Dialect Identification of Spoken North Sámi Language Varieties Using Prosodic Features

Kakouros, Sofoklis; Hiovain, Katri; Vainio, Martti; Šimko, Juraj

2020

http://hdl.handle.net/10138/316167

Kakouros, S, Hiovain, K, Vainio, M & Šimko, J 2020, Dialect Identification of Spoken North Sámi Language Varieties Using Prosodic Features. in Proceedings of 10th International Conference on Speech Prosody 2020 : Communicative and Interactive Prosody. Speech prosody, ISCA, Baixas, pp. 625-629, The 10th International Conference on Speech Prosody, Tokyo, Japan, 25/05/2020. https://doi.org/10.21437/SpeechProsody.2020-128

10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Dialect Identification of Spoken North Sámi Language Varieties Using Prosodic Features

Sofoklis Kakouros[1], *Katri Hiovain*[1], *Martti Vainio*[1], *Juraj Šimko*[1]

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

{sofoklis.kakouros,katri.hiovain,martti.vainio,juraj.simko}@helsinki.fi

Abstract

This work explores the application of various supervised classification approaches using prosodic information for the identification of spoken North Sámi language varieties. Dialects are language varieties that enclose characteristics specific for a given region or community. These characteristics reflect segmental and suprasegmental (prosodic) differences but also high-level properties such as lexical and morphosyntactic. One aspect that is of particular interest and that has not been studied extensively is how the differences in prosody may underpin the potential differences among different dialects. To address this, this work focuses on investigating the standard acoustic prosodic features of energy, fundamental frequency, spectral tilt, duration, and their combinations, using sequential and context-independent supervised classification methods, and evaluated separately over two different units in speech: words and syllables. The primary aim of this work is to gain a better understanding on the role of prosody in identifying among the different language varieties. Our results show that prosodic information holds an important role in distinguishing between the five areal varieties of North Sámi where the inclusion of contextual information for all acoustic prosodic features is critical for the identification of dialects for words and syllables.

Index Terms: prosodic analysis, North Sámi language, dialect comparison, dialect identification, under-resourced languages

1. Introduction

Dialects can be generally defined as language varieties that characterize the form of the language within a specific, typically geographic, region or community (see, e.g., [1]). The perception of dialectal cues in speech is particularly important as it has wide ranging implications in shaping attitudes and signalling the origin of the speaker [2]. Overall, the task of identifying a dialect is similar to the more general problem of language identification, where, however, distinguishing among dialectal varieties introduces more challenges as it involves subtler differences between variants of the same language.

Previous studies have indicated that the perception of language differences seems to be dependent on a complex interplay of segmental (phonemes), suprasegmental (prosody, phonotactics), and higher-level information (such as lexemes) (see, e.g., [3, 4, 5]). The same information seems to be relevant also for dialect identification (see, e.g., [5, 6, 7]). For instance, it has been shown that adults can distinguish between languages that are prosodically similar by making use of either rhythmic timing or pitch information [5].

In automatic language identification (LID), most methods are based on the processing of lexical, phonotactic, and acoustic features (see, e.g., [8]). In general, the task of a LID system is to classify a given spoken utterance into one out of many languages. Similarly, for automatic dialect identification (DID), the task is to classify an utterance among many spoken dialects within a language. Overall, DID has remained relatively unexplored compared to LID, perhaps due to the inherent challenges in DID (dialectal similarities tend to be much higher compared to languages) but also due to the limited data resources available.

Most systems for DID make use of an array of approaches that can be categorized into the following: (i) lexical, (ii) phonotactic, and (iii) acoustic. Lexical and phonotactic techniques are typically based on extracting word or phone representations using, for instance, *n*-gram statistics (see, e.g., [8, 9, 10, 11]). Acoustic approaches utilize acoustic features, such as Mel Frequency Cepstral Coefficients (MFCCs), that are then typically modelled with an i-Vector framework (see, e.g., [12, 13]). Currently, the state-of-the-art methods in DID and LID are based on using deep bottleneck features (BNF) in order to extract frame-level features for i-Vector systems (see, e.g., [13]).

