Differences in acoustically determined sentence stress between native and L2 speakers of Finland Swedish

Heini Kallio[1], Mikko Kuronen[1], Maria Kautonen[2]
[1]Department of Language and Communication Studies, University of Jyväskylä
[2]School of Applied Educational Science and Teacher Education, University of Eastern Finland

Abstract

This study describes a pilot attempt to use acoustically determined sentence stress in distinguishing native and L2 speakers of Finland Swedish at different proficiency levels. The study is part of the DigiTala project that examines and develops automatic tools for spoken L2 assessment.

Stressed syllables were detected in 235 L2 and 30 L1 speech samples using combinations of f0, intensity, and duration. Differences were found in the relative number of stressed syllables between native and L2 speakers using f0 and duration or all three features. Further research with more consistent speech data is needed to establish reliable automatic measures of sentence stress.

Introduction

The production of word and sentence stress is important for intelligibility, comprehensibility, and fluency of L2 speech (Munro, 1995; Wennerström, 2000; Hahn, 2004; Kormos & Dénes, 2004; Trofimovich & Baker, 2006; Heinonen, 2020). Stress features are mainly studied in L2 English and in some cases also integrated in automatic assessment systems of L2 English (e.g., Hsieh et al. 2020). In other languages than English, however, L2 stress is much less studied and seldom integrated in automatic assessment systems (see, however, Wik 2011 on Swedish). This study is a pilot attempt to use acoustic stress measures in distinguishing native and L2 speakers of Finland Swedish at different language proficiency levels. The study is part of the DigiTala[2] project that aims to develop automatic tools for assessing spoken language skills in large-scale, high-stakes contexts (Kautonen & von Zansen, 2020).

Stress features of Finland Swedish

In stress languages, one syllable in a word usually stands out acoustically and perceptually. Moreover, one or more syllables in an utterance stand out as more prominent. Acoustically, these stress-bearing syllables are characterized by an increase in f0, duration, and/or intensity (Lieberman, 1967; Lehiste, 1969; Fant & Kruckenberg, 1994). These parameters combine in a complex and language-dependent manner.

Standard pronunciation of Finland Swedish (FS) differs from Central Standard Swedish (CSS) regarding both word and sentence stress (Tevajärvi, 1982; Vihanta et al., 1990; Hirst & DiCristo, 1998). In CSS, duration and intensity contribute strongly to the production of word stress, while f0 serves as the main cue for lexical pitch accents and primary stress (Engstrand & Krull, 1994; Bruce, 2005). The acute and grave lexical pitch accents that are characteristic for CSS, are absent in FS (Ivars, 2015). This can cause differences in particular in the f0 patterns between the two varieties, realizing in both word and sentence level. For example, in rising contours (e.g., focal peaks of both lexical pitch accents in CSS), the f0 peak is sometimes delayed or spread also to the following syllable (Vihanta et al., 1990; Xu, 1999). In FS, the timing of f0 movements seems to be more constant than in the varieties of Swedish spoken in Sweden (Tevajärvi, 1982; Bruce, 2005). FS speakers are also perceived to produce stressed syllables more often and have weaker stress contrasts than CSS speakers (Vihanta et al., 1990).

Prosody of FS is believed to be affected by Finnish (Helgason et al., 2013) and has therefore

[2] The project is financed by the Academy of Finland 2019–2023 (grant number 3229625).

some similarities with Finnish (Kuronen & Leinonen, 2011), but the linguistic properties of Swedish still define the stress structure of FS. Finnish has, e.g., a fixed word stress while the placement of stress varies in Swedish, which can cause difficulties for the Finnish learners of Swedish.

Assessment of L2 stress features

Language learners can face many difficulties in producing stress in an L2, from word-level to sentence level. Non-native speech often contains more disfluency phenomena than native speech (Cucchiarini et al., 2010), which can result in unintentional and inappropriate stressing of syllables. The stress features of L2 learners' native language can also affect both the placement (Altmann, 2006) and the use of acoustic correlates of stress (Kallio et al., 2020; 2021).

