LUND UNIVERSITY

CENTRE FOR LANGUAGES AND LITERATURE

General Linguistics
Phonetics

WORKING PAPERS

Proceedings of Fonetik 2021

Lund University

Centre for Languages and Literature

General Linguistics
Phonetics

Working Papers

56. 2021

Proceedings of *Fonetik 2021*
Lund, June 8–9, 2021

Edited by Anna Hjortdal and Mikael Roll

Working Papers

Department of Linguistics and Phonetics

Centre for Languages and Literature

Lund University

Box 201

S-221 00 LUND

Sweden

Fax +46 46 222 32 11

https://journals.lub.lu.se/LWPL

This issue was edited by Anna Hjortdal and Mikael Roll

© 2021 The Authors and the Centre for Languages and Literature, Lund University

ISSN 0280-526X

Preface

This volume of the Working Papers in General Linguistics and Phonetics contains the proceedings of the 32nd Swedish Phonetics Conference, FONETIK 2021.

The conference is held from June 8 to June 9, 2021. It will be virtual, transmitted via Zoom, hosted by Lund University, where the organizing committee will pull strings from a control room.

FONETIK 2021 is one in the series of annual conferences for phoneticians and speech scientists in Sweden, which regularly attract participants from Denmark, Finland, Norway, and Estonia, and sometimes from other countries.

There are 11 contributions in this volume, representing a wide variety of phonetic topics: speech perception, acoustic phonetics, speech technology, prosody, neurophonetics, multimodality, sound change, and more. Speakers range from up-and-coming students to established senior professors. Both university staff and participants from other parts of society, companies as well as authorities, are represented.

The virtual format is due to the COVID-19 pandemic, which has afflicted people worldwide for well over a year now. Last year's FONETIK was canceled due to the disease. We thought we could wait for a year, let the dark times pass, and have a physical meeting this year. But, as Publilius Syrus would say, *a person plans one thing, Fate another*. A year later, and still under pandemic restrictions, there was no way we could allow for another lost FONETIK meeting, so we decided to go virtual. To our surprise, we then discovered that the virtual format is not entirely void of appeal. It has led us to the proposal of a new type of session, the *thematic discussion*, which we now try for the first time.

We thank all participants for making FONETIK 2021 an extraordinary occasion that we will long remember as a light in these dark times. We would like to express our special gratitude to all the contributors to the proceedings, who have put their effort into making this event possible even in the face of adverse conditions.

Lund, June 2021

The Organizing Committee

Mikael Roll, Mechtild Tronnier, Gilbert Ambrazaitis, Johan Frid, and Anna Hjortdal

Previous Swedish Phonetics Conferences (from 1986)

I	1986	Uppsala University
II	1988	Lund University
III	1989	KTH Stockholm
IV	1990	Umeå University (Lövånger)
V	1991	Stockholm University
VI	1992	Chalmers and Gothenburg University
VII	1993	Uppsala University
VIII	1994	Lund University (Höör)
—	1995	(XIIIth ICPhS in Stockholm)
IX	1996	KTH Stockholm (Nässlingen)
X	1997	Umeå University
XI	1998	Stockholm University
XII	1999	Gothenburg University
XIII	2000	Skövde University College
XIV	2001	Lund University (Örenäs)
XV	2002	KTH Stockholm
XVI	2003	Umeå University (Lövånger)
XVII	2004	Stockholm University
XVIII	2005	Gothenburg University
XIX	2006	Lund University
XX	2007	KTH Stockholm
XXI	2008	Gothenburg University
XXII	2009	Stockholm University
XXIII	2010	Lund University
XXIV	2011	KTH Stockholm
XXV	2012	Gothenburg University
XXVI	2013	Linköping University
XXVII	2014	Stockholm University
XXVIII	2015	Lund University
XXIX	2016	KTH Stockholm
—	2017	(Interspeech 2017 in Stockholm)
XXX	2018	Gothenburg University
XXXI	2019	Stockholm University
—	2020	Cancelled due to the COVID-19 pandemic
XXXII	2021	Lund University

Contents

Updated 10th June 2021 with an edition of Table 1 in Tånnander & Edlund (2021).

Updated 1st September 2021 with a replacement of Figure 2 in Horne & Roll (2021) and editions to Gosselke Berthelsen (2021).

Pride and prominence

Mattias Heldner[1], Tomas Riad[2], Johan Sundberg[1,3], Marcin Wlodarczak[1], Hatice Zora[1]
[1]Department of Linguistics, Stockholm University
[2]Department of Swedish language and multilingualism, Stockholm University
[3]Speech, Music and Hearing, KTH

Abstract

Given the importance of the entire voice source in prominence expression, this paper aims to explore whether the word accent distinction can be defined by the voice quality dynamics moving beyond the tonal movements. To this end, a list of word accent pairs in Central Swedish were recorded and analysed based on a set of acoustic features extracted from the accelerometer signal. The results indicate that the tonal movements are indeed accompanied by the voice quality dynamics such as intensity, periodicity, harmonic richness and spectral tilt, and suggest that these parameters might contribute to the perception of one vs. two peaks associated with the word accent distinction in this regional variant of Swedish. These results, although based on limited data, are of crucial importance for the designation of voice quality variation as a prosodic feature per se.

Introduction

It is a truth universally acknowledged, that any single conference in the Fonetik series must include at least one presentation about the Scandinavian word accent distinctions. However much is already known about this pride of phoneticians and phonologists in our part of the world, we try to live up to these expectations and hope to add new aspects to the description of the phenomenon.

Word accent distinctions exist in several of the Scandinavian languages and dialects. Word accents allow speakers to differentiate between word pairs like *vreden* [ˈvréːdɛn] 'the door knobs' and *vreden* [ˈvrèːdɛn] 'the wrath' only by the use of accent 1 or accent 2, respectively.

From a phonological point of view, word accents are regarded as tonal phenomena (e.g. Bruce & Hermans, 1999; Riad, 2000a, 2000b, 2006, 2014). From a phonetical point of view, however, prosodic functions such as word accents are increasingly understood as signalled by the dynamics of the entire voice source—of which pitch is just one aspect—and where short-term variations in laryngeal articulation and/or phonatory quality are potentially important qualities (e.g. Esling et al., 2019; Ní Chasaide et al., 2015; Ní Chasaide et al., 2013).

We are only aware of a few observations of voice quality dynamics in relation to Scandinavian word accent distinctions from previous work, and these are all related to creaky voice or *stød* (realized as creaky voice or as a glottal stop). For instance, creaky voice occured more frequently in accent 1 than in accent 2 words in a South Swedish material, where it happened in connection with the pitch fall in the stressed syllable (Svensson Lundmark et al., 2017). Similarly, creaky voice in connection with sharp pitch falls in stressed syllables have also been observed in the variety of Swedish spoken in Eskilstuna west of Stockholm (Riad, 2000a, 2000b, 2009). This so called 'Eskilstuna curl' appears to occur more frequently in accent 1 words, but there are no systematic studies of whether curl really is involved in the word accent distinction. Then, there is the pride of the Danes–stød–where there are many similarities in the distribution of presence of stød and accent 1, and absence of stød and accent 2, especially in simplex forms (e.g. Basbøll, 2005).

While these observations of voice quality dynamics in relation to word accents all concern the closely related aspects low pitch, creaky voice and stød (e.g. Lindblom, 2009), our intuition tells us that word accents may differ also in other respects, including spectral characteristics.

The primary aim of this study is to explore whether the tonal word accent distinction in the Central Swedish regional variant (e.g. the one spoken in the Stockholm area) is accompanied by aspects of voice dynamics related to voice quality. To this end, we will use a set of acoustic features capturing pitch as well as other aspects of voice dynamics to analyse a recording of Claes-Christian Elert's list of word accent pairs

(Elert, 1981). In order to capture signals as close to the voice source as possible, and avoid the influence of the vocal tract present in normal microphone signals, we will instead use the signal from a miniature accelerometer attached to neck (see e.g. Heldner et al., 2018).

The study is a part of a larger project inspired by Ní Chasaide et al. (2015); Ní Chasaide et al. (2013) where we try to demonstrate that voice quality variation should be treated as a prosodic feature in its own right.

Method

Speech material

The speech material consisted of a recording of Claes-Christian Elert's list of 357 word pairs differering in word accent (Elert, 1981) by a male voice talent from the Stockholm region.

The recording was made in the Anechoic chamber in the Phonetics lab at Stockholm University. The signals from an omnidirectional condenser microphone (Sennheiser MKE 2) and from a miniature accelerometer (Knowles BU-21135) were recorded on separate channels on a battery powered Zoom F8N field recorder (48 kHz, 24 bit). The accelerometer was attached to the skin on the neck just below the level of the cricoid cartilage using double sided adhesive disks for electrodes.

All words were produced in citation form. That is, each word included a word accent, a focal (or sentence) accent as well as a boundary tone (Bruce, 1977). Each word was produced once. The majority of words in the list are disyllabic and have primary stress on the first syllable. It appears that all Swedish vowels are represented in the stressed syllables of the words.

After discarding a few mispronounciations and words with more than two syllables, 348 word pairs remained.

Segmentation

In preparation for the extraction of acoustic features, the microphone recording and word list were used to obtain an automatic segmentation on word and segment levels using WebMAUS Basic with Swedish models (Kisler et al., 2017). This automatic segmentation was manually checked and corrected where needed with a special focus on the vowel segments.

Acoustic features

The acoustic feature extraction was limited to the vocalic intervals in the first/stressed and second/unstressed syllables (henceforth V1 and V2) in the words as these appear to be the most relevant regions for exploring voice dynamics. Not only do we expect vowels to be voiced, but they are also regions of relative spectral stability where acoustic properties and changes in acoustic properties (e.g. tonal movements and tonal relations) are most likely to be perceptible (House, 1990). In order to avoid influence of vowel quality (i.e. vowel formants) on the acoustic features, these were extracted from the accelerometer signal where the subglottal resonances are relatively constant. We extracted the following acoustic features:

pitch (in Hz).

intensity (in dB).

degree of periodicity in the signal in terms of Cepstral Peak Prominence Smooth (CPPS, in dB, Hillenbrand & Houde, 1996).

relative amplitude of the fundamental measured as the difference (in dB) between the levels of the first (H1) and second harmonics H2 (e.g. Klatt & Klatt, 1990). H1–H2 is considered a correlate of the closed quotient and has for example been used to distinguish creaky, modal and breathy voice with increasing relative H1 amplitude from creaky to breathy. Note however, that H1–H2 can also be viewed as a measure of spectral slope (in dB/octave) in the lower part of the spectrum.

harmonic richness factor measured as the difference (in dB) between the level of the overtones relative to the fundamental (HRF, Childers & Lee, 1991). HRF is thus another measure of spectral tilt and it has also been used to distinguish creaky, modal and breathy voice with increasingly lower HRF in these modes of phonation. In order to reduce the influence of *Fo* on HRF (e.g. Cortes et al., 2018; Godin & Hansen, 2015), we extracted HRF with a fixed number of harmonics, in our case as the ratio of summed energy in harmonics H2–H10 to the energy in H1.

spectral balance measured as the difference (in dB) between the level in the 1 to 5 kHz band and that below 1 kHz (ALPHA, Frokjaer-Jensen & Prytz, 1976).

spectral tilt in the 1 to 8 kHz band was estimated as the first order Mel-frequency cepstral coefficient (MFCC1, Kakouros et al., 2017; Tsiakoulis et al., 2010).

probability of creaky voice estimated with the creaky voice detection system proposed by Drugman et al. (2014).

All features were extracted every 2 ms using a 50 ms window. Finally, the median of all voiced frames within the vocalic intervals in the first and second syllables of all words was calculated for each acoustic measure.

Results

First, from visual inspection of the pitch tracks it was clear that all of the words included in the analyses displayed the expected tonal contours for citation forms in Central Swedish. That is, with L*HL% for accent 1 and H*LHL% for accent 2. From the perspective of pitch movements happening specifically within the vowels in the stressed (V1) and unstressed (V2) syllables, this meant that the A1 words generally had a rising tone in V1 and a falling tone in V2, while the A2 words had a falling tone in V1 and either a falling or a rising-falling 'hat' pattern in V2.

The most salient difference between the word accents with respect to median pitch in the vowels was the downward pitch jump from stressed to unstressed vowel in accent 1, whereas the vowels had a similar pitch in accent 2 (cf. top left panel in Figure 1). As median pitch ought to give a conservative estimate of perceived pitch at the end of a vowel (d'Alessandro et al., 1998), we can assume that a downward interval from stressed to unstressed vowel is a correlate of accent 1 in citation form, whereas a comparable pitch level in the vowels characterizes accent 2.

Similar patterns, with a downward jump from stressed to unstressed vowel in accent 1 and vowels at more similar levels in accent 2 was found also for intensity, and relative level of the fundamental (H1–H2), see Figure 1. A comparable pattern with larger differences between the vowels in accent 1 than in accent 2 was observed also for degree of periodicity (CPPS). The higher degree of periodicity is consistent both with less breathiness and less creaky voice in the stressed vowel in accent 1. The lower H1–H2 in the unstressed vowel in accent 1 is probably due to a weaker fundamental in line with the weaker intensity in V2. These results are not surprising, similar observations for pitch probably led to the one vs. two peak description of the word accents in previous work (e.g. Gårding, 1977), and the observations regarding intensity, CPPS and H1–H2 will contribute to the impression of one vs. two prominent syllables in accent 1 and 2, respectively.

When we look at the remaining spectral measures in Figure 1, we observe that the word accents were mostly quite similar. The large and negative spectral balance (Alpha) values indicate that the frequency band below 1 kHz (which includes the first subglottal formant around 600 Hz) had considerably more energy than the 1 to 5 kHz band when estimated from the accelerometer signal, but the Alpha values did not mirror the pitch and intensity differences between A1 and A2. The harmonic richness factor values indicates that the energy of the overtones (H2 to H10) was approximately equal (i.e. close to 0) to that of the fundamental (H1) when estimated from the accelerometer signal. This could be due to the influence of the subglottal formant around 500 Hz. There was a slight increase in HRF (< 1dB) from V1 to V2 in A1 and the opposite pattern in A2. The first Mel-frequency cepstrum coefficient (MFCC1) which characterizes spectral tilt up to 8 kHz (although most of the spectral energy above 5 kHz will be absent in the accelerometer signal) showed less tilt (i.e. more high frequency energy) in V2 than in V1 in A1 consistent with the HRF findings and marginal differences between V1 and V2 in A2.

Finally, when we inspect the probability of creaky voice, we find that creak was rare in the recording. However, creak probability was higher in the unstressed vowel in accent 1 than in any of the other vowel positions, that is in connection with the falling low tone in V2.

Discussion

This limited study shows that the tonal word accent distinction in Central Swedish is indeed accompanied by voice quality dynamics, in addition to the tonal movements. The most salient pattern is a marked difference between the stressed and unstressed vowels in accent 1 in several acoustic features, whereas the vowels are more equal in accent 2. The observed voice quality dynamics probably contribute to the perception of one vs. two peaks (or prominent syllables) associated with the word accent distinction in this regional variant of Swedish.

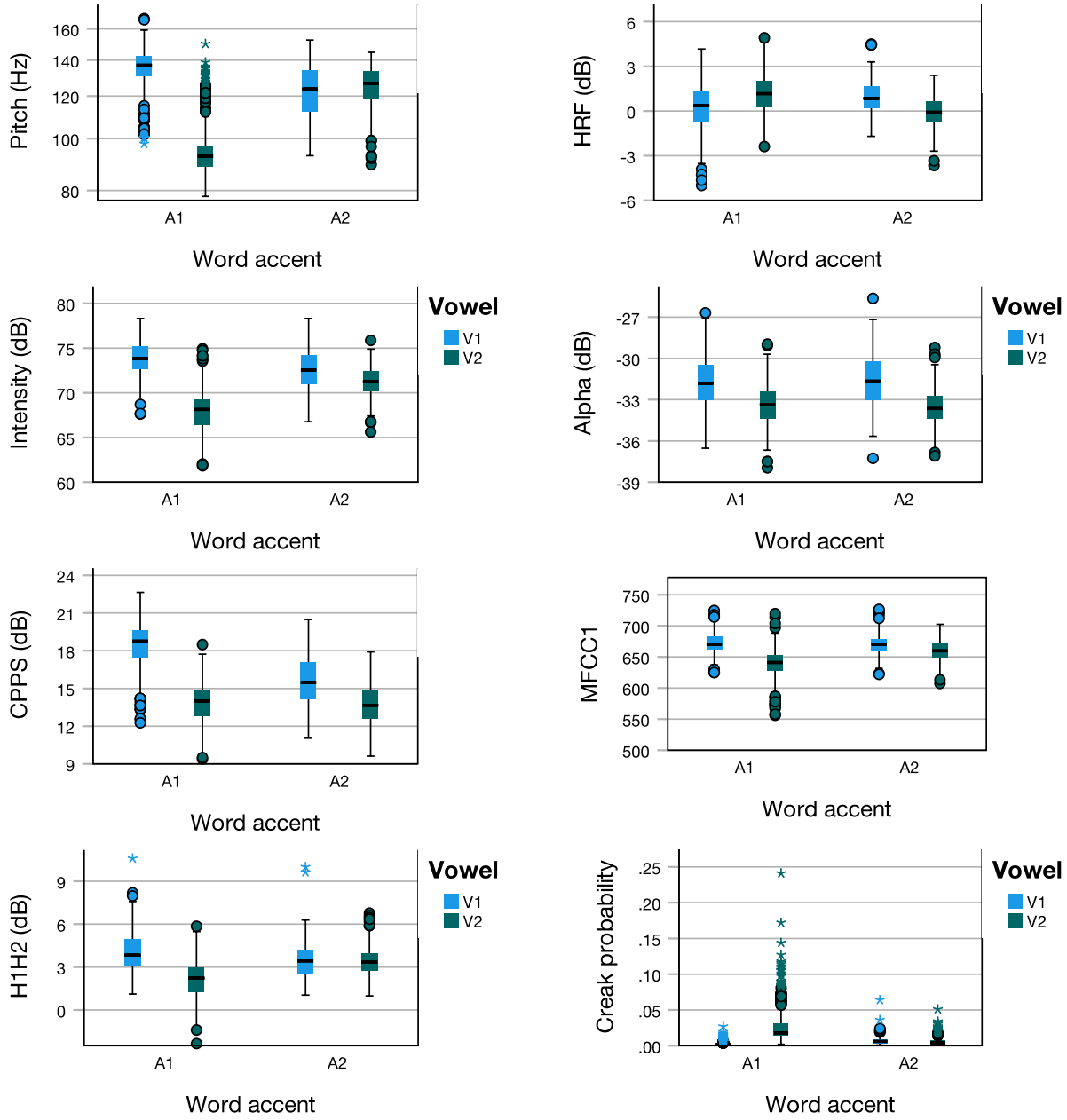


Figure 1. Box plots of acoustic features extracted from the vowels in the first and second syllables (V1 and V2) in words with word accent 1 and 2 (A1 and A2). The plots are based on all words in the speech material.

We would also like to remark that some aspects of voice dynamics are more difficult to measure than others. Indeed, several of the acoustic features used here were developed for analysis of inverse filtered signals where the influence of the vocal tract is removed through manual filtering. Informal tests showed that several of the differences observed in the accelerometer signal disappeared if the features were instead based on a normal microphone signal. While the accelerometer signal has some peculiarities, including the strong but static first subglottal formant, accelerometers clearly facilitate investigations of voice dynamics in larger datasets compared to manual inverse filterning (Heldner et al., 2018).

Finally, the result that we could identify voice quality dynamics involved in the word accent distinction from the accelerometer signal encourages us to explore other prosodic functions such as word and utterance level prominences and turn-taking with the same kind of method.

Acknowledgements

This work was partly funded by the Swedish Research Council project 2019-02932 *Prosodic functions of voice quality dynamics*.

References

Basbøll, H. (2005). *The Phonology of Danish*. Oxford University Press.

Bruce, G. (1977). *Swedish word accents in sentence perspective* [PhD dissertation, Lund University]. Lund.

Bruce, G., & Hermans, B. (1999). Word tone in Germanic languages. In H. van der Hulst (Ed.), *Word Prosodic Systems in the Languages of Europe* (pp. 605–658). Mouton de Gruyter.

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. *The Journal of the Acoustical Society of America*, *90*(5), 2394–2410. https://doi.org/10.1121/1.402044

Cortes, J. P., Espinoza, V. M., Ghassemi, M., Mehta, D. D., Van Stan, J. H., Hillman, R. E., Guttag, J. V., & Zanartu, M. (2018). Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLoS One*, *13*(12), e0209017. https://doi.org/10.1371/journal.pone.0209017

d'Alessandro, C., Rosset, S., & Rossi, J. P. (1998). The pitch of short-duration fundamental frequency glissandos. *Journal of the Acoustical Society of America*, *104*(4), 2339–2348. https://doi.org/10.1121/1.423745

Drugman, T., Kane, J., & Gobl, C. (2014). Data-driven detection and analysis of the patterns of creaky voice. *Computer Speech & Language*, *28*(5), 1233–1253. https://doi.org/10.1016/j.csl.2014.03.002

Elert, C.-C. (1981). Svenska ordpar som skiljs åt av tonaccenten. In *Ljud och ord i svenskan 2* (pp. 59–69). Almqvist&Wiksell International.

Esling, J. H., Moisik, S. R., Benner, A., & Crevier-Buchman, L. (2019). *Voice Quality: The Laryngeal Articulator Model*. Cambridge University Press. https://doi.org/10.1017/9781108696555

Frokjaer-Jensen, B., & Prytz, S. (1976). Registration of voice quality. *Brüel and Kjær Technical Review*, *3*, 3–17.

Godin, K. W., & Hansen, J. H. L. (2015). Physical task stress and speaker variability in voice quality. *EURASIP Journal on Audio, Speech, and Music Processing*, *2015*(1). https://doi.org/10.1186/s13636-015-0072-7

Gårding, E. (1977). *The Scandinavian Word Accents* [PhD dissertation, Lund University]. Lund.

Heldner, M., Wagner, P., & Włodarczak, M. (2018). Deep throat as a source of information. In Å. Abelin (Ed.), *Proceedings Fonetik 2018* (pp. 33–38). University of Gothenburg.

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech Language and Hearing Research*, *39*(2). https://doi.org/10.1044/jshr.3902.311

House, D. (1990). *Tonal Perception in Speech* [PhD dissertation, Lund University]. Lund.

Kakouros, S., Räsänen, O., & Alku, P. (2017). Evaluation of spectral tilt measures for sentence prominence under different noise conditions. In *Proceedings Interspeech 2017* (pp. 3211–3215). ISCA. https://doi.org/10.21437/Interspeech.2017-1237

Kisler, T., Reichel, U., & Schiel, F. (2017). Multilingual processing of speech via web services. *Computer Speech and Language*, *45*, 326–347. https://doi.org/10.1016/j.csl.2017.01.005

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America*, *87*(2), 820–857. https://doi.org/10.1121/1.398894

Lindblom, B. (2009). Laryngeal mechanisms in speech: The contributions of Jan Gauffin. *Logopedics Phoniatrics Vocology*, *34*(4), 149–156. https://doi.org/10.3109/14015430903008772

Ní Chasaide, A., Yanushevskaya, I., & Gobl, C. (2015). Prosody of voice: Declination, sentence mode and interaction with prominence. In *Proceedings of ICPhS 2015*. http://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0476.pdf

Ní Chasaide, A., Yanushevskaya, I., Kane, J., & Gobl, C. (2013). The voice prominence hypothesis: The interplay of f0 and voice source features in accentuation. In *Proceedings of Interspeech 2013* (pp. 3527–3531). https://www.isca-speech.org/archive/archive_papers/interspeech_2013/i13_3527.pdf

Riad, T. (2000a). The origin of Danish stød. In A. Lahiri (Ed.), *Analogy, Levelling, Markedness: Principles of Change in Phonology and Morphology* (pp. 261–300). De Gruyter Mouton. https://doi.org/10.1515/9783110899917.261

Riad, T. (2000b). Stöten som aldrig blev av: Generaliserad accent 2 i Östra Mälardalen. *Folkmålsstudier*, *39*, 319–344.

Riad, T. (2006). Scandinavian accent typology. *STUF – Language Typology and Universals*, *59*(1), 36–55. https://doi.org/10.1524/stuf.2006.59.1.36

Riad, T. (2009). Eskilstuna as the tonal key to Danish. In *Proceedings Fonetik 2009 Stockholm* (pp. 12–17). Department of Linguistics, Stockholm University.

