuous discourse. A TCN trained end-to-end on raw pose substantially outperforms the STS-Net-based model on phrase boundaries, pointing toward sequence-level training objectives as a path to computational modelling of sign language prosody. The STS-Net model is available at `https://github.com/jbeskow/stsnet`

5. References

[1] F. Tavella, V. Schlegel, M. Romeo, A. Galata, and A. Cangelosi, "Wlasl-lex: a dataset for recognising phonological properties in american sign language," *arXiv preprint arXiv:2203.06096*, 2022.

[2] J. Mesch and L. Wallin, "Gloss annotations in the swedish sign language corpus," *International Journal of Corpus Linguistics*, vol. 20, no. 1, pp. 102–120, 2015.

[3] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," in *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2020.

[4] R. A. Potamias, J. Lin, S. Moschoglou, and S. Zafeiriou, "WiLoR: End-to-end 3d hand localisation and reconstruction in-the-wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024.

[6] A. Moryossef, Z. Jiang, M. Müller, S. Ebling, and Y. Goldberg, "Linguistically motivated sign language segmentation," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023, pp. 12 703–12 724.

Initial analyses of Storspigg-TBI: a Swedish multispeaker corpus

Christina Tånnander[1,2], Jim O'Regan and Jens Edlund[1]
[1]KTH Royal Institute of Technology, [2]Swedish Agency for Accessible Media

Storspigg-TBI is a Swedish read-speech corpus comprising 1 000 recordings of nearly identical texts produced by 99 professional narrators [1]. The corpus was created from introductory messages included in talking books and provides a unique resource for investigating variation in read speech while controlling for linguistic content. In this paper, we present initial analyses of the corpus, focusing on inter- and intra-speaker variation in speaking rate. The recordings span approximately 12.7 hours of speech and exhibit substantial variation despite the highly constrained text material.

We examine differences in speaking rate across narrators, as well as within-speaker variation across repeated recordings. Preliminary results indicate considerable inter-speaker variation. The findings demonstrate the value of Storspigg-TBI for studying speech variation and provide a first quantitative characterization of the corpus. Future work will extend the analysis to segmental and prosodic variation and explore applications in speech technology and speech science.

References

[1] C. Tånnander, J. O'Regan, and J. Edlund, 'A shoal of voices: Parallel read speech from professional Swedish narrators', in Proc. of LREC 2026, Palma, Spain, 2026.

Prosodic stability and segmental mobility: Preliminary results from Kronoberg origin speakers living in Stockholm

Erik Åkerdal, Christine Ericsdotter Nordgren & Pétur Helgason

Department of Linguistics, Stockholm University
christine.ericsdotter@su.se

Abstract

The Swedish region of Småland shows great dialect variety, encompassing phonetic and phonologicial features from Västmellansvenska (West Mid Sweden dialects), Östmellansvenska (East Mid Sweden dialects) and Sydsvenska (South Sweden dialects) (ISOF, 2020). As an example, for the /r/ phoneme, we find both dorsal and coronal places of articulation; approximant, fricative and tremulant manners of articulations; and use or no use of the retroflexion rule for coronals (Bruce, 2010, Frid et al, 2010). This contribution presents an ongoing study of the speech of 6 adults with origin in Kronoberg in Småland County, who have lived in Stockholm for at a minimum 4 years (Åkerdal, 2026). The Kronoberg variety is described as a South Sweden Dialect, which differs from the Stockholm variety both from a segmental and a prosodic point of view. Preliminary results on pitch accent and /r/-realisations show that while the adaptation of coronal /r/ and the phonological process of retroflexion is present in all but one speaker, none of the speakers have adapted the Stockholm pitch accent pattern, at least not in read utterance final positions. Their pitch accents show pattern 2B using Gårding's typology (Gårding, 1977), which is characteristic for Götamål rather than Sydsvenska mål. Short qualitative interviews with the speakers reveal their main reasons for speech change are connected to intelligibility rather than social accommodation. The results are discussed with regards to dialect status, language change and language style and styling (Löfström, 2022).

Bruce, G. (2010). *Vår fonetiska geografi: Om svenskans accenter, melodi och uttal*. Lund: Studentlitteratur.

Frid, J., Ericsdotter, C., Engstrand, O., & Fridell, S. (2010). R-ljuden i svenskan. I Ö. Dahl, L.-E. Edlund, L. Wastenson, & M. Elg (Red.), *Språken i Sverige* (s. 66–67). Norstedt.

Gårding, E. (1977). *The Scandinavian word accents*. Lund: LiberLäromedel / Gleerup.

Institutet för språk och folkminnen (ISOF) (7 juni, 2020). *Utforska svenska dialekter.* https://www.isof.se/dialekter/lar-dig-mer-om-svenska-dialekter/utforska-svenska-dialekter

Löfström, M. (2022). *Språklig stil och stajling bland finlandssvenskar i Stockholm: – Ett mobilitetsdialektologiskt perspektiv*. Doctoral thesis, Uppsala universitet: Acta Universitatis Upsaliensis. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-472417

Åkerdal, E (2026). *Hur låter en flyttare?* Bachelor's thesis, Stockholms universitet: Institutionen för lingvistik.

Inter- and intra-dialectal phonetic differences in Danish stød production in Copenhagen and Aarhus

Andrea Siem[1]

[1]Lancaster University
a.siem@lancaster.ac.uk

1. Introduction

Phonetic studies of the phonologically contrastive syllabic prosody *stød* in Danish have focused almost exclusively on the standard dialect, *rigsdansk*, spoken in and around Copenhagen [1], [2], [3] leading to limited descriptions of what the stød is phonetically outside of the standard language variant. This study is an analysis of the under-researched area of dialectal variation in stød production, comparing Copenhagen speakers to speakers from Aarhus, a dialect that has been speculated to have a more tonal variant of the stød under certain phonological conditions [4], [5]. To explore this, this study differentiates two types of minimal pairs with four different phonological stød basis conditions: pair one = no stød vs stød in open syllables with a long vowel (CV: vs CV:ˀ) and pair two = closed monosyllables with stød on either a sonorant consonant or a long vowel (CVCˀ vs CV:ˀC). Two main questions are pursued: 1) Which acoustic and articulatory measures differentiate the stød in each minimal pair within each dialect, and 2) which acoustic measures predict the *stød* differences between dialects?

2. Methods

Acoustic and Electroglottographic data was collected from 10 speakers from Copenhagen (5 male, 5 female) and 11 speakers from Aarhus (6 male, 5 female). Stimuli were sentences presented in standard orthography eliciting one of the four stød basis conditions, each sentence repeated three times in random order using the SpeechRecorder software [6]. Acoustic and Electroglottographic (EGG) data was collected simultaneously via an Audio-Technica AT803 lapel microphone and a Laryngograph® Electroglottograph synched with the audio signal. 30 minimal pairs per speaker were elicited for analysis and all data were manually annotated in Praat [7], creating TextGrids that enabled the extraction of the stød basis portion of each token. The synced signals were separated for acoustic and articulatory analysis. Acoustic parameters f0, Harmonics-to-Noise Ratio (HNR), Cepstral Peak Prominence (CPP), H1-H2 and Intensity were estimated via VoiceSauce [8], [9]. EGG data was transformed to dEGG format and extracted via Praat [7]. Acoustic analysis was done on 3.788 tokens and articulatory analysis was done on 43.666 glottal cycles spread across 3.657 tokens. Statistical analysis was performed in R Studio [10] where Generalised Additive Mixed Models (GAMMs) were fitted to the dynamic trajectories of each acoustic measurement during the stød conditions for each dialect and then with interaction terms between stød and dialect to compare dialects directly. To determine which acoustic parameter predicted the stød best in each dialect, acoustic variable importance was then assessed by splitting the data into each of the minimal pair sets per dialect and fit *random forest* models to respective sets that ranked them in a variable importance hierarchy. GAMMs were also fitted to the articulatory data modelling the Contact Quotient (CQ) calculated from the electroglottographic data as a proportion of vocal fold contact in each cycle from 0 to 1 where 0.50 is modal phonation, higher than 0.50 indicates compressed or creaky phonation, and lower than 0.50 indicates breathy phonation. To determine statistical significance for acoustic and articulatory parameters, likelihood ratio tests were performed comparing full and reduced GAMMs assessing whether dynamic trajectories differed across stød and dialect conditions. The smooth term model predictions were visualized along with difference smooths to assess the nature of the differences.

3. Results

The analysis included 20 separate GAMMs, too many to visualise in this abstract, so Figure 1 and Figure 2 are included as illustrative examples of the GAMMs, Figure 1 for the dialect interaction terms for the stød vs no stød minimal pairs and Figure 2 for the predicted CQ values for each separate dialect in the closed monosyllables with stød on a long vowel vs a sonorant consonant. The acoustic analysis indicated that for the no stød vs stød condition in open syllables with a long vowel (CV: vs CV:ˀ) within the Copenhagen dialect, the presence of stød correlated significantly with f0 and a relative decrease in HNR, CPP and Intensity. Unexpectedly, f0 was higher for stød tokens and H1-H2 was also higher than expected during the stød. For the closed monosyllables with stød on either a sonorant consonant or a long vowel (CVCˀ vs CV:ˀC) HNR, CPP and Intensity were significant in predicting the difference in Copenhagen but f0 and H1-H2 were not. For the Aarhus dialect the CV: vs CV:ˀ pair f0, HNR, CPP, H1-H2 and Intensity were all significant predictors of the stød presence, and these measures decreased towards syllable termination. For the CVCˀ vs CV:ˀC pair f0 and H1-H2 were not significant predictors of the difference of the two stød conditions but Intensity, CPP and HNR were, indicating less signal noise on the CV:ˀC condition. H1-H2, however, did not indicate more glottal constriction to explain increased signal noise in the CVCˀ condition. The factorial interaction GAMMs fitted for direct dialect comparisons indicated that the effect of stød condition differed significantly between dialects on all acoustic parameters for the CV: vs CV:ˀ pair. This was also the case for the CVCˀ vs CV:ˀC pair except for parameter H1-H2 which was not a significant predictor of stød condition in the interaction model. The *random forest* models indicated that the acoustic variable importance differed between the two stød conditions but not between dialects, with f0 being the best predictor of stød type

in CV: vs CV:ˀ syllables and Intensity being the best predictor in CVCˀ vs CV:ˀC syllables for both dialects.

The articulatory analysis of the Contact Quotient (CQ) values indicated that for the Copenhagen speakers, the CV: vs CV:ˀ condition had significantly different values in the model predictions with overall higher values for the stød tokens with the mean rising to 0.55 towards syllable termination compared to the 0.46 mean for non-stød tokens. This indicates more compressed phonation on the stød, but less so than expected if the stød were realised with a prototypical creaky voice. The CVCˀ vs CV:ˀC minimal pair did not have significantly different CQ values. Both had a rising mean trajectory throughout the syllable, ending around a 0.53 CQ. For the Aarhus speakers, the difference between neither minimal pair was predicted by the CQ values. The CV: vs CV:ˀ condition had slightly lower mean CQ values for the non-stød tokens but both trajectories were on the breathy phonation continuum below 0.50. The CVCˀ vs CV:ˀC minimal pair had rising mean CQ values for the CVCˀ syllables ending just below 0.50 and a level CQ mean trajectory at around 0.47 (Figure 2). This indicates that there is absence of creaky phonation in both stød conditions in Aarhus.

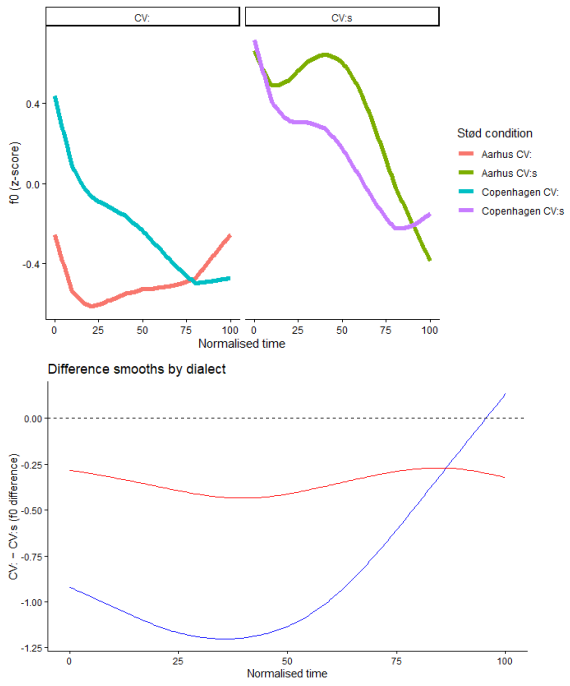


Figure 1: *GAMM-predicted f0 trajectories for the minimal pair CV: vs CV:ˀ (stød indicated with 's') in Aarhus and Copenhagen (top) and the difference smooth by dialect (bottom).*

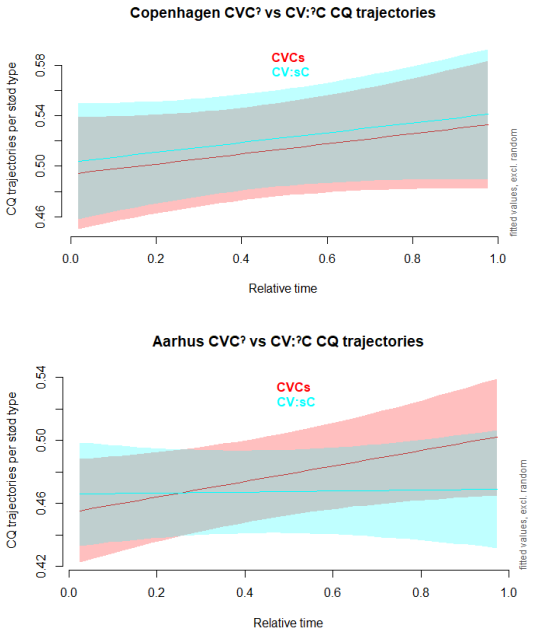


Figure 2: *GAMM-predicted Contact Quotient trajectories for the minimal pair CVCˀ vs CV:ˀC (stød indicated with 's') in Copenhagen (top) and Aarhus (bottom).*

4. Discussion

The results are discussed in the light of 1) previous stød findings [1], [2], [3], 2) findings from cross-linguistic studies on contrastive non-modal phonation [11], [12], [13], [14], [15], [16], [17], [18], 3) the Laryngeal Articulator Model [19], 4) non-modal phonation timing relative to phonetic segments [20], [21], [22], [23], [17], [24], [25] and 5) inter-dialectal variation in phonation and tone in other languages [26].

5. References

[1] P. R. Petersen, "An instrumental investigation of the Danish 'stød'," Annual Report of the Institute of Phonetics, University of Copenhagen, vol. 7, pp. 195–234, 1973. Available: https://tidsskrift.dk/AIRPUC.

[2] E. Fischer-Jørgensen, "Phonetic analysis of the stød in standard Danish," Phonetica, vol. 46, nos. 1–3, pp. 1–59, 1989, doi: 10.1159/000261828.

[3] N. Grønnum, "Three quarters of a century of phonetic research on common Danish stød," Nordic Journal of Linguistics, vol. 46, no. 3, pp. 299–330, 2023, doi: 10.1017/S0332586522000178.

[4] B. Kyst, *Trykgruppens toner i århusiansk regionalsprog*. Unpublished Master's thesis, Aarhus University, 2004.

[5] B. Kyst, "Toner i århusiansk regiolekt," in *Nordisk dialektologi og sociolingvistik*, T. Arboe, Ed. Aarhus: Fællestrykkeriet, Aarhus Universitet, 2007, pp. 238–247.

[6] C. Draxler and K. Jänsch, "SpeechRecorder – a universal platform independent multi-channel audio recording software," in *Proc. International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, 2004. Available: https://www.bas.uni-muenchen.de/Bas/software/speechrecorder/

[7] P. Boersma, D. Weenink, and A. Shchupak, *Praat: doing phonetics by computer* [Computer program], version 6.4.52, 2026. Available: https://praat.org/

[8] Y.-L. Shue, *The voice source in speech production: Data, analysis and models*. Ph.D. dissertation, University of California, Los Angeles, 2010.

[9] Y.-L. Shue, P. Keating, C. Vicenik, and K. Yu, "VoiceSauce: A program for voice analysis," in *Proc. 17th International Congress of Phonetic Sciences (ICPhS XVII)*, Hong Kong, China, 2011, pp. 1846–1849.