L2 learners tend to produce stressed syllables either too frequently (Wennerström, 2000) or too seldom (Kormos & Dénes, 2004). This tendency is found also in Finnish learners of Swedish (Heinonen & Kautonen, 2020). Heinonen and Kautonen (2020) analyzed the sentence stress of Finnish learners of Swedish based on raters' descriptions in pronunciation assessment. The sentence stresses with the lowest ratings were most often described as having too many or too few stressed syllables. Other comments concerned the placement as well as the manner of stress.

Previous studies have found links between stress production and many facets of speaking proficiency, such as intelligibility (Field, 2012; Abelin & Thorén, 2015), comprehensibility (Heinonen, 2020), fluency (Kormos & Dénes, 2004), and proficiency (Kang, 2018; Kallio et al., 2020). The L2 production of Finland Swedish stress has mainly been studied from read speech (Heinonen, 2020; Kallio et al., 2020). Automatic L2 stress detection systems, in turn, focus mainly on word stress (Tepperman & Narayanan, 2005; Ferrer et al., 2015; Yarra, Deshmukh, & Ghosh, 2017). We detect stressed syllables within utterances from short, spontaneous speech samples.

Materials and method

Speech data

The speech samples for this study were taken from a larger speech corpus collected while piloting a computer-aided spoken language test for Swedish as a second language (Karhila et al., 2016). The pilot test was taken by Finnish upper secondary school students (aged 16–17 years) who had studied Swedish as a compulsory subject for 4–7 years. Native Finland Swedish speakers of the same age also participated in the pilot test to obtain reference data for analysis.

The material used in the current study consists of spontaneous narrative speech elicited with several test tasks. In each of the tasks, the speakers had 30 seconds to react to the provided stimulus.

The speech samples were assessed by four expert raters using a holistic six-point scale for overall oral proficiency (A1–C2) and four analytic three-point scales including criteria for fluency, pronunciation, grammar, and vocabulary (Hildén et al., submitted). This study focuses on comparing samples between different proficiency levels.

For the current study the speech data was grouped into four proficiency categories: natives (N=30), B-level samples (N=42), A-level samples (N=163), and below A-level samples (N=30). In total, the current speech data includes 235 L2 samples and 30 samples from native speakers of Finland Swedish.

Analysis

The data was manually annotated to syllable level using the maximum onset principle with the restrictions of Swedish phonotactics (Bruce, 2012: 30–32). A Praat script was used for obtaining f0 and intensity maxima as well as duration of each syllable. All acoustic feature values were then normalized by making them proportional to the mean values of the respective features within a sample using a simple formula *Fsyl / Favg * 100*, where *Fsyl* = syllable feature value and *Favg* = average feature value. Normalizing the values this way is a common procedure that eliminates bias that can arise from, e.g., speaker-specific f0 range and articulation rate (see., e.g., Tepperman & Narayanan, 2005; Ferrer et al., 2015).

Since most stress detection systems focus on word-level, their purpose is usually to pick one syllable with the highest acoustic feature values indicating prominence. However, our goal was to detect all stressed syllables within speech samples of varying length. Since this study does not use machine learning methods, we decided to use an arbitrary threshold in detecting stressed syllables: we categorized syllables as stressed,

when their normalized feature values (*max f0, max intensity*, and *duration*) fall within the highest quartile in that sample (values higher than 75% of the values within sample). We detected syllables this way using the normalized acoustic features separately as well as different combinations of these features (all feature sets are presented in Table 1), and then computed the relative amount of acoustically stressed syllables per sample by dividing the number of stressed syllables with the total number of syllables per sample. The proportion of stressed syllables were then compared between speaker groups (natives, B-level, A-level, and below A-level) using Wilcoxon rank sum test with Bonferroni correction for multiple comparisons.

