INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Comparing languages using hierarchical prosodic analysis

Juraj Šimko[1], *Antti Suni*[1], *Katri Hiovain*[1,2], *Martti Vainio*[1]
[1]University of Helsinki, Finland
[2]University of Tampere, Finland
firstname.secondname@helsinki.fi

Abstract

We present a novel, data-driven approach to assessing mutual similarities and differences among a group of languages, based on purely prosodic characteristics, namely f_0 and energy envelope signals. These signals are decomposed using continuous wavelet transform; the components represent f_0 and energy patterns on three levels of prosodic hierarchy roughly corresponding to syllables, words and phrases. Unigram language models with states derived from a combination of Δ-features obtained from these components are trained and compared using a mutual perplexity measure. In this pilot study we apply this approach to a small corpus of spoken material from seven languages (Estonian, Finnish, Hungarian, German, Swedish, Russian and Slovak) with a rich history of mutual language contacts. We present similarity trees (dendrograms) derived from the models using the hierarchically decomposed prosodic signals separately as well as combined, and compare them with patterns obtained from non-decomposed signals. We show that (1) plausible similarity patterns, reflecting language family relationships and the known contact history can be obtained even from a relatively small data set, and (2) the hierarchical decomposition approach using both f_0 and energy provides the most comprehensive results.

Index Terms: language comparison, prosodic typology, wavelet transform, statistical modelling

1. Introduction

Contacts between neighboring communities leave marks on the languages they speak; the languages in contact often lend each other words, expressions, grammatical and morphological elements as well as their melodic and rhythmic patterns – prosody. Presumably, the dynamics of the transfer differs between these aspects of languages: shared features of grammar and morphology often bear witness to ancient contacts embodied in language family labels, while lexical items keep being exchanged more freely and frequently throughout the history.

Comparing prosodic characteristics of multiple languages is complicated by a multidimensionality of the task; the comparisons need to span and combine intonational and rhythmic properties, resulting stress patterns, as well as suprasegmental features such as presence and realisation of quantity contrast, tonality, etc. Relatedly, development of a coherent prosodic typology is hampered by the lack of a language independent prosodic transcription system [1]. Several existing studies have tried to supplant this shortcoming by adjusting the Autosegmental Metrical approach [2] for grouping languages according to lexical and postlexical intonational features (tone, stress, pitch accent languages) or devising various durational measures [3] to divide them into rhythm classes (mora-, syllable-, stress-timed languages) [4, 5, 6].

In this paper we propose an alternative, purely data-driven approach to assessing mutual similarity among languages from statistical distributions of patterns obtained from hierarchically decomposed f_0 and energy envelope signals. We follow a similar approach to Cummins et al. [7] who, inspired by language recognition techniques, trained recurrent LSTM networks on f_0 and energy contours on a speech corpus containing multiple languages and used the performance of resulting languages models on different languages as a measure of inter-language distance.

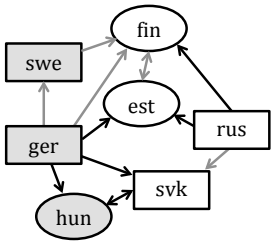

Figure 1: *Family and contact relationships between the analysed languages. Indo-European languages in rectangles (shaded are Germanic, empty are Slavic), Finno-Ugric in ovals, (shaded is Ugric, empty are Finnic). The arrows schematically depict the intensity of historical language contact, the arrows pointing in the prevalent direction of the influence, darker arrows represent more intense contact.*

Our approach presented here differs in several aspects. First, we use much more simple statistical models, namely unigrams on f_0 and energy envelope Δ-features (derivative). Second, more importantly, we train the language models on signals decomposed using Continuous Wavelet Transform (CWT) technique [8]. The individual signal components correspond to intonation and energy patterns pertaining to three levels of prosodic hierarchy, namely syllables, (prosodic) words and phrases. This facilitates, in principle, capturing relationships between changes in f_0 and energy at these hierarchical levels in parallel (e.g., f_0 movement mid-syllable at the beginning of a word towards the end of a phrase). Finally, we use a perplexity measure as a basis for quantifying distances among languages.

We use this approach to compare seven languages from two families: Finno-Ugric Finnish, Estonian and Hungarian and Indo-European German, Russian, Slovak and Swedish. All these languages are spoken in Europe and have a long and complex history of mutual contact, depicted in a broad outline in Fig. 1.

