Mumbling through a wall: Clustering Slavic dialects using hierarchical statistical modeling of prosody

Juraj Šimko[1], Ruprecht von Waldenfels[2,3], Michael Daniel[3], Nina Dobrushina[3], Achim Rabus[4], Antti Suni[1], Katri Hiovain[1] and Martti Vainio[1]

[1]University of Helsinki, Finland; [2]University of Jena, Germany; [3]National Research University Higher School of Economics, Moscow, Russia; [4]University of Freiburg, Germany

Abstract. *Several Slavic dialects and varieties were automatically clustered based solely on prosodic characteristics of spontaneous speech, namely f_0 contours and energy envelope (i.e., the delexicalized information analogous to speech as heard through a wall). A cross-entropy among unigram models derived from these prosodic characteristics was used as a measure of similarity between the dialects and varieties. Our analysis shows that the method can be used to modify groupings of dialects and varieties traditionally based on historical morphology and phonology and/or synchronic isoglosses. Namely, the results expound an influence of majority language on prosodic characteristics of minority languages: a variety of Eastern Slavic Rusyn spoken in Poland is clustered with a Western Slavic Polish dialect rather than with other Eastern Slavic varieties in the corpus.*

Introduction

1. Investigation of similarities in terms of prosody among Slavic varieties spoken over a large geographic area using an automatic procedure [1].
2. Influence of a West Slavic majority language on prosody of an East Slavic language

Speech Material

Informal interviews with speakers of:

- 2 distinct ***Russian*** dialects of Ustja (**UST**) and Rogovatka (**ROG**)
- 2 varieties of ***Rusyn*** spoken in Lemko (**LEM**) in Poland, and in Transcarpathia (**TRA**) in Ukraine
- a ***Polish/Slovak*** dialect spoken in **Spisz** (**SPS**) region of Poland

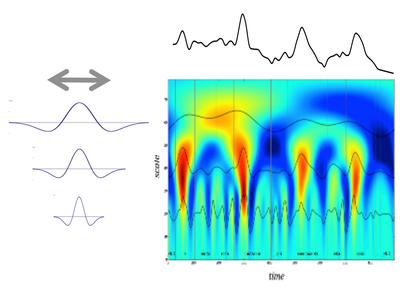
LEM and **SPS** informants are also speakers of West Slavic Polish, others of East Slavic Russian or Ukrainian.

Recordings of 3−5 female speakers born between 1922 and 1952 per dialect.

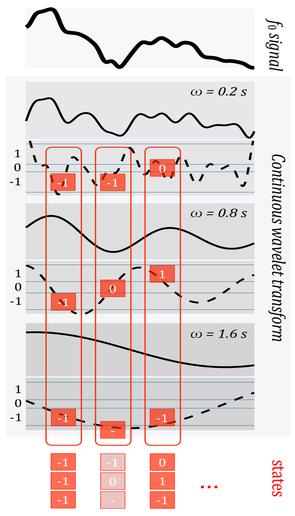
2–4 hours of speech for each variety except of **TRA** Rusyn (about 30 mins).

Methodology

1. f_0 and energy extracted using Praat, (linearly) interpolated and smoothed



2. Continuous wavelet transform of the f_0 and energy signals, three scale functions used with pseudo-frequencies of 200 ms, 800 ms, and 3.2 seconds, corresponding to syllables, words and phrases.



3. Derivatives of the signals (Δ-features) quantized to 3 levels (rising, falling, flat), individually for each speaker. Derivatives exceeding 5th and 95th percentiles were excluded.
   The six quantized signals (3 scales for f0 and 3 scales for energy) were combined, yielding 729 possible states.

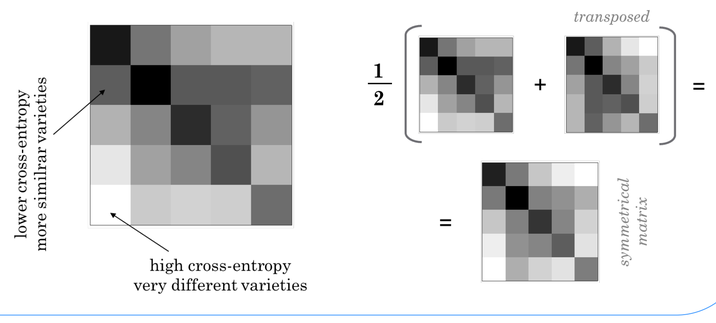
4. Simple unigram models (probabilities of individual states) were calculated for each language:

$$\text{for each state } S, \text{ compute } P_{\text{SPS}}(S), P_{\text{LEM}}(S), P_{\text{TRA}}(S), P_{\text{ROG}}(S), P_{\text{UST}}(S)$$

