9th International Conference on Speech Prosody 2018
13-16 June 2018, Poznań, Poland

Classification of Swedish dialects using a hierarchical prosodic analysis

Marcin Włodarczak[1], Juraj Šimko[2], Antti Suni[2], Martti Vainio[2]

[1] Stockholm University, Sweden
[2] University of Helsinki, Finland

wlodarczak@ling.su.se
{juraj.simko,antti.suni,martti.vainio}@helsinki.fi

Abstract

The present study investigates dialectal variation of Swedish word accents by means of wavelet-based analysis of f_0 and energy. The analysis yields a measure of prosodic similarity between dialects expressed in terms of mutual perplexity of unigram models trained on derivatives of the wavelet-decomposed input features. A comparison of models trained on energy, f_0 and a combination of both features indicates that the energy+f_0 model reaches the highest classification accuracy, in line with the existing descriptions of tonal dialects in terms of the number and timing of pitch peaks with respect to the stressed syllable. At the same time, prosodic similarity between geographically close but typologically distinct dialects suggests an interaction between the traditional distinction between type-1 and type-2 dialects and areal variation, giving rise to northern and southern type-2 dialects (with little difference between 2A and 2B subtypes), and a parallel distinction between 1A and 1B varieties.
Index Terms: word accent, tonal dialect, prosodic typology

1. Introduction

To the delight of Scandinavian phoneticians, Swedish speakers use tonal word accents on top of lexical stress. The distinction consists of two accent types, 'Accent 1' ('acute') and 'Accent 2' ('grave'), associated with mono- or bisyllabic stems, respectively [1]. While the distinction is of limited use in terms of lexical contrast with only around 350 reported minimal pairs [2], it is has been described as highly sensitive to regional variation.

A preliminary description of Swedish tonal dialects was done by Meyer [3, 4], who distinguished three major groups, depending on the number and timing of pitch peaks with respect to the stressed syllable. Meyer's material was subsequently reanalysed by Gårding [5], who proposed five distinct dialects, summarised in Table 1. Briefly, the distinction between types 1 and 2 reflects the number of pitch peaks in Accent 2 (one or two peaks, respectively) while the distinction between types A and B is linked to the pitch peak timing (early or late, respectively). Gårding assigned each type to a specific area: type 1A being spoken in Southern Sweden, 1B in Gotland and Bergslagen, 2A in Central Sweden (Svea dialect region) and 2B in the area between Southern and Central Sweden. In addition, she included type 0, spoken in Finland and in the North of Sweden, which has no contrastive tonal word accents, as well as a subtype of 2A (not included in Table 1), used in Öland.

Bruce [6] compiled an updated map of Gårding's Swedish tonal dialects using a larger number of sites, sampled from the SweDia 2000 corpus [7]. The analysis was based on focally accented phrase-final occurrences of 'dollar' (Accent 1) and 'kronor' (Accent 2) in a sample of older men's speech (the number of analysed speakers was not specified). The distribution of accent realisations resembled to a large extent Gårding's classification of Meyers's material.

Table 1: *Tonal typology of Swedish dialects according to the number and timing of pitch peaks, based on [5].*

Type	Accent 1	Accent 2
0	—	—
	One peak	*One peak*
1A	Early in the stressed syllable	Late in the stressed syllable
1B	Late in the stressed syllable	Early in the post-tonic syllable
	One peak	*Two peaks*
2A	Late in the stressed or early in the post-tonic syllable	One in each syllable
2B	In the post-tonic syllable	One in each syllable

In addition to these manual methods, there have been some attempts at automatic classification of Swedish regional tonal variants, based on low-level acoustic features. For instance, Frid [8] used CART trees to predict, among others, Gårding's dialect in realisations of the words "dollar" and "kronor" from a subset of the SweDia 2000 locations. Three pitch parametrisation methods were used, based on (1) temporal and pitch-related properties of the first fall beginning before the stressed vowel offset, (2) f_0 level at the stressed vowel onset and (3) the Tilt model [9]. The best performing feature set (using method 1) achieved an accuracy of 59% against a random 20% baseline. Given that the same f_0 pattern might correspond to different word accent depending on the dialect and presence of narrow focus, separate classification experiments were conducted for accent type and focus. Overall, focused words and Accent 2 words reached higher accuracy than the other word classes.