Results

Detecting syllables where acoustic feature values fall within the highest quartile in a sample resulted in some speech samples "dropping out": that is, no syllables qualified as stressed using certain feature sets. Table 1 shows how the number of speech samples (with at least one stressed syllable) was reduced when more acoustic features were added to detect stress.

*Table 1. Samples including at least one stressed syllable detected w*ith the respective feature sets. *The total number of original samples was 265.*

Feature set	Samples with stressed syllables
max intensity	264
duration	264
max f0	263
f0-intensity	213
f0-duration	201
intensity-dur	199
f0-int-dur	131

This raised questions about the reasons for the "dropped out" samples. The number of syllables (articulated as parts of words recognized as Swedish) varies considerably between samples as well as speaker groups, ranging from 1 to 92 (see Figure 1). This can cause bias when detecting stressed syllables within samples. Therefore, we first decided to exclude samples with the number of syllables within the lowest quartile. The data was reduced to samples including 9 or more syllables, resulting in 198 samples in total. With this data, the number of samples with stressed syllables was 198 for *max intensity* and *duration*, 197 for *max f0*, 172 for *f0-intensity*, 162 for *f0-duration*, 160 for *intensity-dur*, and 110 for *f0-int-dur*. Excluding samples with less than 9 syllables also reduced the group sizes: the remaining samples included 29 natives, 42 B-level, 123 A-level, and only 4 below A-level samples. Therefore, we decided to exclude the below A-level samples from group comparisons.

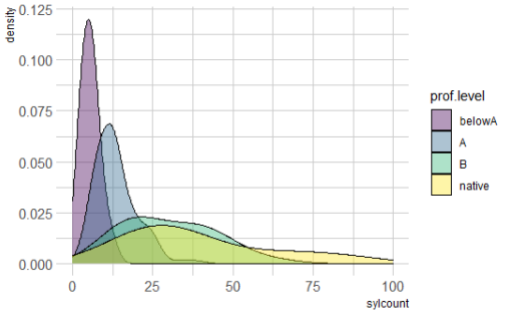

Figure 1. Distribution of syllable counts per speaker group.

The groups were compared using Wilcoxon rank sum test with Bonferroni corrections. The comparisons were done using each of the seven feature sets in detecting stressed syllables. No significant differences in the proportion of stressed syllables were found between the groups using the feature sets *max intensity*, *duration*, *max f0*, *f0-intensity*, and *intensity-duration*. Using the feature set *f0-int-dur*, the proportion of stressed syllables differed significantly between A- and B-level samples ($p < 0.01$) as well as between A-level and native samples ($p < 0.05$), but no significant differences were found between B-level and native samples (Figure 2). Using the feature set *f0-duration*, however, resulted in B-level samples differing from natives ($p < 0.05$) and slightly from A-level samples ($p < 0.1$), while A-level samples did not differ from native samples (Figure 3).

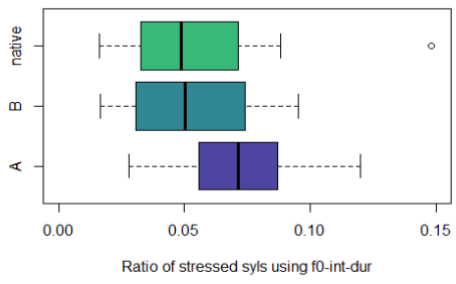

Figure 2. Proportion of stressed syllables in A-level samples (N=59), B-level samples (N=24) and native samples (N=25), when acoustic stress

is based on normalized f0 and intensity maxima and normalized syllable duration.

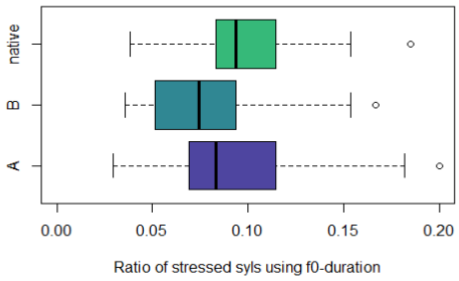

Figure 3. Proportion of stressed syllables in A-level samples (N=92), B-level samples (N=38) and native samples (N=29), when acoustic stress is based on normalized f0 and normalized syllable duration.