Riad, T. (2014). Tonal word accents. In *The Phonology of Swedish* (pp. 181-191). https://doi.org/10.1093/acprof:oso/9780199543571.003.0009

Svensson Lundmark, M., Ambrazaitis, G., & Ewald, O. (2017). Exploring Multidimensionality: Acoustic and Articulatory Correlates of Swedish Word Accents. In *Proceedings Interspeech 2017* (pp. 3236–3240). ISCA. https://doi.org/10.21437/Interspeech.2017-1502

Tsiakoulis, P., Potamianos, A., & Dimitriadis, D. (2010). Spectral Moment Features Augmented by Low Order Cepstral Coefficients for Robust ASR. *IEEE Signal Processing Letters*, *17*(6), 551–554. https://doi.org/10.1109/lsp.2010.2046349

The perceptual importance of falling pitch for speakers from different dialects of Swedish

Sabine Gosselke Berthelsen
Centre for Languages and Literature, Lund University

Abstract

Falling pitch has long been argued to be a key feature in the distinction of Swedish pitch accents. In this paper, neurophysiological and behavioural evidence substantiating the perceptual importance of HL pitch contours at the word level is discussed. When presented with novel words, Swedes, regard-less of dialect, preattentively distinguished meaningful words with falling pitch. Falling pitch was also facilitative for tone mismatch detection. These responses to meaningful HL contours in foreign words were likely based on transfer from the native tongue, thus emphasising the importance of falling pitch in the discrimination of word accents in four Swedish dialect areas. However, while responses to falls were facilitated across dialects, additional dialect-dependent facilitation effects were found.

Introduction

Swedish has two pitch accents, accent 1 and accent 2, which are realised on word stems and affected by affixes (cf. Riad, 1998). Their pitch patterns differ systematically between dialects. Five dialect areas are distinguished in this respect: type 0, type 1A and 1B, and type 2A and 2B (cf. Gårding, 1973). Type 0 does not differentiate accent 1 and 2. The remaining dialect types are originally named with respect to the number of peaks that accent 2 receives in citation form (or focal position). A third word accent type is tentatively included for Gothenburg Swedish: type 3. Here, accent 1 and 2 have both been shown to elicit double peak contours in both focal and non-focal position (e.g., Segerup, 2005).

Word accents clearly differ within dialects, on the one hand, and between dialects, on the other. Yet, it is often argued that this variability is entirely based on timing differences and that all word accents are essentially different encodings of an underlying HL contour (e.g., Bruce, 1983, 2005; Bruce & Gårding, 1978). Thus, regardless of dialect, the onset of the word accent fall is earlier for accent 1 than for accent 2 (cf. *Figure 1*). Besides this crucial difference distinguishing the two word accents, there are further timing differences that distinguish the different dialect types. Thus, the onset of the word accent falls is earlier in type 2 dialects than in type 1 dialects. In type 2 dialects, the HL onset is in fact so early that accent 1 is realised as a low tone ([H]*L). The late onset in type 1 varieties, on the other hand, typically enforces the inclusion of a pitch rise from the previous word's L ([L]*HL). Regarding the A- and B-subtypes within type 1 and type 2, A has an earlier fall than B (Bruce, 1983, 2005). The double-peaked word accents in Gothenburg Swedish do not easily fit into the general word accent typology with respect to timing or pitch movement. Yet, there is a small but important difference in the timing of the fall in the stressed syllable between accent 1 and accent 2 even in this otherwise atypical dialect.

While the HL contour is likely the underlying feature driving word accent distinction within and between dialects overall, the vastly different realisations invite for speculations about the factual perceptual relevance of the fall and whether it is equally important in all dialects. It is conceivable, for example, that listeners from type 2 dialect areas perceive the difference between word accents as a high – low contrast or that speakers of type 1 dialects make use of the rising part of the contour to distinguish words, as the remnant rise precedes the H*L.

A reliable way of investigating the perceptual importance of pitch patterns in a population is to present speakers with a foreign tone system and study acquisition of and interaction with the new tones. Acquired perceptual strategies for the native language (i.e., specialised neural circuits) involuntarily guide the processing of foreign language input, at least initially. Swedish has recently received some attention in this context where it has been shown that Swedish speakers outperform speakers from non-tonal languages in

discrimination or processing of foreign tones (e.g., Burnham et al., 2014; Gosselke Berthelsen, et al., 2020). Interestingly, this is the case when tones are embedded in a linguistic context but does not hold for musical tones.

More closely investigating not simply the advantage of having tones in the native language (L1) but looking at specific pitch patterns, it has further been illustrated that Swedish second language (L2) learners highly accurately identify falling L2 tones (i.e., Mandarin T4; Gao, 2016, 2019). Accuracies for L2 fall-rising tones (T3) are also high. It is not certain whether the latter is due to the presence of the initial fall or related to the fact that the fall-rise contour is reminiscent of the combined pattern for word accent and focus tone (in some varieties). Importantly, high and rising L2 tones (T1 & T2) are recognised with considerably lower accuracy.

This behavioural advantage for falling tones, however, seems to disappear with high intensity training (Gosselke Berthelsen et al., 2021). Under such circumstances, Swedish native speakers did not perform differently in the detection of mismatches related to falls compared to rises and high or low tones. In fact, their behavioural responses were statistically indistinguishable from those of non-tonal participants. Yet, while the behavioural measures were not indicative of a perceptual advantage for falls, a fall-facilitated neurophysiological response was observed just 50 milliseconds after the onset of the tone. This relatively newly discovered response has been associated with a lexical gating process, distinguishing words from nonwords and grammatical from ungrammatical forms (e.g., Herrmann et al., 2009; MacGregor et al., 2012; Shtyrov & Lenzen, 2017). A reduction in this EEG component, argued to be related to eased preconscious processing, was found for falls in meaningful words. This suggests that Swedish listeners are conditioned from their L1 to automatically use falling pitch as a cue to lexicality status. A process, that they preconsciously make use of even in the perception of foreign tones in a meaningful linguistic context.

Somewhat surprisingly maybe, none of the presented studies on Swedish native speakers' identification and processing of foreign tone offer clear information about the participants' dialectal background. Instead, they choose to treat Swedes as a uniform group with respect to word-level pitch experience. A choice that is presumably seen as justified by the suggestion that all word accent realisations are instantiations of HL contours. However, the differences in pitch realisations at the word level as a result of timing differences and interactions with sentence-level pitch (e.g., focus tones) might in fact result in differences in perceptual cues for word discrimination in the different dialects. The present paper presents a first attempt at bridging this gap and investigating if there are dialectal differences concerning perceptually important pitch patterns. For this purpose, behavioural and neurophysiological data from Gosselke Berthelsen et al. (2021) are analysed according to the participants' native dialect and emerging patterns for the lexical gating effect are discussed.

Methods

Participants

Twenty-three right-handed participants (aged 19-29, *M* = 23.7; 12 female) were informants in the present study. They were all native speakers of Swedish and came from three different dialect areas: South Swedish (type 1A), Central Swedish (type 2A, referred to by Bruce (e.g., 2005) as East Swedish), and West Swedish (type 2B). West Swedish was further separated into West Swedish and Gothenburg Swedish (type 3). Participants' dialects were determined in accordance with self-reported native dialect and primary area of residence, *Figure 1*. Note that participants moved within and across dialect borders (number of moves = 1-6; *M* = 2.8), including a move to the Lund area in the south of Sweden, where all participant were residing at the time of testing.

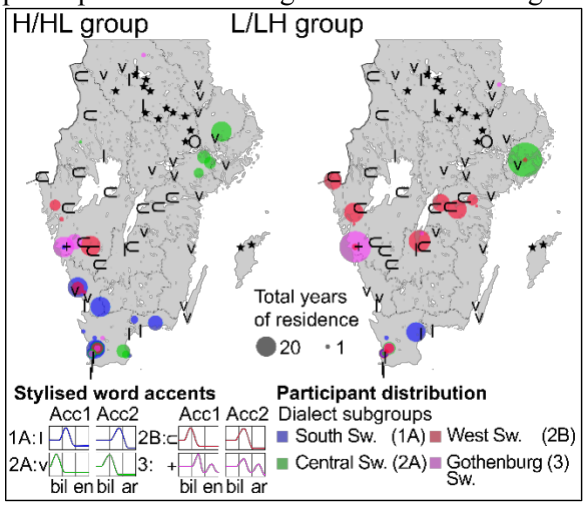


Figure 1. Maps for both groups indicating subgroups' residence distribution as well as Swedish word accent patterns. Word accent distribution adapted from Gårding & Lindblad (1973). Stylised, non-focal word accents modelled based on Bruce (1983) and Segerup (2005).

Participants were divided into two groups. They were tasked with learning novel foreign words with H* and H*L tones (H/HL group) or L* and L*H tones (L/LH group), respectively. Groups were matched for age ($M_{H/HL}$ = 23.4 years; $M_{L/LH}$ = 24.0 years) gender (H/HL: 6 female; L/LH: 6 female), and a large number of other background factors, such as working memory and socioeconomic status. Both learner groups could be divided into four subgroups representing the four dialectal areas, see *Table 1*.

Table 1. Number of participants per dialect.

Dialect	H/HL	L/LH
South Swedish	4	1
Central Swedish	3	3
West Swedish	2	5
Gothenburg Swedish	2	3

Materials

Twenty-four auditory, monosyllabic pseudowords were created for the experiment, *Table 2*. Although Swedish does not have differential pitch on monosyllables (they uniformly bear accent 1, e.g., *bil*$_1$, 'car'), monosyllables were deemed suitable for this investigation as the pitch accent on a monosyllabic stem is systematically altered when suffixes are added (e.g., *bil*$_2$*-ar*, 'cars', *bil*$_2$*-bälte*, 'seatbelt', but *bil*$_1$*-en*, 'the car', cf. Riad, 1998). In the present stimuli, rather than adding pitch-altering suffixes, the tones themselves served as grammatical suprafixes, see procedure below.

Table 2. Pseudowords used in the experiment.

i/u word pairs	a/ɛ word pairs
dif / duf	dap / dɛp
fif / fuf	fap / fɛp
kip / kup	kaf / kɛf
lir / lur	lap / lɛp
sis / sus	sap / sɛp
tip / tup	taf / tɛf

Half of the auditory stimuli were presented as novel words while the other half were presented as nonwords. A pseudoword learning paradigm was selected in order to guarantee equal exposure to the foreign words in which lexicality effects were to be studied. Word division was pseudo-random, such that i/u word pairs and a/ɛ word pairs were used as novel words or nonwords in an equal number of participants. In Praat (Boersma, 2001), four tones (H*, H*L, L*, L*H, see *Figure 2*) were added to the words in the following way: For any given participant, H*/H*L pitch would either be added to the novel words, assigning the participant to the H/HL group, or with nonwords, assigning the participant to the L/LH group. The L*/L*H would then be added to the other set of words: nonwords or novel words, respectively.

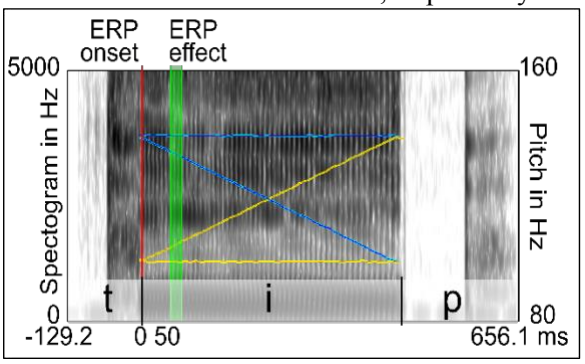


Figure 2. Spectrogram for example word /ti:p/ with the four pitch contours (H, H*L, L*, L*H).*

Procedure

Participants were presented with all words a total of 60 times, i.e. 30 times each on two consecutive days. All novel words were always followed by pictures of people in different professions to assign meaning to the words. The tones were associated with grammatical meaning, such that the difference between H* and H*L in the H/HL group or L* and L*H in the L/LH group would either be associated with grammatical gender or number distinctions (i.e., singular vs plural or feminine vs masculine). This was done to assimilate the function of pitch on Swedish words while keeping an easily controllable, monosyllabic word structure. Participants were told explicitly that the words that they would learn would contain gender and number markers. To uphold the participants' attention and to assess learning progress, in some trials (12.5%), the novel words were presented with a wrong picture. Participants were prompted to indicate incorrectly matched trials by a button press. Behavioural measures documenting the recognition of pictures that incorrectly matched the preceding word's tone will be discussed in the preliminary results.

Electroencephalography

During the experiment, EEG data was recorded from 64 Ag-AgCl EEG electrodes (EASYCAP GmbH, Herrsching, Germany) with a SynAmps^2 amplifier and Curry Neuroimaging Suite 7 software (Compumedics Neuroscan, Victoria, Australia). Two bipolar eye channels (EOG) were included. Left mastoid (M1) served as online reference and

AFz as ground. Sampling rate was 500 Hz and a low pass filter of 200 Hz was run online. Offline, the data was filtered with a 0.01 Hz high pass and a 30 Hz low pass filter. Epochs of 1200 ms were extracted at vowel/tone onset (cf. *Figure 2*) including a 200-ms baseline. Eye artefacts were corrected through independent component analysis (ICA, Jung et al., 2000). Epochs exceeding ±100 μV were excluded (M = 1.09%; SD = 0.72).

Preliminary analysis and results

To investigate the potential effect of dialect in the dataset, the behavioural measures response times and response accuracy were divided into group averages for the different dialects, see *Table 3* and *Table 4*. The H/HL group overall has faster response times (M = 1615 ms) and higher response accuracy (M = 79%) than the L/LH group (M_{RT} = 1889 ms; M_{ACC} = 55 %). Further, mismatches with the H*L pitch pattern (M = 1587 ms) are recognised more quickly than mismatches with any other tone type (M_{H*} = 1644 ms; M_{L*H} = 2018 ms; M_{L*} = 1760 ms). With respect to accuracy, there is virtually no differences between H*L tone mismatch recognition (M = 78.67%) and the detection of mismatches with H* tones (M = 78.68%). The two do, however, clearly differ from L*H mismatches (M = 56%) and L* mismatches M = 55%). Note that Gothenburg Swedish differs from the other dialects and does not favour H*L tones.

Table 3. Response accuracy for recognition of tone mismatches by dialect.

	H/HL group		L/LH group	
Dialect	H*L	H*	L*H	L*
South	64%	59%	34%	22%
Central	95%	94%	46%	49%
West	83%	88%	69%	65%
Gothenburg	73%	74%	75%	81%

Table 4. Response times in milliseconds for recognition of tone mismatches by dialect.

	H/HL group		L/LH group	
Dialect	H*L	H*	L*H	L*
South	1483	1359	1635	1503
Central	1213	1401	2654	2173
West	1295	1657	1741	1737
Gothenburg	2357	2167	2041	1626

Further, to study the participants' neurophysiological responses, effect amplitudes for the significant electrode cluster (FC2, FC4, C1, Cz, C2, CP1, CPz, and CP2) and time window (50-70 ms) from Gosselke Berthelsen et al., (2021) were separated into dialects, see *Table 3*. Mean amplitudes showed a reduced effect for H*L tones when they were part of meaningful new words (H/HL group) regardless of dialect (M = −0.76 μV). This reduction emerged in comparison to both nonwords (M_{L*H} = −1.04 μV; M_{L*} = −0.88 μV) and novel words with H* pitch (M = −0.86 μV). When presented as part of nonwords (L/LH group), the H*L tone (M = −1.06 μV) was not preattentively differentiated from the other tones (novel words: M_{L*H} = −0.98 μV; M_{L*} = −1.00 μV; nonwords: M_{H*} = −1.06 μV). The L/LH group instead showed a tendency towards a global amplitude reduction for newly learned real words (M = −0.99 μV) compared to nonwords (M = −1.06 μV). For Central Swedish speakers (type 2A), in addition to the reduced amplitude for the fall in novel words, there was also a noticeable amplitude reduction for L* and H* tones on novel words.

Table 5. Average amplitude in (μV) of all cluster electrodes in the 50-70 ms time-window for all dialect sub-groups by group and pitch pattern.

HL/H group	Words		Nonwords	
Dialect	**H*L**	H*	L*H	L*
South	**−1.05**	−1.19	−1.15	−1.16
Central	**−0.17**	**−0.19**	−0.82	−0.64
West	**−0.98**	−1.06	−1.25	−1.02
Gothenburg	**−0.83**	−1.00	−0.95	−0.71
LH/L group	**Words**		**Nonwords**	
Dialect	L*H	L*	**H*L**	H*
South	−0.81	−0.90	**−0.78**	−0.88
Central	−0.92	**−0.79**	**−1.09**	−1.07
West	−1.55	−1.57	**−1.70**	−1.66
Gothenburg	−0.65	−0.74	**−0.67**	−0.62

Discussion

The results presented here emphasise the suggested special status of the falling tone at word level for Swedish listeners. Swedes from most dialect areas could most easily identify incorrectly matched pictures when the mismatches were based on H*L tones. More directly measuring the pitch itself, neurophysiological results showed that H*L pitch, when presented on a meaningful word, reduced the early EEG component related to lexical gating. This reduction, arguably related to ease of processing, was found in listeners from all four dialect areas in the current study. This demonstrates the perceptual importance of pitch falls at the word level in Swedish regardless of

dialect, strengthening the claim that it is indeed the fall that primarily distinguishes word accents regardless of their realisation in the complex interplay with timing and sentence level pitch (e.g., Bruce, 1983, 2005; Bruce & Gårding, 1978).

Given its close relation with word stems and affixes in Swedish, it is not surprising that the H*L tone could only serve as a cue for lexicality effects when it was added to meaningful words. When it had no meaning or function, it was too dissimilar from the native tones to affect L2 processing. Thus, in conditions where L* and L*H were associated with novel words, a tone-independent but substantially smaller effect of lexicality was observed. Listeners less successfully learned to use other differences in the speech input (i.e., L* and L*H tones or vowels) in absence of an H*L cue. This is substantiated by the behavioural data which shows reduced mismatch detection accuracy and prolonged response times for learners who had to rely on L* and L*H tones during word acquisition.

An intriguing pattern was observed for the Central Swedish subgroups. In addition to the amplitude reduction for falls, they also showed a reduced amplitude for high and low tones when these were part of meaningful words. Central Swedish (type 2A) is argued to have the earliest fall onset both for accent 1 and accent 2 (cf. Bruce, 1983). As such, accent 1 is realised as a low tone ([H]*L) in pre-focal position and for accent 2, the high tone is prominently associated with the stressed syllable (*HL). The preliminary results presented here indicate that the timing of the fall contour in Central Swedish has an impact on the perceptual prominence of pitch for native speakers of this dialect. To this effect, Central Swedish speakers preconsciously use the pitch fall as well as low and high level tones as perceptual cues for lexical distinction. Thus, response amplitudes for the early EEG component are reduced for all three pitch types, indicating eased processing due to transfer from the native dialect. This is partly corroborated in the behavioural responses. The Central Swedish participants in the H/HL group have by far the highest accuracy and fastest response times overall. This is likely due to the fact that they can preconsciously assess not only the H*L tone but also the H* tone. Further, those Central Swedish participants who learned words with L* and L*H tones have better response accuracy and response times for the L* tones than the L*H tones. However, overall accuracy is comparably low and response times are comparably long in this subgroup.

Interestingly, West Swedish (type 2B) did not share the Central Swedish (type 2A) emphasis on low and high tones. The relatively later onset of the word accent fall in this variety likely diminishes its potential interpretation as a low tone.

Another group of participants that very clearly stood out in the present study was the Gothenburg Swedish variety (here tentatively referred to as type 3). While the expected reduction for the H*L pattern in novel words was found in the neurophysiological data, the largest reductions were evident for level tones in non-words. It is unclear why the reduction would occur in the non-words rather than the words. This confound was not resolved but rather strengthened by the behavioural data. In both Gothenburg Swedish sub-groups, level pitch could more quickly and accurately be used for tone mismatch detection than the H*L and L*H contour tones. In fact, the H*L tone pattern was found to be the least useful cue for mismatch detection. The differences between conditions and between groups were, however, rather small and overall accuracy was high. It is likely that the similarity of the word accent patterns in the Gothenburg in both focal and non-focal position dialect reduces the perceptual importance of the falling pitch movement cue both for preconscious lexicality effects and for mismatch detection. Instead, listeners are likely attentive to other cues. It has previously been argued that vowel length plays an important role Gothenburg Swedish (e.g., Segerup, 2004). It is possible that the listeners in the present study interpreted the level tones as longer and could therefore effectively use them as cues for mismatch detection. However, this does not explain the reverse lexicality effect for level tones (i.e., a reduction for non-words rather than words). Further research on pitch perception in Gothenburg Swedish is thus certainly needed to understand the present data. The 1B dialect type was not included in the present investigation due to lack of participants. However, its relative similarity with type 1A word accents tentatively suggests that the fall would have the same status as it does in the remaining dialects. However, it is of course possible that there are additional perceptual cues. As the dialect area with the latest onset for the word accent falls, the type 1B variety has the longest rise leading up to the H of the HL contour. Hence, rises might potentially play a larger role in this dialect than in the other dialects.

Not mentioned explicitly in the introduction, the two previous studies on Swedish learners

identification of L2 tones carried out by Gao (i.e., Gao 2016, 2019) find slightly different results with respect to the two very accurate patterns (fall vs fall-rise). In one study (2019) the fall is the most accurate pattern while in the other (2016), the fall-rise is slightly more accurate than the fall. In light of the present results, one possible explanation for this minor difference might be the participants' dialect background. It is possible, and indeed to some degree indicated in the paper, that the 2016 study had a large number of participants from type 2B varieties where, in focal position, accent 1 is realised as a fall-rise and accent 2 has a double peak pattern. This might lead to high fall-rise accuracy. The 2019 study, on the other hand, might have been conducted with a larger number of participants from type 1 varieties who are not familiar with double peak word accents from their native dialect, leading to the expected higher accuracy for the fall.

In sum, the current pilot study into the role of dialectal variation for the perceptual salience of pitch cues at the word level found that falling tones on meaningful words boosted preconscious lexicality effects in all the four tested dialectal varieties and possibly beyond. The neurophysiological results emphasise the key function that falling pitch plays in Swedish speakers' word perception overall. Yet, native dialect pitch timings and pitch prominence can affect perceptual salience and potentially sensitise listeners to additional pitch cues or different prosodic cues (such as vowel duration). Thus, while the fall protruded as the most important perceptual cue for speakers of all tested tonal dialects of Swedish, care should be taken when treating Swedes from different dialect backgrounds as a uniform group with respect to word-level pitch processing. More extensive data, including more participants, more dialects, and even a larger variety of tested pitch contours, is needed to corroborate the present preliminary findings and suggestions.

References

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341-345.

Bruce, G. (1983). Accentuation and timing in Swedish. *Folia Linguistica, 17*(1-4), 221-238. doi:10.1515/flin.1983.17.1-4.221

Bruce, G. (2005). Intonational prominence in varieties of Swedish revisited. In S.-A. Jun (Ed.), *Prosodic typology: The phonology of intonation and phrasing*. doi:10.1093/acprof:oso/9780199249633.003.0015

Bruce, G., & Gårding, E. (1978). A prosodic typology for Swedish dialects. *Working Papers. (Phonetics Laboratory Lund University), 17*.

Burnham, D., Kasisopa, B., Reid, A., Luksaneeyanawin, S., Lacerda, F., Attina, V., . . . Webster, D. (2015). Universality and language-specific experience in the perception of lexical tone and pitch. *Applied Psycholinguistics, 36*, 1459-1491. doi:10.1017/S0142716414000496

Gao, M. (2016). Perception of lexical tones by Swedish learners of Mandarin. *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, 33-40.

Gao, M. (2019). Perception of Tone 1-Tone 2 Contrast by Swedish Learners of Mandarin Chinese. *FONETIK 2019*, *27*, 7-12.

Gårding, E. (1973). The Scandinavian word accents. *Working Papers. (Phonetics Laboratory Lund University), 8*, 1-119.

Gårding, E., & Lindblad. (1973). Constancy and variation in Swedish word accent patterns. *Working Papers. (Phonetics Laboratory Lund University), 7*, 36-110.