[10] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2025. Available: https://www.r-project.org/
[11] M. Garellek and P. Keating, "The acoustic consequences of phonation and tone interactions in Jalapa Mazatec," Journal of the International Phonetic Association, vol. 41, no. 2, pp. 185–205, 2011, doi: 10.1017/S0025100311000193.
[12] C. M. Esposito, "The effects of linguistic experience on the perception of phonation," Journal of Phonetics, vol. 38, pp. 306–316, 2010, doi: 10.1016/j.wocn.2010.02.002.
[13] B. Blankenship, "The timing of nonmodal phonation in vowels," Journal of Phonetics, vol. 30, no. 2, pp. 163–191, 2002, doi: 10.1006/jpho.2001.0155.
[14] J. E. Andruski and M. Ratliff, "Phonation types in production of phonological tone: The case of Green Mong," Journal of the International Phonetic Association, vol. 30, nos. 1–2, pp. 37–61, 2000, doi: 10.1017/S0025100300006654.
[15] C. M. Esposito, "An acoustic and electroglottographic study of White Hmong tone and phonation," Journal of Phonetics, vol. 40, no. 3, pp. 466–476, 2012, doi: 10.1016/j.wocn.2012.02.007
[16] H. Avelino, "Acoustic and electroglottographic analyses of nonpathological, nonmodal phonation," Journal of Voice, vol. 24, no. 3, pp. 270–280, 2010, doi: 10.1016/j.jvoice.2008.10.002.
[17] C. DiCanio, "The phonetics of register in Takhian Thong Chong," Journal of the International Phonetic Association, vol. 39, no. 2, pp. 162–188, 2009, doi: 10.1017/S0025100309003946
[18] P. Keating, C. M. Esposito, M. Garellek, S. D. Khan, and J. Kuang, "Phonation contrasts across languages," UCLA Working Papers in Phonetics, vol. 108, pp. 188–202, 2010.
[19] J. H. Esling, S. R. Moisik, A. Benner, and L. Crevier-Buchman, *Voice Quality: The Laryngeal Articulator Model*. Cambridge Studies in Linguistics. Cambridge, United Kingdom: Cambridge University Press, 2019.
[20] S. Hargus, "Deg Xinag word-final glottalized consonants and voice quality," in *The Phonetics and Phonology of Laryngeal Features in Native American Languages*, H. Avelino, M. Coler, and L. Wetzels, Eds. Leiden, Netherlands: Brill, 2016, pp. 71–128.
[21] D. Howe and D. Pulleyblank, "Patterns and timing of glottalisation," Phonology, vol. 18, no. 1, pp. 45–80, 2001, doi: 10.1017/S0952675701004045.
[22] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*. Oxford, United Kingdom: Blackwell, 1996.
[23] D. Silverman, *Phasing and Recoverability*. New York: Garland Press, 1997.
[24] C. Esposito, *Santa Ana del Valle Zapotec phonation*. Master's thesis, University of California, Los Angeles, 2003.
[25] A. Siem, "Variation and dialectal differences in the segmental scope of phonologically contrastive laryngealisation," in *Proc. 20th International Congress of Phonetic Sciences (ICPhS 2023)*, Prague, Czech Republic, Aug. 2023.
[26] D. A. Morrison, "Metrical structure in Scottish Gaelic: Tonal accent, glottalisation and overlength," Phonology, vol. 36, no. 3, pp. 391–432, 2019, doi: 10.1017/S0952675719000184.

On the role of focal lengthening in two different dialects of Swedish

Gilbert Ambrazaitis[1], *Mechtild Tronnier*[2]

[1]Department of Swedish, Linnaeus University, Växjö, Sweden
[2]Centre for Languages and Literature, Lund University, Sweden
gilbert.ambrazaitis@lnu.se, mechtild.tronnier@ling.lu.se

Abstract

We investigate whether Scanian Swedish relies more heavily on durational cues to phrase-level prominence (PLP) than Stockholm Swedish, compensating for its relatively subtle f_o cues for PLP. Results show a clear focal-lengthening effect in both dialects and generally longer stressed syllables in Scanian, but no evidence for enhanced compensatory focal lengthening.

Index Terms: focus, prominence, duration, compensation

1. Introduction

Focus can be signaled prosodically in all Swedish dialects, but it is well known that dialects differ markedly with respect to how intonational, or phrase-level prominence (PLP) is realized phonetically [1]. In this study, we concentrate on Stockholm and Scania Swedish. While PLP in Scanian has traditionally been described as a matter of gradual phonetic adjustments of the accentual HL patterns determined by the lexical pitch accents, in Stockholm Swedish, PLP is realized more categorially through an additional rise in f_o, typically described phonologically as a high tone (H) that is added to the HL word accent patterns [1][2].

A question that arises from this comparison is whether such differences in the tonal realization of PLP would affect the perceptual salience of PLP. Informal observations suggest that perceiving PLP in Scanian can be challenging for speakers of other dialects or second-language learners, although there is evidence that PLP is processed in a very similar way by speakers of Scanian and Stockholm in their respective dialect, at least in a context where PLP is used to mark contrastive focus [3]. Thus, it might well be that speakers of a language or dialect with less salient PLP patterns develop the skill to perceive native PLP distinctions, even if they are phonetically subtle.

However, one might still hypothesize that acoustic cues beyond f_o play a larger role for PLP in Scanian than in Stockholm Swedish, to compensate, at least to some degree, for the weaker tonal cues for PLP. The phonetic realization of PLP is known to involve a multitude of acoustic parameters beyond f_o, such as duration [4], and several spectral characteristics [5]. In this study, we focus on duration and its potential to compensate for f_o cues and compare the magnitude of so-called focal lengthening in Scanian to that in Stockholm Swedish.

On the one hand, previous research on different acoustic cues to PLP suggests an overall positive correlation between f_o and durational cues, implying a certain dependency: for a perceptually stronger PLP, one would typically expect a more salient f_o inflection and a longer duration of the accented syllable. However, previous research also suggests a certain level of independence, or interplay, between duration and f_o [6][7]. The results of this study are expected to shed more light on a potential compensatory interplay between different acoustic cues to prominence, but also to further develop our understanding of the phonetic dimensions of the Swedish prosodic dialect typology.

2. Method and materials

This study is based on recordings made with 24 native speakers of Swedish (two dialects, 6 women and 6 men per dialect). In the present study, we include 575 test sentences from these recordings, comprising 4 individual test sentences (1) uttered in 2 focus conditions with 3 repetitions (= 24 sentences x 24 speakers – 1 missing token = 575 in total). Speakers were asked to read the sentences as appropriate answers to context questions designed to trigger different focus conditions, two of which are included in the present analysis: narrow focus on the medial and on the final noun, underlined in (1).

(1) Boven hade vinet/-er i bilen/-ar.
'The villain had the wine/ wines in the car/ in cars.'

Note: (1) represents one base sentence with four versions related to variations in lexical word accent, which are not further discussed in this paper; word accent is balanced across the two focus conditions in the selected data set.

All words in the 575 recorded sentences were segmented manually at the syllable level using Praat [8] and syllable durations were extracted. Our statistical analysis focuses on the stressed syllable in the two target words. We applied linear mixed modelling to model relative syllables durations (% of sentence duration) considering the predictors Focus condition, Dialect, as well as their interaction, based on a random-effects structure specifying random intercepts for Speaker and random slopes for Focus condition by Speaker. (Considering random effects of Repetition resulted in convergence issues). We used likelihood ratio tests for model comparisons to test whether Focus, Dialect, and the Focus*Dialect interaction term would improve model fit. Modelling was performed in R [9] using the lme4 package [10]. If focal lengthening were stronger in Scanian than in Stockholm Swedish to compensate for the less salient tonal realization of PLP, we would expect a significant contribution of the Focus*Dialect interaction term.

3. Results

Figure 1 presents an overview of the data. The boxplots suggest for both dialects that each target word *(vinet/er* or *bilen/ar*) is prounced longer when in focus than when outside focus. Furthermore, in the case of the medial word (*vinet/er*) focal lengthening appears to affect the stressed syllable more than the following unstressed one.

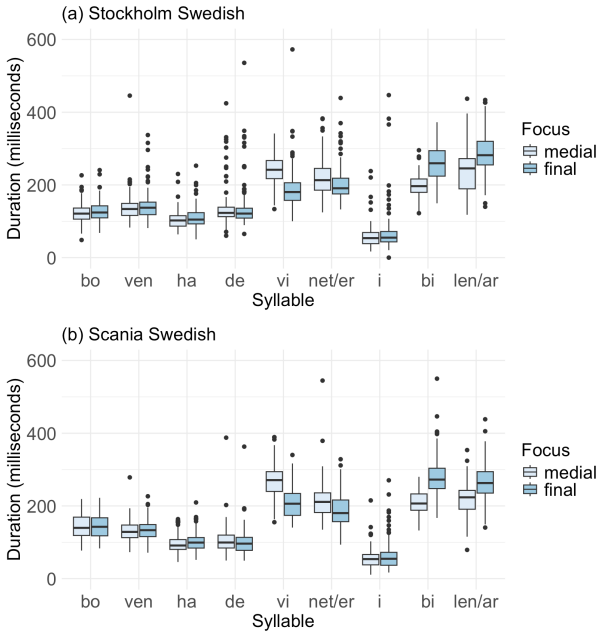


Figure 1: Boxplots summarizing the obtained syllable durations per dialect and focus condition; medial = focus on *vin-et/er*; final = focus on *bi-len/ar*.

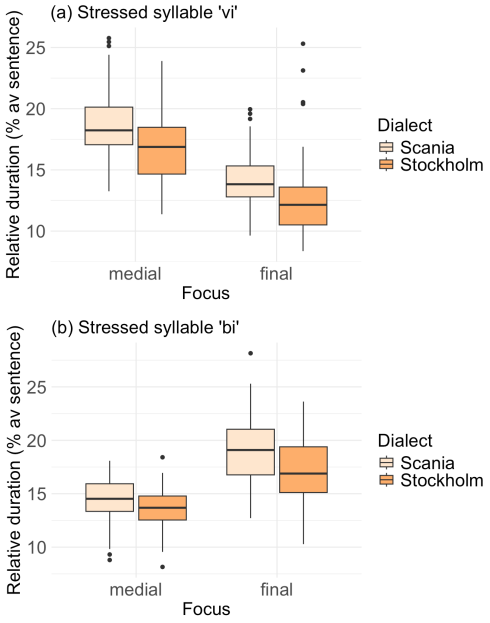


Figure 2: Boxplots summarizing the results for relative syllable durations per dialect and focus condition for the stressed syllable of the medial target word (a) and of the final target word (b); medial = focus on *vin-et/er*; final = focus on *bi-len/ar*.

To visualize potential differences in focal lengthening between the two dialects, we calculated, for each sentence, the relative syllable duration of the stressed syllable for each of the two target words (i.e. *vi* and *bi*) in terms of a percentage of sentence duration; Figure 2 displays these relative syllable durations for *vi* (Fig. 2a) and *bi* (Fig. 2b), respectively, per dialect and focus condition, that is, when the target word is in focus (e.g., *vi* when Focus is medial, in Fig. 2a) and as a reference when not in Focus (e.g., *vi* when Focus is final, in Fig. 2a).

Both Figure 2a and 2b suggest, again, for both dialects, that the stressed syllable of a target word is (relatively) longer when the word is in focus than when outside focus. In addition, they suggest that the stressed syllable of a focused word is slightly longer in Scanian than in Stockholm Swedish. However, this dialect related difference appears to be independent of the Focus condition: the two syllables tend to be relatively longer in Scanian both when the word is in focus and when not.

These observations are supported by our modeling approach: Adding the predictor Focus to our random-effects structure improves model fit significantly (syllable *vi*: $\chi^2(1)= 46.50$***; syllable *bi*: $\chi^2(1)= 40.73$***), and while adding the predictor Dialect to the Focus-model results in further improvement (syllable *vi*: $\chi^2(1)= 12.04$***, syllable *bi*: $\chi^2(1)= 4.05$*), this is not the case for the Focus*Dialect interaction (syllable *vi*: $\chi^2(1)= .13$, $p=.72$; syllable *bi*: $\chi^2(1)= 1.47$, $p=.23$).

4. Conclusion

The present analysis confirms the well-established observation of focal lengthening [4], for two dialects of Swedish, and two different sentence positions (here manifested in two different target words). It also suggests a general trend for longer stressed syllables in Scanian compared to Stockholm Swedish (a difference we will leave unexplained here), but no stronger focal lengthening. Thus, although the tonal marking of PLP is arguable more subtle in Scanian than in Stockholm Swedish, the present results do not suggest any compensatory strategy by means of durational cues.

5. References

[1] G. Bruce and E. Gårding, "A prosodic typology for Swedish dialects," in E. Gårding, G. Bruce and R. Bannert (eds), *Nordic Prosody – Papers from a Symposium*. Lund: Lund University Press,1978, pp. 219-228.

[2] G. Ambrazaitis, J. Frid, and G. Bruce, "Revisiting Southern and Central Swedish intonation from a comparative and functional perspective," in O. Niebuhr (ed.), *Understanding prosody – The role of context, function, and communication*. Berlin: de Gruyter, 2012, pp.135–158.

[3] G. Ambrazaitis, N. Althaus, S. Löhndorf, A. S. H. Romøren, and S. Sayehli, "Contrastive focus comprehension in Swedish adults and children (3-5 years) from two regional varieties: can salient pitch accent provide an advantage for acquisition?", in *Proc. FiNo 2026 – Fonologi i Norden*, Lund, Sweden, Feb. 2026, pp. 25-27.

[4] M. Heldner and E. Strangert, "Temporal effects of focus in Swedish," *Journal of Phonetics,* vol. 29, no. 3, pp. 329–361, 2001.

[5] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in Swedish," *Journal of Phonetics*, vol. 31, no. 1, pp. 39–62, 2003.

[6] A. Arnhold and A.-J. Kyröläinen, "Modelling the interplay of multiple cues in prosodic focus marking," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 8, no. 1:4, pp. 1–25, 2017.

[7] M. Wang, Z. Chen, and L. Hua, "A study on the pattern of focal lengthening in Chinese," in *Proc. 12th ISSP – International Seminar on Speech Production*, virtual conference, Dec. 2020.

[8] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer," computer program, 2018. http://www.praat.org/

[9] R Core Team, "R: A language and environment for statistical computing," *R Foundation for Statistical Computing*, Vienna, Austria, 2012. http://www.R-project.org/

[10] D. M. Bates, M. Maechler, and B. Bolker, "lme4: Linear mixed-effects models using S4 classes," R package version 1.1-15, 2012.

Krake: Interactive preparation of texts for text-to-speech synthesis

Christina Tånnander
Swedish Agency for Accessible Media

We present Krake, a tool for preparing texts for text-to-speech (TTS) synthesis. Krake is developed at the Swedish Agency for Accessible Media (MTM), a government agency whose mission is to adapt texts into accessible formats such as speech and Braille, and to act as a knowledge centre for accessible reading. MTM produces more than 1 500 TTS-based talking books annually, mainly university literature in Swedish and English. During 2026, the agency will extend its use of TTS voices and adapt a wider range of literature using TTS, including popular science and fiction. This places new demands on the production pipeline.

Krake is a JavaScript GUI built on top of Sardin, a speech-oriented text processing system for Swedish [1]. After an automatic pre-analysis, the user can choose to list sentences identified as difficult to read, such as mathematical formulas, dialogue lines, or particle verbs that are commonly mis-stressed by TTS voices. The user then works interactively in Krake to edit these sentences.

In addition to traditional preprocessing of text for TTS, such as text replacements and pronunciation insertions, Krake includes functions for prosodic markup using SSML [2]. These include combinations of prosodic markup for different levels of emphasis, as well as individual features such as pitch and speaking rate. Figure 1 and Figure 2 show examples of the Krake GUI. A demonstration of Krake will be given, showcasing selected functions.

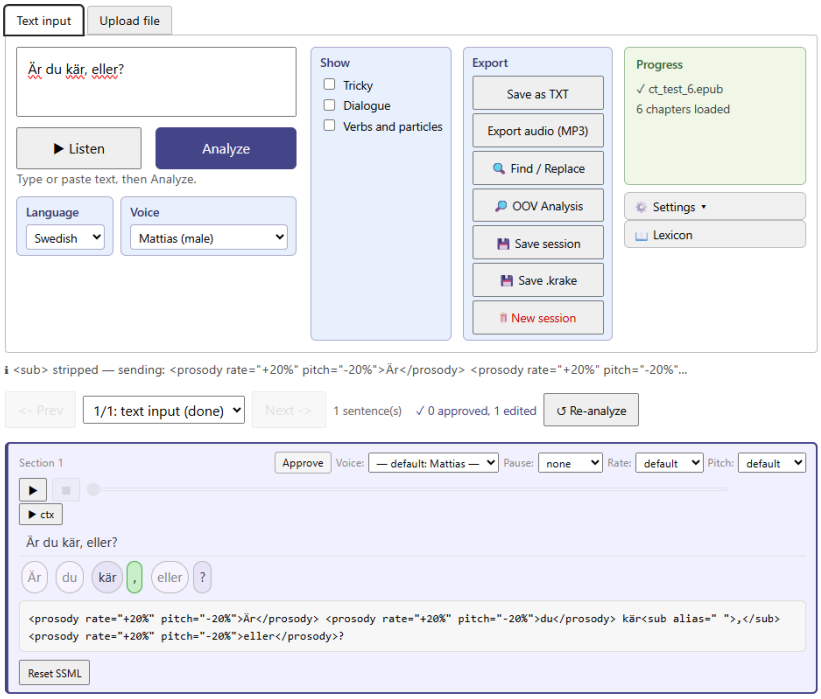


Figure 1. An example of SSML markup of a dialogue line.