More recently, Lidberg and Bromqvist [10] used Gaussian mixture models and convolutional neural networks trained on MFCCs and path signatures of a subset of SweDia 2000 wordlists to distinguish between five geographically defined dialects. The overall accuracy for the Gaussian classifier trained on individual words equalled 61%. Performance further improved (80%) when several single-word classifiers were combined. Dialect classification on spontaneous speech from three regions reach still higher accuracy (88%), which was expected given the reduced number of categories.

10.21437/SpeechProsody.2018-62

In the present paper, we revisit the problem of automatic identification of Swedish tonal dialects using a hierarchical prosodic analysis method. The method was previously used for comparison of languages and was found to capture typological relationships between language families [11]. Specifically, we use unigram models trained on f_0 and energy derivative (Δ) features of f_0 and energy signals decomposed using Continuous Wavelet Transform (CWT). The individual components have been demonstrated to accurately characterise familiar levels of the prosodic hierarchy, such as syllables, words and phrases [12]. Given that the regional variation of Swedish word accents involves both the number of peaks and their timing with respect to the stressed vowel, the hierarchical character of the analysis, preserving the relationships between distinct prosodic levels, is particularly well-suited for this task. The distances between dialects are then expressed in terms of a perplexity-based measure, allowing comparisons at a chosen level of granularity, from regions and dialects to villages to speakers to individual words.

2. Method

2.1. Material

The material from the present study consisted of wordlists recorded as part of the SweDia 2000 corpus of Swedish dialects collected in 107 locations around Sweden and Swedish-speaking areas of Finland in 1999. The wordlists were composed to represent both segmental and prosodic characteristics of the dialects and for this reason are not completely identical across recordings sites.

Given that SweDia has not been annotated for tonal dialect types and since it includes more sites than analysed by Gårding [5], the labels from [8], inferred from neighbouring locations in Gårding, were used. Borderline cases, for which could not be easily assigned to a single category, were excluded from the analysis. Altogether, the analysed material amounted to over 210,122 words from 86 different sites.

2.2. Prosodic analysis and language model comparison

Pitch was extracted with Praat [13] at 10 ms time step, using the two-step extraction procedure described in [14]. The resulting contour was then smoothed (10 Hz bandwidth) and the voiceless frames were interpolated linearly. Energy was obtained from downsampled (8 kHz) waveforms decomposed with wavelet transform (Morlet mother wavelet, $\omega_0 = 3$). Components with pseudoperiods of 0.25, 0.5 and 1 s were then summed to estimate amplitude envelope of the signal.

We have trained a separate unigram model for each SweDia location, using the procedure described in [11]. Briefly, derivatives (Δ) were calculated for 200, 400 and 800 ms pseudoperiod components obtained from the continuous wavelet transform (Morlet mother wavelet, $\omega_0 = 3$) of the energy and f_0 signals in Matlab. The values of the derivatives between the 5th and the 95th percentiles of each speaker were subsequently discretised into an odd number of bins. Three models were tested: models using energy and f_0 components only used a five-bin discretisation, while a combined energy $+ f_0$ model used only three states due to the greater number of parameters to learn. Each of the three bins effectively carries information about the slope of a particular signal component: rising, falling or relatively flat. Finally, the discretised components of all signals at each time point were combined into a single state. States with values falling outside the 5th-95th percentile interval were excluded.

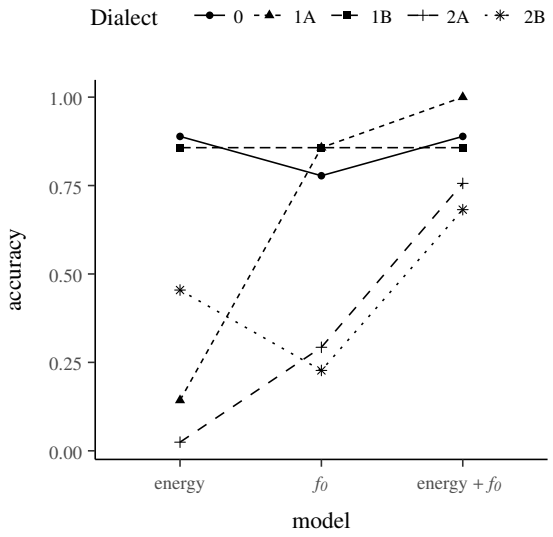

Figure 1: *Prediction accuracy for energy, f_0 and energy + f_0 models grouped by dialect.*

The models were evaluated by calculating perplexity (mean $-\log(p)$) for each word in the material. Subsequently, the individual perplexity values were averaged per-site and collected in a confusion matrix with cell [$site_i$, $site_j$] referring to the mean perplexity of the $site_i$ model across all words from $site_j$. The confusion matrix can be used for deriving mean perplexities for higher-order units. For instance, the mean perplexity of a $dial_n$ model on a $dial_m$ material can be obtained by averaging cells [$site_i$, $site_j$], for all i and j such that $site_i \in dial_n$ and $site_j \in dial_m$, etc. Similarly, the distance between $site_i$ and $site_j$ was calculated by averaging perplexities in [$site_i$, $site_j$] and [$site_j$, $site_i$]. We refer to this measure as *mutual perplexity*.