Gosselke Berthelsen, S., Horne, M., Shtyrov, Y., & Roll, M. (2020). Different neural mechanisms for rapid acquisition of words with grammatical tone in learners from tonal and non-tonal backgrounds: ERP evidence. *Brain Research, 1729*(146614). doi:10.1016/j.brainres.2019.146614

Gosselke Berthelsen, S., Horne, M., Shtyrov, Y., & Roll, M. (2021). Native language experience narrowly shapes pre-attentive foreign tone processing and governs rapid memory trace build-up: An ERP study. *Manuscript submitted*.

Herrmann, B., Maess, B., Hasting, A. S., & Friederici, A. D. (2009). Localization of the syntactic mismatch negativity in the temporal cortex: An MEG study. *NeuroImage, 48*, 590-600. doi:10.1016/j.neuroimage.2009.06.082

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology, 37*, 163-178. doi:10.1111/1469-8986.3720163

MacGregor, L. J., Pulvermüller, F., van Casteren, M., & Shtyrov, Y. (2012). Ultra-rapid access to words in the brain. *Nature Communications, 3*(711), 1:7. doi:10.1038/ncomms1715

Riad, T. (1998). The origin of Scandinavian tone accents. *Diachronica, 15*(1), 63-98. doi:10.1075/dia.15.1.04ria

Segerup, M. (2005). The interaction of word accent and quantity in Gothenburg Swedish. *FONETIK 2005*, 13-16.

Segerup, M. (2004). Gothenburg Swedish word accents: a fine distinction. *FONETIK 2004*, 28-31.

Shtyrov, Y., & Lenzen, M. (2017). First-pass neocortical processing of spoken language takes only 30 msec: Electropysiological evidence. *Cognitive Neuroscience, 8*(1), 24-38. doi:10.1080/17588928.2016.1156663

Segmental durations as a correlate of Swedish word accents: Evidence from Stockholm and Scania Swedish

Gilbert Ambrazaitis[1] & Mechtild Tronnier[2]
[1]Linnaeus University, Växjö
[2]Centre for Languages and Literature, Lund University

Abstract

This study aims to scrutinize the role of segmental duration as a correlate of the two Swedish word accents, asking whether it is a robust correlate, independent of tonal complexity and phrase-level prosodic factors. To this end, we examined segmental durations of Accent 1 and Accent 2 words in two regional dialects, controlling for focus and phrase finality. Recordings from 24 speakers were analysed. The results showed that the vowel in the stressed syllable and the post-vocalic consonant were produced longer with Accent 2 than with Accent 1, irrespective of the speakers' dialect, focus condition and position in the utterance.

Introduction

Swedish exhibits a binary tonal word-accent distinction (Accent 1, Accent 2, henceforth A1 and A2) which is acoustically manifested in the fundamental frequency (f_o) contour. Further acoustic correlates beyond f_o have hardly been acknowledged in the past, although minor differences in segmental durations between A1 and A2 have been observed in some previous studies (e.g., Elert, 1964; Heldner & Strangert, 2001; Svensson Lundmark et al., 2017).

The few previously reported duration data for A1 and A2 in Swedish are, we argue, not well in line with the numerous available reports on tone languages, which suggest that more complex tonal patterns are reflected by longer durations (e.g., Köhnlein, 2015, and references therein). For Swedish, however, previous research has not consequently revealed longer durations where it would be predicted based on tonal complexity. For instance, for Stockholm Swedish, Elert (1964) observed a longer post-stress consonant in A2 than in A1, but at the same time a longer stressed vowel in A1 than in A2, although longer durations would generally seem to be predictable for the more complex H*LH-pattern in A2 than for the (H)L*H in A1. Similar results to those of Elert (1964) were reported by Heldner and Strangert (2001), although they did not explicitly focus on segmental durations as a correlate of word accents. Their aim was instead to examine focal lengthening, which leads us to a possible explanation of the results by Elert (1964).

Thus, possibly, the Swedish word accents *per se* do not actually differ with respect to duration, but the small observable durational differences between A1 and A2 are, instead, a bi-product of phrase-level prosody. In particular, one could hypothesize that focal lengthening has a slightly stronger effect on the segmental material that is aligned with the focus marking H tone, than on remaining segments. This would mean that focal lengthening should have a stronger effect on the stressed syllable in A1 than in A2, and conversely a stronger effect on the post-stress syllable in A2 than in A1. We might thus, under focus, expect a longer stressed vowel in A1 than in A2 and a longer post-vocalic consonant (= onset of the post-stress syllable) in A2, which is what the results by Elert (1964) and Heldner and Strangert (2001) displayed.

Furthermore, if the alignment of the focus marking tone is what determines the segmental scope of focal lengthening, different lengthening patterns for different dialects could be predicted. In Scania Swedish, for instance, focus is typically not encoded through an additional tonal gesture as in Stockholm Swedish, but rather through the tonal pattern determined by the word accents: a H*L fall in A1, and a L*H rise in A2. Thus, for Scanian, our hypothesis would predict that, under focus, the stressed vowel might be lengthened in both A1 and A2, since the focus marking tonal pattern is aligned with the vowel in both cases. In addition, however, one could expect lengthening of the post-vocalic consonant in A2 (only),

because this consonant is typically included in the A2 L*H rise. In line with this prediction are the results by Svensson Lundmark et al. (2017), who found a durational differentiation of A1 and A2 for Scania Swedish, under focus, only in the post-vocalic consonant.

However, for a conclusive test of the hypothesized impact of focus on the durational differentiation between A1 and A2, focus as a factor should be experimentally controlled for (as in Heldner & Strangert, 2001). In this study, we therefore compare segmental durations for A1 and A2 in focal and non-focal conditions. As an additional control factor, we include different positions in the utterance, as it is well known that lengthening is not only triggered by focus but also only by phrase finality (e.g., Heldner & Strangert, 2001). Furthermore, we present two parallel analyses for two dialects of Swedish – Stockholm and Scania – which differ critically in the tonal composition of the word accents, as well as in how focus is marked tonally.

The study

Subjects and materials

The analyses are based on recordings from 24 speakers in total (6 women and 6 men per dialect), and 36 utterances per speaker (12 conditions, 3 repetitions). Speakers were asked to read the following sentences, which were preceded by context questions in order to trigger different focus conditions (varying narrow focus on the three nouns in the phrase):

Boven hade vinet i bilen
'The villain had the wine in the car',
Boven hade viner i bilen
'The villain had wines in the car',
Boven hade vinet i bilar
'The villain had the wine in cars',
Boven hade viner i bilar
'The villain had wines in cars'.

Hereby, the three nouns (villain, wine, car) are associated either with A1 (= def. sg. form) or A2 (= indef. pl.). Only the two last nouns in the phrases (*vinet/viner* and *bilen/bilar*), produced in A1 form and A2 form, were considered. These two nouns occur at different positions in the utterance (*vinet/r* = medial; *bilen/ar* = final), and are tested in focal and non-focal conditions: in pre-focal position (only *vinet/viner* – narrow focus on *bilen/ar),* in focal position (both *vinet/viner* and *bilen/bilar),* and in post-focal position (both *vinet/viner* and *bilen/bilar* – narrow focus either on *boven* or on *vinet/viner).*

Data Analysis

The target words were segmented manually in PRAAT (Boersma & Weenink, 2018). Duration data were extracted for the vowel in the stressed syllable as well as the post-vocalic consonant from both target words (/i/ and /n/ in *vinet/r;* /i/ and /l/ in *bilen/ar*).

Data were analyzed using linear mixed regression models, separately for the two dependent variables (duration of vowel and duration of following consonant), and separately for the two sentence positions (*vinet/er* vs. *bilen/ar*) and dialects. The models included the predictors *word accent (WA)* * *focus* as fixed effects (where '*' denotes an interaction term), and *speaker* as a random effect.

Each of the full models (*WA* * *focus* + (1|*speaker*)) was then compared to two different reduced models using likelihood ratio tests in order to evaluate the significance of the predictor *WA* and its interaction with *focus*.

Results

From Figures 1-8, it can be seen that the results reveal an overall stable and uniform effect of word accent on the duration of the vowel in the stressed syllable (/i:/ in *vin* and *bil*) and the subsequent consonant (either /n/ or /l/), irrespective of focus condition, position in utterance (*vin* vs. *bil*) and dialect: Both segments tend to be slightly longer in A2 than in A1 (except in the /l/ in post-focal *bilen* in Stockholm Swedish). Results of the likelihood ratio tests are included in the captions of Figure 1-8.

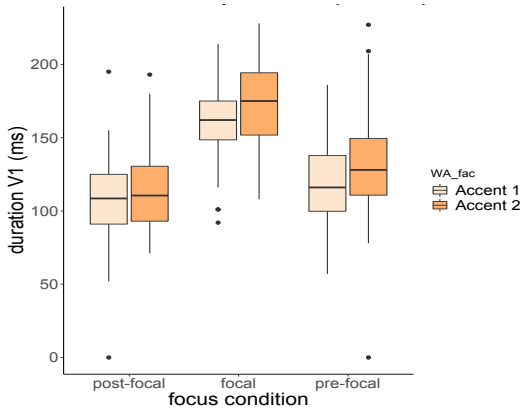


*Figure 1. Duration of the vowel /i/ produced with A1 (in "vinet") and A2 (in "viner") in utterance medial position in Central Swedish (Interact. n.s, WA***, Diff WA =10 ms*).

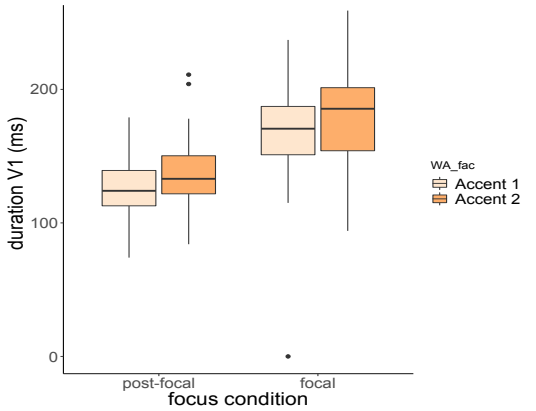


*Figure 2. Duration of the vowel /i/ produced with A1 (in "bilen") and A2 (in "bilar") in utterance final position in Central Swedish (Interact. n.s, WA***, Diff WA =10 ms).*

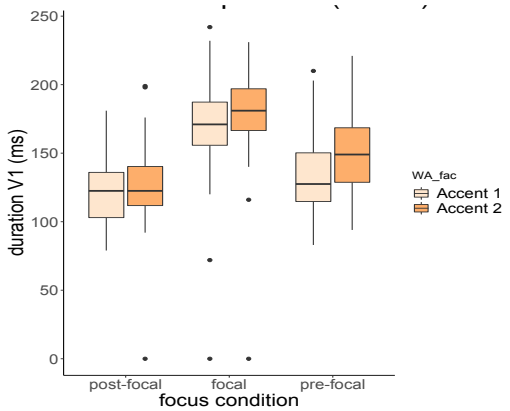


*Figure 5. Duration of the vowel /i/ produced with A1 (in "vinet") and A2 (in "viner") in utterance medial position in South Swedish (Interact. n.s, WA***, Diff WA =14 ms).*

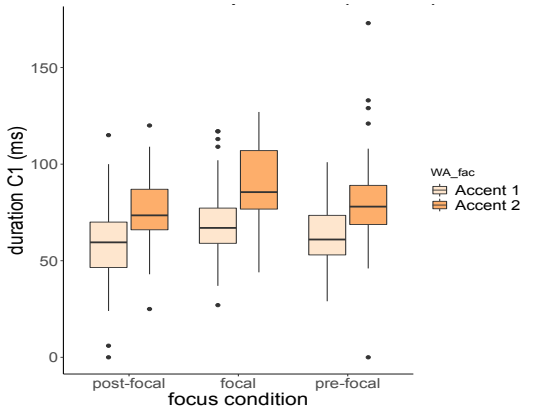


*Figure 3. Duration of the consonant /n/ produced with A1 (in "vinet") and A2 (in "viner") in utterance medial position in Central Swedish (Interact. n.s, WA***, Diff WA =18 ms).*

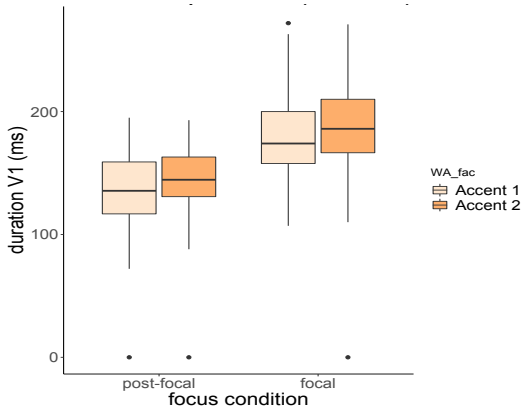


Figure 6 Duration of the vowel /i/ produced with A1 (in "bilen") and A2 (in "bilar") in utterance final position in South Swedish (Interact. n.s, WA, Diff WA =7 ms).*

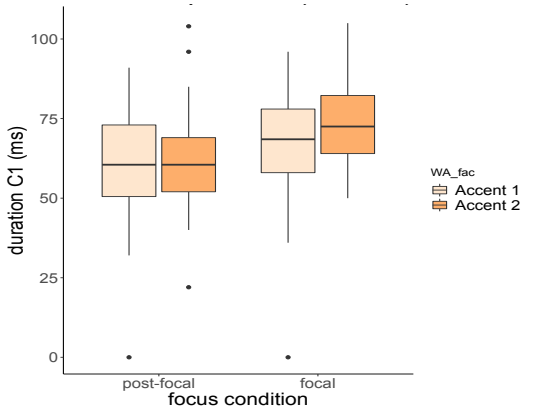


Figure 4. Duration of the consonant /l/ produced with A1 (in "bilen") and A2 (in "bilar") in utterance final position in Central Swedish (Interact., WA***, Diff WA =4 ms).*

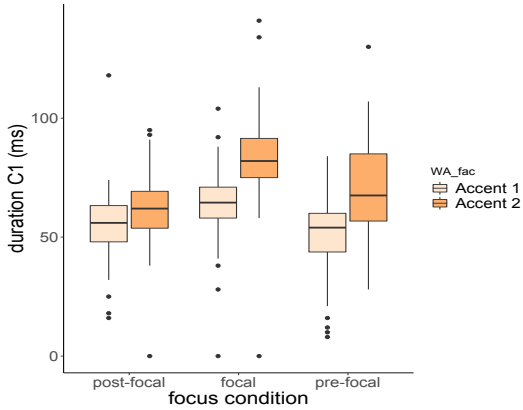


*Figure 7. Duration of the consonant /n/ produced with A1 (in "vinet") and A2 (in "viner") in utterance medial position in South Swedish (Interact. ***, WA***, Diff WA =10 ms).*

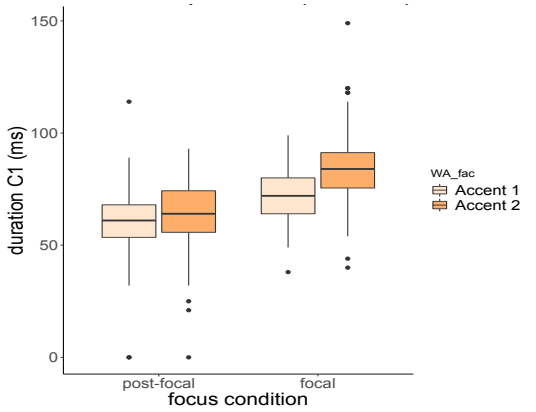


Figure 8. Duration of the consonant /l/ produced with A1 (in "bilen") and A2 (in "bilar") in utterance final position in South Swedish (Interact., WA***, Diff WA =8 ms*).

Discussion and conclusions

This study suggests that segmental durations constitute a systematic and robust correlate of the Swedish word accents, since longer durations were observed for Accent 2 than for Accent 1, irrespective of phrase-level factors. Furthermore, we found equivalent results for two different regional dialects which differ critically with respect to the tonal composition of the word accents. Thus, no obvious relationship of phone duration and tonal composition or complexity can be confirmed by the present results.

Tonal complexity, has, however, been identified as a common predictor of segmental durations in tone languages (e.g., Köhnlein, 2015). On the other hand, it has been argued that diachronic processes can eliminate durational differences originally conditioned by tonal complexity (Köhnlein, 2015). In a similar vein, we suggest, observable durational differences might have explanations beyond tonal complexity. For instance, the longer durations in Accent 2 might be a result of Accent 2 involving a lexical tone (Riad, 2006). This hypothesis could be tested in a future study by means of comparing Accent 2 words with lexical vs. post-lexical tone, as it has been suggested that Accent 2 is post-lexical in compounds (Myrberg & Riad, 2015; Riad, 2006).

Moreover, further phonological factors should be considered in future studies (such as the Swedish quantity distinction, cf. Heldner & Strangert, 2001), as well as further dialects exhibiting yet other variants of tonal patterning of the word accents. Finally, another question for future research is whether the small but significant durational differences between Accent 1 and Accent 2 are perceptually relevant, or if they rather reflect constraints of the speech production process.

Acknowledgments

This study was supported by a grant from Erik Philip-Sörensens stiftelse ("Acoustic dimensions in the melody of Swedish", AkuDim, 2018). The data used were recorded within a previous project granted by the Swedish Research Council (2009–1566). We cordially thank all speakers who volunteered reading weird sentences, as well as the participants of the Fino 2021 – where this work was presented earlier this year – for fruitful questions and comments.

References

Boersma P & Weenink D (2018). Praat: doing phonetics by computer. Computer program. http://www.praat.org/.

Elert C-C (1964). *Phonologic Studies of Quantity in Swedish. Based on Material from Stockholm Speakers*. Stockholm: Almqvist & Wiksell.

Heldner M & Strangert E (2001). Temporal effects of focus in Swedish. *Journal of Phonetics*, 29(3): 329-361. https://doi.org/10.1006/jpho.2001.0143

Köhnlein B (2015). The complex durational relationship of contour tones and level tones: Evidence from diachrony. *Diachronica*, 32(2): 231-267. https://doi.org/10.1075/dia.32.2.03koh

Myrberg S & Riad T (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics*, 38(2): 115–147. https://doi.org/10.1017/S0332586515000177

Riad T (2006). Scandinavian accent typology. *Sprachtypologie und Universalienforschung (STUF)*, 59: 36–55.

Svensson Lundmark M, Ambrazaitis G & Ewald O (2017). Exploring multidimensionality: Acoustic and articulatory correlates of Swedish word accents. *Proc. INTERSPEECH* 2017, Stockholm, Sweden. http://dx.doi.org/10.21437/Interspeech.2017-1502

Stress manipulation in text-to-speech synthesis using speaking rate categories

Christina Tånnander[1,2] and Jens Edlund[1]
[1]Speech, Music and Hearing, KTH
[2]Swedish Agency for Accessible Media, MTM

Abstract

The challenge of controlling prosody in text-to-speech systems (TTS) is as old as TTS itself. The problem is not just to know what the desired stress or intonation patterns are, nor is it limited to knowing how to control specific speech parameters (e.g. durations, amplitude and fundamental frequency). We also need to know the precise speech parameters settings that correspond to a certain stress or intonation pattern – over entire utterances.

We propose that the powerful TTS models afforded by deep neural networks (DNN's), combined with the fact that speech parameters often are correlated and vary in orchestration, allow us to solve at least some stress and intonation parts by influencing a single easy-to-control parameter, rather than detailed control over many parameters.

The paper presents a straightforward method of guiding word durations without recording training material especially for this purpose. The resulting TTS engine is used to produce sentences containing Swedish words that are unstressed in their most common function, but stressed in another common function. The sentences are designed so that it is clear to a listener that the second function is the intended. In these cases, TTS engines often fail and produce an unstressed version.

A group of 20 listeners compared samples that the TTS produced without guidance with samples where it was instructed to slow down the test words. The listeners almost unanimously preferred the latter version. This supports the notion that due to the orchestrated variation of speech characteristics and the strength of modern DNN models, we can provide prosodic guidance to DNN-based TTS systems without having to control every characteristic in detail.

Introduction

If we diminish speech and speaking to a simple one-way information transfer task, stress and intonation may not be as crucial as other speech features. No matter how hard we try to showcase Swedish minimal pairs such as "malen" (noun; Eng. "the moth"; Swedish accent I) and "malen" (past participle, Eng. "ground"; Swedish accent II), in real spoken communication, these are unlikely to cause actual misunderstandings. If intonation and stress were necessary for understanding, then we would not be able to make any sense at all of these minimal pairs in writing. But, in both writing and speech, the context is often sufficient to grasp the intended meaning, unusual intonation patterns notwithstanding.

A more interesting question is how we *perceive* speech with atypical or unexpected stress and intonation. Setting any discussion of dialects and native speaker variation aside, we find atypical stress and intonation in the speech of second language speakers, where it gives an accurate impression of what it is: accented speech. We also find it in text-to-speech synthesis systems (TTS), where it may be a result of text processing that is unaware of the relation between words, or a sound generation system that does not afford control of stress and intonation. Until recently, the general quality of TTS was poor enough that listeners would be implicitly aware that they were listening to a machine rather than a human, and the interpretation of misplaced stress and/or atypical intonation patterns would likely involve this fact. However, modern, DNN-based TTS such as Tacotron (Wang et al., 2017), is hard to tell apart from human speakers, and listeners may face a situation where they hear a perfectly human and native-sounding speaker make unusual intonation and stress choices.

Under these circumstances, misunderstandings may actually occur, but more importantly, the cognitive load involved in listening likely goes up as a result of these surprises. We may safely assume that listeners prefer and are helped by typical, expected stress and intonation patterns.

In this paper, we describe a way of controlling speaking rate by simple means in a DNN TTS, and propose that by controlling per-word speaking rate, we can target the desired stress and intonation patterns. We support this with a limited study in which we target the realisation of two well-known types of homograph pairs in Swedish: stressed verb particles/unstressed prepositions and stressed numerals/unstressed indefinite articles, assuming that the detection of these parts of speech is already solved, and believing that the same method can be applicable also on other phenomena where prominence is involved, such as contrastive focus.

Background

Spoken text

We focus here on *spoken text* (Tånnander & Edlund, 2019), and more specifically text that is read aloud with the assistance of a TTS voice. We assume a text preprocessing step that accurately analyses the structure of the text to be read, so that we can tell for example particle verbs in phrasal verbs apart from other verb + preposition constructs, and that we can tell the indefinite articles “en” and “ett” (Eng. “a/an”) apart from the numerals “en” and “ett” for number 1 (Eng. “one”) (these are homographs in Swedish).

The main driver for the work is practical – we want to be able to read texts aloud with TTS in such a manner that it can be understood with a minimum of cognitive load. The end goal is to increase the accessibility of Swedish texts by improving talking books that for a variety of reasons must be produced with TTS.

Talking books

Talking books are produced by the Swedish Agency for Accessible Media (MTM) for people with vision impairments or reading difficulties in accordance with Section 17 of the Swedish Copyright Act. The target groups studying at university level depend on a fast production of talking textbooks, and MTM produce around 1 000 Swedish and English talking books per year with TTS, as well as more than 100 newspapers (Tånnander, 2018).

Intonation and stress in TTS

Modern DNN-based TTS is trained on large amounts of data and often sound very human-like. The speech it produces reflects the prosodic features of the training data well, but may come out a bit bland, as it tends towards averages in the training data (Wang et al., 2018) Even if the desired prosodic patterns can be assigned a text in theory, the realisation of prosodic features is limited to what can be controlled in the TTS. And whereas more traditional TTS methods (e.g. formant and concatenative TTS) are to some degree transparent and allow for trouble shooting and manual correction of problematic utterances, the inner workings of the DNN are more obscure. This has radically changed the landscape of research into how speech characteristics in TTS can be controlled.

Examples of systems that control speech characteristics include Shechtman & Sorin (2019), who showed that they could control expressiveness and per-sentence speaking rate without losing quality and naturalness. Raitio et al. (2020) included pitch, pitch range, phone durations, energy and spectral tilt in their model, making it possible to control these features per utterance, independently of each other. The evaluation showed a significant decrease in MOS score when slowing down or speeding up the speech. There is a growing number of similar efforts around, but to date, we are not aware of any Swedish DNN-based TTS system that manages sets and generates the appropriate stress and intonation in spoken text.

Method

Assumptions and hypothesis

Assumption 1: Speech characteristics commonly associated with prominence, emphasis and stress include increased phoneme durations, higher fundamental frequency and increased amplitude, and these are correlated. We assume, that at least in some contexts, variations in fundamental frequency and amplitude can to some extent be predicted from variations in phoneme durations.