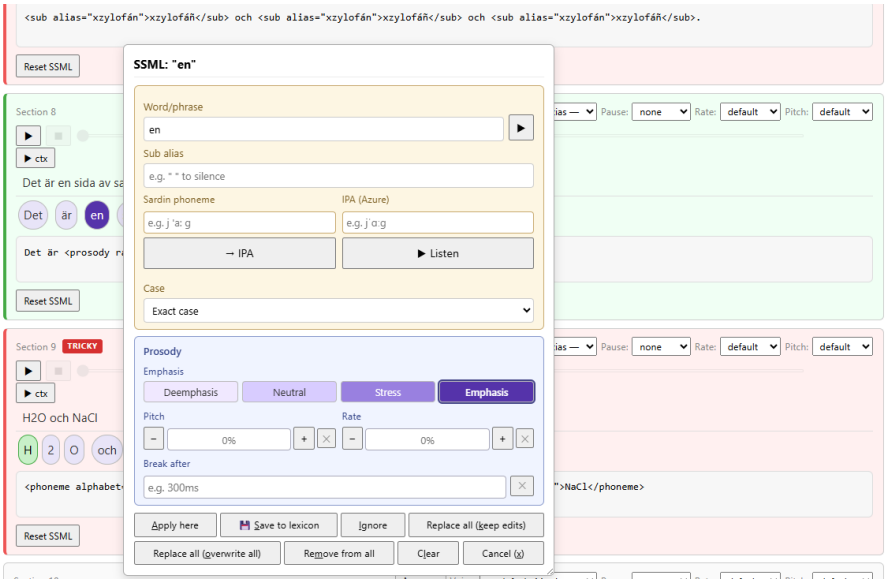


Figure 2. The text editing window.

References

[1] C. Tånnander and J. Edlund, 'Demonstration of Sardin, a Swedish Speech-Oriented Text Processing System', *DHNB*, 2025, doi: 10.5617/dhnbpub.12307.

[2] W3C, 'Speech Synthesis Markup Language (SSML) Version 1.1', W3C, Specification, Oct. 2010. Accessed: Sep. 15, 2021. [Online]. Available: https://www.w3.org/TR/speech-synthesis11/

Exploring rasterised density plots of aligned estimated pitch contours on parallel speech

Jens Edlund & Christina Tånnander

Rasterised density plots of aligned pitch contours estimated from F0 measures provide a compact means of visualising the distribution of prosodic trajectories across large collections of utterances. We have used such visualisations in earlier work (we then called them bitmap clusters), but the effects of alignment and normalisation choices on the resulting density patterns remain largely unexplored. Different temporal anchoring strategies, contour normalisation procedures, and rendering parameters may emphasise or suppress different aspects of the underlying prosodic structure, potentially affecting both interpretation and comparison across datasets.

Here, we explore alignment and normalisation choices for rasterised density plots of aligned estimated pitch contours using Storspigg-TBI, a parallel speech corpus that allows direct comparison of corresponding utterances across speakers. By varying contour alignment and normalisation procedures, we examine how these choices influence the resulting visualisations and the extent to which recurrent prosodic patterns remain stable across conditions.

The results provide methodological guidance for the use of rasterised contour density visualisations in speech and prosody research and illustrate the advantages of parallel speech data for evaluating visualisation techniques.

Effects of embodied pronunciation training on learners' perception of the Swedish length contrast

Federica Raschellà, Frida Splendido, Nadja Althaus, Marieke Hoetjes & Gilbert Ambrazaitis

Presented by Gilbert Ambrazaitis

This demo poster presents work in progress examining the effects of teachers' systematic use of gestures in pronunciation training using a pre- and post-test design (pre-test, training phase, post-test). The training phase was implemented as two instructional videos (training with vs. without gestures), which will be demonstrated during the presentation. We will also present preliminary results focusing on effects on learners' perception of the target phonological contrast, namely vowel length in Swedish.

How to collect jaw-movement data using the non-invasive MARRYS helmet

Donna Erickson (and Oliver Niebuhr)

MARRYS is a non-invasive, mobile wearable system that records jaw lowering bilaterally during speech. In speech communication, these data reveal prominence and rhythmic patterns and can support their assessment, comparison, and teaching across languages. In public speaking, jaw-opening dynamics are associated with perceived competence and passion. In clinical and therapeutic contexts, bilateral recordings may reveal movement asymmetries and help quantify and document changes, for example, after stroke or in Parkinson's disease.

Initiality accent in focus – a matter of scaling and alignment

Frank Kügler[1] *& Sara Myrberg*[2]

[1] Institute of Linguistics, Goethe University Frankfurt, Germany
[2] Centre for Languages and Literature, Lund University, Sweden

kuegler@em.uni-frankfurt.de, sara.myrberg@nordlund.lu.se

Swedish sentence accents have traditionally been analyzed as expressing focus (Bruce, 1977), which typically occurs towards the end of an intonation phrase. In addition, a sentence-level accent may be realized at the beginning of an intonation phrase; this has been termed the 'initiality accent' (Roll, 2009; Myrberg, 2010). The tonal structure of a focus or initiality accent differs across word accents, with L**H* in accent 1 words and H*L*H* in accent 2 words (Myrberg and Riad, 2015), both sharing a rise towards the final H tone. This study investigates focus in sentence-initial position, i.e. a context in which an initiality accent competes with the additional expression of focus. We present evidence from a production study with seven speakers of Central Swedish and from three perception experiments manipulating the scaling and timing of the accentual f0 peak of the H tone, and, in one experiment, the duration of the accented word. The results indicate that both the alignment and the scaling of the f0 peak play significant but independent roles in focus marking. Increased duration in focus emerges as an additional feature in production, but it is not perceptually decisive.

The speech material for the present study consists of sentences containing contrastive ellipsis, as in (1). In (1a), the subject of the matrix clause is in focus, while in (1b), the object is in focus, contrasting with the corresponding word in the elliptical clause, whereas the subject carries an initiality accent. For the production study, we created ten items as in (1) (five items with accent 1 and five items with accent 2). The scaling and alignment of the tones in the H*LH/L*H sequences on the subject, as well as the H*L sequence on the verb, were measured. In addition, the duration of the subject was measured. The results reveal different realizations for (1a) and (1b). Comparing focus realizations with initiality accent realizations, we observe significantly higher scaling, significantly later alignment of the f0 peak (H), and significantly longer word duration in focus (see Figure 1).

Three perception studies investigated whether listeners use the cues found in production to distinguish between initial focus and an initiality accent, using recordings from a female speaker of Stockholm Swedish (the second author). The first two experiments manipulated the alignment of the H tone in four equal steps and the scaling in three equal steps, starting from the mean values obtained in the production study. One set of stimuli was based on originally produced focus accents (1a), while the other was based on originally produced initiality accents (1b). The third experiment used both sets of production stimuli and manipulated the duration of the target words in three steps, ranging from the originally short duration in initiality accent realizations to medium and long durations reflecting focus-related lengthening. For focus stimuli, duration was shortened from originally long to medium and short durations, reflecting initiality accent length. Duration manipulation was carried out separately for phonologically long segments, following previous findings on focus-related duration in Swedish (Heldner and Strangert, 2001). The stimuli were fragments consisting of the subject and the verb from sentences such as those in (1). All three perception studies employed forced-choice sentence completion tasks. Listeners were presented with the two sentence fragments in (1), and looked at the reminder of the sentences, beginning after the verb and continuing to the end of the sentence. Their task was to choose the sentence completion that best matched their interpretation of the initial sentence fragment (either (1a) or (1b)). The expectation was that cues associated with focus realizations (later and higher f0, longer duration) would lead listeners to select the sentence completion correcting the subject (1a). In contrast, cues associated with initiality accent realizations (earlier and lower f0, shorter duration) were expected to lead to the selection of the sentence completion contrasting the object (1b). Across the three experiments, between 17 and 21 listeners completed the task.

Overall, the results show a bias related to stimulus origin, i.e. stimuli originally produced with an initiality accent were significantly more often rated as initiality accents (IA), whereas stimuli originally produced with focus accents were significantly more often rated as focus accents. In addition, the results reveal clear effects of both alignment and scaling in the expected direction: earlier alignment and lower scaling led to significantly more initiality accent ratings, whereas later alignment and higher scaling led to more focus ratings (Figure 2, left panel). The two effects were independent of each other. By contrast, duration had no significant effect (Figure 2, right panel). These findings suggest that production and perception align with respect to tonal alignment and scaling, but not with respect to duration.

In conclusion, the prosodic marking of grammatical structure resembles the marking of sentence accent. The variation observed in the present study indicates that the same phonological unit is used in sentence-initial position. However, the context influences whether speakers realize this unit as an initiality accent or as a focus accent. Listeners are able to distinguish between these two functions. The representation of a single tonal unit implies that it may simultaneously serve both phrasing and prominence-lending functions, similarly to what has been described for other languages such as Hungarian (Langer and Kügler, 2021) and Maltese (Grice, Vella and Bruggeman, 2019), and as proposed for Swedish by Myrberg (2010, 2013). We discuss these findings in terms of a prominence hierarchy reflecting gradient differences between types of prominence, such as corrective focus and all-new focus.

(1) **a. FOCUS ACCENT**

H*L	**H**		H*L				
[^{2}Björn-ar-na]$_{FOC}$	^{2}vakna-de	av	^{2}åska-n,	men	^{2}inte	[^{2}räv-ar-na]$_{FOC}$
bear-PL-DEF	wake up-PAST	by	thunder-DEF	but	not	fox-PL-DEF

'The bears were woken up by the thunder, but not the foxes.'

b. INITIALITY ACCENT

H*L	**H**	H*L					
^{2}Björn-ar-na	^{2}vakna-de	av	[^{2}åska-n]$_{FOC}$,	^{2}inte	av	[^{1}regn-et]$_{FOC}$
bear-PL-DEF	wake up-PAST	by	thunder-DEF	not	by	rain-DEF

'The bears were woken up by the thunder, but not by the rain.'

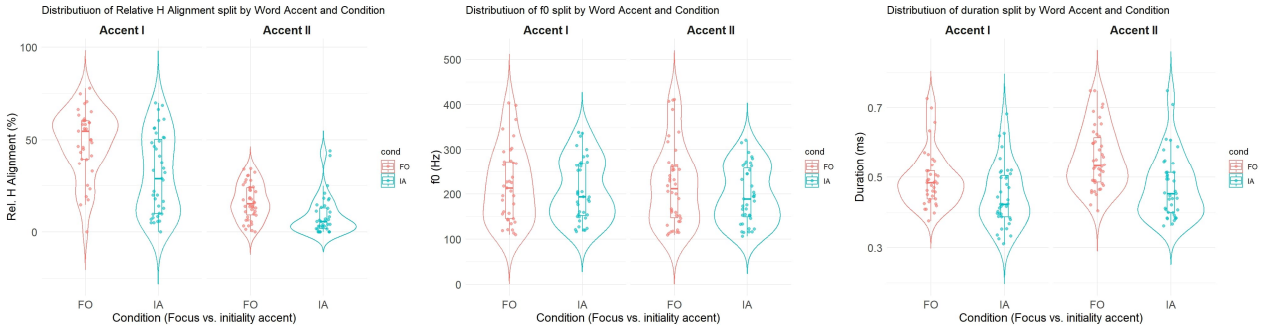


Figure 1. Results from the production study – Alignment (left), scaling (middle) duration (right) split by Accent 1 and Accent 2, and accent condition (focus = red, initiality = blue).

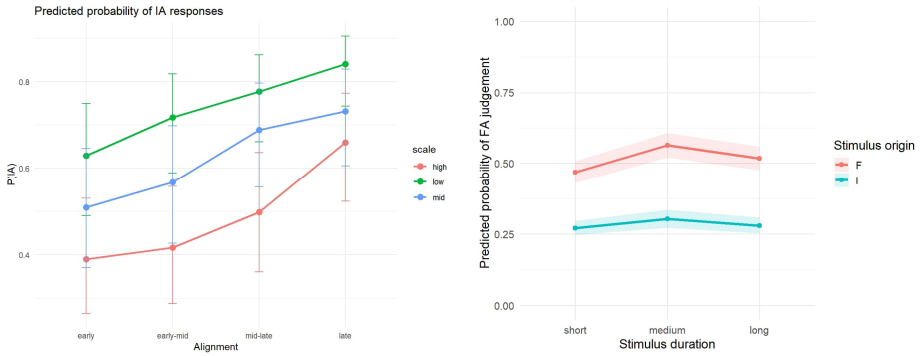


Figure 2. Results of perception studies: left – Experiment 2 with stimuli starting from focus accent production, right – Experiment 3 duration manipulation.

References

Bruce, G. (1977) *Swedish Word Accents in Sentence Perspective*. (Travaux de l'Institut de linguistique de Lund, 12). Lund: Gleerup.

Grice, M., Vella, A. and Bruggeman, A. (2019) 'Stress, pitch accent, and beyond: Intonation in Maltese questions', *Journal of Phonetics*, 76, p. 100913. doi: 10.1016/j.wocn.2019.100913

Heldner, M. and Strangert, E. (2001) 'Temporal effects of focus in Swedish', *Journal of Phonetics*, 29(3), pp. 329–361.

Langer, C. and Kügler, F. (2021) 'Focus and Prosodic Cues in Hungarian Noun Phrases', *Proceedings of 1st International Conference on Tone and Intonation, 1st International Conference on Tone and Intonation*. ISCA, Sonderborg, Denmark, 6-9 December 2021: ISCA Archive, pp. 219–223. doi: 10.21437/TAI.2021-45

Myrberg, S. (2010) *The Intonational Phonology of Stockholm Swedish*. PhD Thesis. Stockholm University. Available at: http://su.diva-portal.org/smash/get/diva2:312940/FULLTEXT02.pdf.

Myrberg, S. (2013) 'Focus type effects on focal accents and boundary tones', *Proceedings of Fonetik 2013*. Linköping: Linköping University, pp. 53–56.

Myrberg, S. and Riad, T. (2015) 'The prosodic hierarchy of Swedish', *Nordic Journal of Linguistics*, 38(2), pp. 115–147. doi: 10.1017/S0332586515000177

Roll, M. (2009) *The neurophysiology of grammatical constraints: ERP studies on the influence of prosody and pragmatics on the processing of syntax and morphology in Swedish*. PhD Thesis. Lund University.

When Context Hears the Wrong Tone:
Contextual Evidence and Decision Dynamics in Swedish Tone Accent Perception

Jinhee Kwon, Mikael Roll

Lund University

Swedish tone accents are lexically distinctive. Accent 1 and Accent 2 form prosodic minimal pairs, making them well-suited for isolating the role of prosody in lexical access. Prior work has shown that tone accents generate local predictions at the word level, reflected in early ERP components such as the Pre-activation Negativity (PrAN) and the N400 (Hjortdal et al., 2024; Kwon & Roll, 2024; Roll, 2015; Roll et al., 2015; Söderström et al., 2017). What remains less understood is how sentence-level semantic context shapes tone accent processing when prosody and meaning come into conflict. This study takes up that question by combining computational estimates of contextual constraint with EEG and behavioural measures.

Participants listened to sentences in which the critical word was a member of a prosodic minimal pair (e.g., ¹stubben 'the stubble' vs. ²stubben 'the stump'). Congruent stimuli consisted of naturally produced sentences where context and target word matched. Incongruent stimuli were created by cross-splicing the target word from one sentence into the carrier sentence of its minimal pair partner, such that the tone accent of the target word conflicted with the semantic expectation established by the context. To quantify contextual constraint independently of the behavioural data, we used a Natural Language Inference model (DeBERTa-v3 XNLI) that estimated, for each trial, how strongly the context supported each interpretation. This gave us two continuous predictors: context alignment (the degree to which the context favoured the heard word over its minimal pair partner) and context support (the absolute plausibility of the heard word given the context). EEG was recorded throughout, with N400 amplitude at the critical word serving as the main neural measure. A hierarchical Drift Diffusion Model (DDM) was then fitted to response times and accuracy from a forced-choice identification task, with trial-wise NLI estimates entered as predictors of drift rate and starting point.

Both context alignment and context support significantly predicted N400 amplitude. Greater contextual support for the heard word was associated with a reduced N400, indicating smoother semantic integration. Notably, the tone accent of the target word itself had no reliable effect on the N400 once contextual predictors were included. This pattern suggests that when context and tone accent are placed in conflict, the electrophysiological response to the incoming word is shaped more by what the listener expects than by the tonal signal they actually hear.