Figures 2 and 3, however, present somewhat conflicting results. With *f0-duration*, the proportion of stressed syllables seems to be higher with native speakers than L2 speakers, while the tendency is reverse with the feature set *f0-int-dur*. We decided to compare the distributions of normalized *max intensity*, *max f0*, and standard deviation of syllable duration between the speaker groups in order to evaluate possible reasons for this unexpected shift in the occurrence of stressed syllables. The distributions for *max intensity* and *max f0* were very similar, A-level group having only slightly less variation in these measures than natives and B-level speakers. However, the rate-normalized standard deviation of syllable durations varies remarkably within A-level group compared to B-level and native speakers, as seen in Figure 4: this indicates, that the A-level speakers cannot be treated as a homogeneous group when it comes to the use of syllable duration as a marker of prominence.

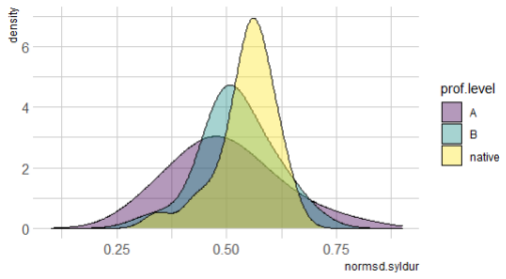

Figure 4. Distribution of rate-normalized standard deviation of syllable duration per speaker group.

Discussion

This study examined whether acoustically detected sentence stress could distinguish native and L2 speakers of Finland Swedish at different proficiency levels. Our goal was to detect stressed syllables from spontaneous utterances using only acoustic measures of f0, intensity, and duration, and compare the proportion of stressed syllables between speaker groups. Using the feature set *f0-int-dur* distinguished A-level speakers from B-level and native speakers, indicating that A-level speakers produced stressed syllables more often than the other groups. However, detecting stress with *f0-duration* provided contradicting results: the B-level speakers differed from native speakers with lower proportion of stressed syllables. Further investigation revealed considerable variance within A-level group in the use of acoustic markers of stress, especially syllable duration (see Figure 4). This indicates that acoustic (sentence) stress measures might not be reliable in assessing L2 Finland Swedish speakers within the lower proficiency levels, and the following discussion should therefore focus only on the differences between B-level and native speakers.

Using *f0-duration* detected significantly fewer stressed syllables from B-level speakers' speech than from native speakers' speech ($p < 0.05$, see Figure 3). Observing the distributions in Figure 4, in turn, reveal that B-level speakers have on average smaller standard deviation of syllable duration than native speakers. This indicates that the syllables of B-level speakers are more even in duration than the ones of native speakers, causing the stress contrasts to weaken. In our previous study on the same data (Kallio et al., submitted), in turn, we found that the higher the standard deviation of syllable durations, the better the fluency ratings of L2 speech. Both results support the ones of Heinonen (2020) regarding the L2 speakers' use of duration in marking stress in FS.

The speech samples varied considerably with regards to length, which raises the question of whether the samples are comparable with respect to stress production. However, reducing the data to samples consisting of 20 or more syllables did not affect the differences between groups, indicating that samples longer than 9 syllables don't notably increase the amount of information on sentence stress production. For further research, however, speech samples of a more even length are recommended to avoid possible bias.