Assumption 2: DNN-based TTS learns, given a character string (e.g. an orthographic, phonemic or phonetic transcription), to produce the acoustic sequence (in actuality a spectrogram, but for our purposes, we will view it as a sound signal) that is maximally typical for what it has previously associated with similar strings.

Assumption 3: Intonation patterns are influenced by phrase stress, and to some extent predictable given that we know where it is located.

We hypothesise, then, that if we can influence syllable durations in DNN-based TTS for constituents that commonly occur in both stressed and unstressed positions, and with different meanings in the training material, we will trigger the TTS to choose a realisation that is appropriate for the stressed variety when we force a longer duration, and for the unstressed variety when we force a shorter duration. In these positions (at least), fundamental frequency and amplitude will follow duration to create realistic stress realisations, and the rest of the utterance will in turn be given a realistic intonation pattern given that stress. In short, we assume that *in DNN-based TTS, per-word control over duration is sufficient to control intonation patterns that are associated with phrase stress*

Training data

Almost 18 hours of a female professional voice was used as training data. The voice was originally recorded for a unit selection voice for non-fiction texts (Tånnander, 2018) and the voice talent was told to read the prompts in a clear and controlled manner, without too much prosodic variation. The training data was represented as phonemic transcriptions, including the Swedish stress markers (accent I, accent II, and secondary stress). Pauses within sentences were marked with '/' and word boundaries with '&'.

The data had been phonemically segmented in the preparation process for the unit selection voice, giving us access to the start and end points for each phoneme. The distribution of syllable lengths is shown in Figure 2. The average syllable duration was 247 ms.

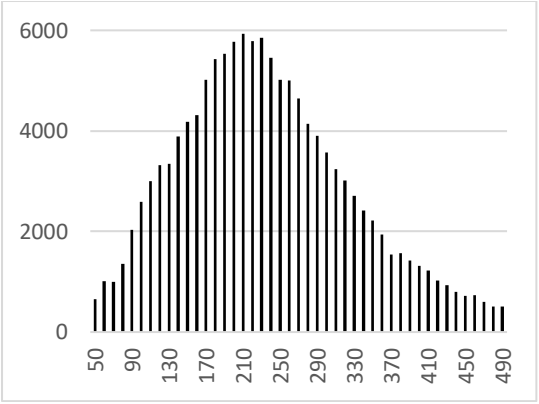


Figure 2. Syllable length (ms) distribution (x axis) and frequencies (y axis) in the training data. Syllable lengths occurring less than 500 times are not shown.

Preprocessing of the training data

All words in the training data were categorized into one of five approximately equal-sized speaking rate categories with around 27 000 words in each, as shown in Table 1.

Table 1. The five speaking rate categories. (1) is the fastest and (5) the slowest speaking rate.

Speaking rate category	Average syllable duration in word (ms)	Occurrences
1	< 150	27 152
2	151- 210	26 075
3	211- 250	28 042
4	251- 315	27 509
5	> 316	27 010

A label for the corresponding group was added to the word tokens in the training data. However, adding extra control features to the metadata makes it necessary to provide the same control features in the input to the TTS system. Here, we only want to control the durations of some of the words in the input, not all of them, so the model must support tokens without durational information as well. As a general fix that does not duplicate data or increase the size of the model, we added the speaking rate categories only to every second utterance in the training data, which in effect trains a model without durational data on half the training material. For the other half, the speaking rate category (1-5) was simply inserted before each word in the metadata with the phonemic representations of the speech. This allows us to guide the speaking rate not only of entire sentences, but of single words.

Note that this change requires no changes to the DNN structure or training framework, it all takes place in the preprocessing of the training data.

TTS training

The voice was trained in a PyTorch framework, and ran 1 500 epochs of decoder training (wav to mel spectrogram) with Tacotron (Wang et al., 2017) and 650 epochs of encoder training (mel spectrogram to wav) with the WaveGlow vocoder (Prenger et al., 2019).

Test data and setup

The test data contained 20 sentences, 10 with a verb particle that could potentially be confused with an unstressed preposition or another function word (see the example sentence in Table 2, and 10 with a numeral (*en* or *ett*, see Table 3),

which could potentially be confused with its homographic counterpart in Swedish, the unstressed indefinite article.

Table 2. Example of a test sentence with a phrasal verb with a verb particle.

Viktor	tänker	hoppa	av	gymnasiet.
Victor	intends to	drop out of		high school.

Table 3. Example of a test sentence with a numeral.

Ett	träd	har	fler	än	ett	löv.
A	tree	has	more	than	one	leaf.

All test sentences are shown in Table 4. Note that sentences where both readings would make sense, but mean different things, were avoided. The goal of this experiment was not to prove that ambiguity exists (in rare cases), and that prosody can be used to disambiguate in these rare cases. We already know this, and although it may be exciting from a research point of view, its value in real-world situations is small. Rather, we want to show that we can create more acceptable readings by guiding the TTS towards the expected prosody using simple means.

Each sentence was synthesized twice, once without any speaking rate categories (**VOICE1**), and once with the speaking rate category 5 before the verb particle or numeral, to signal that the test word should be pronounced in '*slow mode*', that is with longer durations (**VOICE2**). The verbs in the phrasal verb constructions were marked with the speaking rate category 2, to signal that the verb should be destressed. In this way, we hoped to guide the TTS towards a reading in which these words were made more generally more prominent, and be perceived as stressed, with the broader effect that the entire sentence melody was adapted to such a reading.

Experiment

20 native Swedish listeners participated, all employees at the Swedish Agency for Accessible Media, MTM.

A web-based *Comparison Category Rating* (CCR) test was used, in which participants listened to each sentence pair and marked one of them as preferred. The order of the sentence pairs was randomized, as was the order of the **VOICE1** and **VOICE2** within each pair.

*Table 4. The stimuli type, preferences for **VOICE1** and **VOICE2**, proportional preferences for **VOICE2**, and the sentence with the stressed verb particle or numeral in bold.*

Stimulus	Type	VOICE1	VOICE2	%	Sentence
1	VP	3	17	85%	Staten lyckades till slut komma **på** ett sätt att delta i finansieringen.
2	VP	2	18	90%	Jag tänker köra **på** så länge det lönar sig .
3	VP	2	18	90%	Alla visste **om** det utom jag.
4	VP	1	19	95%	Lisa steg **in** i salen.
5	VP	7	13	65%	Pelle gick **fram** och tog trofén.
6	VP	6	14	70%	Karina ska hälsa **på** sin guddotter på söndag.
7	VP	5	15	75%	Vet du **om** att det är torsdag imorgon.
8	VP	5	15	75%	Det kan hända att du inte tycker **om** det.
9	VP	1	19	95%	Pernilla måste lägga **av** med oförskämdheterna.
10	NUM	0	20	100%	Viktor tänker hoppa **av** gymnasiet.
11	NUM	0	20	100%	Allt städades med undantag av **ett** rum.
12	NUM	4	16	80%	Rickard åt bara **en** enda kaka, sen var han nöjd.
13	NUM	0	20	100%	Ett träd har inte bara **ett** löv.
14	NUM	9	11	55%	**Ett** av borden på gräsmattan är rangligt.
15	NUM	2	18	90%	Det är **en** sida av saken.
16	NUM	3	17	85%	Pojken fyller **ett** om två veckor.
17	NUM	0	20	100%	Det var bara **en** person som anmälde sig.
18	NUM	3	17	85%	Ett plus **ett** är lika med två.
19	NUM	15	5	25%	Fyra hästar, tre hönor, **en** kossa och sju grisar.
20	NUM	0	20	100%	Femte paragrafen stycke **ett** till tre.

Results

The readings in which the test word (a verb particle or numeral) was guided towards a slower pace (**VOICE2**) were preferred 332 of 400 times (83%: 82% for numerals and 84% for verb particles). The result for each stimulus is presented in Table 4.

All 20 individual participants preferred **VOICE2**, and the participant that chose **VOICE1** the most times did so in only 7 of 20 comparisons. The participants were free to listen to the stimuli pairs in any order, and were also allowed to listen to them as many times they wanted to, so effects of order are not well controlled for. Regardless, the voice presented last was preferred in 57% of the cases. **VOICE2** scored higher than **VOICE1** in all individual stimuli pairs except one, where **VOICE**1 was preferred 15 of 20 times (stimulus 19). In one other pairwise comparison the result was close: 45% preferred **VOICE1** and 55% **VOICE2.**

Discussion

The results are straightforward: participants preferred **VOICE2**, and there was not much variation among users or stimuli. Only one stimuli pair showed preferences for **VOICE1** (75%), see stimuli 19 in Table 4. This can be explained by the fact that the sentence (*"Four horses, three hens,* ***one/a*** *cow and seven pigs."*) was a mistake on our behalf: it actually works well without treating *'en'* as a numeral. The TTS simply did a better job with that reading of the sentence.

Another stimulus that stands out is number 14, where 55 % preferred **VOICE2** (*"****One*** *of the tables on the lawn is rickety."*). Here, the numeral *'ett'* is sentence initial and followed by a preposition, which gives a hint that 'ett' should be stressed. The Tacotron may just have learned this example from the training data without guidance.

Apart from that, the result is as expected: without guidance, the system produces the statistically most common reading of the test words: unstressed function words and indefinite articles. This produces perfectly comprehensible readings, but with an intonation that is unexpected. With a nudge towards a reading where the test words are lengthened, the system produces an utterance that is more aligned, as a whole, to that lengthening and therefore preferrable to the listeners.

Conclusions and future work

The results show that it is possible to create TTS readings of Swedish verb particles and numerals that are homographs to various function words and indefinite articles, by guiding the TTS model towards a lengthening of these words, and that the resulting overall reading is preferrable to one where no such guidance was provided. We argue that without the lengthening, the more common unstressed reading is likely to be chosen, while the lengthening causes the models to choose a path that is common for slower readings of these words: the more prominent verb particles and numerals.

It should be noted that the differences in stress level for the verb particles and numerals are not isolated. In fact we hope that the proposed method is a viable way to provide prosodic guidance to a DNN-based TTS system without having to control every parameter in detail (something we simply do not know how to do, in part since we do not know what the correct targets are). We believe that the method can be used to set word prominence in other situations, such as emphasising new information.

Acknowledgments

This work is funded in part by the Vinnova funded project *Deep learning based speech synthesis for reading aloud of lengthy and information rich texts in Swedish* (2018-02427). The results will be made more widely accessible through the Swedish Research Council funded national infrastructure Språkbanken Tal (2017-00626).

References

Prenger, R, Valle, R, & Catanzaro, B (2019). Waveglow: A flow-based generative network for speech synthesis. Proc. of ICASSP 2019, 3617–3621. https://doi.org/10.1109/ICASSP.2019.8683143

Raitio, T, Rasipuram, R, & Castellani, D (2020). Controllable neural text-to-speech synthesis using intuitive prosodic features. Proc. of Interspeech 2020, 4432–4436. https://doi.org/10.21437/Interspeech.2020-2861

Shechtman, S, & Sorin, A (2019). Sequence to sequence neural speech synthesis with prosody modification capabilities. Proc. of the 10th ISCA Speech Synthesis Workshop, 275–280. https://doi.org/10.21437/SSW.2019-49

Tånnander, C (2018). Speech Synthesis and evaluation at MTM. Proceedings of Fonetik, 75–80. https://www.ling.su.se/fonetik-2014/tidigare-konferenser

Tånnander, C, & Edlund, J (2019). First steps towards text profiling for speech synthesis. Proc. Digital Humanities in the Nordic Countries 2019 (DHN2019).

Wang, Y, Skerry-Ryan, R J, Stanton, D, Wu, Y, Weiss, R J, Jaitly, N, Yang, Z, Xiao, Y, Chen, Z, Bengio, S, Le, Q, Agiomyrgiannakis, Y, Clark, R, & Saurous, R A (2017b). Tacotron: Towards end-to-end speech synthesis. Proc. of Interspeech 2017, 4006–4010. https://doi.org/10.21437/Interspeech.2017-1452

Wang, Y, Stanton, D, Zhang, Y, Skerry-Ryan, R J, Battenberg, E, Shor, J, Xiao, Y, Ren, F, Jia, Y, & Saurous, R A (2018). Style Tokens: Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. International Conference on Machine Learning, 2018. https://arxiv.org/abs/1803.09017

'superassp' and the need for a sparse signal track data format

Fredrik Karlsson
Department of clinical science & Humlab, Umeå University

Abstract

The digital processing of phonetic data has long been based on the same basic data structures and consequently been stored in essentially the same manner. The transition to structured database-based workflows and the development of more complex analyses challenge the established data structures and storage models. I propose a sparsely defined data track model for structured storing of locally define signal information and suggest an implementation. The use of sparse data created by the amalgamating signal processing library 'superassp' is discussed within a speech database management system.

Background

The digital storage of information for use in phonetic research has remained stable for many years now, and this paper will develop the rationale for extending the existing models with an additional storage type, in order to afford consistant management of speech recordings and derived signals.

With time-aligned signal tracks being the primary format used for storing recordings obtained directly from a speaker or computed as indirect tracks, possibly by windowing of the original signal. While time-aligned, the sampled or derived signal tracks may certainly be multidimensional in nature, but the values are predominately of the same origin in terms of the analysis used. That is, while a spectrum is multimensional, each value shares the same origin, and we may define them continuously throughout a signal and store them as a spectrogram without considering how to label the individual values. This property may not hold in all cases in the future, as we will discuss later on.

Of further interest to us is how the signal is sampled. The sampling of the speech signal at a uniform rate is the predominant model for aligning measurements against the timeline, and while variable rate would be feasible, it could be argued that such an extension would offer little advantage for signal track data. We take the storage of electroglottography data as an illustrative example. Electroglottography is the investigation of the electrical impedance of the larynx as an indirect measure of vocal fold contact area. The impedance is sampled continuously, but of interest is, however, the dynamics of vocal fold closing and opening within a full cycle. The frequency at which the vocal folds describe an entire cycle varies over time when speaking, and the time frame for which derived measures (e.g., the closed quotient) offer a valid description of the activity will therefore vary in length. If sampled at a modern sampling rate, however (44100 Hz or above), the time resolution of glottal events will be > 70 samples per glottal cycle even for speakers with a high (600 Hz) f0. Since the output of analysis for a glottal cycle is relatively simple it may still be efficiently stored as identical values repeated throughout all analysis frames within the glottal cycle. An electropalatography signal may similarly be sampled relatively infrequently, but resampled to fit the timeline of an audio recording of speech since the number of stored channels is reasonably small and the sampling rate is known. It is observed, therefore, that variable rate low-frequency information may be shoehorned into a continuous track model without much loss in efficiency, provided that the data is relatively simple and analysis windows do not overlap.

Speech information with potential overlap and no periodicity

It has long been recognised that storage of information attached to the speech signal by a human as an annotation in some form requires a storage form that is separate from signal track. The reasons for differentiating between these two information forms are of particular interest for the argument developed in this paper. First, it is observed that the need for a separate data format

for human added annotation is driven by the fact that the timing of where the response is started to be defined has no periodicity, and further that the portion of the timeline of a recording for which the where the response is valid is highly variable. In short, we can rarely predict exactly where a human would like to add a note related to their perception of a signal, and exactly where the perception ends. While it would be possible to encode the human perceptions as being present in a portion of a continuously defined signal track, it has long been recognized that a single start end end point noted in a text file along with a textual note (label) is a more effient encoding.

Second, human attached information may very well partially or completely overlap in relation to the recording session timeline. While one could have opted to find the solution to the problem of storing this information also in variable-rate signal track formats and multiple tracks (or multiple signal files), such an implementation would have been wasteful, and the solution has instead been the storage of human transcription data as time slices with a text label that is placed in different tiers in a collection.

One could note that the results of human perceptual experiments that use part of a speech signal as stimulus could be also stored in the same format, as they could be aligned with the timeline as soon as one could compute a time of a response. However, as perceptual experiment captures responses of many participants that likely will overlap, separate tiers will have to be constructed for to store participants' responses, which would make management unwieldy. It is observed that while new types of speech information may be shoehorned into one of the well established storage models within a database context, we are likely better served by acknowledging the lack of fit and reconsider our thinking of the data we store.

Sparsely defined signals and when we may need them

We have seen that we currently possess a data format for working efficiently with regularly defined signals, which may be multidimensional but remain sequential in their arrangement, and different formats for storing sparsly defined or data with overlap, but which handled multidimensionality less flexibly. I will illustrate that developing analysis techniques may result in data that, while possible to shoehorn into the best fitting of these two models, may demand a separate mode of storage when striving to take advantage of the possibility of structured and repeatable analysis offered by speech database management systems.

The VoiceAnalysisToolbox is an analysis package aimed at establishing possible acoustic markers of voice production deterioration due to a disease. It takes a single prolonged vowel as an input and computes 339 acoustic measures, including variants of jitter/shimmer, MFCCs, and their first and second-order derivatives, summarises a wavelet decomposition of the pitch track, computes closed quotient of the vocal fold cycle from inverse filtering, and so on. The analysis is implemented in compiled Matlab code and takes from 10 to 60 times the duration of the signal to complete. We have no evidence that the output of the procedure is identical across all samples of the same speaker, so we may assume that a well-defined research project would want to analyse either multiple prolonged vowels for the same speaker, multiple portions of the prolonged vowel in, or both.

So, after having up to 3 minutes to complete an analysis of a single prolonged vowel, where do we store the output so that it may be retrieved efficiently from a database later on? Having only 339 output values, the output itself could relatively easily be stored in manner similar to a spectrogram, but as the analysed regions may both vary in size and overlap, the management of the unpredictably many signal files may prevent efficient use of the information later on. Further, as the information that requires storage may me of different natures (e.g. MFCCs and wavelet features mixed) a labelling strategy may be required which signal files usually do not have. Therefore, the user will be burdened with the additional task of carefully indexing when retrieving the required information.

Similarly, the storage model that we now use for holding transcriptions is not well suited for storing the data. In their current form, files that we now use for holding human annotations (transcriptions) associated with a speech signal are not well suited for holding arrays of 339 values. Further, the serialization of values into a form that travels well between programming languages while remaining efficient becomes an issue, as does the task of making sure that the specifications of data stores and considered comparable actually does contain the same order columns and in the same order.

Listing 1. The definition of a table that holds slices. For each slice, the start and end sample of the slice needs to be defined, along with a checksum of the file content ('hash). The columns holding measurements to be stored for the slice are indicated as "[…]". The sample rate of the signal file is also required, which ensures that the sample number can be converted to time if needed.

```
CREATE TABLE slices (
    `start_sample` INTEGER NOT NULL,
    `end_sample` INTEGER NOT NULL,
    `samplerate` INTEGER NOT NULL,
    `hash` TEXT NOT NULL,
    […]
    PRIMARY KEY (start_sample, end_sample, sha)
);
```

The implementation of sparse signals superassp

'superassp' is an R speech signal processing library aimed at bringing together algorithm implementations of various sources and mold them so that the output could be used consistently by a speech database management system. It was originally contrived as an extension of the wrassp R package for use within the EMU Speech Database Management System (Emu). The `superassp` package aims to bring together a very heterogeneous collection of signal processing algorithms implemented in R (wrassp; Bombien, Winkelmann, Scheffers, 2021), C, Praat, Matlab™[1] and Python under a common interface. Regardless of origin, most algorithms will be made to produce an SSFF track when repackaged for use within superassp. For the output of the VoiceAnalysisToolbox procedure, however, a sparse signal track format is implemented in a way that deliberately makes it incompatible with the SSFF format and precludes direct use within the Emu system. Instead, the sparse track format is implemented as an SQLite database file. The choice of this database file format is arbitrary, but it may be observed that it is arguably the most used database, easily accessible across platforms and programming languages, and is reasonably performant.

In the SQLite format, the sparse collection of possibly overlapping slices for which data needs to be stored receives the simple definition presented in Listing 1. The simple definition allows for one sparse slice file to hold multiple slices with the same specifications (that is, the same set of measures). Identically structured measurements from partially overlapping slices may be inserted into the same sparse slice file, but the slices are uniquely defined by their start and end sample and the content of the file it was computed from, and duplicate definitions cannot be inserted. Instead, the application needs to determine at insert time how this situation should be handled. It is likely that most insert operations will be (SQL) `INSERT OR REPLACE` statements so that revised measurements will overwrite previously stored information for the slice.

It should be noted that, however, that slices are unique only for particular file content. Thus, it is possible to store the result of multiple assessments of the same portion of a signal timeline but with altered signal content (such as a preceding filtering operation) in the same sparse slice file, which may prove convenient in studies of algorithm robustness.

The use of sparse signals in speech databases

An implementation of sparse signals is, to my knowledge, not available in a speech database system currently. As 'superassp' is being developed as a complementary library for use in the Emu SDMS (Winkelmann, Harrington, & Jänsch, 2017), we will consider some aspects of sparse signals in that context. Emu has a client-server architecture, with the transcription interface being implemented as a web application and transcriptions transferred as JSON files. Since sparse signals are found above to be share properties with transcription elements in terms of time specification while demanding higher dimensionality, the mode of transfer and visualization of the data to the user should be considered. In Listing 2, a tentative JSON format for transferring a sparse signal containing parts of a voice report for two portions of the speech signal to the client is illustrated. How the data should be presented is, of course, up to the developer of the receiving client. In the context of the Emu SDMS, one could note that 2D panel of the web client could be set up to show the data efficiently to the user. In the case of overlapping slices, such as in Listing 2, multiple colors may

[1] We are currently investigating how we can expose algorithms implemented in Matlab™ within our license agreement and will do so if possible.

be used to differentiate between the first and additional slices. As the user moves the cursor into the prolonged vowel, the data for the larger slice is shown intially, and as the user moves the cursor into the region also covered by the second slice, additional and differently colored data points may be injected into the display. The 2D panel of the Emu web client is already well equipped to display just some columns of a specification and leave some undisplayed. Therefore, we see that just a portion of a slice, such as some MFCCs that are particular interest for the analys, may be displayed to the user when available in an efficient manner.

Listing 2. An example of selected measures of two temporally overlapping voice reports in JSON format for transfer between applications.

```
[
  {
    "start_sample": 44100,
    "end_sample": 132300,
    "data": {
      "Jitter (local)": 0.401,
      "Jitter (rap)": 0.171,
      "Jitter (ppq5)": 0.209,
      "Jitter (ddp)": 0.514,
      "Shimmer (local)": 2.131,
      "Shimmer (apq3)": 1.028,
      "Shimmer (apq5)": 1.347,
      "Shimmer (dda)": 3.084
    }
  },
  {
    "start_sample": 66150,
    "end_sample": 110250,
    "data": {
      "Jitter (local)": 0.469,
      "Jitter (rap)": 1.207,
      "Jitter (ppq5)": 0.233,
      "Jitter (ddp)": 0.624,
      "Shimmer (local)": 2.341,
      "Shimmer (apq3)": 1.207,
      "Shimmer (apq5)": 1.456,
      "Shimmer (dda)": 3.622
    }
  }
]
```

Conclusion

I have argued that acoustic data that are sparsely defined, potentially overlapping in time of definitiona but which may be multidimensional in nature, do not confidently fit in the data formats traditionally used in speech research and in speech databases. I have illustrated how they might be thought of in storage by presenting the implementation considered for the 'superassp' library. While visualization of exessivelly overlapping sparsely defined multidimensional data to the user will likely offer several challenges, I argue that a meaningful presentation is feasible for at least a subset of the data, depending on the fascilities made available by the receiving client software.

Acknowledgments

The research presented here has been enabled by the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

References

Bombien L, Winkelmann R, Scheffers M (2021). *wrassp: an R wrapper to the ASSP Library*. R package version 1.0.0.

Winkelmann, R., Harrington, J., & Jänsch, K. (2017). EMU-SDMS: Advanced speech database management and analysis in R. *Computer Speech & Language*, *45*(Supplement C), 392–410. doi: 10.1016/j.csl.2017.01.002

Tsanas, A. (2012). *Accurate telemonitoring of Parkinson's disease symptom severity using nonlinear speech signal processing and statistical machine learning*. Doctoral dissertation (DPhil), University of Oxford.

Phonemic and subphonemic cues in prediction: Evidence from ERP, eye-tracking and Danish words with and without stødbasis

Anna Hjortdal & Mikael Roll
Centre for languages and literature, Lund University

Abstract

The brain is constantly trying to predict the future and phonological and prosodic cues are used to anticipate forthcoming information. Even cues on the subphonemic level such as vowel transitions, nasalisation and assimilation across word boundaries are useful in anticipating upcoming speech. In event-related potential (ERP) studies examining subphonemic and lexical/phonological mismatches, only the latter yielded N400 effects, an ERP component associated with lexical prediction error. The results indicate that phonetic cues are resolved prelexically. However, subphonemic cues still seem to be used in prediction as evidenced by valid cues yielding faster fixations in eye-tracking studies and invalid cues modulating P600 amplitudes, indicating structural violations and context updating.