The DDM results told a consistent story. Context alignment was the strongest predictor of drift rate, meaning that the speed of evidence accumulation toward a lexical decision tracked

how well the context aligned with the perceived tone accent. Context prior additionally influenced the starting point of evidence accumulation, reflecting a pre-stimulus bias toward the more contextually probable interpretation. As with the ERP data, tone accent identity did not independently modulate drift rate when contextual factors were in the model.

Together, these findings suggest that semantic context boosts neural responses and guides lexical decisions toward the contextually expected tone accent — not by sharpening perception of the tonal signal itself, but by accelerating the accumulation of evidence in its favour. Tone accent perception, at least under conditions of contextual conflict, looks less like straightforward bottom-up decoding and more like a probabilistic inference process in which sentence-level meaning carries considerable weight. More broadly, the results add to a growing body of evidence that suprasegmental features are not processed in isolation but are deeply embedded in — and at times overridden by — the expectations that context generates.

Hjortdal, A., Frid, J., Novén, M., & Roll, M. (2024). Swift Prosodic Modulation of Lexical Access: Brain Potentials From Three North Germanic Language Varieties. *Journal of Speech, Language, and Hearing Research*, *67*(2), 400–414. https://doi.org/10.1044/2023_JSLHR-23-00193

Kwon, J., & Roll, M. (2024). Neural semantic effects of tone accents. *NeuroReport*, *35*(13), 868. https://doi.org/10.1097/WNR.0000000000002077

Roll, M. (2015). A neurolinguistic study of South Swedish word accents: Electrical brain potentials in nouns and verbs. *Nordic Journal of Linguistics*, *38*(2), 149–162. https://doi.org/10.1017/S0332586515000189

Roll, M., Söderström, P., Mannfolk, P., Shtyrov, Y., Johansson, M., Van Westen, D., & Horne, M. (2015). Word tones cueing morphosyntactic structure: Neuroanatomical substrates and activation time-course assessed by EEG and fMRI. *Brain and Language*, *150*, 14–21. https://doi.org/10.1016/j.bandl.2015.07.009

Söderström, P., Horne, M., & Roll, M. (2017). Stem Tones Pre-activate Suffixes in the Brain. *Journal of Psycholinguistic Research*, *46*(2), 271–280. https://doi.org/10.1007/s10936-016-9434-2

Timing or Privativity? Perceptual Evidence for the Phonological Representation of Urban East Norwegian Tonal Accents

Manuel Lipstein[1], Corinna Langer[1], and Frank Kügler[1]

[1]Institute of Linguistics, Goethe University Frankfurt, Germany
lipstein@lingua.uni-frankfurt.de, langer@lingua.uni-frankfurt.de, kuegler@em.uni-frankfurt.de

Our study investigates the phonological representation of the two tonal accents (Accent 1 and Accent 2) of Urban East Norwegian (UEN). Their phonological make-up remains a subject of debate, centred around two well-established competing hypotheses (cf. Kelly & Smiljanić, 2017, 2018; Köhnlein, 2020). On one hand, the Timing Hypothesis (Bruce, 1977 on Stockholm Swedish) proposes that both accents share the same underlying tonal sequence (i. e., a HL contour) and differ solely in the temporal alignment of the tones with the prominent, stressed syllable. From this assumption, it follows that Accent 1 shows an early f_0 fall while Accent 2 shows a late f_0 fall (Ambrazaitis & Bruce, 2006; Bruce, 1977; Kelly & Smiljanić, 2018). On the other hand, the Privativity Hypothesis (Kristoffersen, 2000 on UEN; Riad, 1998 on Stockholm Swedish) posits that only one of the two accents is tonally specified. Accordingly, Accent 2 is proposed to be marked by a lexical H tone and Accent 1 to be toneless, such that the two accents have a distinct tonal make-up (Fretheim, 1992; Fretheim & Nilsen, 1991; Gussenhoven, 2004; Kelly & Smiljanić, 2017).

Recent perception studies have provided empirical support for the Timing Hypothesis in different Scandinavian varieties, including general East Norwegian (van Dommelen & Nilsen, 2003), Trøndersk Norwegian (Kelly & Smiljanić, 2018), and South Swedish (Ambrazaitis & Bruce, 2006). For UEN, experimental results suggest that not only f_0 cues but possibly also syllable duration and intensity are relevant for perception (Nicholson & Teig, 2004). However, UEN remains under-researched regarding perceptual evidence for the two hypotheses, while it has been repeatedly described as aligning to the Privativity Hypothesis based on production data (Gussenhoven, 2004; Kristoffersen, 2006a, 2006b; Wetterlin 2010). Therefore, the present study aims to determine whether perception data aligns with a timing approach or whether it corroborates the privative analysis of UEN.

To test the two hypotheses, three online perception experiments were conducted via SoSci Survey (Leiner, 2024). The stimuli consisted of natural speech, spoken by a female UEN speaker from Oslo. The materials comprised seven sentence pairs that varied in their semantic and pragmatic contexts. Each pair was segmentally identical but contained one tonally ambiguous target word, illustrated by example (1) with the minimal pair *[1/2]bøkene* 'the books/beeches'. Across the three experiments, both f_0 alignment and f_0 height of the original target words were systematically manipulated in several manipulation steps (M1 to M5/M6, with M0 as the unmanipulated recording) using Praat (Boersma & Weenink, 2023). Participants (n = 27, 20, 29, respectively) performed a two-alternative forced-choice picture selection task after listening to each stimulus. They clicked on one of two provided pictures that corresponded to the tonal minimal pair in the stimulus. The experimental results were analysed with logistic regression models, using the *rms* package (Harrell, 2023) for R (R Core Team, 2023) in RStudio (RStudio Team, 2022), evaluating the overall fit via Nagelkerke's R^2 and the Concordance index (C). To account for instances of perfect data separation, post-hoc comparisons were conducted using estimated marginal means (*emmeans* package; Lenth & Piaskowski, 2026) derived from bias-reduced models (*brglm2* package; Kosmidis, 2026). Dunnett's adjustment was applied for within-accent contrasts, and the Bonferroni correction for across-accent differences.

(1)	*I*	*sitt*	*nye*	*kunstprosjekt*	*skjærer*	*hun*	*i*	*[1]bøkene*	/	*[2]bøkene*
	in	her	new	art.project	cuts	she	in	books.the		beeches.the
	med	*en*	*stor*	*kniv.*						
	with	a	big	knife						

'In her new art project, she cuts in the books/beeches with a big knife.'

Experiment 1 tested the Timing Hypothesis by manipulating the relative temporal alignment of the f_0 maximum and minimum contour (i. e., the H and L tones) from the recorded early Accent 1 and late Accent 2. The original contour was moved in four steps of 17 ms each from the left to the right for Accent 1 and vice versa for Accent 2 (see Figure 1). The model results (see Figure 5; $\chi^2(43) = 1297.07$, $p < .01$, $R^2 = .58$, $C = .90$) strongly support the Timing Hypothesis, since delaying the HL contour of Accent 1 stimuli led to a significant shift towards Accent 2 identification ($p < .01$ for M1 to M5), while advancing the HL contour of Accent 2 stimuli significantly led to Accent 1 categorisation ($p < .01$ for M1 to M5). This bidirectional categorical change suggests that the relative alignment of the f_0 trajectory is a primary cue for the accent contrast. However, an early salient f_0 fall might additionally enhance Accent 1 identification, while a significantly higher f_0 maximum (= max.) might constitute an additional salient cue for Accent 2 identification.

Experiment 2 evaluated the Privativity Hypothesis by isolating the role of the initial lexical H tone. In Subexperiment 2a, the f_0 contour was interpolated in four manipulation steps between the H% boundary tone of the preceding word and the lexical L tone, which was averaged between the recorded Accent 1 and Accent 2 L tones in terms of both f_0 height and alignment (see Figure 2). The results show that f_0 trajectory interpolation of recorded Accent 2 led to Accent 1 categorisation ($p < .01$ for M1 to M5), while interpolation of recorded Accent 1 did not cause a shift in discrimination ($p > .01$ for M3 to M5); with a main effect of accent ($\beta = 6.21$, $p < .01$). In Subexperiment 2b, the f_0 max. was replaced entirely with an L plateau to eliminate any potential H target, while preserving the temporal alignment of the original stimuli (see Figure 3). Manipulated items of Subexperiment 2b were consistently classified as Accent 1 for

both Accent 1 and Accent 2 originals ($p < .01$). The model results of both subexperiments (see Figure 6; $\chi^2(37) = 874.38$, $p < .01$, $R^2 = .59$, $C = .92$) indicate that the absence of a distinct f_0 max. on the stressed syllable persistently led to Accent 1 identification, while its presence led to Accent 2 identification. Thus, the findings corroborate the Privativity Hypothesis, suggesting that Accent 1 is characterised by the absence of a lexical H tone and Accent 2 by its presence.

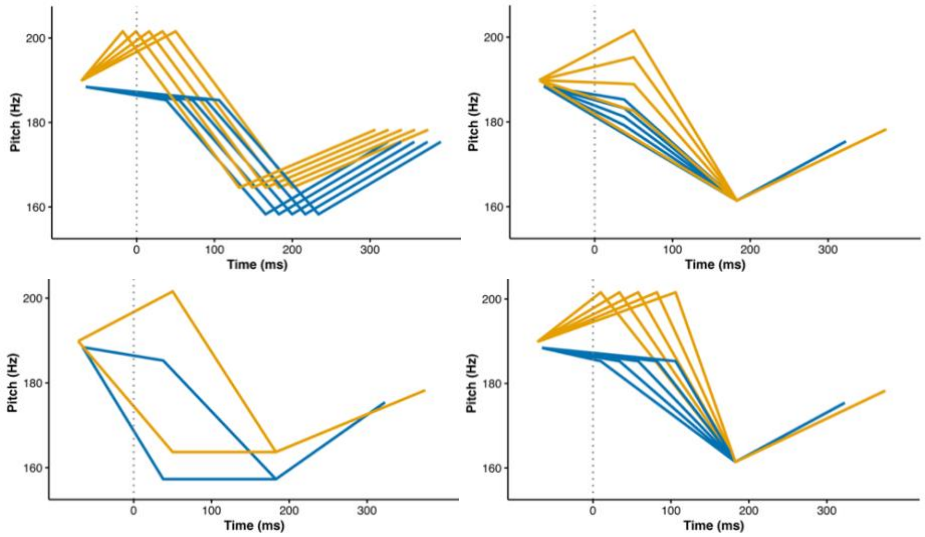


Fig. 1 (top left), Fig. 2 (top right), Fig. 3 (bottom left), and Fig. 4 (bottom right):

Manipulated stimuli for Experiments 1, 2a, 2b, and 3, respectively. Manipulation of the f_0 contour of recorded Accent 1 (blue) and Accent 2 (orange).

Experiment 3 investigated a variant of the Timing Hypothesis, namely whether the absolute timing of the f_0 max. is more decisive than the height of the f_0 max. as a perceptual cue. To test this, only the timing of the f_0 max. was manipulated in five steps of 22.5 ms each. The two original f_0 max. heights were left unaltered, while the subsequent L tone was maintained at an averaged accent-neutral f_0 height and alignment (see Figure 4). The model results (see Figure 7; $\chi^2(45) = 603.46$, $p < .01$, $R^2 = .29$, $C = .78$) reveal a complex interplay of perceptual cues, as both the height of the f_0 max. as well as its alignment or the resulting slope of the fall significantly affected identification. Primarily, f_0 height manipulation showed the strongest effects: The lower f_0 level led to Accent 1 identification and the higher f_0 level to Accent 2 identification ($\beta = 2.97$, $p < .01$). Secondarily, an earlier f_0 max. and a shallower slope of the fall enhance Accent 1 identification ($p < .01$ for M1 to M3), while a later f_0 max. and steeper slope enhance Accent 2 identification ($p < .01$ for M2 to M5). Unlike in Experiment 1, moving the f_0 max. did not cause a categorical shift; instead, the height of the f_0 max. is the decisive cue for accent discrimination, in line with Experiment 2. This supports the Privativity Hypothesis and suggests a weaker version of the Timing Hypothesis: The temporal alignment of the tones with the syllable structure does have an effect on accent identification, but this cue is not as salient as the f_0 max. height.

In conclusion, the results from the three experiments indicate that the accentual contrast of UEN involves an interplay of perceptual cues, ultimately supporting both the Timing and Privativity Hypotheses. For the Timing Hypothesis, the critical cue is the alignment of the f_0 max. (or of the high turning point, cf. Kelly & Smiljanić, 2018), where a late f_0 fall leads to Accent 2 identification (Exp. 1, 3) and an early f_0 fall leads to Accent 1 categorisation (Exp. 1, 2b, 3). For the Privativity Hypothesis, the decisive cue is the perceived presence or absence of the lexical H tone: an f_0 max. located on the stressed vowel is perceived as H and leads to Accent 2 identification (Exp. 1, 3), whereas all other instances are not perceived as H, resulting in Accent 1 identification (all Exp.). We will discuss how this theoretical tension may be resolved by a hybrid or alternative model in order to capture the empirical facts of UEN. Such a model must also account for the phonological significance of f_0 scaling: why the Accent 2 H peak prominently exceeds the H% boundary tone, and why Accent 2 exhibits higher global f_0 levels across both H and L targets. Finally, further experimental investigation into these phonological representations is essential in order to account for the typological variation of tonal systems across Scandinavian varieties.

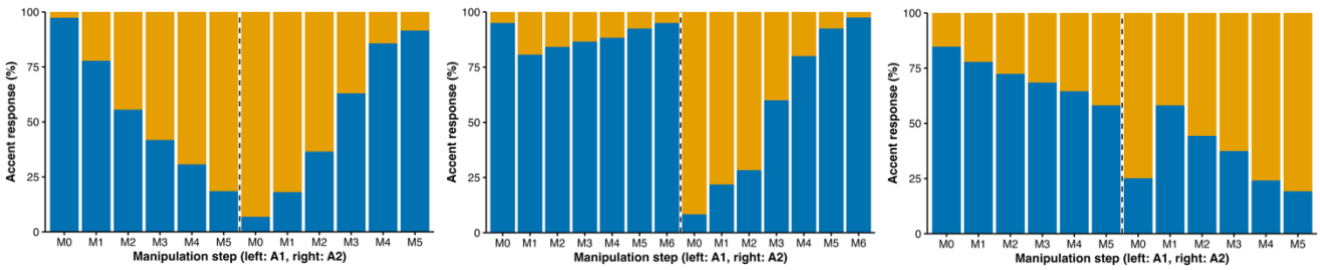


Fig. 5 (left), Fig. 6 (centre), and Fig. 7 (right): Results for Experiments 1, 2, and 3, respectively. The plots show the participants' response rate (in %) of 'Accent 1' (blue) and 'Accent 2' (orange) for the combination of original accent (A1, A2) and manipulation step (M1 to M5/M6; M0 is the unmanipulated recording).

References

Ambrazaitis, Gilbert & Gösta Bruce. 2006. Perception of South Swedish word accents. *Working Papers (Dept. of Linguistics & Phonetics, Lund University)* 52. 5–8.

Boersma, Paul & David Weenink. 2023. *Praat: Doing phonetics by computer* [Computer programme], version 6.4.01. http://www.praat.org/.

Bruce, Gösta. 1977. *Swedish word accents in sentence perspective*. Gleerup.

van Dommelen, Wim A. & Randi A. Nilsen. 2003. Toneme realization in two East Norwegian dialects. *PHONUM (Reports in Phonetics, University of Umeå)* 9. 21–24.

Fretheim, Thorstein. 1992. Themehood, rhemehood and Norwegian focus structure. *Folia Linguistica* 26. 111–150.

Fretheim, Thorstein & Randi Alice Nilsen. 1991. In defense of [+-foc]. In *ESCOL 90: Proceedings of the Seventh Eastern Conference on Linguistics*, 102–111. Ohio State University.

Gussenhoven, Carlos. 2004. *The phonology of tone and intonation* (Research Surveys in Linguistics). Cambridge University Press.

Harrell, Frank E., Jr. 2023. *Regression modeling strategies* [R package], version 6.8-0. https://hbiostat.org/R/rms/.

Kelly, Niamh E. & Rajka Smiljanić. 2017. The effect of focus and phrase position on East Norwegian lexical tone accents. *Phonetica* 74. 193–230.

Kelly, Niamh E. & Rajka Smiljanić. 2018. Perception of the lexical accent contrast in one variety of East Norwegian. *Language and Speech* 61(3). 339–357.

Köhnlein, Björn. 2020. Tone accent in North and West Germanic. In Michael T. Putnam & B. Richard Page (eds.), *The Cambridge handbook of Germanic linguistics* (Cambridge Handbooks in Language and Linguistics), 143–166. Cambridge University Press.

Kosmidis, Ioannis. 2026. *brglm2: Bias Reduction in Generalized Linear Models*. R package version 1.1.0. https://CRAN.R-project.org/package=brglm2.

Kristoffersen, Gjert. 2000. *The phonology of Norwegian* (The Phonology of the World's Languages). Oxford University Press.

Kristoffersen, Gjert. 2006a. Markedness in Urban East Norwegian tonal accent. *Nordic Journal of Linguistics* 29. 95–135.