Using different combinations of syllable f0 and intensity maxima and duration failed to detect stressed syllables in some samples. Possible reasons why the highest quartile of f0, intensity, and duration values do not fall on the same syllables were examined with auditive analysis on 4 native and 18 B-level speech samples. Our observations were mostly associated with the f0 and included so-called list intonation (caused by the task assignment), exaggerated or rising word and phrase intonation, emulating CSS (and in one case, Norwegian) word accents, and lack of f0 variation leading to monotonous speech. Some samples included a considerable amount of unmodal voice or were simply of bad signal quality. The observations indicate that, despite being tonally balanced, f0 is an important cue in signalling sentence stress in FS. Moreover, the f0 peak in FS is very likely positioned on the stressed syllable, while in CSS it can be delayed to or realized in the following unstressed syllable (see, e.g., Vihanta et al., 1990) – thus the speakers who use CSS word accents in our data ended up as outliers.

Our method of categorizing stressed syllables based on acoustic feature values within the highest quartile is arbitrary and doesn't perfectly acknowledge the relative nature of stress. When categorizing stress, in longer utterances in particular, the best method would be to make acoustic features values proportional to neighbouring syllables, such as in Yarra et al. (2017) and Kallio et al. (2020). Further, our method was not validated with perceived stress markings of native speakers of the target language. Regardless of these drawbacks, significant differences were found between native and B-level speakers.

We conclude that scrutinizing the realizations of sentence stress in L2, in spontaneous speech in particular, could benefit the automatic assessment of L2 speaking proficiency, but the stress detection methods as well as the type of speech data should be considered carefully. For example, the production of sentence stress requires the ability to produce longer stretches of connected speech, which can be difficult for low level L2 speakers whose speech is often characterized by many disfluencies and pronunciation problems. It would thus be relevant to include sentence stress only to the assessment of speakers at B-level or higher proficiency. Moreover, f0 seems to be an important cue for sentence stress (alongside duration), which should be taken into account with appropriate test tasks and recording quality, but also with the use of different language varieties.

References

Abelin, Å & Thorén, B (2015). What affects recognition most – wrong word stress or wrong word accent? M Svensson Lundmark m.fl. (Eds.), *Proceedings from Fonetik 2015 Lund, June 8–10 2015*. Working Papers 55. Lund: Lund University, 7–10.

Altmann, H (2006). *The perception and production of second language stress: A cross-linguistic experimental study*. DE, USA: University of Delaware Newark.

Bruce, G (2005). Intonational prominence in varieties of Swedish revisited. *Prosodic Typology: The phonology of intonation and phrasing*, 410–429.

Bruce, G (2012). Allmän och svensk prosodi. Lund: Studentlitteratur.

Cucchiarini, C, Doremalen, J, & Strik, H (2010). Fluency in non-native read and spontaneous speech. In DiSS-LPSS Joint Workshop 2010, Tokyo, Japan.

Engstrand, O, & Krull, D (1994). Durational correlates of quantity in Swedish, Finnish and Estonian: Cross-language evidence for a theory of adaptive dispersion. Phonetica, 51(1–3), 80–91.

Fant, G & Kruckenberg, A (1994). Notes on stress and word accent in Swedish. *Proceedings of the International Symposium on Prosody*. 1994, Yokohama, Japan.

Ferrer, L, Bratt, H, Richey, C, Franco, H, Abrash, V, & Precoda, K (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. Speech Communication, 69, 31–45.

Field, J (2012). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly* 39 (3), 399–423.

Hahn, L D (2004). Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals. TESOL quarterly, 38(2), 201–223.

Heinonen, H (2020). *Uttal och dess begriplighet i finskspråkiga gymnasisters L2-svenska*. JYU Dissertations 296. Jyväskylä: Jyväskylä University.

Heinonen, H & Kautonen, M (2020). L2-ruotsin lausepainon sanallinen kuvaaminen – menetelmällisiä avauksia ääntämisen arviointiin. S Grasz, T Keisanen, F Oloff, M Rauniomaa, I Rautiainen & M Siromaa (red.), *Menetelmällisiä käänteitä soveltavassa kielentutkimuksessa – Methodological Turns in Applied Language Studies. AFinLAs årsbok* 2020. Suomen soveltavan kielitieteen yhdistyksen julkaisuja n:o 78. Jyväskylä: Suomen soveltavan kielitieteen yhdistys ry, 71–94.