Overall, in the context of LID and DID, prosody has not been studied extensively, perhaps due to the inherent challenges in modelling the dynamic nature of prosodic phenomena that entail temporal variations across long sequences in speech. To address this, in this work we examine the prosodic differences between the five language varieties of North Sámi and determine how important is the role of prosody in distinguishing among the five dialects. Our target is not to improve over existing state-of-the-art DID systems, but rather to gain an understanding of the importance and relevance of prosody in dialect identification but also to potentially identify important prosodic factors to consider in the development of future systems. Next, the North Sámi language and its dialects are briefly presented.

1.1. The North Sámi language and dialects

The North Sámi language is an indigenous and endangered language with speakers covering a geographic area spanning the northern regions of three countries: Finland, Norway, and Sweden. North Sámi is the most widely spoken among the Sámi languages with approximately 25 000 speakers [14, 15]. As a minority language, practically all North Sámi speakers are bilingual in Sámi and the majority language of the country they live in. However, North Sámi is a majority language in only two regions, in Kautokeino and Karasjoki in Norway [14].

North Sámi belongs to the Uralic language family and is related to, for instance, Finnish, Estonian, and Hungarian. There are 10 Sámi languages altogether [16] that differ from each other in many linguistic aspects such as phonology, morphology, syntax, and lexicon. In general, adjacent Sámi languages can be mutually intelligible.

The North Sámi language is further divided into three main dialect groups: Sea Sámi (northernmost coastal area in Norway), Finnmark Sámi (northern Finland and Norway), and Torne Sámi (west from Finnmark Sámi speaking area, and

10.21437/SpeechProsody.2020-128

Table 1: *Number of speakers, number of recordings, and overall duration of recordings for each variety in the extended DigiSami corpus. Light gray denotes the recordings from Finland and dark gray those from Norway.*

Location/Varieties	Speakers (female)	Recordings	Minutes
Inari (**sin**)	4 (2)	245	43:54
Ivalo (**siv**)	7 (6)	403	52:58
Utsjoki (**sut**)	7 (5)	496	67:13
Kautokeino (**skt**)	9 (4)	609	91:23
Karasjoki (**skr**)	5 (3)	305	47:31

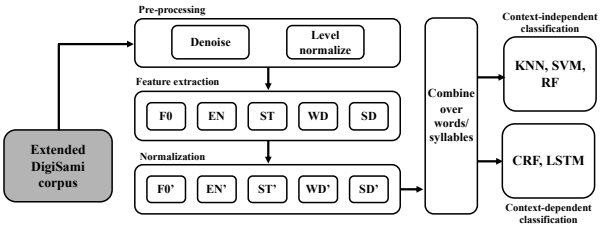

Figure 1: *Processing pipeline for the signals.*

northernmost parts of Sweden) [15]. The data used in this work represents the Finnmark North Sámi variety. The Finnmark North Sámi dialect group is still traditionally divided into Western and Eastern subdialects [15]. The speech data used for this work represents both of these subdialects and speakers from both Norway and Finland. This means that in addition to the dialectal differences (mostly concerning phonetic and [morpho]phonological features), there are two majority languages, Finnish and Norwegian, which are in constant language contact with the Sámi languages [14].

In previous works on North Sámi language varieties, differences between the five dialects have been identified using i-vectors [17] and evidence of their typological differences have been observed [18]. In the present study we investigate the potential prosodic differences among the five regional varieties of North Sámi and explore whether the geographical dispersion of North Sámi also underpins disparities in the production of North Sámi that are reflected in the acoustic prosodic characteristics of North Sámi speakers. Moreover, we investigate the features and units of analysis that hold a role in discriminating between the dialects.

2. Materials and Methods

The material used in this study consists of North Sámi continuous speech from recordings of speakers from five geographically distinct locations. Analysis of the data is based on the extraction of standard acoustic features that are commonly utilized for speech analysis (see, e.g., [19]).