Kristoffersen, Gjert. 2006b. Tonal melodies and tonal alignment. Unpublished manuscript. [Paper based on a talk at Nordic Prosody IX, Lund University, 2004].

Leiner, Dominik J. 2024. *SoSci Survey*, version 3.5.01. https://www.soscisurvey.de.

Lenth, Russell V. & Julia Piaskowski. 2026. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 2.0.3. https://rvlenth.github.io/emmeans/.

Nicholson, Hannele & Andreas Hilmo Teig. 2004. How to tell beans from farmers: Cues to the perception of pitch accent in whispered Norwegian. *Nordlyd* 31(2). 315–325.

Riad, Tomas. 1998. Towards a Scandinavian accent typology. In W. Kehrein & R. Wiese (eds.), *Phonology and Morphology of the Germanic Languages,* 77–112. Max Niemeyer Verlag.

R Core Team. 2023. *R: A language and environment for statistical computing*, version 4.3.1. R Foundation for Statistical Computing. https://www.R-project.org/.

RStudio Team. 2022. *RStudio: Integrated development environment for R*, version 2023.06.1+524. RStudio, PBC. http://www.rstudio.com/.

Wetterlin, Allison. 2010. *Tonal accents in Norwegian: Phonology, morphology and lexical specification*. De Gruyter.

The functional status of pitch accents in Standard Lithuanian: Phonetic interaction with other prosodic elements

Evaldas Švageris

[1]Associate Professor, Institute of Applied Linguistics, Vilnius University
evaldas.svageris@flf.vu.lt.

The functional status of pitch accents in Standard Lithuanian remains a subject of ongoing debate (Dogil & Möhler 1998; Girdenis 2014; Hualde & Riad 2014). In practice, the pitch accent distinction is clearly observable only in diphthongs containing open, low vowels. From a phonological perspective, however, a prosodic opposition cannot be restricted to long syllables of a specific phonemic structure; this suggests that the notion of pitch accent could be abandoned altogether. Nevertheless, as long as the dephonologization of these elements remains an open question, it is premature to ignore them entirely.

Additionally, there is a lack of studies examining Lithuanian pitch accents from an interactional perspective. There is no clear understanding of how phonetic cues to word stress and sentence intonation are integrated with, yet kept distinct from, these accents. While this mechanism remains unclear, we cannot be certain about the functionality of pitch accents in modern Lithuanian. It is important to recall that the distinctive correlates of prosodic elements—word stress, pitch accents, and intonation—are linked to phonation, i.e., vocal fold activity (Titze 2000; Zhang 2016; Xu 2019). Essentially, one may expect syllables with varying prosodic weights to be phonated with different levels of both effort and control. Following this logic, a new model (Švageris 2023; 2025) is proposed that treats prosodic structure as a dynamical system. In this model, different prosodic elements modify the F0 trajectory through distinct time derivatives: F0 acceleration (a) and F0 jerk (j) (Eager 2016). Sentence intonation functions as a baseline F0 acceleration—whether positive, negative, or near-zero—that determines the global direction of the tonal contour. Stress and pitch accent modulate the local temporal distribution of that acceleration (the F0 jerk). In this respect, stress differentiates syllables syntagmatically, whereas pitch accent differentiates them paradigmatically. Syllable duration serves as the physical medium through which these dynamic F0 profiles are realized. In terms of classical mechanics, intonation regulates the overall force vector (global F0 change), while stress and pitch accent regulate local impulses; thus, the final prosodic contour is the time integral of their combined effects (Švageris 2023; 2025).

To test this, a multivariate analysis of variance (MANOVA) was conducted, revealing a rigid hierarchy of prosodic elements in Standard Lithuanian. Word and phrase stress, alongside intonation, emerge as the dominant factors influencing the distribution of acoustic parameters (duration, F0 range, and F0 jerk)—effectively controlling the activation and deactivation of phonation. While pitch accents occasionally reach statistical significance, they occupy the lowest hierarchical position and exhibit the weakest effect sizes.

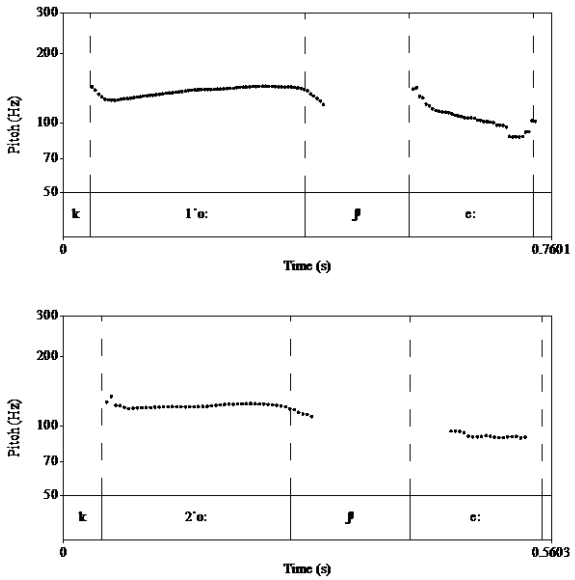


Figure 1: *F0 contours for* [1ˈko:ʃʲe:] *(above) and* [2ˈko:ʃʲe:] *(below) (declarative intonation, +focus)*

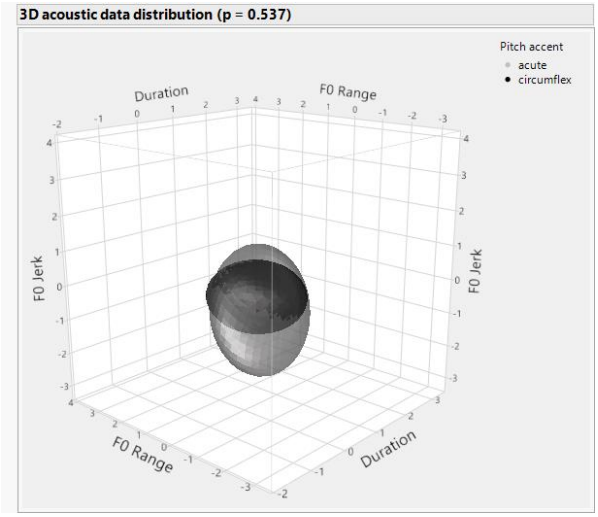


Figure 2: *Pitch accent effect on acoustic data distribution in* [1ˈko:ʃʲe:] *and* [2ˈko:ʃʲe:] *(declarative intonation, +focus)*

Consequently, the acoustic realization of pitch accents is directly contingent upon the phonetic context and their relationship with other prosodic factors. Statistically insignificant or weak interactions between syllable accents, intonation type, and focus indicate that these elements lack

universal, context-resistant properties. Some phonetic differences observed in certain diphthongs, primarily driven by inherent vowel quantity, do not correlate with independent tonal oppositions or distinct dynamic F0 profiles. This lack of robust F0 distinction points to a further diminishing of the functional weight of pitch accents in Standard Lithuanian.

References

[1] G. Dogil and G. Möhler, "Phonetic Invariance and Phonological Stability: Lithuanian Pitch Accents," *In 5th International Conference on Spoken Language Processing, Sydney, Australia*, pp. 75–86, 1998.

[2] D. Eager, 2016, "Beyond velocity and acceleration: jerk, snap, and higher derivatives," *European Journal of Physics*, vol. 37, pp. 1–11, 2016.

[3] A. Girdenis, *Theoretical Foundations of Lithuanian Phonology*. Vilnius. 2014.

[4] J. I. Hualde and T. Riad, "Word accent and intonation in Baltic," *Speech Prosody*, vol. 7, pp. 668–672, 2014.

[5] E. Švageris, "The phonetic interaction of prosodic elements in the Baltic languages: a tentative theoretical model," *Baltistica*, vol. 58, no. 2, pp. 195–224, 2023.

[6] E. Švageris, "A multifactorial analysis of phonetic word stress correlates in Standard Lithuanian: verification of the prosodic element interaction model," *Baltistica, vol* 60, no. 2, pp. 323–354, 2025.

[7] I. Titze, *Principles of Voice Production*. Iowa City: National Center for Voice and Speech. 2000

[8] Y. Xu, "Prosody, Tone and Intonation," *Handbook of Phonetics*, 314–356. New York: Routledge, 2019.

Prosodic realisation of Information Structure under language contact: a study of Livonian-Latvian bilinguals

Tuuli Tuisk (University of Tartu), Eva Liina Asu (University of Tartu), Heete Sahkai (Institute of the Estonian Language)

This paper studies prosodic expression of Information Structure (IS) in the production of Livonian-Latvian bilingual speakers. Livonian is a critically endangered indigenous language of Latvia that belongs to the Finnic branch of the Uralic language family, being closely related to Estonian. Despite the fact that Livonian is no longer used as a main language of communication, the Livonian community is active in preserving and developing Livonian heritage and the language plays a key role in this process [1]. As all Livonian speakers are bilingual, Livonian as spoken today is strongly influenced by Latvian pronunciation. Therefore, contact-induced prosodic realisation is of particular interest in this context.

The current study seeks to find out which prosodic and syntactic strategies are used by Livionian-Latvian bilingual speakers to express different types of focus in their two languages. The prosodic realisation of three focus structures (subject focus, object focus and sentence focus) in terms of prominence strength (as reflected in pitch excursion and word duration changes) and prosodic phrasing are analysed, as well as preferred word order. The results are compared to a similar study on Estonian [2].

The data were elicited using the design and materials of the "Who does what" production task of the "Questionnaire on Information Structure" [3]. The task is part of a series of tasks designed for a broad typological study of different means of expressing IS, and consists of answering questions about pictures. Informants are thus not presented with preformed sentences but are prompted to produce the sentences themselves, which permits to study not only the prosodic aspects of IS but also e.g. the syntactic ones.

For the purposes of the current study three types of focus were elicited with single-event pictures, e.g. the picture of a man drinking coke. In order to elicit different focus types the same picture was shown three times accompanied by different questions. For instance, subject focus (SF) was elicited with the question "Who is drinking coke?", object focus (OF) with the question "What is the man drinking?", and sentence focus (SEF) with the question "What is happening?".

The data was recorded from four speakers in their forties, 2 women and 2 men, who are all fluent in Livonian although their dominant language is Latvian. The speakers completed the task in both languages on the same day. The audio recordings were annotated manually using the speech analysis software Praat [4]. The focused constituents of a phrase were distinguished from the non-focused parts. Focus placement was analysed as expressed by word order, pitch, and duration of the constituent.

Preliminary findings (at the time of abstract submission) show that the preferred word order for all focus types in both Livonian and Latvian utterances was the SVO-word order. Only one speaker used the inverted constituent order (OVS) to express object focus, but did not do this consistently. This is different from Estonian where the inverted word-order optionally marks subject focus.

As to prosodic phrasing, the subject was most frequently separately phrased (i.e. followed by a pause) in the case of subject focus. The subject was similarly phrased separately in the case of object focus, whereas in such cases it was often produced with rising intonation.

It can be tentatively concluded that prosodic expression of IS in Livonian is similar to that of Estonian with respect to prosodic phrasing but there are differences in the marking of IS with word order. Results from the Latvian data are somewhat inconclusive, as there was a lot of variation.

References

[1] V. Ernštreits and G. Kļava, "Taking the Livonians into the Digital Space," *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries* (DHN 2020), 2021, pp. 26–37.
[2] H. Sahkai, M.-L. Kalvik and M. Mihkla, "Prosodic effects of information structure in Estonian," *Nordic prosody: Proceedings of the XIth conference*, Tartu, 2012, pp. 323–332.
[3] S. Skopeteas, I. Fiedler, S. Hellmuth, A. Schwarz, R. Stoel, G. Fanselow, C. Féry, and M. Krifka, "Questionnaire on Information Structure: Reference Manual," *Interdisciplinary studies on information structure*, Vol. 4, 2006. Universitätsverlag Potsdam. Retrieved from <https://publishup.uni-potsdam.de/opus4-ubp/frontdoor/index/index/docId/1145>
[4] P. Boersma and D. Weenink, *Praat: doing phonetics by computer*. Computer program [version 6.4.65], 2026.

What motivates word stress adaptation in Estonian?

Heete Sahkai, Liisi Piits, Meelis Mihkla, Hille Pajupuu, Liis Ermus
Institute of the Estonian Language

Estonian, like many other Uralic languages, has fixed initial primary word stress by default. However, unlike for instance in Finnish, non-initial primary stress is wide-spread in more recent loanwords, probably due to stronger influence from neighbouring Indo-European languages [1]; cf. also [2]. Estonian non-initial primary stress is considered to be an unstable phenomenon, as there is taken to be a general tendency for stress to shift to the initial syllable in the course of loanword adaptation, or to vary between the initial and the non-initial syllable [3], [4], [5], [6]. Recent findings suggest however that stress shift and variation are almost entirely limited to a single type of non-initially stressed loanwords, namely, trisyllabic nouns with final stress [7], despite the fact that non-initial primary stress is frequent also in other positions and word types. While stress shift in such words is clearly determined by the number of syllables in the nominative, it takes place at the level of the lemma, affecting the entire paradigm independently of the syllable count of inflected forms.

The limited domain of stress adaptation in Estonian raises the question what causes this limitation. In our talk, we will attempt to start answering this question by examining the prosody of the previously primary-stressed syllable in words that have undergone stress adaptation. We focus on a type of finally stressed trisyllabic loanwords that have given rise to two types of adaptations in terms of lexicographic representation. The examined type of finally stressed loanwords is exemplified in (1). All trisyllabic finally stressed loanwords can be taken to consist of a secondary-stressed disyllabic foot followed by a primary-stressed monosyllabic overlong (Q3) foot. The examined type of words are additionally characterised by a long vowel and a short coda consonant in the primary-stressed syllable.

(1) *meditsiin* ‘medicine’: $_{\omega}(_{Ft}(_{,}\text{me.tit})_{Ft}\ _{Ft}(\text{ˈsi::n})_{Ft})_{\omega}$

The type of non-initially stressed loanwords illustrated in (1) have given rise to two types of adaptations. The first type is exemplified in (2). In this type, the long vowel of the final syllable has shortened. Since Estonian trisyllabic words are taken to constitute a single trisyllabic foot (e.g. [8]), it can be assumed that, in addition to stress shift, these words have undergone a change in prosodic structure whereby the final Q3 foot has become an unstressed syllable, causing the two-footed structure to be replaced by a single trisyllabic foot. In the second type of adaptations, exemplified in (3), primary stress has shifted to the initial syllable, but the final syllable is still represented as having a long vowel and overlong quantity. This suggests an analysis whereby the primary and secondary stress have switched places, but the foot and syllable structure have remained unchanged.

(2) Type 1 adaptation
restoran ‘restaurant’: $_{\omega}(_{Ft}(\text{ˈres.to.rɑn})_{Ft})_{\omega}$

(3) Type 2 adaptation
marginaal ‘margin’: $_{\omega}(_{Ft}(\text{ˈmɑr.ki})_{Ft}\ _{Ft}(_{,}\text{nɑ::l})_{Ft})_{\omega}$

The existence of these two types of adaptations suggests two potential scenarios of adaptation. The first possibility is that adaptation is a complex process consisting of stress shift to the initial syllable and, simultaneously, simplification of foot structure, which takes time to be reflected in lexicographic representation. In this scenario, type 1 and type 2 adaptations would be prosodically identical, differing only in orthography and lexicographic description. This scenario could explain the limitation of stress adaptation to trisyllabic words: it is not an independent tendency per se but part of a more general process of prosodic simplification that can only take place in this context. The second possibility is that adaptation is a gradual process where stress shift is followed, at a later stage, by change in foot structure. Under this scenario, the different lexicographic representations reflect different stages of the adaptation process. In this case, the limitation of stress adaptation to trisyllabic words would remain unexplained.

In our talk, we will present a pilot corpus study that aimed to investigate which of these two scenarios is more likely. The study compared the two types of adaptations in (2) and (3) to each other and to two control conditions. The first control condition represented native trisyllabic words that constitute a single foot, as in (4). The second control condition aimed to approximate trisyllabic words ending with a secondary-stressed monosyllabic Q3 foot with a long vowel. Since this type of simplex words do not otherwise naturally occur in Estonian, we used compounds of the type illustrated in (5), headed by the first component.

(4) Control condition 1
pisarad ‘tears’: $_{\omega}(_{Ft}(\text{ˈpi.sa.rɑt})_{Ft})_{\omega}$

(5) Control condition 2
elulaad ‘lifestyle’: $_{\omega}(_{\omega}(_{Ft}(\text{ˈe.lu})_{Ft})_{\omega}\ _{\omega}(_{Ft}(\text{ˈlɑːt})_{Ft})_{\omega})_{\omega}$

Each of the four conditions contained 50 words (five different words, each represented by ten instances), extracted from the Estonian Public Broadcasting’s Radio Corpus [9].