Helgason, P, Ringen, C & Suomi, K (2013). Swedish Quantity: Central Standard Swedish and Fenno-Swedish. *Journal of Phonetics* 41 (6), 534–545.

Hildén, R, von Zansen A, Kautonen, M, Huhta, A, Kuronen, M, & Kurimo, M (submitted). Paving the way for an online tool for teaching, learning and

assessing L2 Swedish and Finnish learners' speaking.

Hirst, D, & Di Cristo, A (1998). A survey of intonation systems. Intonation systems: A survey of twenty languages, 1–44.

Hsieh, C-N, Zechner, K & Xi, X (2020). Features Measuring Fluency and Pronunciation. I: Zechner, K & Evanini, K (red.). *Automated speaking assessment: Using language technologies to score spontaneous speech.* New York: Routledge. 101–122.

Ivars, A-M (2015). *Dialekter och småstadsspråk.* Svenskan i Finland – i dag och i går I:1. Skrifter utgivna av Svenska litteratursällskapet i Finland Nr 798. Helsingfors: Svenska litteratursällskapet i Finland.

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

Kallio, H, Suni, A, & Šimko, J (2021, in press). Fluency-related temporal features and syllable prominence as prosodic proficiency predictors for learners of English with different language backgrounds. *Language and Speech*.

Kallio, H, Kautonen, M, & Kuronen, M (submitted). Prosody and fluency of Finland Swedish as a second language: investigating global parameters for automated speaking assessment.

Kang, O, & Johnson, D (2018). The roles of suprasegmental features in predicting English oral proficiency with an automated system. *Language Assessment Quarterly*, 15(2): 150–168.

Karhila, R, Rouhe, A, Smit, P, Mansikkaniemi, A, Kallio, H, Lindroos, E, Hildén, R, Vainio, M, & Kurimo, M, (2016). Digitala: An augmented test and review process prototype for high-stakes spoken foreign language examination. In *Interspeech 2016*: 784–785.

Kautonen, M, & von Zansen, A (2020). DigiTala research project: Automatic speech recognition in assessing L2 speaking. *Kieli, koulutus ja yhteiskunta*, 11(4).

Kormos, J, & Dénes, M (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32(2),* 145–164.

Kuronen, M & Leinonen, K (2011). *Historiska och nya perspektiv på svenskan i Finland.* Tampere: Tampereen Yliopistopaino.

Lehiste, I, 1969. *Suprasegmentals*. M.I.T. Press.

Lieberman, P, 1967. *Intonation, perception, and language.* MIT Research Monograph.

Munro, M J, 1995. Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition 17* (1), 17–34.

Tepperman, J, & Narayanan, S (2005). Automatic syllable stress detection using prosodicfeatures for pronunciation evaluation of language learners. In *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005*: 1, 937–940.

Tevajärvi, K (1982). Intonation in Finland-Swedish: Word and sentence stress in the Helsinki dialect. *Working papers 22*, 175–180. Lund: Lund University, Department of Linguistics.

Trofimovich, P, Baker, W, 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition 28* (1), 1–30.

Wennerström, A (2000). The role of intonation in second language fluency. *In Perspectives on fluency,* 102-127. University of Michigan.

Vihanta, V V, Leinonen, K & Pitkänen, A J (1990). On Rhytmic Features in Finland-Swedish and Sweden-Swedish. I: K. Wiik & I. Raimo (Eds.), *Nordic Prosody V: Papers from A Symposium.* Åbo: Åbo universitet, 325–350.

Wik, P (2011). *The Virtual Language Teacher: Models and applications for language learning using embodied conversational agents.* Doctoral dissertation, KTH Royal Institute of Technology.

Yarra, C, Deshmukh, O D, & Ghosh, P K (2017). Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing*, 5845–5849.

Xu, Y, 1999. F0 peak delay: When, where and why it occurs. In: *International Congress of Phonetic Sciences 1999*, 1881–1884.