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.

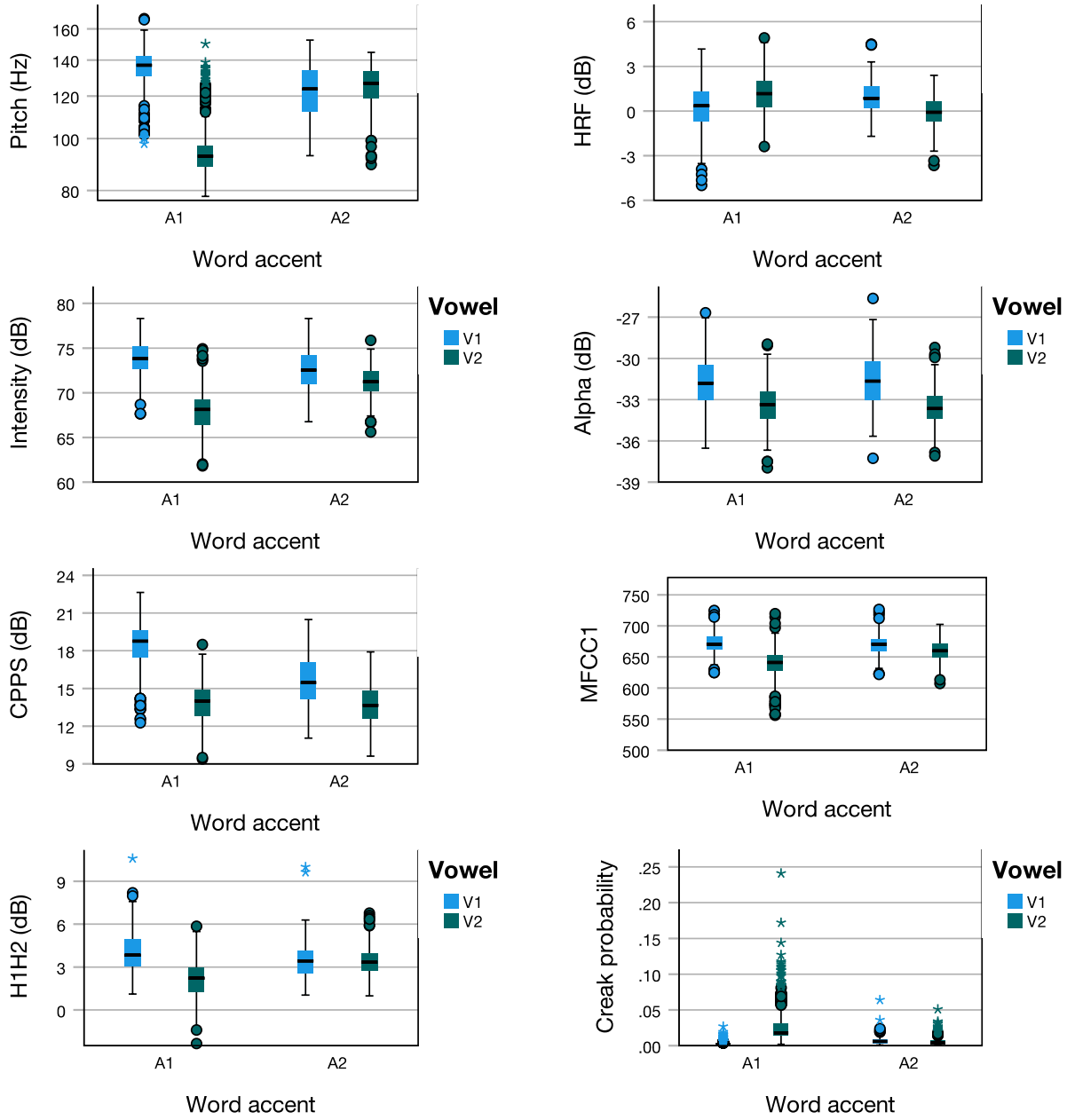

Figure 1. Box plots of acoustic features extracted from the vowels in the first and second syllables (V1 and V2) in words with word accent 1 and 2 (A1 and A2). The plots are based on all words in the speech material.

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/ic

phs-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