In addition, for each site we predict dialect type by selecting the category with the lowest per-row mean perplexity.

3. Results

Using both f_0 and energy, the dialect type has been predicted correctly in 78% of sites (67 out of 86 cases), against a 48% majority baseline. By comparison, energy- and f_0-only models achieved much lower accuracy, 30% and 42%, respectively. Figure 1, shows the same results grouped by dialect type. With the exception of 0, 1B and 2B dialects, addition of f_0 improves prediction accuracy over energy-only models, with further improvement when the two feature sets are combined. Inspection of confusion matrices (not included here) revealed that for 2B dialects, inclusion of f_0 resulted in increased confusion with 1A and 2A dialects, which, however, is outweighed by the cumulative improvement of energy and f_0. 0-type dialects achieve high accuracy for all three models (with a slight deterioration for the f_0-only model due to a misclassification of a single village). 1B dialects were completely unaffected by the feature set used.

Notably, the same feature set is quite good at predicting region (Götaland, Svealand, Norrland, Finland). Here, the accuracy of the energy $+ f_0$ model equals 72% (against a majority baseline of 36%). The model trained on f_0 achieves somewhat lower accuracy of 55% (but substantially higher than in case of

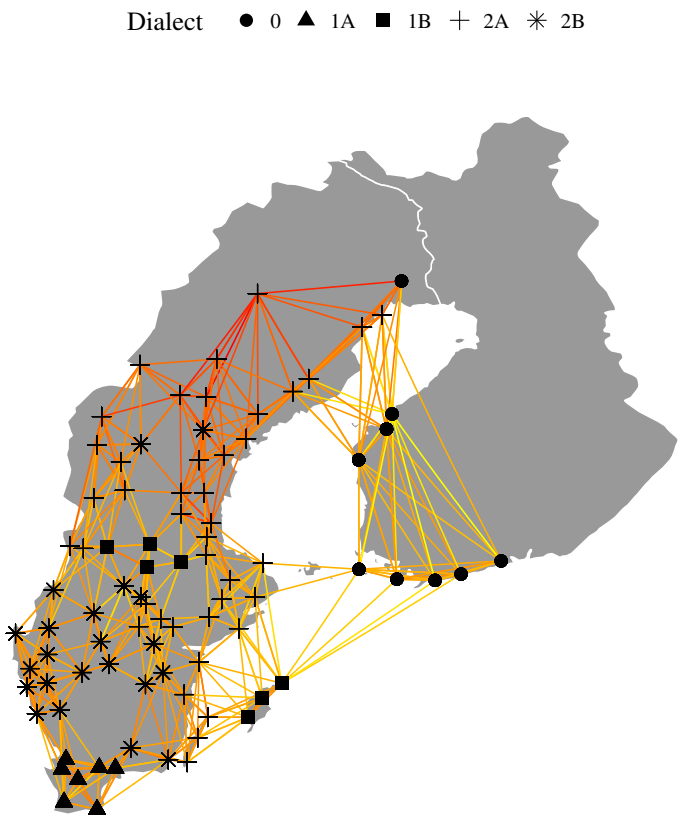

Figure 2: *Mutual perplexities for eight nearest neighbours of each recording site, grouped by tonal dialect. Red indicates low perplexity (and thus higher degree of prosodic similarity) and yellow indicates high perplexity.*

dialect type), with energy-only model performing worse than the majority baseline (34%).

Taken together, these results pose a question about an interaction between dialect type and region. To investigate this further, in Figure 2 we plot mutual perplexity for eight nearest neighbours of each SweDia site overlayed on a map of Sweden and Finland. Perhaps the most striking feature of that plot is lack of clear-cut dialect boundaries, which would manifest themselves as high-perplexity (low prosodic similarity) links between neighbouring locations belonging to different dialectal varieties. Instead, geographically close villages from different dialects show a high degree of similarity. For instance, all Norrland dialects (2A, 2B and even 0) form one prosodic cluster. By contrast, the variety of the 2A dialect spoken in Svealand seems to differ from its Norrland counterpart, forming (together with 1B) an area of relatively high-perplexity cutting across Central Sweden. Southern dialects and the Finnish Swedish dialects also form low-perplexity groupings.