Introduction

In recent years, the idea that prediction plays a vital role in language processing has won terrain (DeLong et al., 2005; Kutas et al., 2011). Within this framework, linguistic items are pre-activated before being perceived. The predictions are based on 'subjective Bayesian probability' which is the believed probability of certain events based on previous experiences (Bar, 2007; Friston, 2005). Predictions are carried from cognitively higher to lower levels. If predicted features fail to manifest themselves, prediction errors are carried back to higher levels (Rao & Ballard, 1999).

Several studies have found that lexical, phonological and prosodic structures such as whole words (DeLong et al., 2005; León-Cabrera et al., 2017), word-initial phonemes (Roll et al., 2017; Söderström et al., 2016), word tones (Roll, 2015; Roll et al., 2017; Roll et al., 2015; Söderström et al., 2016; Söderström et al., 2017b; Söderström et al., 2017a) and syntactically-related tones (Söderström et al., 2018) can function as cues to the input string. Evidence for this comes, among other things, from event-related potential (ERP) studies showing that unexpected structures yield ERP components associated with prediction error such as the N400 and P600. Further, more predictively useful cues produce an electrically more negative pre-activation negativity (PrAN) deflection, a component modulated by the predictive strength of phonological cues (Roll et al., 2017; Söderström et al., 2016). Even cues on the subphonemic level appear to be used in prediction. Response time, eye-tracking, event-related potential (ERP) and magnetoencephalographic (MEG) studies have shown that listeners are sensitive to fine-grained acoustic differences on the subphonemic level (Archibald & Joanisse, 2011; Flagg et al., 2006; Grosvald & Corinna, 2009; Martin & Bunnell, 1981, 1982; McQueen et al., 1999; Mitterer & Blomert, 2003; Streeter & Nigro, 1979; Warren & Marslen-Wilson, 1987) and such cues appear to be used actively to predict upcoming structures (Beddor et al., 2013; Dahan et al., 2001b; Salverda et al., 2014). In the following, we wil review the literature on how subphonemic (phonetic, coarticulatory) cues affect spoken word recognition and prediction and discuss subphonemic versus phonological cues. Further, we will discuss potential implications for how the prosodic features Swedish word accents and Danish stød are to be understood.

Phonetic cues

Listeners are sensitive to subtle phonetic cues. 'Subcategorical phonetic mismatches', as they were termed by Whalen (1984), are phonetic cues spliced into a new environment in which they conflict with existing cues but are not enough to change phonemic identity. In a lexical decision task with English speakers, Streeter and Nigro (1979) found that response times were

faster when VC formant transitions were valid than invalid, even when there was no difference in the intelligibility of words. Whalen (1991) obtained similar results for lexical decisions but observed that effects of subcategorical mismatches were reduced in auditory naming tasks.

Marslen-Wilson and Warren (1994) and McQueen et al. (1999) examined the influence of the lexical status of invalid cues for speakers of English and Dutch respectively. In both experiments, CVC words were cross-spliced so that some stimulus words had valid coarticulatory cues, some had formant transitions from a lexical competitor word (e.g. English *jog* with formant transitions from *job*) while others had formant transitions from a non-word (e.g. *jog* with formant transitions from *jod*). Words with valid cues were identified faster than those with invalid cues, but there was no difference in response latencies between the two different invalid conditions. However, in an eye-tracking study, Dahan et al. (2001b) found that participants were the slowest at fixating on a target picture when transitions came from a competitor word, somewhat faster when they came from a non-word while the fastest response times were found for words with valid transitions. This was interpreted as evidence for lexical competition, the activation of a word also depending on activation of competing candidates. Thus, activation of the target word would be inhibited by activation of a competitor, or, from a predictive processing point of view, activation of a competitor by mismatched phonetic cues yielded misguided predictions, thus slowing down fixations.

Similar results have been obtained for other types of coarticulation. In studies with speakers of English, mismatched presence or absence of nasalisation during a vowel preceding an oral or nasal consonant yielded longer response times (Fowler & Brown, 2000) and, for oral consonants invalidly cued by a nasal vowel, a delay in neuromagnetic activity in a magnetoencephalographic (MEG) study (Flagg et al., 2006). Nasal consonants following oral vowels were also delayed, but not significantly, probably reflecting that in English, nasalised vowels are stronger predictors of an upcoming nasal consonant than are oral vowels of oral consonants because a nasal following an oral vowel is still viable. In an eye-tracking study, also with speakers of English, Beddor et al. (2013) found that listeners fixated on images with nasal consonants (e.g. *send* rather than *said*) faster when nasalisation started already during the vowel. When eye movement programming delay was taken into account, listeners started fixating on target words before the onset of a nasal consonant, indicating that listeners take advantage of coarticulatory cues to predict what is further down the input string.

Listeners are also sensitive to coarticulation across syllable (Martin & Bunnell, 1981, 1982) and word boundaries (Gow, 2003). Martin and Bunnell (1981, 1982) found that vowel-vowel subcategorical phonetic mismatches across syllables yielded longer response times. They speculated that information already heard was used to predict outlines of the signal yet to come. In a combined behavioural and mismatch negativity (MMN) ERP study, Grosvald and Corinna (2009) observed that listeners were sensitive to vowel-to-vowel coarticulation across three intervening segments. Some listeners were sensitive to coarticulation across as much as five segments. Salverda et al. (2014) showed that coarticulation during English definite articles allowed listeners to fixate more rapidly on a target word, suggesting that short and phonetically reduced function words such as *a* and *the* play an important role in facilitating processing of following content words in English.

Listeners appear to continuously make use of what cues are available. In an eye-tracking study, McMurray et al. (2008) found that early (voice onset time (VOT) and formant transition slope) and late (vowel length) cues to voicing and manner contrasts in English modulated the probability of eye movements to pictures of target and competitor words as these cues became available. Even children are sensitive to phonetic cues (Cross & Joanisse, 2018; Paquette-Smith et al., 2016; Zamuner et al., 2016). In an eye-tracking study, Paquette-Smith et al. (2016) found that English-speaking 2-year-old children are sensitive to subphonemic mismatches, but word recognition was less disrupted than for phonemic mismatches. The findings indicate that listeners use coarticulatory cues to constantly make and update predictions. The next question is whether segments invalidly cued by subphonemic cues, such as subcategorical phonetic mismatches, produce the same neural error signals that have been observed for phonological cues.

Subphonemic mismatches and prediction error

In a combined response time and ERP study with speakers of English, Archibald and Joanisse (2011) examined lexical, phonemic and coarticulatory mismatches. Participants looked at colour stock photographs on a screen and listened to stimuli. They were asked to answer whether the picture and word matched by pressing buttons on a keypad. Stimulus words came in five conditions: 1) coarticulatory and lexical match (e.g. picture of a hat, heard *h^{a}at*), 2) coarticulatory and lexical mismatch (e.g. picture of a hat, heard *h^{o}ot*), 3) coarticulatory match and lexical mismatch (e.g. picture of a hat, heard *h^{a}ot*), 4) lexical match and coarticulatory mismatch (e.g. picture of a hat, heard *h^{o}at*) and 5) unrelated (e.g. picture of ship, heard *h^{a}at*). They found that only lexical mismatches yielded N400 effects, an ERP effect associated with lexical prediction error (DeLong et al., 2005) while both conditions yielded an increased negativity between 230 and 310 ms after word onset, interpreted as a phonological mismatch negativity (PMN). The PMN is modulated by pre-lexical phonological processing, although its status as a separate component is debated due to inconsistencies in reported topography, timing and sensitivity (Lewendon et al., 2020). No differences in response times were reported, except for the unrelated condition. Archibald and Joanisse (2011) interpreted the findings as evidence that subphonemic information does not influence word-level selection but is processed at the prelexical level.

Hjortdal and Roll (submitted) examined how phonetic, coarticulatory cues and contrastive, phonological cues interact in a combined response time and ERP study. Effects from the two phases of the Danish creaky voice feature, stød, were isolated in a cross-splicing design. The first phase displays phonetic differences in e.g. pitch while the second phase, realised as creaky voice, is the phonological locus of stød (Basbøll, 2014). Stød can distinguish word meanings but is also associated with specific morphological structures (Basbøll, 2005). For instance, monosyllabic nouns which have stødbasis and thus support stød have stød in definite singular, but lose stød when pluralised with *-e*. The presence or absence of stød during a stem can therefore cue upcoming information. Words were cross-spliced, occurring in eight different conditions in singular and plural and cued by 1) valid phonetic and stød/non-stød cues, 2) valid phonetic but invalid stød/non-stød cues, 3) invalid phonetic but valid stød/non-stød cues and 4) invalid phonetic and stød/non-stød cues. Words lacking stødbasis, which do not support stød due to sonority constraints, were included as controls. Such words do not attain stød when they engage in morphological constructions which, in words with stødbasis, would lead to stød/non-stød alternations. However, words without stødbasis still display small and consistent phonetic differences. Invalid phonological stød/non-stød cues yielded lower response accuracy and increased response times, in line with previous findings (Clausen & Kristensen, 2015). Phonetic cues in the control condition did not affect neither accuracy nor response times while phonetic cues in the stød condition only affected response times in the absence of a valid phonological cue. Invalid phonological cues yielded N400 and P600 effects while invalid phonetic cues led to no such effects in the stød condition. In the control condition, singular suffixes invalidly cued by plural stems yielded a P600 effect.

Pre-activation negativity

While the ERP components N400 and P600 have been associated with prediction error, the pre-activation negativity (PrAN) has been interpreted as an index of the actual pre-activation of linguistic information (word endings or syntactic structure) and is modulated by the predictive strength of phonological cues (Roll, 2015; Roll et al., 2015; Söderström et al., 2016). PrAN amplitudes are higher for word beginnings (i.e. the first 2-3 phonemes of a word) with highly frequent continuations and few lexical competitors. Also, Swedish accent 1, which is associated with much fewer continuations than accent 2, produces more negative PrAN amplitudes, indicating that PrAN reflects the degree to which continuations can be predicted (Roll et al., 2017; Söderström et al., 2016). PrAN has been identified for prosodic cues such as word initial fragments, word tones and syntactically-related tones (Roll, 2015; Roll et al., 2017; Roll et al., 2015; Söderström et al., 2016; Söderström et al., 2017b; Söderström et al., 2017a).

Studies using fMRI have shown increased activity for pre-activation (PrAN) in the primary auditory cortex and surrounding areas between

70 and 150 ms after stimulus onset (Roll et al., 2015; Söderström et al., 2018), which is thought to reflect stronger activation of more predictively useful forms. Later, after 200 ms, PrAN correlates with activity in Broca's area. This activation has been interpreted as reflecting selection through inhibition of irrelevant candidates (Roll et al., 2017; Roll et al., 2015; Söderström et al., 2017b). In the Hjortdal and Roll (submitted) study with Danish stød, plural first phases yielded early PrAN effects while stød yielded late PrAN amplitudes, the latter reflecting that stød occurs under more constrained conditions than non-stød and thus is a better predictor.

Discussion

In both ERP studies examining phonological/lexical and subphonemic cues (Archibald & Joanisse, 2011; Hjortdal & Roll, submitted), N400 modulations were reported for lexical/phonological mismatches while subphonemic mismatches yielded no such effects. This could be interpreted as support for the proposal that subphonemic cues are processed prelexically without constraining lexical processing (Archibald & Joanisse, 2011). While the N400 is modulated by differences on the lexical level, the P600 has been reported for violations in form and structure (Osterhout & Holcomb, 1992; Rodriguez-Fornells et al., 2001; Roll et al., 2010; Sassenhagen et al., 2014). The component can be understood as context updating in terms of morphological and syntactic structure (Sassenhagen et al., 2014). N400 effects have been reported for stimulus words differing from a target image by just one phoneme. For instance, Desroches et al. (2009) and Archibald and Joanisse (2011) reported N400 modulations when participants looked at e.g. an image of a cone while hearing *comb* or looking at an image of a hat while hearing *hot.* Further, an N400 effect was reported when the Danish creaky voice feature stød/non-stød did not match definite singular/indefinite plural suffixes.

It might be that N400 effects occur when a word strongly inhibited or perhaps even ruled out from lexical competition, e.g. due to a speech error or for experimental reasons, is eventually recognised. Such a re-entrance might show up in the ERP signal as an N400 effect, a signal of lexical prediction error (DeLong et al., 2005). Subphonemic phonetic mismatches, on the other hand, might not lead to words dropping out completely from lexical competition. However, phonetic cues still appear to be used in prediction, as evidenced by eye-tracking studies showing that valid phonetic cues yield faster fixations on target words (Beddor et al., 2013; Dahan et al., 2001b; Salverda et al., 2014) and the P600 component reported in Hjortdal and Roll (submitted), indicating prediction error and the PrAN, indicating that phonetic cues were used for prediction.

If this interpretation is correct, it has implications for how the roles of Swedish word accents and Danish stød should be interpreted. Swedish words have either a high or a low tone on the stem. Like Danish stød, the tones (word accents) can distinguish meanings but can also be induced by suffixes and thus cue word endings (Riad, 2014; Rischel, 1963; Roll et al., 2010). P600 effects have been reported for Swedish word accents invalidly cuing suffixes (Gosselke Berthelsen et al., 2018; Roll, 2015; Roll et al., 2010; Roll et al., 2013; Roll et al., 2015) while N400 effects have only rarely been reported (Gosselke Berthelsen et al., 2018). It might be that only Danish stød, and not Swedish word accents, takes itself into what is traditionally known as the lexically contrastive level. As mentioned above, stød as well as word accents can distinguish word meanings as well as function as a cue to upcoming information. However, the phonetically reduced structure of Danish might give stød an even more prominent role. In spontaneous speech, a final schwa is assimilated to the preceding vowel or consonant, often resulting in almost identical singular and plural forms, e.g. *land* [lanˀ] 'country' and *lande* [lann̩] 'countries' (Grønnum, 2005). However, if the singular has stød, the forms are kept apart (Basbøll et al., 2011). The N400 modulation reported for stød/non-stød-suffix mismatches could therefore indicate that stød is becoming lexicalized for monosyllabic singular nouns in Danish – and more so than Swedish word accents. Further support for this interpretation comes from the finding that for at least one participant in the stød study, the presence or absence of stød, rather than suffixes, appeared to be the principal factor in determining whether a test word was singular or plural (Hjortdal, 2021).

To sum up, listeners do seem to use subphonemic cues to predict upcoming information. Such cues appear to be handled prelexically without constraining lexical processing, as proposed by Archibald and Joanisse (2011). Further, Swedish word accents

would also appear to be dealt with prelexically, only rarely making their way to the lexical level, while an N400 effect suggests that Danish stød does make its way to the lexical level.

Acknowledgements

This work was supported by Knut and Alice Wallenberg Foundation (grant number 2018.0454), Swedish Research Council (grant number 2018.00632), Crafoord Foundation (grant number 2017.0006) and Marcus and Alice Wallenberg Foundation (grant number 2018.0021).

References

Archibald, L M, & Joanisse, M F (2011). Electrophysiological responses to coarticulatory and word level miscues. *J Exp Psychol Hum Percept Perform*, *37*(4), 1275-1291.

Bar, M (2007). The proactive brain: using analogies and associations to generate predictions. *TRENDS in Cognitive Sciences*, *11*(7), 280-289.

Basbøll, H (2005). *The Phonology of Danish*. Oxford University Press.

Basbøll, H (2014). Danish stød as evidence for grammaticalization of suffixal positions in word structure. *Acta Linguistica Hafniensia*, *46*(2), 137-158.

Basbøll, H, Kjærbæk, L, & Lambertsen, C (2011). The Danish noun plural landscape. *Acta Linguistica Hafniensia*, *43*(2), 81-105.

Beddor, P S, McGowan, K B, Boland, J E, Coetzee, A W, & Brasher, A (2013). The time course of perception of coarticulation. *J Acoust Soc Am*, *133*(4), 2350-2366.

Clausen, S, & Kristensen, L B (2015). The cognitive status of stød. *Nordic Journal of Linguistics*, *38*(2).

Cross, A M, & Joanisse, M F (2018). Eyetracking of coarticulatory cue responses in children and adults. *Language, Cognition and Neuroscience*, *33*(10), 1315-1324.

Dahan, D, Magnuson, J S, Tanenhaus, M K, & Hogan, E M (2001b). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5/6), 507-534.

DeLong, K A, Urbach, T P, & Kutas, M (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, *8*(8), 1117-1121.

Desroches, A S, Newman, R L, & Joanisse, M F (2009). Investigating the time course of spoken word recognition: electrophysiological evidence for the influences of phonological similarity. *J Cogn Neurosci*, *21*(10), 1893-1906.

Flagg, E J, Oram Cardy, J E, & Roberts, T P (2006). MEG detects neural consequences of anomalous nasalization in vowel-consonant pairs. *Neurosci Lett*, *397*(3), 263-268.

Fowler, C A, & Brown, J M (2000). Perceptual parsing of acoustic consequences of velum lowering from information for vowels. *Percept Psychophys*, *62*(1), 21-32.

Friston, K (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B Biological Sciences*, *360*, 815-836.

Gosselke Berthelsen, S, Horne, M, Brännström, J, Shtyrov, Y, & Roll, M (2018). Neural processing of morphosyntactic tonal cues in second-language learners. *Journal of Neurolinguistics*, *45*, 60-78.

Gow, D W, Jr. (2003). Feature parsing: feature cue mapping in spoken word recognition. *Percept Psychophys*, *65*(4), 575-590.

Grønnum, N (2005). *Fonetik og Fonologi* (3rd ed.). Akademisk Forlag.

Grosvald, M, & Corinna, D (2009). Perception of long-distance coarticulation: An event-related potential and behavioral study. *Applied Psycholinguistics*, *33*, 55-82.

Hjortdal, A (2021). Danish stød as a cue to upcoming suffixes: an ERP and response time study. *Bachelor thesis, Lund University*.

Hjortdal, A, & Roll, M (submitted). Phonetic and phonological cues to prediction: Neurophysiology of Danish stød.

Kutas, M, DeLong, K A, & Smith, N J (2011). A Look around at What Lies Ahead: Prediction and Predictibility in Language Procesing. In M. Bar (Ed.), *Predictions in the Brain: Using our Past to Generate a Future*. Oxford University Press.

León-Cabrera, P, Rodríguez-Fornells, A, & Morís, J (2017). Electrophysiological correlates of semantic anticipation during speech comprehension. *Neuropsychologia*, *99*, 326-334.

Lewendon, J, Mortimore, L, & Egan, C (2020). The Phonological Mapping (Mismatch) Negativity: History, Inconsistency, and Future Direction. *Front Psychol*, *11*, 1967.

Marslen-Wilson, W, & Warren, P (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. *Psychol Rev*, *101*(4), 653-675.

Martin, J G, & Bunnell, H T (1981). Perception of anticipatory coarticulation effects in /stri, stru/. *J Acoust Soc Am*, *69*(2), 559-567.

Martin, J G, & Bunnell, H T (1982). Perception of anticipatory coarticulation effects in vowel-stop consonant-bowel sequences. *J Exp Psychol Hum Percept Perform*, *8*(3), 473-488.

McMurray, B, Clayards, M A, Tanenhaus, M K, & Aslin, R N (2008). Tracking the time course of phonetic cue integration during spoken word recognition. *Psychon Bull Rev*, *15*(6), 1064-1071.

McQueen, J M, Norris, D, & Cutler, A (1999). Lexical in?uence in phonetic decision making: evidence from subcategorical mismatches. *Journal of Experimental Psychology: Human Perception and Performance*, *25*, 1363-1389.

Mitterer, H, & Blomert, L (2003). Coping with phonological assimilation in speech perception: evidence for early compensation. *Percept Psychophys*, *65*(6), 956-969.

Osterhout, L, & Holcomb, P J (1992). Event-Related Brain Potentials Elicited by Syntactic Anomaly. *Journal of Memory and Language*, *31*, 785-806.

Paquette-Smith, M, Fecher, N, & Johnson, E K (2016). Two-year-olds' sensitivity to subphonemic mismatch during online spoken word recognition. *Atten Percept Psychophys*, *78*(8), 2329-2340.

Rao, R P N, & Ballard, D H (1999). Predictive Coding in the Visual Cortex: a Functional Interpretation of Some Extra-classical Receptive-field Effects. *Nature Neuroscience*, *2*(1), 79-87.

Riad, T (2014). *Phonology of Swedish*. Oxford University Press.

Rischel, J (1963). Morphemic Tone and Word Tone in Eastern Norwegian. *Phonetica*, *10*, 154-164.

Rodriguez-Fornells, A, Clahsen, H, Lleo, C, Zaake, W, & Munte, T F (2001). Event-related brain responses to morphological violations in Catalan. *Brain Res Cogn Brain Res*, *11*(1), 47-58.

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.

Roll, M, Horne, M, & Lindgren, M (2010). Word accents and morphology – ERPs of Swedish word processing. *Brain Research*, *1330*, 114-123.

Roll, M, Söderström, P, Frid, J, Mannfolk, P, & Horne, M (2017). Forehearing words: Pre-activation of word endings at word onset. *Neuroscience Letters*, *658*, 57-61.

Roll, M, Söderström, P, & Horne, M (2013). Word-stem tones cue suffixes in the brain. *Brain Research*, *1520*, 116-120.

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.

Salverda, A P, Kleinschmidt, D, & Tanenhaus, M K (2014). Immediate effects of anticipatory coarticulation in spoken-word recognition. *J Mem Lang*, *71*(1), 145-163.

Sassenhagen, J, Schlesewsky, M, & Bornkessel-Schlesewsky, I (2014). The P600-as-P3 hypothesis revisited: single-trial analyses reveal that the late EEG positivity following linguistically deviant material is reaction time aligned. *Brain Lang*, *137*, 29-39.

Söderström, P, Horne, M, Frid, J, & Roll, M (2016). Pre-Activation Negativity (PrAN) in Brain Potentials to Unfolding Words. *Frontiers in Human Neuroscience*, *10*, 1-11.

Söderström, P, Horne, M, Mannfolk, P, van Westen, D, & Roll, M (2017b). Tone-grammar association within words: Concurrent ERP and fMRI show rapid neural pre-activation and involvement of left inferior frontal gyrus during pseudoword processing. *Brain and Language*, *174*, 119-126.

Söderström, P, Horne, M, Mannfolk, P, van Westen, D, & Roll, M (2018). Rapid syntactic pre- activation in Broca's area: Concurrent electrophysiological and haemodynamic recordings. *Brain Research*, *1697*, 76-82.

Söderström, P, Horne, M, & Roll, M (2017a). Stem Tones Pre-activate Suffixes in the Brain. *Journal of Psycholinguistic Research*, *46*, 271-280.

Streeter, L A, & Nigro, G N (1979). The role of medial consonant transitions in word perception. *J Acoust Soc Am*, *65*(6), 1533-1541.

Warren, P, & Marslen-Wilson, W (1987). Continuous uptake of acoustic cues in spoken word recognition. *Percept Psychophys*, *41*(3), 262-275.

Whalen, D H (1984). Subcategorical phonetic mismatches slow phonetic judgments. *Percept Psychophys*, *35*(1), 49-64.

Whalen, D H (1991). Subcategorical phonetic mismatches and lexical access. *Percept Psychophys*, *50*(4), 351-360.

Zamuner, T S, Moore, C, & Desmeules-Trudel, F (2016). Toddlers' sensitivity to within-word coarticulation during spoken word recognition: Developmental differences in lexical competition. *J Exp Child Psychol*, *152*, 136-148.

Primary auditory cortex's vowel representation

Andrea Fingerhut[1] & Mikael Roll[2]
[1]Medical Radiation Physics, Lund University
[2]Centre for Languages and Literature, Lund University

Abstract

Sound frequencies are represented in the primary auditory cortex (PAC) in a tonotopic structure which can be measured using functional magnetic resonance imaging (fMRI). A previous study has investigated vowels [a] and [i] and seen a correlation between the vowels' activation and the activation of simple tones corresponding to the vowels' formant frequencies. Other vowels have not yet been studied. In this study, we are investigating [a], [ɛ], [i], and [u] and compared those with activation of simple tones corresponding to their formant frequencies. This is ongoing work, and only five volunteers have participated. The preliminary results vary from high correlation between areas activated to low or no correlation. More data has to be collected to draw any further conclusions.