2.1. The extended DigiSami corpus

The DigiSami corpus [20] is a collection of read speech recordings from five locations that have been traditionally inhabited by the Sámi: Inari, Ivalo, and Utsjoki in Northern Finland, and Kautokeino and Karasjoki in Northern Norway. The task of the speakers was to read aloud Wikipedia articles about Sámi languages and traditional costumes. The corpus consists of speech data from 25 (16 female) native North Sámi speakers (16-65 years of age) recorded using a 4-channel portable recorder (Roland Edirol R-4 Pro) and condenser mini-lavalier microphones (AKG C 417L). The data also contain annotations at the word, sentence, and phrase level that were produced with a combination of manual annotations and forced alignment (using WebMAUS Basic, see [21, 22]). For a more detailed description of the data collection process and corpus, see [23, 17, 20].

To extend the original DigiSami corpus [20] (due to the limited duration of the corpus recordings), additional recordings were collected from seven native Sámi speakers (4 female, age range 22-64) at two locations, namely, at Oulu University (Finland) and Kautokeino (Norway) (see Table 1 for an overview). Specifically, two speakers were recorded at Oulu University (originally from Utsjoki and Ivalo), and five speakers at Kautokeino (four originally from Kautokeino and one from Karasjoki). The recording process was the same as in the DigiSami corpus [20]: the speakers read aloud the same Wikipedia articles in North Sámi (with the inclusion of some additional sentences). For the recordings, a Zoom H2n portable recorder with five built-in microphones (using Mid-Side recording mode) was used. The new data were annotated with the same procedure as in the original DigiSami corpus to make the data compatible with the DigiSami corpus (e.g., annotations at the word, sentence, and phrase level). This resulted, in total, to 32 speakers and approximately 5 hours of speech data for the *extended DigiSami corpus*.

In the experiments, a 4-fold evaluation procedure was used where speakers from each dialect where split into four groups to form batches of approximately equal duration in minutes – no speaker occurring at the same time in the training and test sets. Although this resulted in balanced speech data within each dialectal group, this was not the case across dialectal groups where the batch duration was different for each dialect (see Table 1). However, given the current data, this type of data division was deemed optimal for the current experiments. The 4-fold evaluation process was repeated five times (with different training and test sets) and the final results were averaged.

2.2. Feature extraction

Extraction of the acoustic features from the speech signals first involved a pre-processing stage that was followed by computation of the features and normalization.

2.2.1. Pre-processing

As majority of the speech recordings were carried out in non-ideal conditions, thus introducing varying types of noise as well as irregular sound levels across the recordings, all signals were first passed through a pre-processing pipeline. Following manual inspection of the signals it was observed that the noise component present within individual recordings was relatively stationary. Therefore, spectral noise subtraction was utilized to denoise the signals [24]. Specifically, each recording was first denoised (using spectral noise subtraction in Praat [25]) and subsequently level normalized (see also Fig. 1). This process led to an improved subjective quality of the recordings and also resulted in higher signal-to-noise ratio (SNR).

2.2.2. F0, energy, spectral tilt, and duration

Following pre-processing of the speech data, each recording was initially downsampled to 8 kHz and, subsequently, features were computed using a 25-ms window and 5-ms frame

shift. F0 was computed using the YAAPT pitch tracking algorithm [26], signal energy was computed according to Eq. (1) (where x denotes the signal, t the current sample, and w the frame length; see, e.g., [19]), and spectral tilt by computing the mel frequency cepstral coefficients (MFCCs) and taking the first (C1) MFCC (see, e.g., [27, 28]). Word durations were extracted from the corpus annotations. In addition to words, syllable segmentations were estimated from the recordings using a signal envelope-based harmonic oscillator algorithm [29]. This segmentation method has been shown to compare well with other syllabification methods (see [29] for more details and comparison). In the present study, the oscillator was set to a centre frequency of 5 Hz with critical damping (Q = 0.5).

$$EN(t) = \sum_{\tau=-\frac{w}{2}}^{\tau=\frac{w}{2}-1} |x(t+\tau)|^2 \quad (1)$$