This contact history, historic as well as ancient, is presumably reflected in prosodic characteristics of these languages (see, e.g, [9, 10, 11, 12, 13, 14]). One of those presumed to be relevant for the present work is, for example, a word stress pattern: while the Finno-Ugric languages and Slovak have fixed, word-initial stress, the lexical stress in the remaining languages is not fixed. Swedish, unlike the other languages considered here, has contrastive lexical tones. Also, several of the languages (Finno-Ugric and Slovak) have fully-blown phonolog-

http://dx.doi.org/10.21437/Interspeech.2017-1044

ical quantity systems (and, at least in Finnish and Estonian, the quantity contrast is co-signalled by pitch movement [19, 20]), the quantity contrast of the other languages (Swedish and German) is more limited and linked to vowel quality; the Russian is said not to have phonological quantity whatsoever. Prosody is of course involved in signalling other aspects such as phrasal boundaries, sentence modality, etc.

The aim of this work is to verify whether simple language technology methods can capture some of these similarities between the studied languages, and therefore help reveal underlying typological relationships. Also, we address a broader question of whether, and to what extent, can prosody alone encode language contact and family relationships.

2. Methodology

2.1. Material

Small spoken material corpora from the analysed languages were used (see Tab. 1). For Russian, we used the first 10 sentences of the Phonetically balanced text from CoRuSS corpus [15]. For the remaining languages, a version of the North Wind and the Sun story was used; Estonian recordings come from the Estonian North Wind and the Sun Corpus [16], German from the Bavarian Archive for Speech Signals (BAS) and Hungarian from a material used in a previous study [17]. The Finnish, Slovak and Swedish corpora were recorded for the present study.

Table 1: *Number of speakers, sentences in the text and the overall duration of the given language corpus.*

Language	Spkrs (female)	Sentences	Seconds
Estonian (**est**)	6 (3)	8	207
Finnish (**fin**)	7 (3)	6	226
German (**ger**)	9 (4)	5	349
Hungarian (**hun**)	6 (3)	7	213
Russian (**rus**)	5 (5)	10	178
Slovak (**svk**)	6 (3)	6	176
Swedish (**swe**)	4 (2)	5	138

2.2. Prosodic analysis and language model comparison

For each sentence, the f_0-contour was extracted using the standard Praat pitch extraction routine with time step of 10 ms and pitch range of 120–320 and 80–250 Hz for female and male speakers, respectively [18]. The unvoiced intervals were subsequently (linearly) interpolated and the resulting contour was smoothed (10 Hz bandwidth). Signal envelope (energy) was calculated for each utterance (and sampled at the same time points as f_0) as follows. First, the waveform, downsampled to 8 kHz, was decomposed using wavelet transform (Morlet mother wavelet, $\omega_0 = 3$) to components with pseudo-frequencies of 0.25, 0.5, 1, and 2 s. Subsequently, the energy signals obtained for each component were summed to yield a close approximation of signal energy envelope.

The f_0 and energy signals obtained this way were decomposed using continuos wavelet transform (Morlet mother wavelet, $\omega_0 = 2$) to three components with pseudo-frequencies of 200 and 800 ms and 1.6 s, roughly corresponding to syllable, word and phrase durations, respectively (see the solid lines in the shaded box in Fig. 2). A derivative (Δ-feature) was calculated for each component (dashed curves in Fig. 2). For each speaker, all the derivative values were collected, and divided to

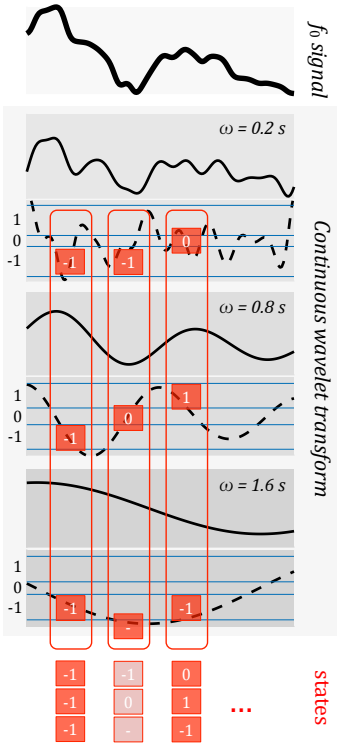

Figure 2: *Converting a signal to states depicting prosodic hierarchy, see text for details.*

an odd number of bins with equal number of elements using appropriate speaker-dependent percentiles (the bin boundaries are shown as horizontal lines in Fig. 2). The values below 5th and above 95th percentiles were marked as inadmissible for subsequent state calculation.