Introduction

The primary auditory cortex (PAC) is the first instance of sound processing in the brain's cortex. In PAC, different neurons are more or less sensitive to specific sound frequencies, and their frequency sensitivity is spatially dependent. Gradients have been found spanning from higher frequencies to lower to higher again, creating a mirror-symmetric structure along the axis of Heschl's gyrus (Formisano et al., 2009). This representation of frequencies in the brain, called 'tonotopy,' can be seen in Figure 1.

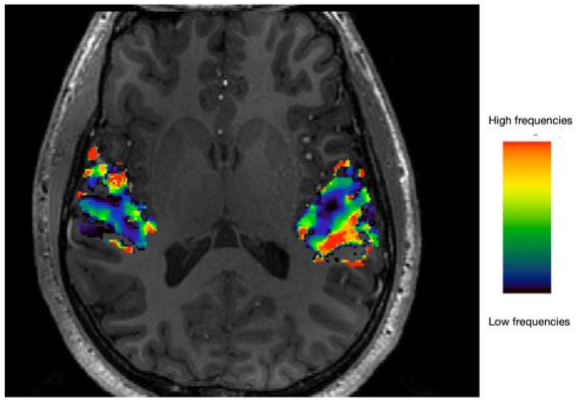


Figure 1: Figure shows a mirror-symmetric tonotopy map. Lower frequencies can be seen in the middle, surrounded by higher frequencies.

Vowels invovle resonance peaks around different frequencies on a frequency spectrum, called 'formants.' First and second formants are required to distinguish different vowels. Fisher (2019) studied vowels [i] and [ɑ], and found that they activate regions in the tonotopic structure of PAC corresponding to their formant frequencies. However, to more firmly establish a relation between vowel formants and the tonotopic structure of PAC, more than two vowels should be tested. Therefore, in this project, we are investigating the effects of the vowels [ɑ], [ɛ], [i], and [u].

Using high resolution functional magnetic resonance imaging (fMRI), a comparison between the activation pattern for simple tone and vowel activations can be made. FMRI uses MRI to study brain activation. It takes advantage of the difference in blood oxygen level to separate different states, commonly between performing a task and resting. In our case, the task is to listen to sounds, and rest is silence. By repeating sound and silence multiple times in a block paradigm, the brain areas connected to the specific task can be distinguished. This design was chosen for its robustness but is, on the other hand, very time-consuming.

Method

To map out the tonotopic structure, high resolution and fast fMRI are needed, and imaging parameters had to be optimized to reach a submillimeter resolution (0.9 mm isotropic resolution) while keeping a short imaging time (2s per volume). A 7 Tesla Philips Achieva scanner was used for this purpose. The same imaging parameters were used for both the vowel and tonotopy measurements.

So far, the auditory processing of five volunteers has been measured. They are all right-handed native speakers of Central Swedish.

Twelve simple tones of different frequencies were selected for the tonotopy measurement to match the vowel formants (300, 375, 525, 600, 675, 825, 1575, 1650, 2400, 2700, 2850, and 3300 Hz). The tones were jittered with ± 75 Hz, which has been shown to increase activation and reduce exhaustion. Each stimulus had a duration of 8s and was repeated four times. The interstimulus interval was 8s. The order of the frequencies was randomized, and the sound level was adjusted for the individual volunteers.

Three formants for every vowel were included (Table 1), as seen for [ɑ] in Figure 2. To simplify the comparison, we created synthetic vowel stimuli (complex tones, F0 = 75 Hz) in Praat that were as acoustically simple as possible but were perceived as vowels. The formant frequencies for Central Swedish in Kuronen (2000) were selected. For the complex tones to be recognizable as vowels, we had to include two harmonics surrounding the formant frequencies (formant frequency ± 75 Hz). The vowels were chosen to maximize the difference between formant 1 and 2. All the formant frequencies can be seen in table 1. We tested both a standard spectral envelope and keeping formant frequencies at the same amplitude as F0 and found little difference in acceptability of the vowels. Therefore, to increase activation for the formants, we kept their amplitude at the same level as F0.

The same block paradigm structure was used for the vowel measurement, with the same duration and number of repetitions. A task was added during this measurement where the volunteers had to answer which vowel they heard. This was both to keep their attention and to make sure that they perceived the correct vowel.

Table 1: Table shows formants 1, 2 and 3 for vowels [a], [ɛ], [i] and [u].

Vowel	F1 [Hz]	F2 [Hz]	F3 [Hz]
[ɑ]	525	825	2475
[ɛ]	600	1650	2700
[i]	300	2400	3300
[u]	300	675	2700

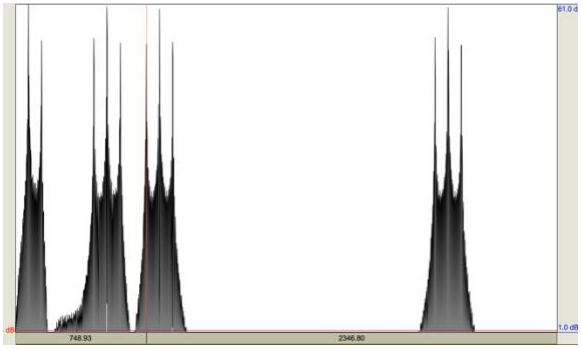


Figure 2: Figure shows frequency spectrum for vowel [a]. The fundamental frequency and the three formants can be seen with the ± 75 Hz frequency surrounding them.

For every volunteer, a tonotopy map was created by giving every voxel in the image a frequency based on which frequency had the highest activation (best frequency). Formant-specific region of interests were created for every vowel by only including voxels with a best frequency matching one of the formants. This gave a template for every vowel showing, if the hypothesis is correct, the expected vowel-activated area. The images were masked based on Harvard-Oxford cortical structural atlases, where Heschl's Gyrus was selected. The mean t-value in the overlap between the format-specific regions of interest and vowel t-maps was calculated.

Results and discussion

The results have not been consistent. We had a higher overlap between the expected formant-specific regions of interest (vowel template) and vowels for one volunteer and no or small overlap for the rest. Table 2 shows the volunteer with the highest overlap between vowel and the right template maps created from the tonotopy measurement. Two of the vowels, [ɑ], [ɛ] had the highest mean t-value in the overlap with the right vowel template. Vowel [i] had the same mean t-value in the overlap between the template for vowel [i] and [u] and vowel [u] had the highest mean t-value in the overlap with the template for vowel [ɛ]. The inconsistency in the results could of course be because the hypothesis of tonotopic vowel representation might be incorrect, but the effect is tiny, and we need to collect more data to either discard or conclude anything. Heschl's Gyrus is also small, and a little head movement or signal displacement can lead to wrong interpretations, although this can partly be solved by using robust preprocessing. The method is still

under development and data will be collected during the coming period.

Table 2: Table mean t-value in overlapping area between template region of interested and vowels activation for [a], [ɛ], [i] and [u].

Template	[ɑ]	[ɛ]	[i]	[u]
[ɑ]	2.6	2.4	2.4	2.3
[ɛ]	2.7	2.9	2.5	2.6
[i]	2.2	2.0	2.4	2.4
[u]	2.6	2.7	2.3	2.4

Acknowledgements

This work was supported by Knut and Alice Wallenberg Foundation (grant number 2018.0454), Swedish Research Council (grant number 2018.00632), Crafoord Foundation (grant number 2017.0006) and Marcus and Alice Wallenberg Foundation (grant number 2018.0021).

References

Fisher J, Dick F, Levy D & Wilson S (2019). Neural representation of vowel formants in tonotopic auditory cortex. *Neuroimage*, 178: 574-582.

Formisano E, Kim D, Salle F, Moortele P, Ugurbil K & Goebel R (2003). Mirror-symmetric Tonotopic Maps In Human Auditory Cortex. *Neuron*. 40: 859-869.

Kuronen, M (2000). Vokaluttalets akustik i sverigesvenska, finlandssvenska och finska. *Studia Philologica Jyvdsfyldensia*, 49. Dissertation, University of Jyvaskyla.

Perceiving head movements in news readings: Evidence from web-based auditory vs. audio-visual prominence ratings

Gilbert Ambrazaitis[1], Johan Frid[2], & David House[3]
[1]Linnaeus University, Växjö
[2]Lund University Humanities Lab
[3]KTH (Royal Institute of Technology)

Abstract

Previous research has shown that visual information can be integrated in the perception of prominence, but the available evidence stems mostly from controlled experimental settings, often making use of synthetic stimuli. The present study provides evidence from spontaneously produced head gestures that occurred in Swedish television news readings. Materials were rated for word prominence by 85 adult volunteers in a between-subjects design (audio-visual vs. audio-only ratings) using a crowd-sourcing approach. Accented words accompanied by a head movement were perceived as more prominent than accented words lacking a head movement. Crucially, the difference in perceived prominence level between words with and without a head gesture was found to be larger in an audio-visual rating condition compared to an audio-only condition. The results suggest that visual prominence signals are integrated in speech processing even in a relatively uncontrolled, naturalistic setting, such as watching the news.

Introduction

Spoken language is essentially an audio-visual, or multimodal phenomenon, comprising audible, acoustic information and visible, kinematic information, both concerning articulatory (e.g., lip) movements (Dohen & Loevenbruck, 2009; Scarborough et al., 2009) as well as gestures, produced, for instance, with the limbs, the fingers, the torso, the head, or the eyebrows (e.g., Kendon, 2004; McNeill, 2005). In particular, speech and gesture have been shown to converge in the production of prominence, as pitch accents and stressed syllables are regularly co-produced and temporally aligned with gestures (e.g., Alexanderson et al., 2013; Ambrazaitis et al., 2020; Esteve-Gibert & Prieto, 2013; Esteve-Gibert et al., 2017; Leonard & Cummins, 2011; Loehr, 2012; Swerts & Krahmer, 2010, Yasinnik et al., 2004).

Following a standard account of gesture classification going back to McNeill (1992), we can distinguish between iconic, metaphoric, deictic, and beat gestures (or rather, dimensions of gestures, e.g., Shattuck-Hufnagel & Prieto, 2019), where beat gestures are assumed to signal prominence. In this study we focus on head movements and their role as prominence cues in speech perception. We thus refer to these movements as beat gestures or 'head beats'.

Previous studies have shown that visually perceived beat gestures are integrated in speech perception in various ways (e.g., Al Moubayed et al., 2011; Wang et al., 2013). For instance, seeing a head movement may improve speech intelligibility (Al Moubayed et al., 2011). Furthermore, there is evidence suggesting that visually perceived gestures can contribute to perceived prominence (e.g., Krahmer & Swerts, 2007; Prieto et al., 2015). However, to our knowledge, most studies have been restricted to experimental settings typically using stimuli where the audio and video are presented separately or are non-congruent (e.g., Dohen & Loevenbruck, 2009; House et al., 2001; Krahmer & Swerts, 2007; Scarborough et al., 2009) or where carefully controlled synthetic stimuli are used (House et al., 2001; Prieto et al., 2015).

Although results from experimental settings are informative, they should be validated by means of testing spontaneously produced gestures from ecologically valid settings. A recent example for this approach is the study by Jiménez-Bravo & Marrero-Aguiar (2020), who collected audio-only and audio-visual promi-

nence ratings for a sample of spontaneous speech taken from a Spanish television talent show. Their results showed significantly more prominence marks in the audio-visual than in the audio-only condition.

The present study continues this line of research asking how spontaneously produced gestures (albeit in a very special genre: news readings) contribute to prominence perception. To this end, we collected prominence ratings using a web-based set-up and a crowd-sourcing approach, where participants rated the words in a sample of news readings from home, via their personal computer or mobile phone, almost as if they actually watched the news. Thus, not only the material used, but also the rating situation can be ascribed a high level of ecological validity. While in Jiménez-Bravo & Marrero-Aguiar's (2020) data, head gestures were combined with manual gestures in a majority of the cases, our data (news readings) generally lack manual gestures, which enables us to focus strictly on the role of head beats.

Methods

Sixteen short video clips from Swedish television news broadcasts were rated, in a between-subjects design, by 44 participants in an audio-visual condition, and 41 participants in an audio-only condition. Ratings were collected using a web-based set-up. Each word was to be rated as either non-prominent, moderately prominent, or strongly prominent, by means of clicking the word in question until the desired prominence level was encoded though a specific colour (see below for details).

The audio-visual speech sample

The clips were between 4 and 7 seconds long and contained 13 words on average (218 words in total), ranging from 8 to 19 words. The clips comprise speech of five different speakers (news anchors) and were taken from a larger corpus (see Ambrazaitis & House, 2017) that had previously been annotated for head movements (binary absence/presence decision per word), as well as for so-called 'big' pitch accents in Swedish (Myrberg & Riad, 2015) also known as the 'sentence accent' or the 'focal accent'.

Data collection

The set-up /rating procedure

Data collection was performed using a custom-made web page implemented in JavaScript, jQuery and the jQuery Simple Presentation plugin. We used the HTML5 software solution stack, particularly making use of the <video> tag, which facilitates web-based video playback considerably. The web page guided the participant through an instruction phase and a training phase. Then, the data collection proper consisted of 16 rating tasks (16 clips), described in detail below. The order of clips to be rated was randomized for each participant. When the test was finished, all the data was sent to a sheet in Google docs.

The rating task

Each clip was rated using a GUI including a video-player (in the audio-visual condition) or an audio-player (in the audio-only condition), an orthographic representation of the text of the clip, as well as a *Nästa* 'Next' button. The text was presented word-by-word in equally-sized boxes. The boxes were to be used as buttons for the prominence rating: A click with the mouse (or the touch screen) changed the colour of the box, which would turn YELLOW (prominence level 1) after one click, RED (prominence level 2) after another click, and neutral again after a third click.

A clip presentation always started with a still video and a 'Start' button. When that button was clicked, the clip was played automatically two times, without any break in between and without the option to pause the video/audio. During this initial presentation, the rating buttons (incl. the orthographic representations) were hidden. Participants in the audio-visual condition were instructed to carefully look at the video during this double screening. This was done in order to ensure that the participants' first impression of the clip and its prominence relations would be based on the full audio-visual input. After this initial phase, the text buttons along with usual video playing controls appeared. The participant was then free to play the video/audio again as often as necessary, making use of pausing or playing smaller parts if desired, and to rate all words using the text buttons. When satisfied, the participant clicked the 'Next' button to reach the next clip.

Participants

Volunteers were recruited via social media and e-mail. They were offered a (digital) cinema ticket for their participation. A total of 85 adult native Swedish volunteers participated in the study: 44 in the audio-visual condition, and 41 in the audio-only condition. All raters were native Swedish adult volunteers with no reported hearing impairment and normal or corrected sight. They were encouraged to conduct the rating in a silent surrounding.

Analysis

The collected prominence ratings were analysed in two steps. First, four heuristic measures were calculated per rater in order to explore overall rating behaviour as a function of rating condition (audio-visual vs. audio-only). These were:

(1) $\bar{x}_w$ (average word prominence): sum of all ratings (for all 218 words) divided by number of words (218)
(2) $\%W_{pr}$ (percentage of prominent words): number of all words that were marked either moderately or strongly prominent, divided by number of words (218), multiplied by 100
(3) $\%W_{st}$ (percentage of strong words): number of all words marked strongly prominent, divided by number of words (218), multiplied by 100
(4) $\%W_{st/pr}$ (relation between prominence levels): number of all words marked strongly prominent, divided by number of words marked either moderately or strongly prominent, multiplied by 100

The effect of rating condition on these measures was assessed by means of independent samples t-tests.

In a second step, we explored whether the rating condition would affect specifically words that were produced with a head gesture. To this end, all words in our sample were classified as either being realized with a big accent (BA) and a head beat (HB), a BA only, or neither (using the available head beat annotations from Ambrazaitis & House, 2017). Table 1 displays token frequencies for these three categories in the selected data set. We then calculated the following three measures for each rater:

(5) $\bar{x}_{noBA}$, $\bar{x}_{BA}$, and $\bar{x}_{BAHB}$ (average word prominence): sum of all ratings (for all words within each category according to Tab. 1) divided by *n* according to Tab. 1

The measures described in (5) were evaluated by means of linear mixed effects regression models. In particular, we tested how well average word ratings are predicted by (a) the rating condition and (b) the cumulative addition of multimodal prominence markers (+BA+HB). The latter predictor was modelled as a three-level (within-subjects) fixed effect *MMP* (*multimodal prominence*: noBA, BA, BAHB). The former predictor was modelled as a two-level (between subjects) fixed effect *AV_condition* (audio-only, audio-visual). We included an interaction term for *MMP* and *AV_condition,* and *rater* (intercepts only) as a random effect. The full model (*MMP* * *AV_condition* + (1|*rater*)) was then compared to three different reduced models using likelihood ratio tests in order to evaluate the significance of the two predictors *MMP* and *AV_condition* as well as their interaction (see Tab. 2).

Table 1. Frequencies of occurrence of words with 'big accent' (BA), with BA and a 'head beat' (BAHB), and without BA (noBA).

noBA	BA	BAHB	Total
148	22	48	218

Finally, the following difference measure was calculated for each rater:

(6) $\bar{x}_{\Delta HB} = \bar{x}_{BAHB} - \bar{x}_{BA}$

Just as for measure (1-4), the effect of the rating condition on measure (6) was assessed using an independent samples t-test.

All modelling was done in R (R Core Team, 2012) using the lmer function from the lme4 package (Bates et al., 2012). R^2-values were obtained using the function r.squaredGLMM from the MuMIn package (Barton, 2020), and χ^2-tests and t-tests were performed using the t.test and the anova function from the stats package (R Core Team, 2012).

Results and discussion

Figure 1 displays the results for the four heuristic measures defined to explore overall rating behaviour. For none of these measures did the results suggest an effect of the rating condition (audio-only vs. audio-visual; $\bar{x}_w$: $t = .11$, $df = 83$, $p = .91$; $\%W_{pr}$: $t = -.41$, $df = 83$, $p = .69$; $\%W_{st}$: $t = .98$, $df = 83$, $p = .33$; $\%W_{st/pr}$: $t = 1.30$, $df = 83$, $p = .20$). That is, raters did, on average, not assign significantly more or higher prominence ratings

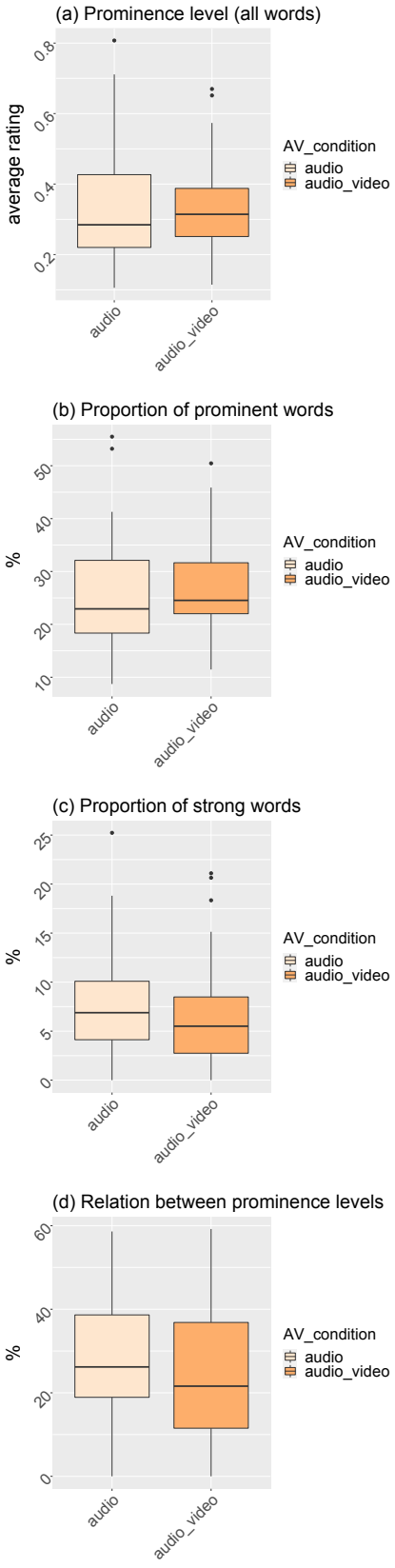
to words in the audio-visual condition compared to the audio-only condition. A possible interpretation of these results is that visual information did not significantly add to perceived prominence. However, it might also be the case that our four heuristic measures were too coarse to capture effects of visual prominence cues. If, for instance, a visible head beat in fact makes a word more prominent, but the effect is rather small, then it might not have a significant impact on the overall word prominence average (taken across words with and without head beats).

Therefore, in a second step, we distinguished between words produced with and words produced without an accompanying head beat. Figure 2 (next page) displays the distribution of average prominence ratings per multimodal prominence constellation (*MMP*) for the audio-only and the audio-visual condition. It can be seen – for both rating conditions – that words with big accent and head beat tend to receive higher prominence ratings than words with a big accent only, and words lacking a big accent are generally rated low. This predictive value of *MMP* is highly significant (Tab. 3). Furthermore, Table 2 shows that our models account for up to 83.2% of the observed variability in prominence ratings, whereby the contribution of *MMP* is decisive: Models reach an R^2_c of around 83% only if *MMP* is included, and irrespective of other factors or interactions present.

That is, words accompanied by a head beat tend to be rated stronger even if no visual information on the head beat is available (in the audio-only condition). This is well in line with the results from the (audio-only) ratings obtained by Swerts & Krahmer (2010), suggesting that words with head beats are also produced with stronger acoustic prominence cues. A tendency for larger accentual big-accent rises as a function of accompanying head (and eyebrow) beats has indeed been observed in an ongoing study (Ambrazaitis & House, submitted).

The rating condition, however, had no predictive value in the linear mixed models (Tab. 2 & 3). Notably, the results did not even reveal a significant interaction between *AV_condition* and *MMP*, although Figure 2 suggests a tendency for such an interaction: For BA words, the plot suggests a tendency for slightly lower ratings in the audio-visual condition, while when a head beat is present (BAHB), a slight trend for higher ratings is seen when the visual modality is available.

Figure 1. Boxplots for the four measures of overall prominence rating behavior comparing the audio-only and the audiovisual condition: (a) $\bar{x}_w$, (b) $\%W_{pr}$, (c) $\%W_{st}$, (d) $\%W_{st/pr}$ (see 1-4 for explanations).



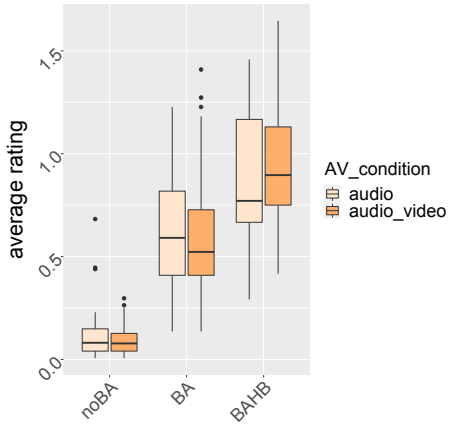


Figure 2. Boxplots of average prominence ratings collected in an audio-only and in an audio-visual condition for words realized without any big accent, with a big accent only, or a big accent and a head beat ($\bar{x}_{noBA}$, $\bar{x}_{BA}$, and $\bar{x}_{BAHB}$).

Table 2. Model fit for all full and reduced models measured using R^2. R^2_m = marginal R^2 measuring the amount of variation described by the fixed factors; R^2_c = conditional R^2 measuring the amount of variation described by the entire model including random-effects factor; '' denotes the interaction between factors, as oppose to '+'.*

	Model	R^2_m	R^2_c
Full	MMP * AV_con. + (1|rater)	.635	.832
Red.1	MMP + AV_con. + (1|rater)	.632	.828
Red.2	MMP + (1|rater)	.632	.828
Red.3	AV_con. + (1|rater)	.000	.000

Table 3. Results of likelihood ratio tests comparing full and reduced models.

Model comparison	Effect tested	χ^2	*df*	*p*
Full vs. Red.1	*interaction*	3.93	2	.14
Red.1 vs. Red.2	*AV_con.*	.01	1	.92
Red.1 vs. Red.3	*MMP*	322.37	2	.000

Figure 3 scrutinizes this relation between BA and BAHB words, displaying the mean differences in prominence between BAHB and BA words. The figure suggests that a slightly greater distinction is made between words with and without head beats in the audio-visual compared to the audio-only condition, and this trend is significant (t = -2.5459, df = 83, p = .01). This result suggests that the visual perception of head beats indeed adds to perceived prominence.