2.2.3. Normalization

To account for inter- and intra-talker variation as well as the relationship of some features with their subjective perception, all raw feature values were normalized. Specifically, energy was logarithmically normalized (e.g., [30]), F0 was semitone normalized relative to the median F0 for each speaker according to Eq. (2), spectral tilt was z-score normalized per speaker, and durations (words/syllables) were logarithmically normalized.

$$F0'(t) = 12 \cdot \log_2 \left(\frac{F0(t)}{F0_{median}} \right) \quad (2)$$

2.3. Statistical descriptors

Five statistical descriptors were computed over words and syllables utilizing the normalized feature values (except for word/syllable duration as descriptors cannot be computed for single values). Specifically, the *mean*, *standard deviation*, *minimum*, *maximum*, and *range* (defined as the difference between the maximum and minimum feature value during a word/syllable) were computed for all data. Words and syllables were selected as the level for analysis as they provide a good basis for the evaluation of potential differences in the data.

2.4. Supervised classification

To explore the potential importance of contextual information in the classification of the five dialectal varieties, both sequence classification and context-independent classification methods were utilized. Specifically, three standard context-independent classification algorithms were selected: (i) *k*-nearest neighbours (kNN), (ii) support vector machines (SVM), and (iii) random forests (RF). These algorithms do not make use of the sequence of their inputs. To account for input ordering, as prosodic information is expected to extend across segmental content to temporal segments that spread further than a single segment and might span words and whole utterances, two sequence classifiers were used: (i) linear chain conditional random fields (CRF) and (ii) long short-term memory (LSTM) recurrent neural networks.

The hyperparameters for all classifiers were optimized for best performance in a subset of the test data. Specifically, for the context-independent classifiers and for *k*NN, the number of *k* was set to 10 after heuristically searching for the best *k*, for SVM, a radial basis function (RBF) kernel was used with C = 100 and σ = 12.0790, and for RF, a classifier with 50 decision trees was utilized. For the sequence classifiers, CRFs were trained using belief propagation and L2 regularization. The maximum number of iterations for CRF training was set to 100 (ensuring training error convergence) and the regularization term to 1. Finally, for the LSTMs, a unidirectional architecture with one hidden LSTM layer containing 128 cells was used and was trained forwards with delays from 0 to 10 frames. Training was carried out with a mini-batch size of 128, 200 epochs, and learning rate of 0.1. At the output layer, a softmax activation function was used to provide the class probability values.

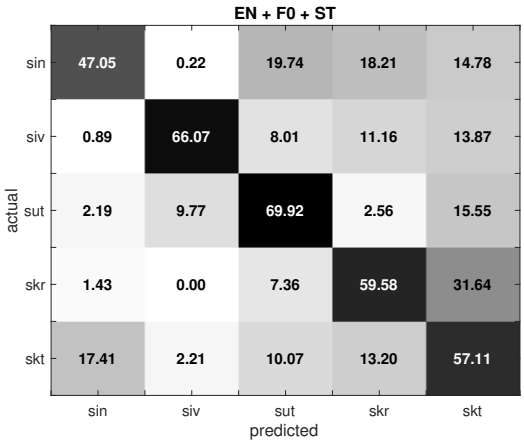

Figure 2: *Confusion matrix for the combination of energy, F0, and spectral tilt, over words for the CRF classifier.*

2.5. Evaluation

Classification performance was measured both at the word- and syllable-level separately for all classifiers. For each reference unit (word/syllable) the label of the original dialect was compared to the hypotheses provided by the classifiers. Performance was measured in terms of the unweighted average recall (UAR) and classification accuracy.

3. Results

Experiments were run in a 4-fold evaluation setup (see section 2.1) for energy, F0, spectral tilt, duration, and all possible combinations, with features computed over words and syllables separately, and trained using *k*NNs, SVMs, RFs, CRFs, and LSTM networks. A second setup was also tested with the same evaluation procedure, but with the difference that training and testing vectors included the two previous and two forthcoming (word/syllable) vectors to account for contextual dependencies in the data.