As shown in Fig. 2, the bin labels for all components and all signals considered in evaluation (see below) were combined to form a state for every time point. The states for which any value fell outside the admissible 5–95th percentile range was excluded from evaluation (see the middle one of the three depicted states in Fig. 2).

For each language, a unigram language model was calculated depicting likelihood of occurrence of each particular state in all utterances of the given language in the corpus. We evaluate the models using only f_0 and energy contours, respectively, and a combined model using both the f_0 and energy Δ-features. Also, we evaluated a model using the Δ-features calculated directly from the f_0 and energy signals (combined) without the prior wavelet decomposition. For the wavelet based models we have limited the number of discretization bins to 3; in effect simply depicting whether the signal component is going up, down or staying relatively constant. For the non-wavelet approach with considerably fewer possible states we use 11 bins.

Finally, for each sentence in the corpus (the same as used for the model training) we calculated the perplexity value for each language model. (Perplexity depicts the "surprise," the mean –log-probability, of the model exposed to the given set of states.) A confusion matrix was calculated with each [$lang_1$, $lang_2$] cell depicting the average perplexity of the $lang_1$ model (on the x-axis in Fig. 3) across all sentences in the $lang_2$ (y-axis in Fig. 3). The dendrogram depicting similarity among the language models was subsequently drawn.

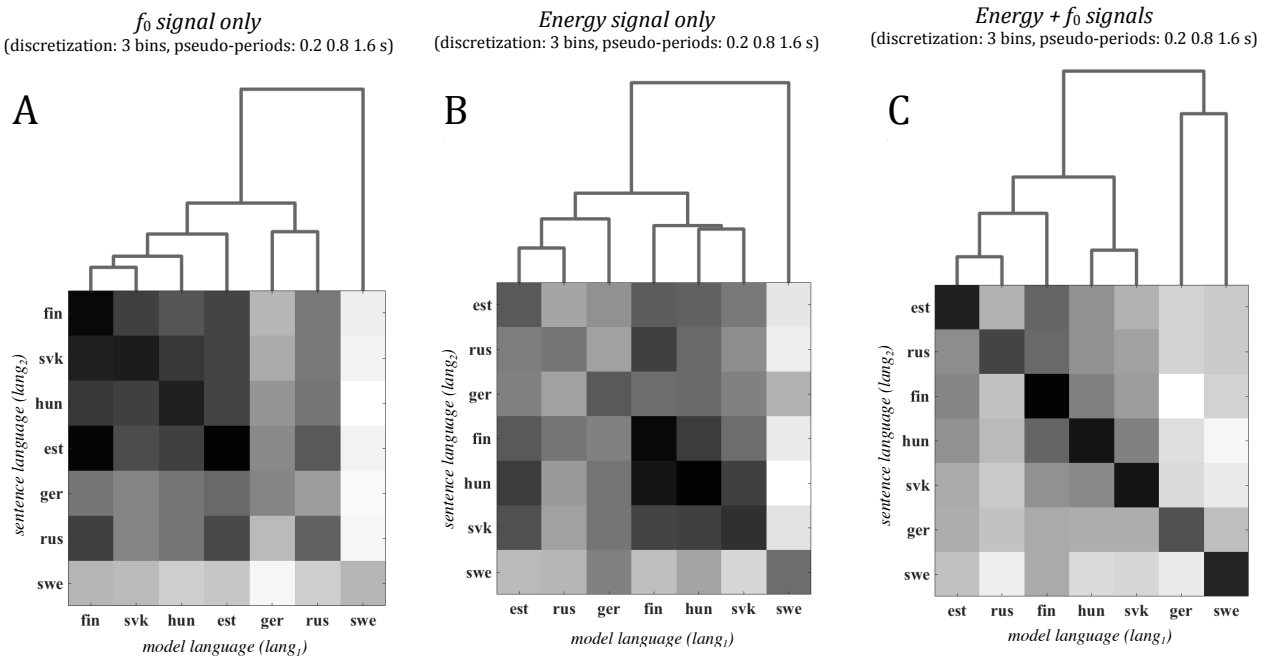

Figure 3: *Clustering the languages and confusion matrix obtained using wavelet-decomposed signals.*