We compared the normalised duration of the third-syllable vowel (V3, /ɑ/ in all instances) of the words in the four conditions. We assumed that the scenario of a complex adaptation involving both stress shift and simplification of foot structure would be supported by a result where the duration of V3 in both types of adaptations resembles the duration of V3 in the first control condition. The scenario of gradual adaptation in contrast was expected to be supported by a result where type 1 adapations resemble control condition 1, and type 2 adaptations resemble control condition 2.

Post hoc tests after Kruskal-Wallis test showed a significant difference between all conditions except type 2 adaptations and control condition 2, supporting the gradual adaptation scenario. However, both these conditions, unlike the remaining two conditions, showed a wide distribution of values, suggesting a need for more data.

References

[1] Karl Pajusalu 2022. Prosody. In Bakró-Nagy, Marianne, Johanna Laakso, and Elena Skribnik (eds), *The Oxford Guide to the Uralic Languages*, Oxford: Oxford University Press, 868-878.

[2] Yoonjung Kang 2010. Tutorial overview: Suprasegmental adaptation in loanwords. *Lingua*, 120(9), 2295-2310, doi:10.1016/j.lingua.2010.02.015.

[3] Mati Hint 1973. *Eesti keele sõnafonoloogia. Rõhusüsteemi fonoloogia ja morfofonoloogia põhiprobleemid* [Estonian word phonology. Phonology of the stress system and main issues in morphophonology]. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.

[4] Tiiu Erelt 2002. *Eesti keelekorraldus* [Estonian language planning]. Tallinn: Eesti Keele Sihtasutus.

[5] Tiina Paet 2023. *Võõrainese kinnistumine eesti keeles: keelekorralduslik ja leksikograafiline vaade* [Fixation of Loanwords and Foreign Words in Estonian from the Perspective of Standardization and Lexicography]. Dissertationes philolgiae estonicae Universitatis Tartuensis 51. Tartu: Tartu Ülikooli Kirjastus.

[6] Maire Raadik (ed.) 2018. *Eesti õigekeelsussõnaraamat ÕS 2018* [Estonian orthographic dictionary 2018]. Tallinn: Eesti Keele Sihtasutus.

[7] Heete Sahkai, Liisi Piits, Meelis Mihkla, Liis Ermus, Hille Pajupuu to appear. Pearõhu asukoht ja seda mõjutavad tegurid eesti keele võõrsõnades [The location of primary stress and its factors in recent loanwords in Estonian].

[8] Ilse Lehiste 1997. The structure of trisyllabic words. In: Ilse Lehiste & Jaan Ross (eds.), *Estonian prosody: Papers from a symposium*. Tallinn: Institute of the Estonian Language, 149-164.

[9] Pärtel Lippus, Tanel Alumäe, Siim Orasmaa, Katrin Tsepelina, Liina Lindström 2023. *Eesti Rahvusringhäälingu raadiosaadete korpus*. DOI: https://doi.org/10.23673/re-441

Calculating prosodic boundaries from articulatory data

Donna Erickson[1], Johan Frid[2] and Malin Svensson Lundmark[2]
[1]Haskins Laboratories, USA, [2]Lund University, Sweden,
Ericksondonna2000@gmail.com, johan.frid@humlab.lu.se, malin.svenssonlundmark@gmail.com

This paper reports on a method to calculate prosodic boundaries from articulatory data based on Fujimura's Converter Distributor (C/D) Model (Fujimura 2003, Erickson 2025). A central hypothesis of this model is that the articulatory basis of prosody is related to the amount of jaw displacement for each syllable in a spoken utterance. Previous research has shown that jaw lowering patternings reflect the prosodic prominence organization of the utterance (e.g. Erickson, Svensson Lundmark and Huang 2026) and that these patterns are language specific (e.g. Erickson and Niebuhr 2023). This paper reports on an articulatory experiment that shows that not only patterns of prominence are manifested in jaw lowering patterns but also it is the strength/amount of jaw lowering which dictates the location and size of the morpheme, syllable, word and phrase boundaries in an utterance (Fujimura 2003, Erickson 2025).

We report on simultaneous electromagnetic articulographic (EMA) and acoustic recordings of four American English speakers who replied to the question, "Did the fat cat sit with Matt?" The questions were presented on a powerpoint display to elicit broad and narrow focus answers. Since jaw displacement varies as a function of vowel height (e.g., Menezes and Erickson 2013, Williams et al. 2013), the content words in the utterance have the same low /ae/ vowel. The recordings were made using 3D EMA (Carstens AG501) at the Lund University Humanities Lab. Sensors were placed on the lip, tongue (front, mid and back referred to as TT, TM, and TB, respectively) and lower medial incisors (LI) to track mandible motion.

The articulatory analysis of the recordings was done using the software, C/D_BoundaryCalculator (created by the second author), following the procedure described in Erickson et al. (2015), Erickson (2025). The first step determines the syllable magnitude of each content word, which is represented in the model as a syllable pulse. The height of the syllable pulse is equivalent to the amount of jaw lowering for that syllable, in our case, monosyllabic word. The location of the syllable pulse, determined by the velocity of the onset and coda articulators, is located at the midpoint between the maximum velocity of the release of the onset articulator and the start of the coda articulator. For the sentence, *The fat cat sat with Matt*, the crucial articulators for the words are LL and TT for onset and coda of *fat*, TB and TT for *cat,* TT and TT for *sat*, LL and TT for *with* and LL and TT for *Matt*. The syllable pulse is then placed between the two landmark velocity points. The height of the pulse reflects the prominence value/syllable magnitude of each word. While vowel dependent variation in jaw displacement will ultimately require neutralization for broader cross vowel comparisons, the present pilot study deliberately postpones this step in order to first evaluate the viability of the proposed boundary calculation procedure'To determine location/duration of prosodic boundaries, an abstract syllable duration is calculated based on an algorithm to construct isosceles triangles around each syllable pulse such that only one set of adjacent edges of the triangles touches (Erickson et al. 2015). A sample C/D_BoundaryCalculator display of the sentence is shown in Figure 1: broad focus is on the left side and narrow focus on *CAT* on the right side. The triangles with zero duration decide the angle for the rest of the triangles in the same utterance by the speaker.

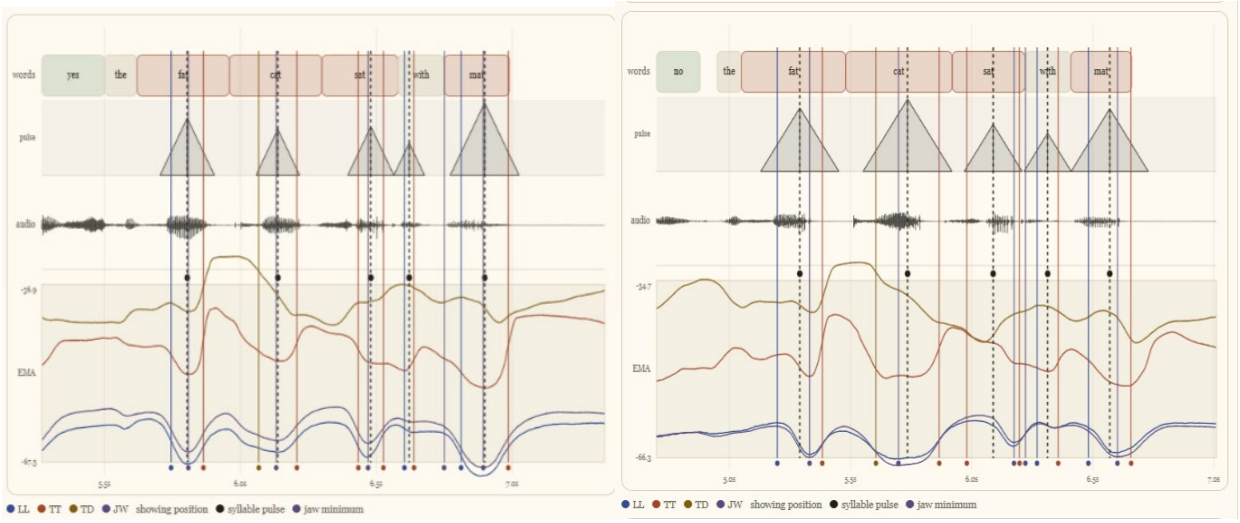


Figure 1: C/D_BoundaryCalculator display for the utterances, *the fat cat sat with Matt* (broad focus) on left, and *the fat CAT sat with Matt* (narrow focus on *CAT*) on right. The top panel shows the words, *(the)fat cat sat with Matt*, the next panel shows the syllable triangles for each monosyllabic word, the next panel, the acoustic wave form, followed by articulatory data position tracings: TD, TT, and LL, and LI (Jaw), respectively. The jaw minimum (max jaw lowering) is represented by the dotted vertical line, while the syllable pulse, located midway between the max velocity of the syllable onset and coda articulators for each word, is represented by the solid vertical line. Boundary strengths are manifested as duration gaps between syllable triangles.

As for the syllable magnitude patterns (i.e., prominence patterns), we see a pattern of s-w (strong-weak) for the broad focus utterance with the strongest prominence on *Matt*, while for the narrow focus *CAT*, we see a different prominence pattern with the largest prominence on *CAT*. Notice also a possible effect on timing of syllable pulse with maximum jaw lowering due to whether a word has narrow focus or broad focus. For the broad focus utterance (left side of Fig. 1), the syllable pulse and maximum jaw lowering tend to align for the most part. However, for the utterance with narrow focus on CAT (right side of Fig. 1), the syllable pulse occurs after the jaw minimum, while for *fat* and *mat*, the pulse occurs before the jaw minimum. This needs to be explored further.

As for the boundary strengths, these are manifested as duration gaps between syllable triangles. For the broad focus utterance, it is if the speaker placed a boundary before and after each content word, *fat*, *cat* and *Matt*, but grouped *sat with* together. For narrow focus on *CAT*, we see boundaries before and after *CAT,* but *sat with Matt* is grouped as one phrase. Informal perception assessments corroborate these phrase and prominence patterns. Preliminary investigation suggests that boundary durations increase before and after narrow focus. Statistical results will be reported in the presentation.

Fine tuning for calculating prosodic boundaries from articulatory data will ultimately require neutralization for broader cross vowel comparisons; the present pilot study, however, deliberately postpones this step in order to first evaluate the viability of the proposed boundary calculation procedure. Also, that speech rate gradually slows down especially for phrase final position needs to be examined.

This pilot study reports on an approach to explore the hypothesis that the amount of jaw lowering not only corresponds to how prominent the syllable is, (i..e, the syllable magnitude), but it also "dictates" the size and location of the prosodic boundaries. Furthermore, it suggests that timing of maximum jaw lowering varies as a function of both the type of focus (broad vs narrow) as well as whether a syllable (word) is narrowly focused. Furthermore, it suggests that timing of maximum jaw lowering varies as a function of both focus type (broad vs. narrow) and whether a syllable is narrowly focused. The newly developed C/D_BoundaryCalculator tool provides a means for systematically investigating these relationships across languages and dialects in future work, Work with different languages and dialects is needed to further test this approach for calculating prosodic boundaries from articulatory data. Work with different languages and dialects is needed to further test this approach for calculating prosodic boundaries from articulatory data.

Key words: articulation, prosodic boundaries, prominence, jaw lowering, C/D model

References

Erickson, D. (2025). The C/D Model and the Effect of Prosodic Structure on Articulation. *Languages*, . https://doi.org/10.3390/languages10120298

Erickson, D., Kim, J., Kawahara, S., Wilson, I., Menezes, C.,Suemitsu, A., and Moore, J. (2015) Bridging articulation a perception: The C/D model and contrastive emphasis. International Congress of Phonetic Sciences 2015.

Erickson, D. and Niebuhr, O. (2023). Articulation of prosody and rhythm: Some possible applications to language teaching, *Studies in Laboratory Phonology. Language Science Press (langsci-press.org),* pp.1-45. *DOI: 10.2478/9788366675728-001.*

Erickson, D., Svensson Lundmark, M., & Huang, T. (2026) Jaw opening patterns and their correspondence with syllable stress patterns. In Lars Meyer & Antje Strauss (Eds.) *Rhythms of Speech and Language*. Chapter 2.3. Rhythms of Speech and Language *Physiology, Cognition, Culture*, pp. 21 – 43*,* DOI: https://doi.org/10.1017/9781009295888.004] Cambridge University Press.

Fujimura, O. (2000). The C/D model and prosodic control of articulatory behavior. *Phonetica*, 57, 128-138.

Menezes, C., and Erickson, D. (2013) Intrinsic variations in jaw deviation in English vowels. *Proc. of International Congress of Acoustics*. *Proceedings of Meetings on Acoustics, Vol. 19,* 060253.

Williams, J.C., Erickson, D., Ozaki, Y., Suemitsu, A. Minematsu, N., Fujimura, O. (2013).Neutralizing differences in jaw displacement for English vowels, *Proc. of International Congress of Acoustics*. POMA **19**, 060268.

Speaker transition patterns in conversations with delay

Qiang Xia[1]*, Emer Gilmartin*[2]*, Marcin Włodarczak*[1]
[1]*Department of Linguistics, Stockholm University, Sweden*
[2]*Trinity College, Dublin*
[qiang.xia, marcin.wlodarczak]@ling.su.se, gilmare@tcd.ie

Conversational turn-taking patterns have been studied for decades. With the increasing prevalence of digitally mediated communication, more challenges are raised for the classic understanding of speaker transition patterns that takes face-to-face conversations as the interaction baseline. Unlike the latter, remote conversations are characterised by unavoidable latency in signal transmission, which has been found to be disruptive to conversational turn-taking [1]. Previous studies have reported that speakers in Zoom conversations tend to have longer transition times and fewer floor changes, compared with face-to-face interactions [2, 3, 4]. Although these findings demonstrate that delay influences conversational behaviours, it remains unclear to what extent the pattern of speaker transitions is affected by delays in remote communication.

At the same time, traditional approaches to analysing conversational turn-taking may not adequately capture these effects. Most of the literature defines a floor transition simply as any change between speakers, regardless of whether it is pragmatically meaningful or how short it is. This assumption overlooks the complexity of conversational turn-taking, as not every single floor transition necessarily constitutes a speaker change. For example, backchannelling frequently overlaps with ongoing speech turns but does not typically demonstrate the speaker's intention to take the floor. Treating all talk spurts, however brief, as speech turns, seems to be problematic in depicting speaker transition patterns in conversational turn-taking. Hence, several researchers have proposed a more fine-grained approach to describe the floor states and their transition patterns. They found that more than half of speaker transitions are completed in multiple rounds of floor changes rather than a single change, when using stricter criteria as indication of having the floor (e.g., a longer stretch of speech of at least one second) [5, 6, 7].

This study aims to examine if and how speakers adapt their transition patterns to delays in conversations. To address this question, we analysed spontaneous conversations under 0, 500, 1000 and 1500 ms audio delay conditions. The introduced delay was held constant throughout each of the conditions (i.e., no fluctuating delays or jitter were introduced in the routing system). Seven pairs of Swedish native speakers performed a spot-the-difference task (Diapix task [8], 10 minutes each), and seven pairs completed a Survival task (12 minutes each), under the four delay conditions, respectively. In total, 16409 speaker transitions in 10.55 hours of conversation recordings were analysed.

A similar paradigm to that in [5, 6, 7] was used to segment and label floor states. Only solo speech of at least one second in duration is seen as indicating floor possession, while short intervals of speech and silence that occur in between these stretches are concatenated into a series of intervening intervals. Floor transition can either be between-speaker (BST) or within-speaker (WST), depending on whether the solo speech at the two ends is produced by the same speaker or not. A transition can be further characterised by the number of intervening intervals needed to complete the floor transition. As shown in Figure 1, Speaker A's first and last speech chunks are longer than one second and are hence treated as instances of floor possession. The short talk spurts in between form seven intervening intervals that connect a within-speaker floor transition, with silences represented by X.

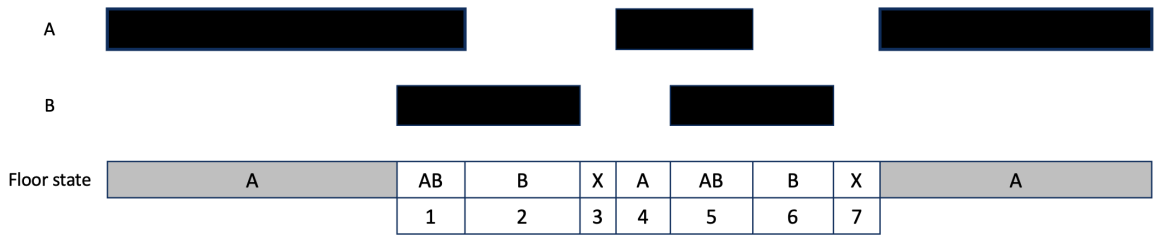


Figure 1: Example of a between-speaker transition with seven intervening intervals (A_AB_B_X_A_AB_B_X_A, where AB represents an overlap and X represents global silence). The first two rows represent speakers' utterances. The third row represents the floor state with solo-speech longer than one second marked in grey.