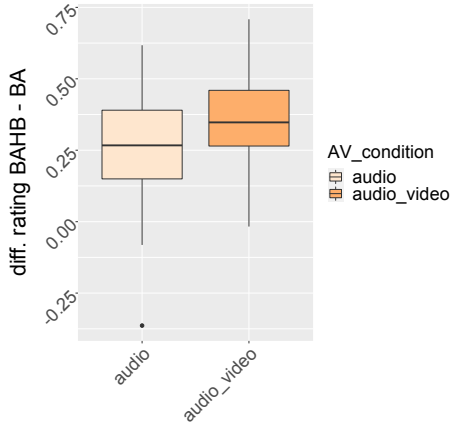


Figure 3. Boxplots for the difference measure ($\bar{x}_{\Delta HB}$) comparing the audio-only and the audio-visual condition.

Conclusions

According to a recent proposal by Holler and Levinson (2019), multimodal information can be expected to support spoken language processing in natural communicative settings, rather than to impede processing. We therefore should explicitly look for, and expect to find, evidence for audio-visual integration in the perception of ecologically valid speech samples. However, few previous studies on the impact of visually perceived gestures on perceived prominence have tested spontaneous speech or spontaneously produced gestures. Furthermore, little is known about the individual contributions of different gestural articulators (Jiménez-Bravo & Marrero-Aguiar, 2020).

The present study has provided novel evidence based on spontaneous gestures, albeit produced with non-spontaneous speech (news readings). However, the choice of this speech genre enabled us to isolated head beats from manual gestures. The results suggest that the visual perception of head beats can add to perceived prominence, although the observed effect of the availability of the visual modality was relatively small – much smaller than the effect of acoustic prominence cues.

The relatively small predictive power of the rating condition may have several explanations. First, not only the speech material was uncontrolled, but also the rating set-up. We did not, in this study, control for factors such as screen size or viewport size in the audio-visual condition (but see Ambrazaitis et al., 2019), nor did we have any possibility to control to what degree the participants really looked at the video. Moreover, and most importantly, in an

uncontrolled setting, a multitude of multimodal signals are processed simultaneously (Holler & Levinson, 2019). Given these circumstances, it is noteworthy that the impact of the visual modality on the perception of prominence is robust enough to be measurable in ecologically valid data using relatively uncontrolled rating setting.

Acknowledgements

This work was supported by two grants from the Swedish Research Council (VR-2017-02140 and VR-2013-2003).

References

Alexanderson S, House D & Beskow J (2013). Aspects of co-occurring syllables and head nods in spontaneous dialogue. In: *Proc of the 12th International Conference on Auditory-Visual Speech Processing (AVSP2013)*. Annecy, France.

Al Moubayed S, Beskow J, Granström B & House D (2011). Audio-visual prosody: Perception, detection, and synthesis of prominence. In: A Esposito, A M Esposito, R Martone, V C Müller & G Scarpetta, eds, *Toward autonomous, adaptive, and context-aware multimodal interfaces. Theoretical and practical issues. Lecture Notes in Computer Science, 6456*. Berlin, Heidelberg: Springer, 55-71.

Ambrazaitis G, Frid J & House D (2019). Multimodal prominence ratings: Effects of screen size and audio device. In *Book of abstracts from the 6th European and 9th Nordic Symposium on Multimodal Communication (MMSYM 2019)*. Leuven, Belgium, 2-3.

Ambrazaitis G & House D (2017). Multimodal prominences: Exploring the patterning and usage of focal pitch accents, head beats and eyebrow beats in Swedish television news readings. *Speech Communication,* 95: 100-113.

Ambrazaitis G & House D (submitted). Probing effects of lexical prosody on speech-gesture integration in prominence production by Swedish news presenters.

Ambrazaitis G, Zellers M & House D (2020). Compounds in interaction: Patterns of synchronization between manual gestures and lexically stressed syllables in spontaneous Swedish. In: *Proc of Gesture and Speech in Interaction (GESPIN2020).* Stockholm, Sweden.

Barton K (2020). *MuMIn: Multi-Model Inference.* R package version 1.43.17.

Bates D M, Maechler M & Bolker B (2012). *lme4: Linear mixed-effects models using S4 classes.* R package version 1.1-15.

Dohen M & Loevenbruck H (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Language and Speech,* 52: 177-206.

Esteve-Gibert N, Borràs-Comes J, Asor E, Swerts M & Prieto P. (2017). The timing of head movements: The role of prosodic heads and edges. *The Journal of the Acoustical Society of America*, 141(6): 4727-4739.

Esteve-Gibert N & Prieto P (2013). Prosodic structure shapes the temporal realization of intonation and manual gesture movements. *Journal of Speech, Language, and Hearing Research,* 56(3): 850–864.

Holler J & Levinson S C (2019). Multimodal Language Processing in Human Communication. *Trends in Cognitive Sciences,* 23(8): 639-652.

House D, Beskow J & Granström B (2001). Timing and interaction of visual cues for prominence in audiovisual speech perception. In: *Proc of Eurospeech 2001*. Aalborg, Denmark, 387-390.

Jiménez-Bravo M & Marrero-Aguiar V (2020). Multimodal perception of prominence in spontaneous speech: A methodological proposal using mixed models and AIC. *Speech Communication,* 124: 28-45.

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

Krahmer E & Swerts M (2007). The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. *Journal of Memory and Language,* 57(3): 396–414.

Leonard T & Cummins F (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes,* 26(10): 1457–1471.

Loehr D (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture. *Laboratory Phonology,* 3(1): 71–89.

McNeill D (1992*). Hand and mind: what gestures reveal about thought.* Chicago: University of Chicago Press.

McNeill D (2005). *Gesture and thought.* Chicago: University of Chicago Press.

Myrberg S & Riad T (2015). The prosodic hierarchy of Swedish. *Nordic Journal of Linguistics,* 38(2): 115–147.

Prieto P, Puglesi C, Borràs-Comes J, Arroyo E & Blat J (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent. *Journal of Phonetics,* 49(1): 41–54.

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

Scarborough R, Keating P, Mattys S L, Cho T, Alwan A & Auer E T (2009). Optical phonetics and visual perception of lexical and phrasal stress in English. *Language and Speech,* 52: 135–175.

Shattuck-Hufnagel S, Prieto P (2019). Dimensionalizing co-speech gestures. In*: Proc. of the 19th International Congress of Phonetic Sciences*. Melbourne, Australia, 1490-1494.

Swerts M & Krahmer E (2010). Visual prosody of newsreaders: effects of information structure, emotional content and intended audience on facial expressions. *Journal of Phonetics,* 38(2): 197–206.

Wang L & Chu M (2013). The role of beat gesture and pitch accent in semantic processing: an ERP study. *Neuropsychologia,* 51(13): 2847–2855.

Yasinnik Y, Renwick M & Shattuck-Hufnagel S (2004). The timing of speech-accompanied gestures with respect to prosody. In: *Proc. of From Sound to Sense*. Cambridge, MA, USA, 97–102.

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.

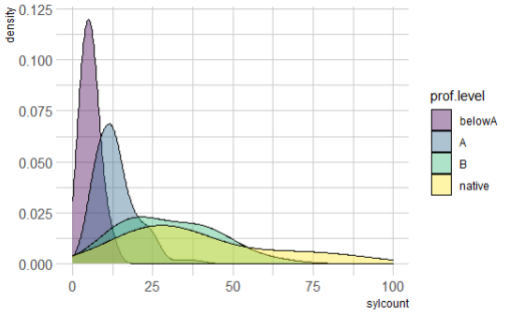


Figure 1. Distribution of syllable counts per speaker group.

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).

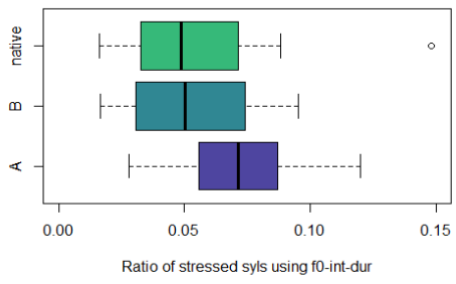


Figure 2. Proportion of stressed syllables in A-level samples (N=59), B-level samples (N=24) and native samples (N=25), when acoustic stress

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

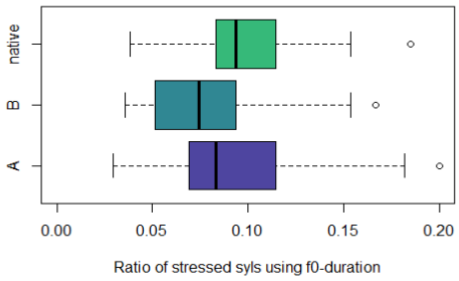


Figure 3. Proportion of stressed syllables in A-level samples (N=92), B-level samples (N=38) and native samples (N=29), when acoustic stress is based on normalized f0 and normalized syllable duration.

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.

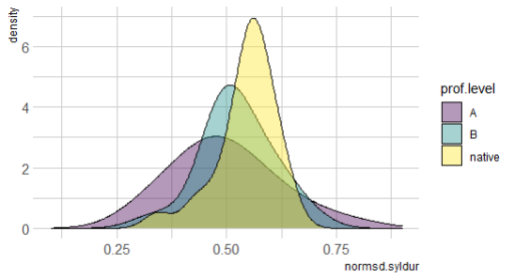


Figure 4. Distribution of rate-normalized standard deviation of syllable duration per speaker group.

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.

Dating the new open TRAP sound change in South East England

Sidney Wood
Retired, formerly at Universities of Lund and Umeå, Sweden

Abstract

The new open TRAP sound change was reported in the 1980s, in Received Pronunciation. The earliest example found by Fabricius in 2007 was born in 1926, still Received Pronunciation. Jones and Gimson reported only the earlier closer TRAP. The earliest regional example of new open TRAP, born in 1866 in Kent, was reported by Wood in 2017. Four groups of informants were studied, RP and regional. The earliest RP speaker found with new open TRAP was born in 1857, one hundred years before it was noticed. This raises questions like how did Jones manage to miss it? (or ignore it?). Suggested explanations include inadequate vowel theory (the Bell vowel model) and experimental methods (feeling the Bell vowel locations).

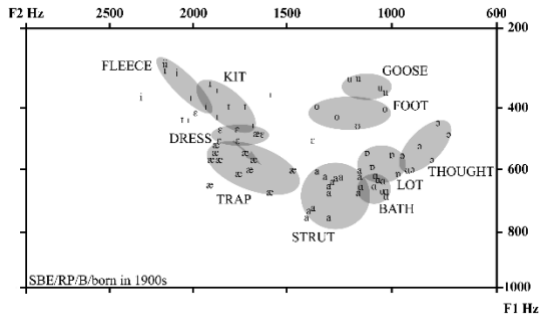


Figure 1. A selection of RP vowels by informant B, illustrating the earlier closer TRAP with lower F1 at 489-682 Hz, DRESS and KIT compressed towards FLEECE.

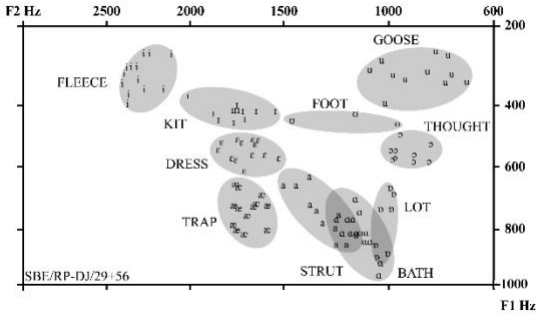


Figure 2. A selection of RP vowels by Daniel Jones (1881-1967), analysed from Jones (1929, 1956), illustrating the new open TRAP (higher F1 668-825 Hz, no compression of DRESS and KIT).

The Problem

This article is concerned with the recently observed sound change in south-eastern England that opened[1] TRAP[2] pronunciation both in non-regional Received Pronunciation[3] (RP) and in regional accents. The timbre of the earlier closer

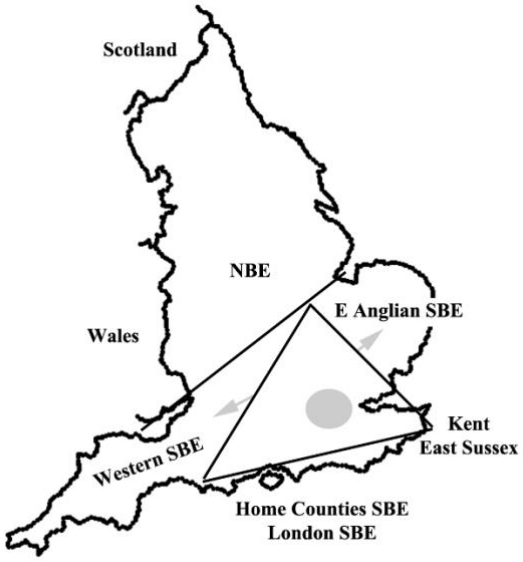


Figure 3. The stylized isogloss between Northern British English (NBE) and Southern British English (SBE). Four regiolects of SBE: East Anglian, London, Home Counties (triangle) and Western.

TRAP was IPA [æ][4] (Fig. 1), while new open TRAP is open [æ] (Fig. 2). The criteria for distinguishing new open TRAP (F1 higher than about 600 Hz and no compression of DRESS and KIT towards FLEECE), are defined in §3). The regional Kentish TRAP sound change had commenced by the 1860s (Wood, 2017). It was also spreading through the home counties (the region surrounding London and bounded by East

Sussex, Hampshire and Northamptonshire, Fig. 3). It has been noticed in RP since the 1980s (Wells, 1982), while the earliest RP speaker with open TRAP reported by Fabricius (2007) was born in 1926. Did Fabricius catch it starting in RP during the 1920s, or was it already in progress? Figure 2 suggests it was (Daniel Jones, phonetician, professor, born in 1881). Early examples like this are considered controversial, born long before anyone showed any awareness of this sound change in RP. Jones never mentioned this sound change himself, always describing only the earlier *closer RP* TRAP timbre (1908, 1918, all editions to the 1960s). An additional recording of Jones has recently been analysed by Przedlacka & Ashby (2019). Their vowel diagram (their Fig. 1) also shows Jones' open TRAP at high F1. Controversial as they may be, it was examples like this, together with regional examples from SE England, that first prompted doubts about the dating of this sound change. The status of new *open* TRAP in earlier RP was investigated by analysing the spectra of TRAP instances taken from recordings of RP speakers in two sequences preceding Fabricius' confirmation: (i) an RP-speaking group born in 1850-1899 (the controversial period), and (ii) a second RP group born in 1900-1930 (the non-controversial period up to Fabricius' dating). For comparison, published formant data from regional Kentish SBE speakers born in 1860-1895 (Wood, 2017), and recordings of six speakers of 20th century Home Counties SBE (HCSBE) born in 1900-1960 were also included.

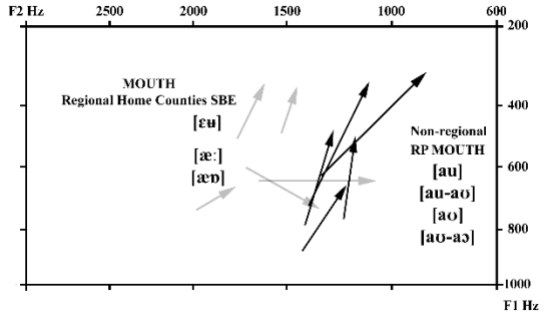


Figure 4. Pronunciations of the MOUTH diphthong in S. E. England: regional accents (left) and non-regional RP (right).

Accents of S. E. England

Southern British English (SBE) is the regional dialect spoken right across southern England from Norfolk to Cornwall, south of an isogloss from the Wash to the Severn Estuary, distinguished from neighbouring Northern British English (NBE) by the TRAP-BATH split and the FOOT-STRUT split (Wales, 2006; Britain, 2012). There are several SBE regiolects, generalized by Wells (1982§§4.2-4.3) from east to south-west as East Anglian, London, Home Counties, and Western. The accent of interest here is Home Counties SBE (HCSBE). Ellis (1889) referred to the earlier dialect of Kent and East Sussex as *Eastern Southern English*, already giving way to contact with London pronunciation. For the present study, these regional accents are distinguished as Kentish SBE (19th century, transitional) and Home Counties SBE (HCSBE, 20th century) following Wood (2017).

RP and regional HCSBE are identified by referring to their respective vowel phoneme systems. RP was described by Jones (1909, 1918)), other revisions are provided by Gimson (1962) and Wells (1982). HCSBE was stigmatized until about the 1960s and very little had been published on it until Rosewarne (1984) mentioned Estuary English. Consequently, Wells (1982:§4.3.1) found little to report on home counties pronunciation. Since then this region has attracted more attention, summarized by Jansen & Amos (2020). Wood (2017) described a shibboleth that has distinguished regional HCSBE from non-regional RP at least since the late 19th century. This concerns the respective pronunciations of the MOUTH lexical set (Fig. 4) that had taken different paths through the Great Vowel Shift: the "polite educated" community (Cooper, 1687) shifted MOUTH down the back vowels to current RP [aʊ-ɑʊ] (Luick, 1896; Jespersen, 1909, Wolfe, 1973), while the "provincial" community shifted MOUTH down the front vowels to regional [ɛʉ] (Britain, 2008; Cooper, 1687), eventually to *open* [æː~æɒ] in London and the South East (Wells, 1982; Wood, 2017), while [ɛʉ] still occurs in the South West. Table 1 lists some differences between RP and the non-regional accents.

Table 1. Some pronunciation differences between accents of South East England ('x>y' reads 'x becoming y').

Accent	TRAP	BATH	LOT	THOUGHT	GOAT	MOUTH
19th, 20th c RP	close (open?)	[ɑː]	[ɒ]	[ɔː]	[ou>əʊ]	[aʊ>ɑʊ]
19th c. Kent	close or open	[aː]	[ɑ>ɔ]	[ɔː>oː]	[ou>aʊ]	[ɛʉ>æɒ]
20th c. HCSBE	open	[ɑː]	[ɔ]	[oː]	[aʊ]	[ɛʉ>æɒ]

Criteria for open TRAP

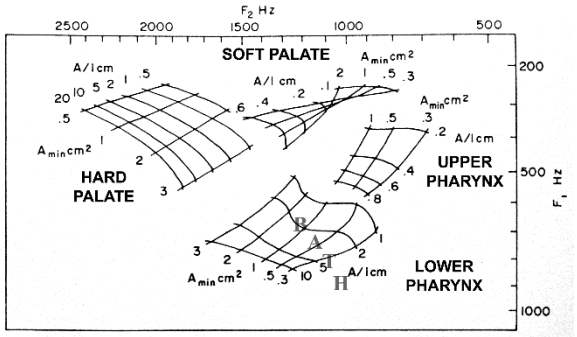


Figure 5. Vowel F1 and F2 frequencies obtained by varying the degree of constriction (A_{min}) and degree of mouth opening (A/l) at each of four constriction locations. Superimposed: Daniel Jones' open back BATH vowel zone. Adapted from Wood (1979).

Essner (1947) and Joos (1948) would have expected high F1 for open vowels. But how high? Figure 5 shows that a vowel timbre designated as *unrounded open* (Jones' [ɑː]-like BATH) would be articulated with a low pharyngeal constriction together with a large mouth opening, yielding F1 higher than about 600 Hz. This means that open is an extended zone, with F1 varying from about 600 to 800 Hz or beyond. The actual frequency of the boundary appears to vary both between and within individual speakers (not illustrated here). Between-speaker variation might reflect differences of vocal tract size, shorter vocal tracts yielding higher boundary frequencies. Within-speaker variation might reflect varying speaking style with varied degree of mouth opening yielding continuous dynamic variation of the boundary.

The first criterion, then, for open TRAP is F1 higher than about 600 Hz

The timbre of the earlier close TRAP was IPA [æ]. Four phonemes had to share the non-open front region since close TRAP encroached on the F1 region around 500 Hz, with consequent compression of DRESS and KIT towards, and even alongside, FLEECE. However, F1 of close TRAP also extended into the open F1 region above 600 Hz (clearly seen in Fig. 6, with the majority of TRAP instances having F1 within 500-600 Hz while about 25% extend into the open region at 600-700 Hz). Similar mixes of closer and open instances have been found for other recordings of old closer RP TRAP, including the linguist J R Firth described by Przedlacka & Ashby (2019), easily seen in their diagram.

A second criterion for open TRAP is, consequently, no compression of DRESS and KIT towards FLEECE.

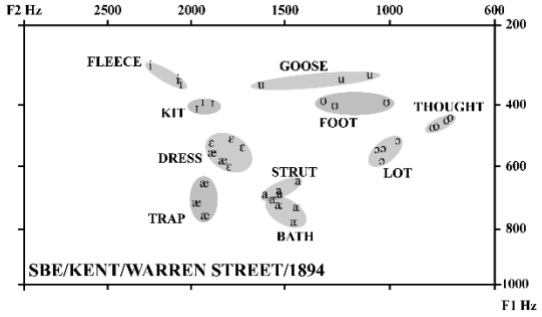


Figure 6. Late 19th c. Kent: new open TRAP, and the last example of [aː]*-like BATH.*

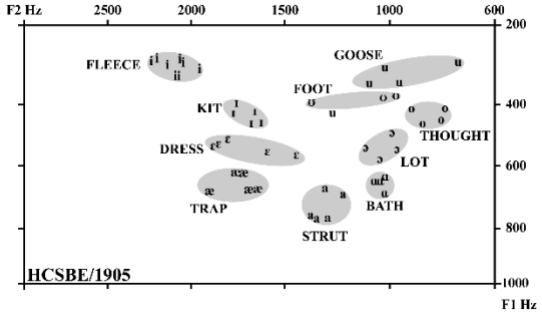


Figure 7. 20th c. HCSBE: new open TRAP, and the earliest regional example of [ɑː]*-like BATH.*

Table 2. *Results for new open and old close* TRAP *in the four groups of south-eastern accents (§5).*

Informant	TRAP	Born	TRAP F1; F2 Hz
19th c. RP			
R. Baden-Powell	*open*	1857	680-800; 1713-2059
Stanley Baldwin	*open*	1867	641-785; 1717-1994
Richard Paget	*open*	1869	608-812; 1722-1932
W S Maugham	*open*	1874	592-687; 1730-1968
Daniel Jones	*open*	1881	668-825; 1560-1776
H. MacMillan	*open*	1894	602-657; 1329-1509
N. Chamberlain	*closer*	1864	F1 352-713
RP 1901-1930			
A	*open*	1900-1910	601-708; 1516-2005
C	*open*	1900-1910	720-818; 1413-1749
D	*open*	1910-1920	612-744; 1475-1610
E	*open*	1910-1920	602-701; 1546-1839
F	*open*	1920-1930	580-728; 1507-1722
G	*open*	1920-1930	629-730; 1509-1814
H	*open*	1920-1930	565-710; 1491-1664
B	*closer*	1900-1910	F1 489-682
19th c. Kent			
I (H. G. Wells)	*open*	1866	595-727; 1849-2018
L	*open*	1880	601-665; 1495-1811
M	*open*	1881	633-680; 1442-1798
N	*open*	1881	633-744; 1566-1946
O	*open*	1888	582-713; 1443-1545
P	*open*	1894	562-759; 1835-1977
J	*closer*	1868	F1 483-538
K	*closer*	1880	F1 514-578
20th c. HCSBE			
Q	*open*	1905	602-691; 1629-1930
R	*open*	1909	696-724; 1971-2055
S	*open*	1928	608-717; 1664-1756
T	*open*	1940	634-662; 1430-1549
U	*open*	1955	639-749; 1505-1576
V	*open*	1958	671-788; 1479-1640

The story of TRAP and BATH in the S. E. in recent centuries

Before the 17th century the TRAP and BATH sets were both pronounced together with the same open vowel, [a], (Wyld, 1936:196-205). Then TRAP was fronted and raised to a closer location near DRESS while BATH remained at [a]. At the same time, DRESS and KIT were displaced towards FLEECE, ensuring the spectral contrast between TRAP and DRESS. That was the TRAP-BATH split, unique to SBE in the British Isles. In some cases, such as regional London Cockney, the fronting and raising of TRAP went as far as [ɛ], still there in the 1830s-1930s (Matthews, 1938:Chapt. 1) and in the 1950s (Sivertsen, 1960:Chapt. 3.3). Since the late 19th century, TRAP has been changing again in the South East, to open [æ]. Figure 6 shows an example of new open TRAP from late 19th century regional Kentish SBE.