3. Results

Fig. 3 shows the confusion matrices and derived dendrograms for language models using f_0 only, energy only, and combination of both f_0 and energy features, respectively; all signal are decomposed with continuous wavelet transform with 3 scales corresponding to pseudo-frequencies of 0.4, 0.8 and 1.6 s.

The confusion matrices depict the average perplexity of the language models with sentences in different languages, the brighter the shade the higher the perplexity. Unsurprisingly, the models are generally the least perplexed with the sentences from their own language, the very sentences they were trained on. Note, however, several exceptions, for example the energy-feature **est** model that finds **hun** and **svk** sentences less perplexing than its "own" **est** ones (the first column in the matrix in Fig. 3B). Also, for the f_0-based models, **est** sentences yield less perplexity for the **fin** model than for the **est** one (the fourth row in Fig. 3A). The combined f_0+energy models provide more robust "language recognition" ability – in terms of perplexity with their own *versus* other language material – than the single feature models. This is likely due to the considerably lower number of possible states accounted for by the latter (27) compared to the former ($27^2 = 729$).

Each set of features produces somewhat different distances among the languages (language models). The f_0 only features (Fig. 3A) group **fin** with **svk**, and then with **hun** and **est**. This cluster is then grouped with a **ger**–**rus** pair. **swe** is found quite different from the other languages.

The energy only approach also groups **fin**, **hun** and **svk**, the last two are slightly more similar to each other than either is to **fin**. **est** is found similar to **rus** and is consequently moved to the **rus**–**ger** cluster. Once again, **swe** is different.

The combination of f_0 and energy features presents somewhat different picture. **est**, once again close to **rus**, is grouped with the fellow Finnic language, **fin**. This cluster is further grouped with the **hun**–**svk** branch. The Germanic languages, **ger** and **swe**, are grouped together and found rather distinct from the other languages.

For comparison, in Fig. 4 we present the dendrograms obtained using directly the f_0 and energy Δ-features, without wavelet-based decomposition (the corresponding confusion matrices are not shown). The signal derivatives were discretized instead of the wavelet components, and used directly to train the unigram language models. The numbers of discretization bins were chosen to approximate the complexity of the wavelet-based models, i.e., 5 yielding 25 states (close to 27 of the single feature wavelet models) and 27 (729 states, the same as in the combined wavelet approach).

The two resulting dendrograms are neither identical with each other, nor with the corresponding picture (using the same prosodic signals) in Fig. 3C. In all three, **est** and **rus** are grouped closely together, and **swe** is different from most languages. The grouping of **est**, **rus**, **hun** and **svk** closely together derived from 5-bin discretization can also be seen in Fig. 3C. **fin**, however, is grouped with **est** and **rus** in both 27-bin based comparison and the wavelet-based approach, but with **swe** – and far from **est**–**rus** branch – when 5 discretization bins are used.

4. Discussion

Before discussing the possible interpretations of the results, we need to address some potential issues with the reported work. First, the corpus used for this pilot project is very small, moreover, all speakers of the same language utter the same set of sentences. This somewhat limits the scope of the modelled prosodic variation for each language and increases the danger of overfitting; nevertheless, our approach provides a meaningful clustering of languages despite these limitations.

Second, to warrant generalizability of the language models, we would ideally want to train the models on a different data set than that subsequently used for comparison. The main reason we abandoned this standard machine-learning principle is the diminutive size of our multi-lingual corpus[1]. Also, it is not