Results from the first setup indicate that both the word and syllable context is very important for the classification of the five dialects of Sámi (see Fig. 2 and Fig. 3). In particular, for the context-independent classifiers (for words), performance for F0 and EN is near the random baseline (20%), and for ST, it is higher, at approximately 35%. When context information is introduced with the sequence classifiers, the features of F0 and EN get an increase in performance, both reaching with CRFs a UAR of approximately 34% and for ST 50.8%. Similarly, for the overall best feature combination of EN, F0, and ST, context-independent classifiers reached a performance between 35-40% whereas context-dependent reached up to 60%. Performance for both words and syllables is similar with the exception that

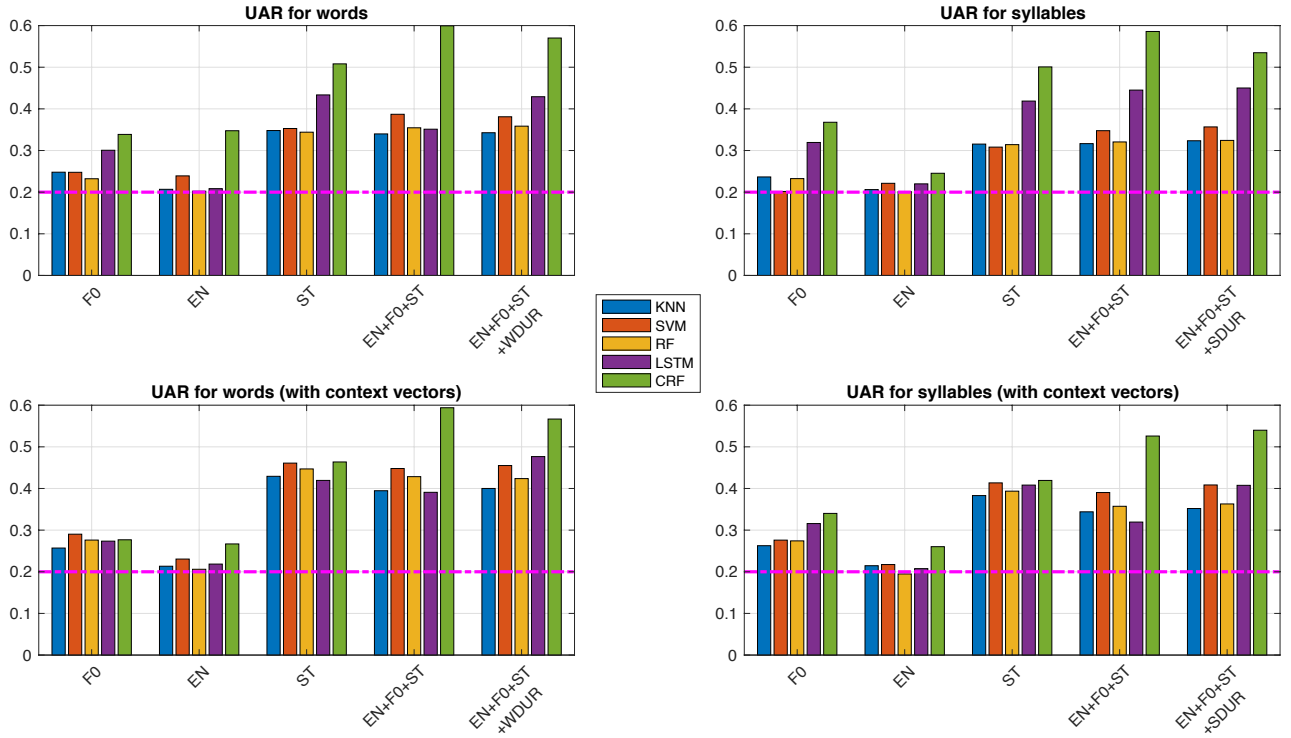

Figure 3: *UAR for all classifiers over syllables and words. Horizontal dash dotted line denotes the random baseline performance. Top panel: results with context-independent vectors. Bottom panel: results with context vectors.*

the LSTM network seems to perform better over syllables.