The BATH set, left behind at [a] after the split, was eventually lengthened to [aː] and finally retracted to [ɑː]. This final BATH change is usually dated to the 18th century for RP (Hickey, 2020), seen in Figs. 1-2. Figure 6 shows BATH still at [aː] in late 19th century Kent, the last example in the data available to this study. Figure 7 shows the earliest HCSBE example of [ɑː]-like BATH.

Speech samples

Speech recordings were collected for four groups of SBE informants representing the south-eastern SBE accents. They were found randomly online, often in the BBC Archive or the British Library. 19th century RP informants were prominent public figures and they are identified. Anonymity is preserved for all other informants.

SBE Group 1 consists of seven RP speakers born in the 19th century.

SBE Group 2 consists of nine RP speakers born in 1900-1930, referred to as A-H. Recordings were taken from broadcast interviews made in the 1950s or 1980s.

SBE Group 3 consists of eight speakers of 19th century Kentish SBE, referred to as I to P. The recording by H G Wells is a BBC broadcast. The remaining 19th century Kentish recordings were made by the Survey of English Dialects (Orton & Dieth 1962), available online at the British Library. Their formant data is cited from Wood (2017).

SBE Group 4 consists of six speakers of 20th century HCSBE, referred to as Q to V, born in 1905-1958. The recordings were found online at university websites or in the BBC Millennium or BBC Voices collections at the British Library.

Formant analysis

Vowel formants were analysed using *Praat* (Boersma & Weeninck, 2017), FFT slices from narrowband spectrograms offering the best solution for some recordings, while linear predictive formant tracking was successful for the others.

Only fully prominent exemplars of vowels were analysed, taken from focally accented syllables (to minimize spectral effects of vowel reduction). Formants were taken at the moment where the vocoid segment was least affected by adjacent consonants (to minimize spectral effects from coarticulation), determined by observing CV and VC formant transitions on spectrograms.

Results

Table 2 records the results for all informants in each group. The ratios of old close TRAP were 1 of 7, 1 of 8 and 2 of 8 in the first three groups. The earliest RP example of new open TRAP in this data set was born in 1857.

Conclusions

New open TRAP was happening in RP as early as the 1850s, a century before anyone reported it. How did they come to miss it for so long?

One possible reason is their absolute belief in the Bell vowel model (or Passy's version), that assigned the difference between open and close TRAP entirely to tongue location. They would not have been aware that F1 for pharyngeal vowels is controlled by the mouth opening and F2 by tongue location (Fig. 5).

A second possible reason is their claimed ability to feel the tongue positios of the Bell vowel model (first expressed by Sweet, 1877:18). Sadly, Bell's tongue positions for vowels were just not there, waiting to be felt. Eventually, this ability to feel tongue positions for vowels was linked to the neurophysiological function of *proprioception* (for example, by Catford, 1981), although that is not how proprioception works, as any comprehensive handbook of neurophysiology, such as Siegel & Sapru (2015:253), will make clear. These authors explain that proprioception is only partly conscious, allowing awareness of some articulator positions or movements, but also partly nonconscious, precluding awareness of other articulator positions or movements. In conscious proprioception, receptors in joint capsules provide sensory information to the cerebral cortex, enabling awareness of kinesthesia. Only the mandibular joints would be relevant for speech, enabling awareness of jaw position. That had always been a parameter of the ancient model since the time of Panini. Nonconscious proprioception, however, arises from muscle spindles and Golgi tendon organs and is passed to subcortical motor centres solely for internal control of movement. Consequently, no sensations are available to the cerebral cortex from this background activity working silently and efficiently during vowel production.

Notes

[1]For the present study, the terms *tongue height* and *backness* follow the usage of the past 150 years (A. M. Bell 1867:15-16, 71, Sweet 1877, D. Jones 1932, IPA 1999), but advisedly because the Bell vowel model has never been validated, but was seriously compromised on numerous occasions (Wood, 1982).

[2]Expressions like TRAP are keywords for what J. C. Wells (1982:§2.2) calls lexical sets

representing vowels that participated in various sound changes in English, with different outcomes in different dialects. They are more useful than phoneme notation when pronunciations are changing, or where accents differ.

[3]The expression *Received Pronunciation* (*RP*) is preferred here for the sake of continuity with earlier literature although other rival synonyms have come into use for this accent, especially *General British English* (*GBE*) (Lewis 1972) and *Standard Southern British English* (*SSBE*) (IPA 1999). SSBE is especially unsuitable as it is also increasingly being used as a nickname for regional HCSBE, the "new standard" (like Lindsey 2019:4). The phonology of non-regional RP nevertheless belongs typologically to SBE.

[4]Close [æ] refers to IPA [æ], the timbre of old close TRAP. The timbre of open TRAP was never given an IPA identity and is referred to here as open [æ].

References

Bell, A. M. 1867. *Visible Speech*. London: Methuen.

Boersma, P. & D. Weeningk. 2017. *Praat: Doing phonetics by computer* (versions 5 and 6). Online at http://www.fon.hum.uva.nl/praat/, accessed 2014 - 2017.

Britain, D. 2008. On the wrong track? A non-standard history of non-standard /au/ in English. Essex Research Reports in Linguistics 57:33-77.

Britain, D. 2012. English in England. In Raymond Hickey (Ed.), Areal Features of the Anglophone World, Berlin, Mouten de Gruyter, 23-52.

Catford, J. C. 1981. Observations on the recent history of vowel classification. In R. E. Asher and J. A. Henderson, *Towards a History of Phonetics*, Edinburgh University Press, 19-32

Cooper, C. 1687. *The English Teacher*. London. Edited by B. Sundby, 1953, *Christopher Cooper's English Teacher (1687)*, Lund (Gleerup).

Ellis, A. 1889. *On Early English Prununciation*. Teubner, Vol 5.

Essner, C. 1947. Recherches sur la structure des voyelles orales. *Archives Néerlandaises de Phonétique Expérimentale* 20:40-77.

Fabricius, A: H. 2007. Variation and change in the trap and strut vowels of RP; a real time comparison of five acoustic data sets. *Journal of the International Phonetic Association* 37:293-320.

Gimson, A. C. 1962. *An Introduction to the Pronunciation of English*. London, Arnold.

Hickey, R. 2020. On the recent history of low vowels in English. *English Language and Linguistics* 24: 545-567.

IPA. 1999. *Handbook of the International Phonetic Association*. Cambridge University Press.

Jansen, S and J. Amos, (eds). 2020. *English in the South of England*. Special issue of *English Today* 36, no. 3.

Jespersen, O. 1909. *A Modern English Grammar*. Part 1, *Sounds and Spellings*. Heidelberg, Carl Winter.

Jones, D. 1909. *The Pronunciation of English*. London: Cambridge University Press.

Jones, D. 1917. Experimental phonetics and its utility to the linguist. *Nature* 100:96-98

Jones, D. 1918. *An Outline of English Phonetics*. Leipzig: Teubner (1st edition).

Jones, D. 1929. *A voyage*. Linguaphone English disc EC18E. http://sounds.bl.uk/Arts-literature-and-performance/Early-spoken-word-recordings. Retrieved 2014.

Jones, D. 1932. *An Outline of English Phonetics*. Leipzig: Teubner (3rd revised edition).

Jones, D. 1956. *Cardinal Vowels*. London: Linguaphone Disc ENG 252-5.

Joos, M. 1948. *Acoustic Phonetics*. Language Monograph 23, supplement to *Language* 24.

Lewis, J. W. 1972. *A Concise Pronouncing Dictionary of British and American English*. London: Oxford University Press.

Lindsey, G. 2019. *English after RP*. Palgrave Macmillan.

Luick, K. 1896. Untersuchungen zur Englischen Lautgeschichte. Strassburg, Trubner.

Matthews, W. 1938. *Cockney Past and Present*. London, Routledge.

Orton, H & E. Dieth. 1962. *Survey of English Dialects: Introduction*. Leeds: Arnold.

Przedlacka, J. and M. Ashby. 2019. Comparing the Received Pronunciation of J. R. Firth and Daniel Jones: a sociophonetic perspective. *Journal of the International Phonetic Association* 49:381-400.

Siegel, A. & H. Sapru. 2015. *Essential Neuroscience*. Baltimore: Lippincott Williams & Wilkins. 3rd edition.

Sivertsen, E. 1960. *Cockney Phonology*. Oslo, University of Oslo.

Sweet, H. 1877. *Handbook of Phonetics*. London, MacMillan.

Wales, K. 2006. *Northern English*. Cambridge University Press.

Wells, J. C. 1982. *Accents of English*. Cambridge University Press (3 vols).

Wolfe, P. M. 1973. *Linguistic Change and the Great Vowel Shift in English*. University of California Press.

Wood, S. A. J. 1979. A radiographic analysis of constriction locations for vowels. *Journal of Phonetics* 7:25−43.

Wood, S. A. J. 1982. *X-Ray and Model Studies of Vowel Articulation*. *Working Papers* 23, Department of Linguistics, Lund University.

Wood, S. A. J. 2017. A spectrographic study of sound changes in nineteenth century Kent. In Tsudzuki, M. and M. Taniguchi (eds), *A Festschrift for Jack Windsor Lewis*. *Journal of the English Society of Japan* 21: 215-246.

Wyld, H. C. 1936. *A History of Modern Colloquial English*. Oxford: Blackwell, 3rd revised edition.

Question intonation in Southern Swedish

Merle Horne & Mikael Roll
Centre for Languages and Literature, Lund University

Abstract

Swedish has been generally assumed not to have any well-defined question intonation. However, with respect to Southern Swedish, Lindblad & Gårding (1973) present data showing final rises in polar (Y/N) questions. Gårding (1979) also assumes a broader and higher F0 range in Y/N questions than in statements, particularly on the final focussed word. The present study investigates the extent to which these question cues occur in a material involving spontaneous dialogues led by a speaker of Southern Swedish (Malmö). Results for utterances with question syntax show final rises in 16% of the Y/N questions and 3% of the Wh-questions. For utterances without question syntax, 79% were associated with final rises. Comparing F0 on the first prosodic word of questions, both Y/N- and Wh-questions showed a significantly higher F0 level (ca. 2-3 ST higher) than statements. A further comparison of the F0 level on the most prominent word following the first prosodic word in questions and statements showed that Y/N questions had a higher F0 on the most prominent word than statements. However, Wh-questions did not differ significantly from statements in that respect. Thus initial F0-level appears to be a strong prosodic cue distinguishing between questions (both Y/N questions and Wh-questions) and statements in the Southern Swedish material examined. F0 level on the most prominent word following the first prosodic word is also a reliable prosodic cue distinguishing Y/N questions from statements and Wh-questions, but not for distinguishing between Y/N questions and Wh-questions.

Introduction and background

It has been assumed that Swedish does not have any general prosodic cues to questions. Syntactic cues can distinguish questions from statements, and thus the use of intonational cues such as high boundary tones is not commonly found. Final rises have, however, been seen to occur in some contexts. House (2004, 2005) observed that Wh-questions in Central Swedish sometimes are associated with a final rise. In the material House examined (200 Wh-questions beginning with *vad* 'what' and *vem* 'who' extracted from human-computer dialogues), 22% of the Wh-questions had a final rise. Rises appeared most often in children's questions (32%), women had somewhat fewer rises (27%), and men, the fewest rises (17%). The rises were also observed to occur mostly on final focused words and were characterized by a delayed focal peak. Gårding & Lindblad (1973) and Gårding (1979) investigated differences in statement and Y/N question intonation in different Swedish dialects. They compared words uttered as statements and (echo) questions elicited from speakers from four dialect areas: Skåne (Southern Sweden), Stockholm (Central Sweden), Gotland, and Götaland (between Southern and Central Sweden). Results from their study showed that Y/N questions in Southern Swedish exhibited a final rise which resembled that in other Germanic languages (Gårding 1977). The rise occurred on the final syllable (see Figure 1). Gårding (1979) further assumes that intonation in Swedish Y/N questions is also characterized by an overall wider F0 register in comparison with statements, where the local widening of the F0 range is thought to be the most important cue to questions associated with the sentence (focal) accent (see Figure 2). In a perception study, House (2003) showed that, in

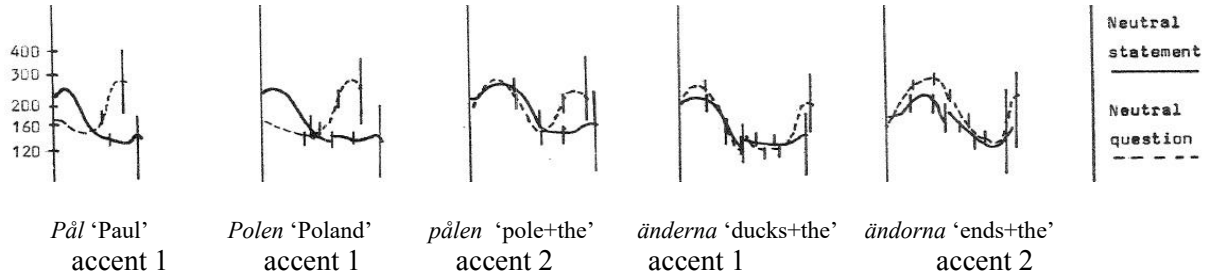


Figure 1. Figure adapted from Gårding & Lindblad (1973) showing differences between statement and question intonation on final focused words in Southern Swedish. Question intonation is characterized by a final rise. The vertical lines represent boundaries between segments.

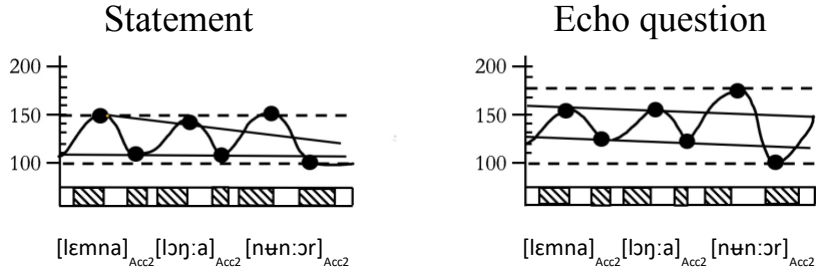


Figure 2. Figure adapted from Gårding (1979) showing statement (left) and echo question (right) intonation in Southern Swedish. Question intonation is characterized by a wider F0 register, most clearly visible on the rightmost prominent word in the utterance Lämna långa nunnor *'Leave tall nuns'.*

Central Swedish, a combination of raised F0 range and delayed utterance-final focal peak is an effective cue to interrogative mode in echo questions.

In order to investigate the extent to which intonational cues to questions occur in a more extensive material, we examined statements and questions (both Y/N questions and Wh-questions) in dialogues led by a speaker from Skåne (Malmö).

Method

Two spontaneous/extemporaneous dialogues between a well-known male Southern Swedish journalist (Sven Melander) and two different political leaders (Jonas Sjöstedt and Jimmy Åkesson) were used. The dialogues contained approximately 2 hours of speech in total. All Y/N questions (n=38) and Wh-questions (n=37) produced by the journalist were extracted from the dialogues. To compare question intonation to statements, 37 statements were also extracted from the material. The Y/N questions had Verb-Subject word order and the statements had Subject-Verb word order. The subjects were pronominal in 100% of the Y/N questions and in 95% of the statements. In order to compare the extent of final rises in questions with question syntax (wh-words, Subject-Verb inversion) and in utterances functioning as questions but without question syntax, we also extracted 24 utterances that were introduced as Y/N questions by the interviewer.

Presence of final rises in Y/N questions, Wh-questions, and utterances without question syntax was registered. In addition, to obtain a measure which could reflect differences in register width between statements and questions, the maximal F0 on the first prosodic word (PW) in all utterances and the maximal F0 on the prosodically most prominent word following the first prosodic word was also measured. In Y/N questions, the F0 on the first prosodic word corresponded to the highest F0 on the PW consisting of V-Pro and in statements on the PW consisting of Pro-V. In the 2 cases where the subject was non-pronominal, the F0 was measured on the subject noun. In Wh-questions, the first PW corresponded to the Wh-word. Examples of the three sentence types are given in Figures 3-5.

Statistical analyses were carried out using two-sided independent samples t-tests. P values were Bonferroni-corrected for multiple comparisons. An F0 difference of 1.5 ST was taken as a threshold for assuming a significant difference

between the categories. This threshold is based on results in Rietveld & Gussenhoven (1985), suggesting that a difference of 1.5 ST is sufficient to cause a difference in the perception of prominence.

Results

Final Rises

Rises in utterances with question syntax

In the Southern Swedish questions with question syntax, rises occurred in 6 Y/N questions (16%) and 1 Wh-question (3%).

Rises in utterances without question syntax

In the utterances without question syntax (n=24), 19 (79%) exhibited final rises. Examples are given in Figure 6 and Figure 7.

F0 register

First PW

The mean F0 on the first PW in Y/N questions (10.3 ST) and Wh-questions (11.1 ST) was significantly higher than in statements (7.6 ST), $p<0.001$. There was no significant difference in F0 on the first PW between Y/N questions and Wh-questions, $p=0.478$.

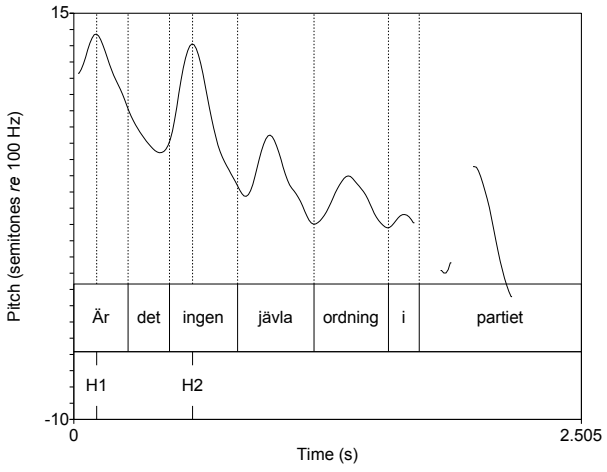


Figure 3. Example of a Y/N question: Är det ingen jävla ordning i partiet? *'Is there no damn discipline in the party?' H1 and H2 indicate the points where F0 was measured in the first prosodic word and the following most prominent word, respectively.*

Most prominent word after first PW

The mean F0 height on the most prominent word following the first PW was significantly greater in Y/N questions (12.6 ST) than in Wh-questions (10.9 ST), $p=0.033$, and statements (10.0 ST), $p=0.003$. There was no significant difference in F0 between Wh-questions and statements in this position, $p=0.290$.

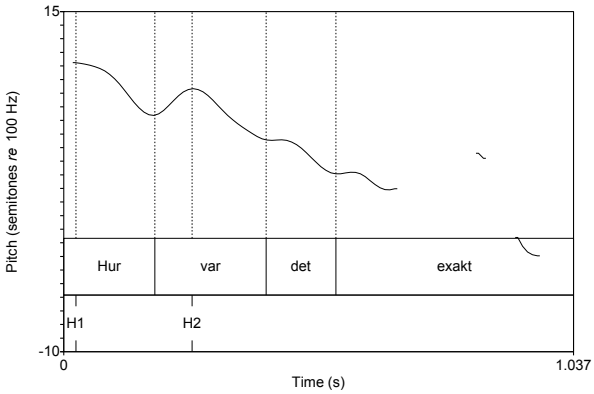


Figure 4. Example of a Wh-question: Hur var det exakt? *'How exactly was that?' H1 and H2 indicate the points where F0 was measured in the first prosodic word and the following most prominent word, respectively.*

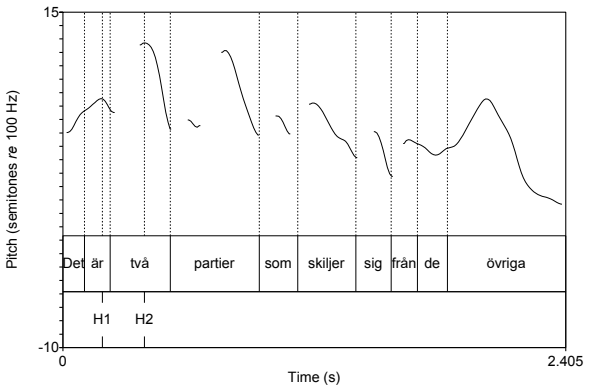


Figure 5. Example of a statement: Det är två partier som skiljer sig från de övriga. *'There are two parties that are different from the others'. H1 and H2 indicate the points where F0 was measured in the first prosodic word and the following most prominent word, respectively.*

Discussion

The results presented here show that question intonation in Southern Swedish can be realized through a number of different prosodic cues. Firstly, in line with previous results for Central Swedish (House (2004, 2005)), they illustrate the

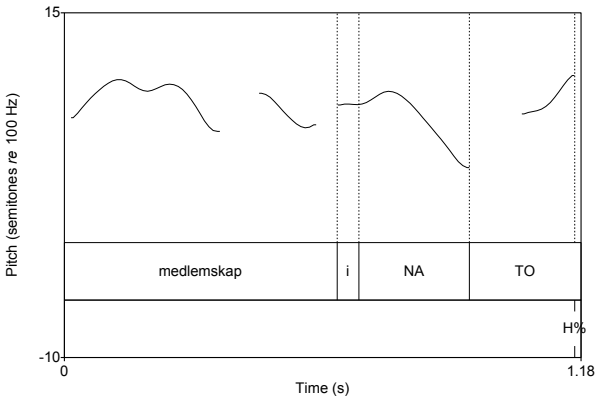


Figure 6. Example of a rise (H%) on the final syllable of NATO (accent 1) in the phrase medlemskap i NATO? *'membership in NATO?'*

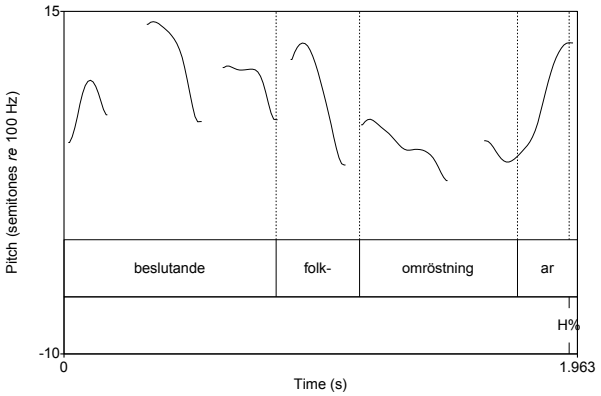


Figure 7. Example of a H% on the final syllable of the phrase beslutande folkomröstningar? *(accent 2) 'decisive referendums?'*

non-obligatoriness of final rises as cues to questions. They also show that the use of final rises in Y/N questions is greater in utterances that do not have question syntax. The results further indicate that syntactic differences interact with differences in F0 register at the onset of questions and statements. In line with the model for Swedish intonation presented in Gårding (1979), a higher register was observed for questions than statements, not only for Y/N questions as studied by Gårding, but also for Wh-questions. Moreover, in agreement with Gårding (1979), the wider register extended from the onset to the following most prominent (focussed) word, but only in Y/N questions. In Wh-questions, this continued heightened register was not found.

These findings on the relation between syntax and prosody are in line with approaches to grammar such as Rizzi & Bocci (2017), where operators representing different speech act types are assumed to be part of the syntactic representation. The register differences seen in the data studied here could be assumed to be intonational reflexes of these illocutionary force operators for Y/N questions and Wh-questions.

The differences in F0 height associated with prominences to the right of the first prosodic word can be thought to reflect the scope of the question operator. Whereas the question operator in Y/N questions is assumed to dominate focus, and have scope over the whole sentence, in Wh-questions, it is assumed to overlap with focus on the Wh-word (Rizzi & Bocci (2017)).

Acknowledgements

This work was supported by Knut and Alice Wallenberg Foundation (grant number 2018.0454), Swedish Research Council (grant number 2018.00632), Crafoord Foundation (grant number 2017.0006) and Marcus and Alice Wallenberg Foundation (grant number 2018.0021).

References

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

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

Gårding, E & Lindblad, P (1973). Consistency and variation in Swedish word accent patterns. *Working papers* (Dept. of Linguistics and Phonetics, Lund University) 7: 36-110.

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

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

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

Rietveld, T & Gussenhoven, C (1985). On the relation between pitch excursion size and prominence, *Journal of phonetics* 13: 299-308.

Rizzi, L & Bocci, G (2017). Left periphery of the clause primarily illustrated for Italian. In: M Everaert & H van Riemsdijk, *The Wiley Blackwell companion to syntax*, 2nd Ed., New Jersey: John Wiley & Sons, 1-30.

LUND
UNIVERSITY