Our results showed that, in line with previous studies, more than half of speaker transitions (56-64%) is completed in more than one silent or overlapping interval for both tasks, while between-speaker changes with a single gap or overlap constitute about 15.6% of all speaker transitions in Diapix tasks and 10.6% in Survival tasks on average, corresponding to 38.9% of all BST in Diapix tasks and 33.9% in Survival tasks. With delay increasing, the percentage of BST completed with one intervening interval decreased (black bars in Figure 2), suggesting that the pattern of speaker changes became more complex in terms of the number of intervals it took to achieve a speaker change. In comparison, the percentage of WST with one intervening interval increased (grey bars in Figure 2), indicating that speakers tend to retain the conversational floor more when delay is present. In addition, the analysis of WSTs with three intervening intervals showed that the involvement of the other speaker exhibited a decreasing tendency when the delay increased. In other words, speakers may be backchannelling less when a longer delay is introduced (see the left panel in Figure 3). Interestingly, BST with

three intervening intervals exhibited different speaker transition patterns in the Diapix and Survival tasks. In Diapix tasks, speakers overlapped less when longer delays (1000 and 1500 ms) were introduced, while in Survival tasks, delays resulted in more overlap. The above results may be explained by the increasing difficulty in maintaining the "one party at a time" pattern [9] due to the introduced delays. Speakers therefore switched to a less interactive way of communication (e.g., less backchannelling, fewer floor transitions) [10].

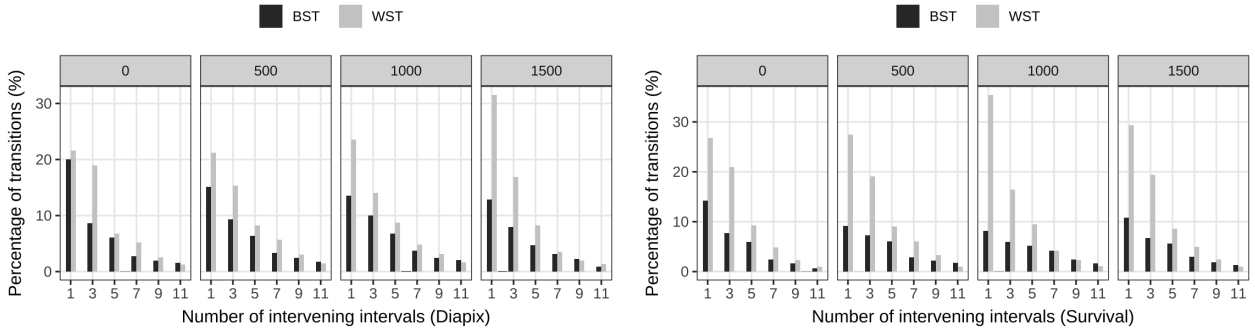


Figure 2: Frequency distribution of speaker transitions in the Diapix (left) and Survival tasks (right) under different delay conditions (in ms) depending on the number of intervening intervals.

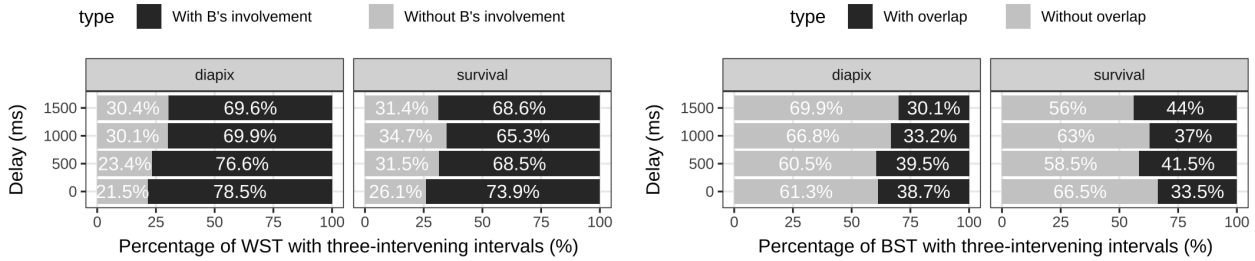


Figure 3: Percentage of WST (left) and BST (right) with three intervening intervals in the Diapix and Survival tasks under different delay conditions (in ms).

References

[1] Julie E. Boland, Pedro Fonseca, Ilana Mermelstein, and Myles Williamson. Zoom disrupts the rhythm of conversation. *Journal of Experimental Psychology: General*, 151(6):1272–1282, 2021.

[2] Ying Tian, Siyun Liu, and Jianying Wang. A corpus study on the difference of turn-taking in online audio, online video, and face-to-face conversation. *Language and Speech*, 67(3), June 2023.

[3] Qiang Xia. Does Zoom affect conversational interactivity? Turn-taking in free and task-based conversations. In Alexander Haselow and Gunther Kaltenböck, editors, *Language in Social Interaction*, pages 23–46. Walter de Gruyter, Berlin, 2025.

[4] Julie E. Boland. Turn transition time in the map task: visual cues and electronic transmission delays. *Language, Cognition and Neuroscience*, 40(5):727–742, May 2025.

[5] Emer Gilmartin. *Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis*. Thesis, Trinity College Dublin. School of Computer Science & Statistics. Discipline of Computer Science, 2021.

[6] Emer Gilmartin, Kätlin Aare, Maria O'Reilly, and Marcin Wlodarczak. Between and within speaker transitions in multiparty conversation. In *Proc. Speech Prosody 2020*, pages 799–803. ISCA, May 2020.

[7] Qiang Xia, Emer Gilmartin, and Marcin Włodarczak. Speaker transition patterns in german: A comparison between task-based and casual conversation in face-to-face and remote conversation. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Trento, Italy, September 2024.

[8] Rachel Baker and Valerie Hazan. DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770, 2011.

[9] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

[10] Deborah Tannen. Toward a Theory of Conversational Style: The Machine-Gun Question. Southwest Educational Development Laboratory, Austin, Texas, March 1980.

Investigating traces of language contact using speech embeddings

Tuukka Törö, Fionn Ó Dubhairle, Antti Suni, Juraj Šimko
University of Helsinki, Finland

Keywords: language phylogeny, language contact, speech embeddings

Past research into the Icelandic and Faroese languages have demonstrated a notable layer of vocabulary loaned from Old Irish due to movements during the Viking Age (Sigurðsson, 2022). Recent studies have also uncovered a more substantial Goidelic-associated genetic heritage among these populations than previously assumed (McKenna, 2013), as well as more higher-status Norse Gaels in these regions than traditionally believed. Given the significant degree of historic Goidelic-speaking presence in Iceland and the Faroes, combined with the demonstrable influence on the lexicons, phonetic and prosodic influence from Goidelic is also plausible.

However, historical linguistic contact research has largely been limited to lexical loans. While a universally accepted phenomenon, much of the discussion around phonetic influence has been determined by the subjective discretion of the listener, especially when the language contact in question lacks unbroken chronological continuity or dialect continua across physical space.

Recently, self-supervised speech models' (S3Ms) have broadened the scope of speech technology to thousands of new languages (Pratap et al., 2024) and their latent spaces have emerged as a novel tool to analyse language and dialect relationships (Hsieh et al., 2024; Törö et al., 2025a). Previous studies have used S3Ms to analyse how geography and language family relationships condition language and dialect level variation, typically aligning with established genealogical relations or areal phenomena (Törö et al., 2025b).

The present study examines relationships between Icelandic and Faroese with multiple geographically relevant languages in the latent space of an S3M utilizing various corpora: Common Voice, MMS for Scottish Gaelic, Swedia 2000 for Swedish, NB Tale for Norwegian, and Ravnursson corpus for Faroese.

Figure 1 presents a T-SNE plot derived from S3M embeddings for the analysed languages. It shows a continuum from Continental Scandinavian languages to Icelandic and Faroese, as well as the link between the latter and Celtic languages Irish, Scottish Gaelic, and Breton. Given the nature of the S3M embeddings, the depicted links reflect relationships among the languages along multiple dimensions including phonology, phonetic detail, phonotactics, and, presumably, prosody.

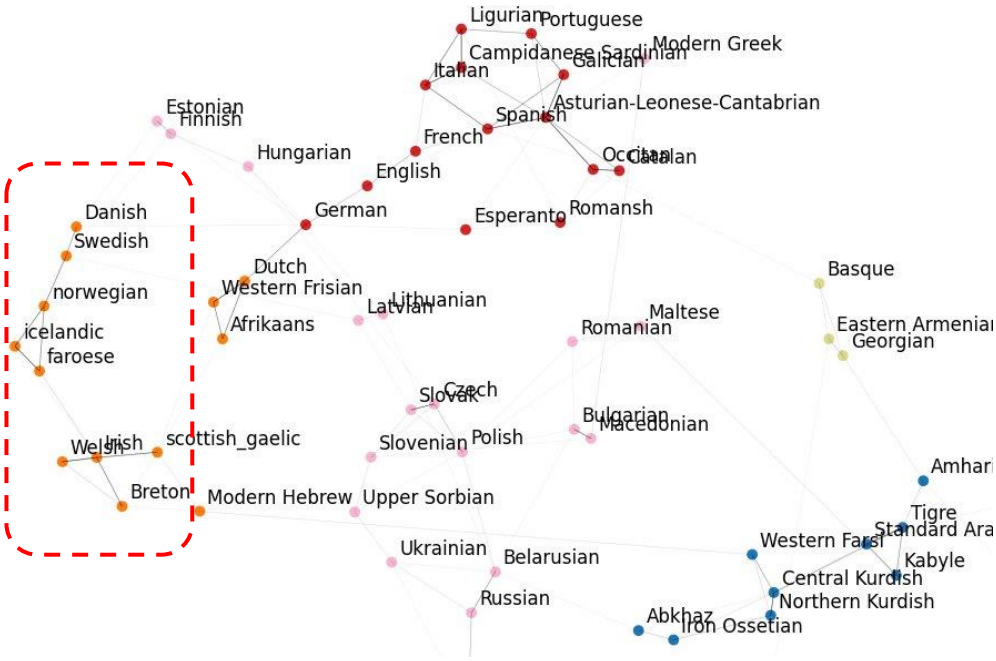


Figure 1. *T-SNE plot of Icelandic and Faroese situated in their geographic and genealogical context derived from S3M embeddings. The cluster of interest is highlighted on the left-hand side of the figure.*

In this work we investigate the degree to which the relationships captured by the S3M embeddings are influenced by prosodic phenomena present in the modern Scandinavian and Celtic languages. Ablation experiments such as signal manipulations along prosodic dimensions (e.g., flattening or exaggerating f_0 contours) and their consequences on S3M-based inter-language distances will be used as the primary methodological approach for the analyses.

References:

Hsieh, S. K., Tseng, Y. H., Lian, D. C., & Wang, C. W. (2024). Self-supervised learning for Formosan speech representation and linguistic phylogeny. Frontiers in Language Sciences, 3, 1338684.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., ... & Auli, M. (2024). Scaling speech technology to 1,000+ languages. Journal of Machine Learning Research, 25(97), 1-52.

Sigurðsson, G. (2022). Gaelic Whispers: What the Icelanders Remembered of Their Irish Past. 10.1163/9789004505339_010.

McKenna, E. (2013). The origin of continental pre-aspiration in Gàidhlig, Icelandic and Faroese: a discussion. University of Aberdeen Press.

Törö, T., Suni, A., & Šimko, J. (2025a). Neighbors and relatives: How do speech embeddings reflect linguistic connections across the world?. PLoS One, 20(8), e0330755.

Törö, T., Suni, A., Dihingia, L., Šimko, J., & Sarmah, P. (2025b). Exploring Dialects with Speech Embeddings: Insights from Two Speech Databases in Assamese and Finnish. (O-COCOSDA) (pp. 1-6). IEEE.

Predicting Word-Level Prominence in Swedish News Speech with Self-Supervised Speech Representations

Johan Frid[1], Gilbert Ambrazaitis[2], David House[3]

[1]Lund University Humanities Lab, Lund, Sweden
[2]Linneaus University, Växjö, Sweden
[3]KTH, Stockholm, Sweden
johan.frid@humlab.lu.se, gilbert.ambrazaitis@lnu.se, davidh@kth.se

Automatic prediction of prosody has a long research history using both rule-based and machine learning approaches [1,2,3]. Recent work has increasingly explored pretrained self-supervised speech models such as wav2vec 2.0 [4] as general-purpose representations for prosodic analysis. Such models have shown promising results for detecting pitch events and intonation phrase boundaries in broadcast speech [5,6], and related work has suggested that pitch-sensitive information from self-supervised models can also benefit automatic speech recognition [7]. A remaining question is whether these representations adequately encode not only local pitch structure, but also the perceptual salience of words in spoken utterances - that is: perceptual prominence (henceforth, prominence).

This question is practically relevant because modern automatic transcription systems, including models such as Whisper, can now produce useful text and word timing for many speech recordings, but they do not include prosodic cues, which in turn can encode for instance paralinguistic or pragmatic information, such as information structure. A transcript therefore captures what was said, while leaving underspecified how the utterance was delivered, and thus how the utterance is to be interpreted in the context of the actual discourse. We address this gap by investigating automatic prediction of continuous word-level prominence ratings in Swedish news speech. The task is framed as a complement to automatic speech recognition. Given spoken audio and word-level timing, the system estimates how prominent each word is.

Our data [8] consist of approximately 130 audio files, totaling about 20 minutes of speech from five speakers reading Swedish news material in a homogeneous dialect region. The target labels are mean per-word prominence ratings on a continuous 0-2 scale, aggregated from mass crowdsourcing with more than 20 raters per file. To evaluate speaker-independent generalization, models were assessed with Leave-One-Speaker-Out cross-validation.

We compare two pretrained Wav2Vec 2.0 backbones: a generic English model (facebook/wav2vec2-base-960h) and a Swedish-specific model (KBLab/wav2vec2-large-voxrex-swedish). We further compare three levels of architectural complexity. The baseline uses frozen Wav2Vec 2.0 frame embeddings with mean pooling over each word interval and log duration. The second configuration adds learned attention pooling, max pooling, and a target-dependent weighted MSE objective that upweights higher-prominence targets. The third configuration adds explicit prosodic features, including RMS mean, RMS maximum, spectral centroid, and quadratic pitch-shape coefficients describing curvature, slope and height.

Results show that the Swedish-specific VoxRex backbone consistently outperforms the generic Wav2Vec 2.0 baseline. With the bare configuration, VoxRex reaches a mean Pearson correlation of 0.7288 and MSE of 0.0394, compared with 0.6877 and 0.0438 for the generic model. Adding attention, max pooling, and weighted loss substantially improves VoxRex performance to 0.7957 correlation and 0.0331 MSE. Adding explicit pitch-shape and acoustic scalar features yields the lowest MSE (0.0311) and a correlation of 0.7987. For the generic model, explicit prosodic features are more important, improving correlation from 0.7046 in the enhanced configuration to 0.7238 with pitch-shape features. An example of the model's performance is in Figure 1.

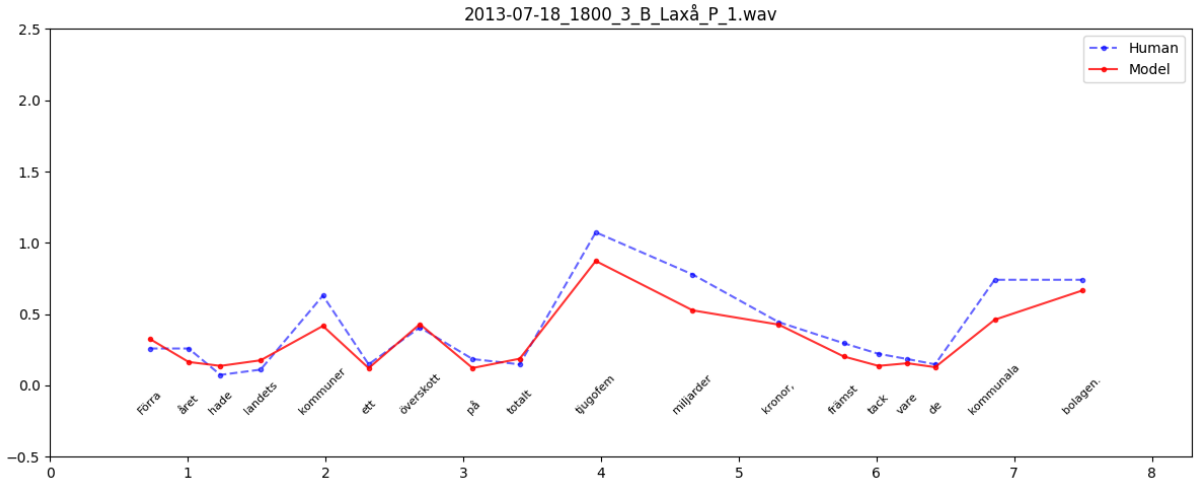


Figure 1: Word-level prominence contour for one sentence: human mean ratings (blue) and model predictions (red).