In the second setup, context information was included in the feature vectors. The results provide further support for the importance of acoustic prosodic context in identifying North Sámi dialects. In this case, the performance improved for the context-independent classification methods, where, for instance, for SVM and words, F0 reached a UAR of around 30% and ST 46%. EN did not change in performance and the best overall UAR was 58% for a combination of EN, F0, and ST with CRFs.

Finally, observing the individual classification performance of the five dialects (Fig. 2) for the best feature combination of EN, F0, and ST for CRFs, it can be seen that it ranges between 47.05% and 69.92%. In particular, the lowest overall performance was for Inari and the highest for the Utsjoki dialect. In general, the performance across the dialects varied with the different classifiers but mostly with the different features and feature combinations utilized. These differences in the performance of prosodic features for the distinct dialects likely reflect dialectal differences in the use of prosodic cues.

4. Discussion and Conclusions

The results presented in this work provide evidence that prosodic information is important for the identification of the differences among the five North Sámi dialects. In particular, from the individual features' perspective, F0, EN, ST, and duration (word/syllable) did not reach high performance when considered independently of their context. However, when context information was introduced, their capacity to discriminate the dialects improved considerably, with the overall best individual feature performance reached for ST (50.8%) and the best combination for EN, F0, and ST (60%). Both the syllable- and word-level experiments provided good performance, indicating that in the current task, beyond the selection of the unit of analysis (words/syllables), the inclusion of longer temporal contexts is likely a more important factor. The good performance of ST, a measure that quantifies the relative contribution of high versus low frequency bands of the spectrum[31, 27], is indicative that ST carries relevant prosodic information for the dialects. However, ST also introduces some segmental information that might have helped in improving performance.

Overall, both the Finnish Sámi language varieties (69.6% for Utsjoki, 66.1% for Ivalo, 47.1% for Inari) and Norwegian Sámi (59.6% for Karasjoki, 57.1% for Kautokeino) clustered well. It is noteworthy that identifying dialects seems to be a difficult task also for human listeners. For instance, across four regional varieties of Dutch, listeners were able to correctly identify 90% the country of origin, 60% the region, and 40% the province [32, 33].

In future work, it will be interesting to extend the experiments to include more sequence classification methods that take into account forward and backward states (bidirectional LSTM networks) and include more data in the evaluations. In addition, beyond the acoustic prosodic investigations, it would be also of interest to include linguistic data in the analysis (phonotactics and lexical analysis).

5. Acknowledgements

This study was funded by the Academy of Finland project *Digital Language Typology: Mining from the Surface to the Core* (project no. 12933481). We also thank Kristiina Jokinen and the DigiSami project for the permission to use their speech data.

6. References

[1] M. Sumner and A. G. Samuel, "The effect of experience on the perception and representation of dialect variants," *Journal of Memory and Language*, vol. 60, no. 4, pp. 487–501, 2009.

[2] D. R. Preston and G. C. Robinson, "Dialect perception and attitudes to variation," *Language in Society*, vol. 36, p. 133, 2005.

[3] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904–1911, 2007.

[4] M. Barkat and I. Vasilescu, "From perceptual designs to linguistic typology and automatic language identification: Overview and perspectives," in *Seventh European Conference on Speech Communication and Technology*, 2001.

[5] C. Vicenik and M. Sundara, "The role of intonation in language and dialect discrimination by adults," *Journal of Phonetics*, vol. 41, no. 5, pp. 297–306, 2013.

[6] C. G. Clopper and D. B. Pisoni, "Some acoustic cues for the perceptual categorization of american english regional dialects," *journal of phonetics*, vol. 32, no. 1, pp. 111–140, 2004.

[7] C. G. Clopper and A. R. Bradlow, "Native and non-native perceptual dialect similarity spaces," *Proceedings of ICPhS XVI*, pp. 665–668, 2007.