In order to verify whether regional effects do in fact to some extent override (or mediate) dialectal classifications, we calculated mean mutual perplexity values for all region-dialect combinations and input the resulting distance matrix into a hierarchical clustering procedure, using the *hclust* function in R. The resulting dendrogram, depicted in Figure 3, indeed supports this hypothesis. Namely, Norrland dialects (including 0, which lacks the word accent distinction) cluster first, followed by Götaland's 2A and 2B varieties and Svealand 2A type. Notably, with the exception of Norrland-0 (represented by only one village) this part of the tree groups all occurrences of type-2 dialects. The highest-level links include Svealand-2A, Finland, as well as the 1B dialects spoken in Götaland and Svealand, the last two forming a separate branch.

4. Discussion and conclusions

The present paper presents an attempt at automatic description of the Swedish tonal dialects using a CWT-based hierarchical analysis of prosody, previously applied to classification of a typologically and genealogically varied sample of languages. By contrast, this study deals with what is a relatively a subtle variation of a prosodic parameter, spoken over a well-defined and limited territory. The difficulty of this task is indeed reflected in greater model perplexity values, indicating a greater degree of prosodic similarity between the analysed types.

Nevertheless, the results clearly demonstrate that inclusion of f_0 improves the overall accuracy of dialect classification over a model trained on energy-based features only, and combining the two features offers further improvement. This is expected given that the dialectal variation of Swedish word accents is predominantly characterised by the pattern of f_0 and its timing with respect to syllabic boundaries. Indeed, it is only when both energy and f_0 are used together that the overall accuracy ex-

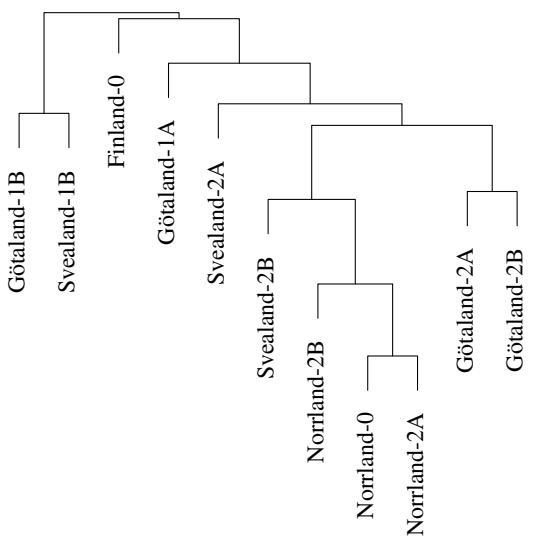

Figure 3: *Dendrogram of region and dialect type combinations based on mutual perplexity.*

ceeds the majority baseline. The combined effect of energy+f_0 can also be observed on the level of individual dialects, although it seems that the gain is greater for some dialects (1A) than for others (1B).

While the accuracy was calculated on the same data on which the models were trained, it should be noted that the training was not done with the particular task of dialect classification in mind. Rather, the models express general sequential characteristics of intensity and f_0 co-variation in the analysed material. In addition, the predicted label is based on the perplexity of an *averaged* dialect model rather than on the perplexity of a particular site. A formal evaluation of the results on unseen data is left for future research.

Notably, the models used in this work have a rather limited memory, represented by signal deltas. That the nevertheless manage to capture the variability of energy and f_0 as well as their relationship related to dialectal variability of tonal contrasts in Swedish testifies to the relative robustness of these effects. In addition, the material used was also more varied than in similar studies, which used models for specific words [8, 10]. By contrast, our wordlists were not entirely identical across the recording sites and were not controlled for the presence of monosyllabic words, where the dialectal differences are neutralised.

Given that comparable classification accuracy was achieved for regional variation (Götaland, Norrland, Svealand and Finland), we subsequently investigated the question to what extent dialectal variation is mediated by territorial proximity. In other words, we were interested whether speech material from different but geographically close dialects exhibits some degree of prosodic similarity. Mutual perplexity of neighbouring sites plotted in Figure 2 suggests that the familiar dialect labels are indeed at least partly overridden by geographical proximity. Specifically, the analysis employed in this work did not reveal presence of clear-cut boundaries between dialects characterised by low within-group and high between-group values of mutual perplexity. Rather, the results indicate regions of relatively high prosodic similarity (low mutual perplexity) in Northern and Southern Sweden with a high-perplexity belt cutting across Central Sweden.