[1] In order to test the generalizability and mitigate the differences be-

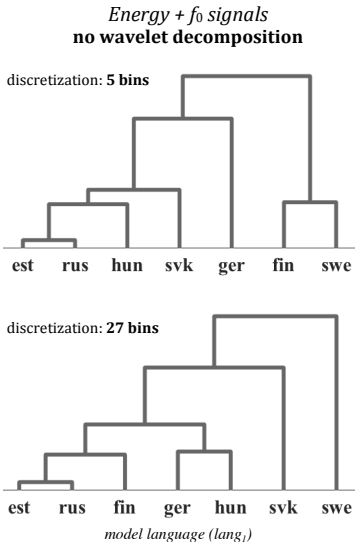

Figure 4: *Clustering the languages obtained using non-decomposed f_0 and energy signals.*

our aim at this stage to design a fully-fledged language recognition system based on prosody. In fact, the language models only take into account the positive examples from their own language; the language comparison is performed on the full corpus that, for each individual model, predominantly consists of the other-language material not used for its training. Nevertheless, when testing our approach on larger corpora in the future we will divide the data to training and testing subsets; this will help us to assess the potential usability of our approach for building, for example, a prosody-based language recognition system.

Third, in contrast with, e.g., language recognition task, in our enterprise of comparing languages based on their prosodic characteristics we lack a clearly defined ground truth, that is an independently given correct grouping of the tested languages. The lack of such ground truth is in fact related to the pioneering nature of the present investigation. In any case, we need to evaluate our results using the known language family relationships, suprasegmental and prosodic characteristics of the languages (such as quantity systems, stress patterns) and the history of mutual language contacts outlined in Introduction.

Despite these limitations, our approach yields meaningful results that largely reflect the known prosodic characteristics of the tested languages. In most cases, the quantity languages with word-initial lexical stress – **est**, **fin**, **hun** and **svk** – are found mutually closer to each other than to other languages (quantity contrast, as shown at least for **est** and **fin**, is co-signalled by pitch patterns [19, 20]). This grouping is most prominent for the f_0 wavelet-based model (Fig. 3A) with **fin–svk–hun–est** forming a similarity group (in this order), subsequently joined by **ger–rus** branch (languages with non-fixed stress pattern), all these languages found dissimilar to **swe**, the only language with tonal patterns. Wavelet-decomposed energy envelope (Fig. 3B) keeps the **fin–svk–hun** group, but finds **est** similar to **rus** and both these languages closer to **ger** than to the previous group.

The wavelet-based approach using both f_0 and energy envelope (Fig. 3C) presents probably the most comprehensive picture. Only in this case, the Germanic **ger** and **swe** are grouped together, with the remaining languages forming a separate branch. Within this branch, **hun** and **svk** with the long contact history and many documented prosodic similarities (despite different language families) form a sub-branch. Another sub-branch contains **est**, **rus** and **fin**. Grouping **fin** and **est** close together based on many shared prosodic characteristics, language affinity and the history of contact is expected. The closeness of **est** and **rus** in the dendrogram (repeated in all other results except f_0-based wavelet approach) is, in our opinion, more difficult to justify. Despite the very long history of language contact between these two languages (e.g., most Estonians born before 1980, and many younger ones, would have good command of Russian) we find this result spurious: there are simply too many differences in prosodic characteristics between these languages[2]. In any case, this persistent finding deserves further investigation in the future.

The dendrograms produced directly from the f_0 and energy signals (Fig 4) deviate further from our expectations than the wavelet-based ones. **est** and **rus** are still found very similar, but neither quantity / word-initial stress languages nor Germanic languages are grouped together consistently. Similarly to the approaches using the single signal (f_0 or energy, respectively) **swe** is kept further apart (grouped with **fin** in one case), **ger**, **svk** and **hun** form an intermediate group with differing similarity judgements depending on sensitivity of analysis (number of discretization bins).

Overall, the presented results support the viability of our approach. The wavelet decomposition of the signals allows for statistical evaluation of f_0 and energy envelope movement distribution patterns on multiple hierarchical levels (syllables, words, phrases). Perhaps even more importantly, it takes into account the mutual interdependencies between these patterns across the hierarchical levels; our results suggest that this type of hierarchical decomposition is beneficial for the classifying (and perhaps also language-recognition) task. Also, in line with the findings of [7], the results indicate that although f_0-based Δ-features alone serve very well for the language comparison task (cf. [21]), combining them with energy envelope features seems to further improve classification results.