[8] S. Khurana, M. Najafian, A. M. Ali, T. Al Hanai, Y. Belinkov, and J. R. Glass, "Qmdis: Qcri-mit advanced dialect identification system." in *Proceedings of Interspeech*, 2017, pp. 2591–2595.

[9] M. Najafian, S. Khurana, S. Shan, A. Ali, and J. Glass, "Exploiting convolutional neural networks for phonotactic based dialect identification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5174–5178.

[10] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "ivector approach to phonotactic language recognition," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[11] A. Ali, N. Dehak, P. Cardinal, S. Khurana, S. H. Yella, J. Glass, P. Bell, and S. Renals, "Automatic dialect detection in arabic broadcast speech," *arXiv preprint arXiv:1509.06928*, 2015.

[12] Q. Zhang and J. H. Hansen, "Language/dialect recognition based on unsupervised deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 5, pp. 873–882, 2018.

[13] T. N. Trong, V. Hautamäki, and K. A. Lee, "Deep language: a comprehensive deep learning approach to end-to-end language recognition," in *Odyssey: the Speaker and Language Recognition Workshop*, 2016.

[14] A. Aikio, L. Arola, and N. Kunnas, "Variation in North Saami," *Globalising sociolinguistics: Challenging and expanding theory*, pp. 243–255, 2015.

[15] P. Sammallahti, *The Saami languages: an introduction*. Davvi girji, 1998.

[16] U. M. Kulonen, I. Seurujarvi-Kari, and R. Pulkkinen, "The Saami - a cultural encyclopedia," *Globalising sociolinguistics: Challenging and expanding theory*, 2005.

[17] K. Jokinen, T. N. Trong, and V. Hautamäki, "Variation in spoken North Sámi language." in *Proceedings of Interspeech*, 2016, pp. 3299–3303.

[18] K. Hiovain, A. S. Suni, J. Simko, and M. T. Vainio, "Mapping areal variation and majority language influence in North Sámi using hierarchical prosodic analysis," in *Proceedings of the 9th International Conference on Speech Prosody*. International Speech Communication Association, 2018.

[19] S. Kakouros and O. Räsänen, "3PRO–an unsupervised method for the automatic detection of sentence prominence in speech," *Speech Communication*, vol. 82, pp. 67–84, 2016.

[20] K. Jokinen, K. Hiovain, N. Laxström, I. Rauhala, and G. Wilcock, "Digisami and digital natives: Interaction technology for the North Saami language," in *Dialogues with Social Robots, Proceedings of the International Workshop on Spoken Dialogue Systems (IWSDS-2016)*. Springer, 2017, pp. 3–19.

[21] F. Schiel, "Automatic phonetic transcription of non-prompted speech," in *Proceedings of ICPhS*, 1999, pp. 607–610.

[22] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[23] K. Jokinen, "Open-domain interaction and online content in the Saami language." in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 2014, pp. 517–522.

[24] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.

[25] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]. version 5.3.13," *retrieved from http://www.praat.org/*, 2012.

[26] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008.

[27] S. Kakouros, O. Räsänen, and P. Alku, "Comparison of spectral tilt measures for sentence prominence in speecheffects of dimensionality and adverse noise conditions," *Speech Communication*, vol. 103, pp. 11–26, 2018.

[28] S. Kakouros, O. Räsanen, and P. Alku, "Evaluation of spectral tilt measures for sentence prominence under different noise conditions." in *Proceedings of Interspeech*, 2017, pp. 3211–3215.

[29] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[30] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *Bell System Technical Journal*, vol. 12, no. 4, pp. 377–430, 1933.

[31] A. M. Sluijter and V. J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *The Journal of the Acoustical society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[32] R. Van Bezooijen and C. Gooskens, "Identification of language varieties: The contribution of different linguistic levels," *Journal of language and social psychology*, vol. 18, no. 1, pp. 31–48, 1999.

[33] E. A. McCullough, C. G. Clopper, and L. Wagner, "Regional dialect perception across the lifespan: Identification and discrimination," *Language and speech*, vol. 62, pp. 115–136, 2019.