We further evaluated a binary classification variant (prominent vs. non-prominent), deriving labels via either Gaussian mixture models (GMMs) or weighted majority voting. The GMM-based labels yielded 0.85 accuracy and 0.84 macro-F1, while weighted-majority labels yielded 0.87 accuracy and 0.80 macro-F1.

These results suggest that language-specific self-supervised speech representations are particularly effective for small-data prominence prediction, plausibly because they encode aspects of Swedish prosodic structure that are less available to a generic English model. Explicit pitch-shape features contribute to the language-independent case, which has implications for languages for which no language-specific model exists. Furthermore, the results appear to be valid across multiple speakers. While not directly comparable, earlier work on Swedish read speech [3] reported syllable-level accuracy of 0.86, for a single speaker.

We conclude that word-level prominence prediction can provide a useful prosodic layer on top of automatic transcription. In this view, systems such as Whisper supply text and timing, while prominence prediction supplies a complementary estimate of emphasis. This enables enriched transcripts that better represent not only lexical content, but also the prosodic organization of spoken discourse. Such estimates of prominence structure might be useful also for higher-level automatic text analysis, such as content analysis.

References

[1] Rosenberg A & Hirschberg J (2009) Detecting pitch accents at the word, syllable and vowel level. *Proc. HLT-NAACL 2009*, pp. 81-84
[2] House D, Bruce G, Eriksson L & Lacerda F (1988) Recognition of prosodic categories in Swedish: Rule implementation. *Working Papers 33* (Dept. of Linguistics, Lund University), pp. 153-161
[3] Al Moubayed S & Beskow J (2010) Prominence Detection in Swedish Using Syllable Correlates. *Proc. Interspeech 2010*, pp. 1784-1787
[4] Baevski A, Zhou H, Mohamed A & Auli M (2020) wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proc. NIPS 2020*, pp. 12449-12460
[5] Kunešová M & Řezáčková M (2022) Detection of prosodic boundaries in speech using wav2vec 2.0. *International Conference on Text, Speech, and Dialogue 2022*, pp. 377–388
[6] Zhai W & Hasegawa-Johnson M (2023) Wav2ToBI: a new approach to automatic ToBI transcription. *Proc. Interspeech 2023*, 2748-2752
[7] Sasu D & Schluter N (2025) Pitch Accent Detection improves Pretrained Automatic Speech Recognition. *Proc. Interspeech 2025*, 4733-4737
[8] Ambrazaitis G, Frid J & House D (2025). Measuring the effects of pitch accent realization and the patterning of head and eyebrow gestures on perceived multimodal prominence. *Proc. ISGS 2025*

Speech Timing and Heteroclinic Networks

Michael O'Dell[1]

[1]Department of Digital Humanities, University of Helsinki, Finland

michael.odell@helsinki.fi

Abstract

For many years dynamical systems theory has played an important role in the investigation of phonetic processes and speech timing. In recent decades there has been a great deal of research in dynamical systems theory concerning so-called (noisy, stable) *heteroclinic networks* and their close relatives, *excitable network attractors* (see [1] for an excellent short review). Basically, a heteroclinic network is a dynamical system with a collection of *metastable* states (attracting in some directions and repelling in others) connected by trajectories in phase space. These systems are also called *winnerless competition* models [2]—in a sense, "soft" categories (neighborhoods of the metastable states) are in competition, but any "win" can only be temporary.

Such systems have several advantages [3, 4], not least of which is the fact that they are capable of (re)producing different sequences of behavior based on small changes in their parameters. That is, they have the ability to combine *reproducibility and flexibility of transient behavior* [5], which is one reason heteroclinic network theory has proved useful in robotics, for example [6]. In addition, they bridge the gap between between classical discrete computation models of sequential categories and continuous-time dynamical systems [7, 8]: "Such network attractors ... lend themselves to modelling discrete-state computations with continuous inputs, and can sometimes be thought of as a hybrid model between classical discrete computation and continuous-time dynamical systems." [1] Since they are dynamical systems, heteroclinic networks are also well suited for investigating the timing of observed behavior [9, 10, 11, 12]. In this respect, they are also more flexible than simple Hidden Markov Models (HMMs) because heteroclinic networks can exhibit non-Markovian behavior [13, 14]. In spite of these advantages, the theory of heteroclinic networks has not yet found wide application in phonetics (but see, e.g. [15]).

In phonetics, heteroclinic networks may help in investigating the connection between phonotactics (syntagmatic and paradigmatic structure) and the timing of roughly sequential categories (cf. Articulatory Phonology). For instance, in recent years *hierarchical* heteroclinic networks have often been the object of investigation [16, 17, 18, 19, 20], because so many types of behavior are hierarchically organized, and of course speech is no exception. There now exist several general algorithms for constructing an explicit system of differential equations (containing a network attractor) corresponding to any given finite state diagram (directed graph) [21, 22, 8, 23]. Several methods have also been devised for learning network structure based on data [18, 24]. These tools can be used to investigate the impact of network architecture on timing and durational aspects of speech. Finnish, with its extensive phonological use of quantity, offers a rich source of data for such investigation.

At the conference, a very brief introduction to heteroclinic network theory will be presented, as well as results and discussion of preliminary investigations into applying the theory to speech timing, focusing on quantity in Finnish.

Index Terms: dynamical systems, network attractors, speech timing, Finnish quantity

1. 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] V. S. Afraimovich, M. I. Rabinovich, and P. Varona, "Heteroclinic contours in neural ensembles and the winnerless competition principle," *International Journal of Bifurcation and Chaos*, vol. 14, no. 4, 2004.

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

[4] C. M. Postlethwaite, P. Ashwin, and M. Egbert, "A continuous time dynamical Turing machine," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 4, pp. 6503–6515, 2025.

[5] 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, p. e1000072, 2008.

[6] N. A. Rouse, A. D. Horchler, H. J. Chiel, and K. A. Daltorio, "Stable heteroclinic channels as a decision-making model: Overcoming low signal-to-noise ratio with mutual inhibition," *Bioinspiration & Biomimetics*, vol. 20, no. 036004, 2025.

[7] Y. Bakhtin, "Noisy heteroclinic networks," *Probability Theory and Related Fields*, vol. 150, no. 1–2, pp. 1–42, 2011.

[8] 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.

[9] 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.

[10] M. Aravind and H. Meyer-Ortmanns, "On relaxation times of heteroclinic dynamics," *Chaos*, vol. 33, no. 10, 2023.

[11] R. Latorre, P. Varona, and M. I. Rabinovich, "Rhythmic control of oscillatory sequential dynamics in heteroclinic motifs," *Neurocomputing*, vol. 331, pp. 108–120, 2019.

[12] M. Voit and H. Meyer-Ortmanns, "Predicting the separation of time scales in a heteroclininc network," *Applied Mathematics and Nonlinear Sciences*, vol. 4, no. 2, pp. 279–288, 2019.

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

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

[15] I. B. Yildiz, K. von Kriegstein, and S. J. Kiebel, "From birdsong to human speech recognition: Bayesian inference on a hierarchy of nonlinear dynamical systems," *PLoS Computational Biology*, vol. 9, no. 9, pp. e1 003 219 (1–16), 2013.

[16] V. S. Afraimovich, T. R. Young, and M. I. Rabinovich, "Hierarchical heteroclinics in dynamical model of cognitive processes: Chunking," *International Journal of Bifurcation and Chaos*, vol. 24, no. 10, p. 1450132, 2014.

[17] M. I. Rabinovich, P. Varona, I. Tristan, and V. S. Afraimovich, "Chunking dynamics: Heteroclinics in mind," *Frontiers in Computational Neuroscience*, vol. 8, no. 22, pp. 1–10, 2014.

[18] J. Fonollosa, E. Neftci, and M. I. Rabinovich, "Learning of chunking sequences in cognition and behavior," *PLoS Computational Biology*, vol. 11, no. 11, 2015.

[19] M. Voit and H. Meyer-Ortmanns, "A hierarchical heteroclinic network: Controlling the time evolution along its paths," *The European Physical Journal Special Topics*, vol. 227, pp. 1101–1115, 2018.

[20] ——, "Dynamics of nested, self-similar winnerless competition in time and space," *Physical Review Research*, vol. 1, no. 023008, 2019.

[21] P. Ashwin and C. Postlethwaite, "On designing heteroclinic networks from graphs," *Physica D: Nonlinear Phenomena*, vol. 265, pp. 26–39, 2013.

[22] ——, "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.

[23] M. Field, "Heteroclinic networks in homogeneous and heterogeneous identical cell systems," *Journal of Nonlinear Science*, vol. 25, no. 3, pp. 779–813, 2015.

[24] M. Voit and H. Meyer-Ortmanns, "Dynamical inference of simple heteroclinic networks," *Frontiers in Applied Mathematics and Statistics*, vol. 5, no. 63, 2019.

Prosodic Dynamics in Finnish-Speaking Autistic Preadolescents

Ida-Lotta Myllylä *(Department of Languages, University of Helsinki)*
Martti Vainio *(Department of Digital Humanities, University of Helsinki)*
Mari Wiklund *(Department of Languages, University of Helsinki)*

Speech prosody in autism is frequently characterized by distinctive acoustic features (see e.g., Fusaroli et al., 2017, 2022), yet these are often interpreted through an individual-deficit lens rather than an interactional or relational framework. Grounded in a neurodiversity-affirming standpoint (see e.g., Pellicano & den Houting, 2022) and the Double Empathy problem (Milton, 2012; Mitchell et al., 2021), this presentation synthesizes findings from two complementary studies investigating the production, interactional deployment, and listener perception of prosody in Finnish-speaking autistic preadolescents and matched non-autistic comparison speakers.

The first study (Myllylä et al., upcoming) examines conversational dynamics, focusing on within-turn silent pauses and pitch variability during spontaneous multi-party group interaction. Quantitative acoustic-prosodic analysis reveals that while autistic speakers produce substantially longer and durationally more variable within-turn silent pauses, their overall pause rates and syntactic placement align symmetrically with comparison peers. Furthermore, the autistic speakers exhibit greater fundamental frequency (f_0) variability and a significantly broader inventory of pause-adjacent non-lexical vocal events, such as glottal stops, breath cues, and click-like features, that can potentially serve as floor-monitoring and turn-holding cues within their local conversational ecologies.

The second study (Myllylä et al., 2026) pivots to a perceptual framework, investigating how prosodic typicality in autism is perceived by non-autistic listeners and how these judgments map onto objective acoustic parameters. Utilizing low-pass filtered speech to isolate suprasegmental features, 53 listeners rated 150 utterances for prosodic typicality. Mixed-effects modelling revealed no significant main effect of diagnostic group on typicality ratings. Instead, perceived typicality was driven by an interaction between specific idiosyncratic acoustic configurations and listener-specific expectations, pointing towards the notion that prosodic typicality is not a straightforward marker of neurodivergence.

Taken together, these studies illustrate that prosody in autism constitutes a cohesive, highly functional communicative style. When production is contextualized within supportive participation frameworks, distinct prosodic profiles function effectively without causing interactional breakdown. Concurrently, perceptual outcomes underscore that communicative friction arises from a bidirectional mismatch in inferential calibration rather than a unilateral speaker deficit. By linking

interactive production with experimental perception, this work broadens cross-linguistic understandings of prosodic diversity in autism.

References

Fusaroli, R., Grossman, R., Bilenberg, N., Cantio, C., Jepsen, J. R. M., & Weed, E. (2022). 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

Milton, D. E. M. (2012). On the ontological status of autism: The "double empathy problem." *Disability and Society*, *27*(6), 883–887. https://doi.org/10.1080/09687599.2012.710008

Mitchell, P., Sheppard, E., & Cassidy, S. (2021). Autism and the double empathy problem: Implications for development and mental health. *British Journal of Developmental Psychology*, *39*(1), 1–18. https://doi.org/10.1111/bjdp.12350

Myllylä, I.-L., Wiklund, M., Haakana, V., Vainio, L., Saalasti, S., & Vainio, M. (2026). Perceived prosodic typicality in Finnish preadolescents: Findings from speakers on the autism spectrum. *Clinical Linguistics & Phonetics*, 1–26. https://doi.org/10.1080/02699206.2026.2642726

Myllylä, I.-L., Vainio, M & Wiklund, M., (upcoming) Exploring silent pauses and turn-holding in Finnish-speaking preadolescents on the autism spectrum

Pellicano, E., & den Houting, J. (2022). Annual Research Review: Shifting from 'normal science' to neurodiversity in autism science. In *Journal of Child Psychology and Psychiatry and Allied Disciplines* (Vol. 63, Number 4, pp. 381–396). John Wiley and Sons Inc. https://doi.org/10.1111/jcpp.13534

Speech Tasks and Prosodic Features in Early Parkinson's Disease: Insights from Automatic Speech Analysis

Anna Slocinska[a], Farhad Javanmardi[b], Paavo Alku[b], Brent Wilson[a], Nelly Penttilä[a]

a. *Faculty of Social Sciences, Tampere University, Tampere, Finland*
b. *Department of Information and Communications Engineering, Aalto University, Espoo, Finland*

Speech impairments in Parkinson's disease (PD) often affects prosody, resulting in reduced pitch, altered intonation and stress patterns, monoloudness and monotonous voice. Impaired prosody leads to disruptions in interactions among people, since it becomes difficult or impossible to perceive emotions and intentions of a Parkinson's individual. Detecting these subtle prosodic changes in early-stage PD remains challenging, and relatively little is known about which speech tasks best reveal such impairments and could be used in speech therapy.

This study examined how speech tasks with different prosodic demands contribute to the automatic discrimination between speakers with early-stage PD and healthy controls. The dataset consisted of recordings from 49 speakers (24 PD speakers and 25 healthy controls) performing multiple speaking tasks, including sustained phonation, word repetition, sentence repetition, normal reading, prosodic reading, diadochokinesia, and semi-spontaneous speech. Automatic speech recognition experiments were conducted separately for each task using several acoustic representations, including MFCCs, mel spectrograms, and transformer-based wav2vec2 features.

The results demonstrated clear task-dependent differences. Tasks involving connected speech and expressive variation generally resulted in higher accuracy than simple phonatory tasks. In particular, prosodic reading showed strong results with wav2vec2-based representations, suggesting that tasks involving prosodic reading may be especially sensitive to PD related speech impairment. Sentence repetition also showed promising performance, outperforming word repetition in several conditions. Prosodic reading consistently demonstrated strong performance and outperformed or closely matched other speech tasks in detecting subtle prosodic impairments associated with PD.

The findings suggest that prosodically rich speech tasks may better reveal subtle speech changes associated with early-stage PD, particularly reduced prosodic flexibility. From a clinical perspective, prosodic reading tasks may provide valuable information about early-stage PD impairments and could complement more traditional assessment methods.

References

1. Jankovic, J. (2008). Parkinson's disease: Clinical features and diagnosis. Journal of Neurology, Neurosurgery, and Psychiatry, 79(4), 368–376. https://doi.org/10.1136/jnnp.2007.131045
2. Järvinen, K. (2017). Voice characteristics in speaking a foreign language. A study of voice in Finnish and English as L1 and L2. Academic dissertation. Tampere University. 978-952-03-0424-9
3. Javanmardi, F., Kadiri, S.R, Alku, P. (2024) Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, vol. 158, article no. 103047. https://doi.org/10.1016/j.specom.2024.103047
4. Liu, Y., Reddy, M. K., Penttilä, N., Ihalainen, T., Alku, P., & Räsänen, O. (2023). Automatic Assessment of Parkinson's Disease Using Speech Representations of Phonation and Articulation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 242–255.https://doi.org/10.1109/TASLP.2022.3212829
5. Miller, N. (2017). Communication changes in Parkinson's disease. Practical Neurology, 17(4), 266–274. https://doi.org/10.1136/practneurol-2017-001635
6. Moro-Velazquez, L., Gomez-Garcia, J. A., Arias-Londoño, J. D., Dehak, N., & Godino-Llorente, J. I. (2021). Advances in Parkinson's Disease detection and assessment using voice and speech: A review of the articulatory and phonatory aspects. Biomedical Signal Processing and Control, 66, 102418. https://doi.org/10.1016/j.bspc.2021.102418
7. Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., & Nöth, E. (2016). Automatic detection of Parkinson's disease in running speech spoken in three different languages. The Journal of the Acoustical Society of America, 139(1), 481–500 https://doi.org/10.1121/1.4939739
8. Penttilä, N., Tavi, L., Hyppönen, M., Rontu, K., Rantala, L., & Werner, S. (2022). Prosodic features in Finnish-speaking adults with Parkinson´s disease. Clinical Linguistics & Phonetics, 1–16. https://doi.org/10.1080/02699206.2022.2081612
9. Sapir, S. Multiple factors are involved in the dysarthria associated with Parkinson's disease: a review with implications for clinical practice and research. (2014) J. Speech Lang. Hear Res. 57, 1330–1343. https://doi.org/10.1044/2014_JSLHR-S-13-0039