The present pilot work will be expanded in several directions in the future. More (and more diverse) languages, with considerably larger corpora will be incorporated. Also, in the future work we will explore a more complex and state-of-the-art language modelling techniques, primarily (deep) recurrent networks. Finally, a method of direct, semiautomatic identification of the prosodic patterns primarily responsible for the detected differences between the languages – in essence the patterns yielding the highest perplexity in another language – is currently under preparation.

5. Acknowledgements

This work was funded by the Academy of Finland DLT project (No. 12933481). We are greatly indebted to Pärtel Lippus, Katalin Mády, Klára Vicsi, Tatiana Kachkovskaia, Daniil Kocharov, Štefan Beňuš and Marcin Włodarczak for their help with recording and sharing the speech material.

tween number of speakers for different languages, we also trained the models on randomly selected subsets consisting of three speakers for each language. The language comparison was done on the full corpus. Despite some variability, the clustering results–not reported here–were not considerably different from the reported ones.

[2]Given the lack of the ground truth brings a temptation to justify almost any possible pattern as a result of the complex mutual contact history and/or family-relationships within the small group of languages. We don't want to fall into this trap.

6. References

[1] D. Hirst and A. Di Cristo, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.

[2] C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the 1992 International Conference on Spoken Language Processing, ICSLP*, 1992, pp. 12–16.

[3] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.

[4] S.-A. Jun, *Prosodic typology: The phonology of intonation and phrasing*. Oxford University Press on Demand, 2006, vol. 1.

[5] L. M. Hyman, "Word-prosodic typology," *Phonology*, pp. 225–257, 2006.

[6] D. Gil, "A prosodic typology of language," *Folia Linguistica*, vol. 20, no. 1-2, pp. 165–232, 1986.

[7] F. Cummins, F. Gers, and J. Schmidhuber, "Automatic discrimination among languages based on prosody alone," Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland, Tech. Rep., 1999.

[8] A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," *Computer Speech & Language*, 2016.

[9] M. Erelt, T. Erelt, and K. Ross, *Eesti keele käsiraamat*. Eesti keele sihtasutus, 1997.

[10] A. Hakulinen, M. Vilkuna, R. Korhonen, V. Koivisto, R. T. Heinonen, and I. Alho, *Iso suomen kielioppi*. Helsinki: Suomalaisen Kirjallisuuden Seura, 2004.

[11] K. Mády, U. Reichel, and Š. Beňuš, "Accentual phrases in Slovak and Hungarian," in *Proc. of the 7th international conference on Speech Prosody*, Dublin, Ireland, 2014, pp. 752–756.

[12] Š. Beňuš and K. Mády, "Stress and phonemic length in the perception of Slovak vowels," in *Proc. of the 6th international conference on Speech Prosody*, Shanghai, China, 2012.

[13] D. Jones and D. Ward, *The phonetics of Russian*. Cambridge University Press, 2010.

[14] G. Bruce and O. Engstrand, "The phonetic profile of Swedish," *Sprachtypologie und Universalienforschung*, vol. 59, no. 1, p. 12, 2006.

[15] T. Kachkovskaia, D. Kocharov, P. Skrelin, and N. Volskaya, "CoRuSS — a new prosodically annotated corpus of russian spontaneous speech," in *Proc 10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

[16] P. Lippus. (2016) Estonian North Wind and the Sun Corpus v.1.0.3. [Online]. Available: https://doi.org/10.15155/1-00-0000-0000-0000-00129L

[17] K. Vicsi, V. Imre, and K. Mészáros, "Voice disorder detection on the basis of continuous speech," in *5th European Conference of the International Federation for Medical and Biological Engineering*. Springer, 2011, pp. 86–89.

[18] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 6.0. 18)[computer program]," 2016.

[19] M. Vainio, J. Järvikivi, D. Aalto, and A. Suni, "Phonetic tone signals phonological quantity and word structure," *The Journal of the Acoustical Society of America*, vol. 128, no. 3, pp. 1313–1321, 2010.

[20] P. Lippus, E. L. Asu, P. Teras, and T. Tuisk, "Quantity-related variation of duration, pitch and vowel quality in spontaneous estonian," *Journal of Phonetics*, vol. 41, no. 1, pp. 17–28, 2013.

[21] A. E. Thymé-Gobbel and S. E. Hutchins, "On using prosodic cues in automatic language identification," in *International Conference on Spoken Language Processing*, vol. 3, 1996, pp. 1768–1772.