Converging conclusions are suggested by the results of the hierarchical clustering in Figure 3, where the right-hand side of the dendrogram corresponds almost exclusively to type-2 dialects, with additional regional sub-groupings of Götaland and Norrland dialects (including the lone instance of Norrland-0). In particular, it seems that the distinction between 2A and 2B varieties plays a secondary role with no discernible clusters of either type. By contrast, the 1B dialect spoken in Götaland and Svealand are grouped together and are relatively distant from Götaland's 1A dialect. Predictably, Finland with its neutralisation of tonal word accents forms a separate branch in the tree.

Altogether, the results point towards a strong interaction between the traditional distinction between type-1 and type-2 dialects mediated by regional variation, giving rise to northern and southern type-2 dialects, and a parallel distinction between 1A and 1B varieties. The present study thus sheds new light on the typology of Swedish word accents found in literature. Notably, these insights were arrived at by means of a fully automatic analysis, allowing processing of large amounts of speech data, exceeding by several orders of magnitude the data set sizes found in conventional dialectological descriptions. In particular, unlike the parametrisation procedures used by Frid [8], the method does not require any knowledge of the segmental or suprasegmental structure of the data. Instead, the relevant landmarks are inferred from the energy signal.

In future work, we are planning to repeat the analysis on the other subsets of the SweDia data, namely the "prosody" set, consisting of words 'dollar', 'kronor' and 'D-mark' elicited under different focus conditions, and the "spontaneous" sets, involving unscripted conversations with an interviewer. While the former data set is particularly well suited for our purposes, preliminary analysis has revealed that it might involve too few repetitions for the models to generalise across individual sites.

5. Acknowledgements

This work was funded in part by a collaboration grant between Stockholm University and the University of Helsinki and in part by the Swedish Research Council project 2014-1072 *Andning i samtal* (*Breathing in conversation*) to the first author, and the Academy of Finland DLT project (No. 12933481) to the second author.

The authors would like to thank Johan Frid for sharing his dialect labels for the SweDia 2000 corpus.

6. References

[1] T. Riad, *The phonology of Swedish*. Oxford: Oxford University Press, 2014.

[2] C.-C. Elert, "Tonality in Swedish: Rules and a list of minimal pairs," in *Studies for Einar Haugen*, E. S. Frichow, G. Kaaren, N. Hasselmo, and W. O'Neil, Eds. The Hague: Mouton, 1971, pp. 151–173.

[3] E. A. Meyer, *Die Intonation im Schwedischen. Die Svea-mundarten*. Helsingfors: Fritzes, 1937.

[4] ——, "Die Intonation in Schwedischen. II: Die norrländischen Mundarten," *Stockholm Studies in Scandinavian Philology*, vol. 11, 1954.

[5] E. Gårding, *The Scandinavian word accents*. Lund: Gleerup, 1977.

[6] G. Bruce, *Vår fonetiska geografi. Om svenskans accenter, melodi och uttal*. Lund: Studentliteratur AB, 2010.

[7] A. Eriksson, "SweDia 2000: A Swedish dialect database," in *Babylonic Confusion Resolved: Proceedings of the Nordic Symposium on the Comparison of Spoken Languages*, ser. Copenhagen Working Papers in LSP, P. J. Henrichsen, Ed., no. 1, 2004, pp. 33–48.

[8] J. Frid, "Lexical and acoustic modelling of Swedish prosody," Ph.D. dissertation, Lund University, Lund, 2003.

[9] P. Taylor, "Analysis and synthesis of intonation using the Tilt model," *The Journal of the Acoustical Society of America*, vol. 107, no. 3, pp. 1697–1714, 2000.

[10] D. Lidberg and V. Blomqvist, "Swedish dialect classification using artificial neural networks and gaussian mixture models," Master's thesis, Chalmers University of Technology, Gothenburg, 2017.

[11] J. Šimko, A. Suni, K. Hiovain, and M. Vainio, "Comparing languages using hierarchical prosodic analysis," in *Proceedings of Interspeech 2017*, Stockholm, Sweden, 2017, pp. 1213–1217.

[12] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, vol. 45, pp. 123–136, 2017.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.0.24)," Computer program, 2017, http://www.praat.org/.

[14] C. De Looze and D. Hirst, "Detecting changes in key and range for the automatic modelling of intonation," in *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 2008, pp. 135–138.