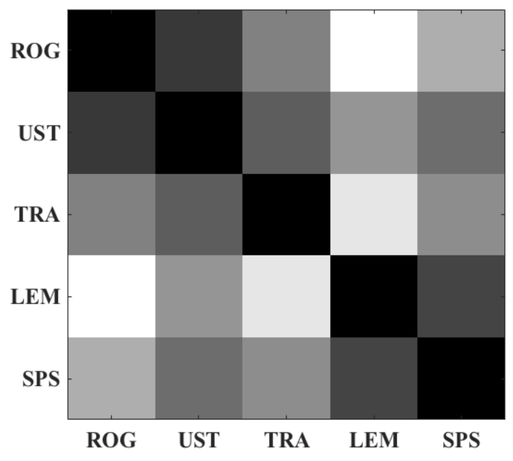
5. For each pair of varieties, a cross-entropy between the respective models was used as a measure of prosodic similarity

$$-\sum_{i=1}^{N} P_{\text{LAN1}}(S_i) \log_2 P_{\text{LAN2}}(S_i)$$

6. A mutual cross-entropy of two models was calculated as a mean of the two cross-entropy measures between the models

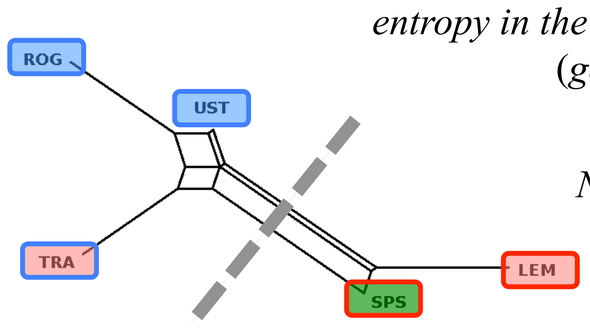


Results



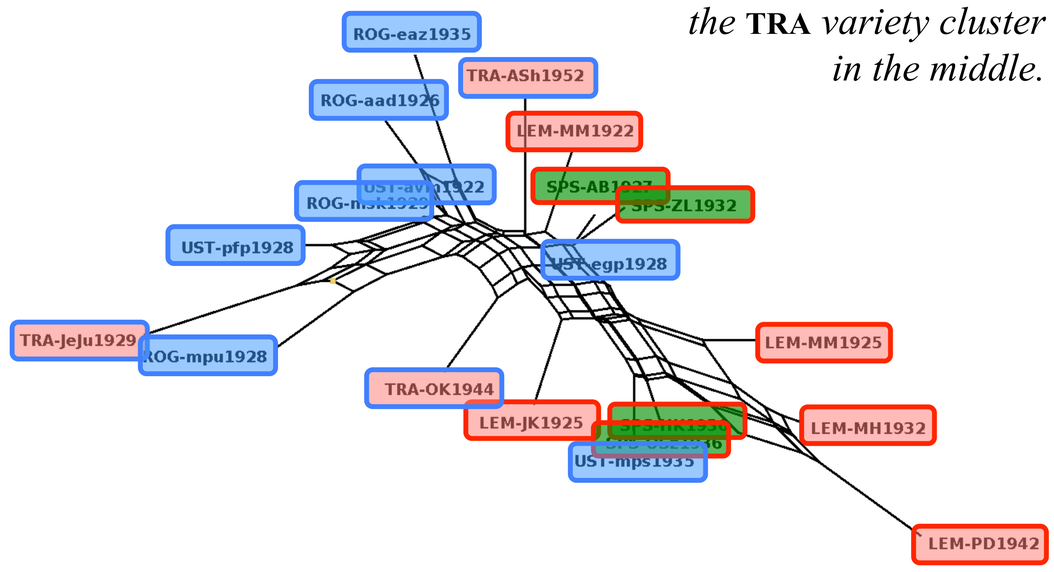
The symmetrized confusion matrix among the language models with cross-entropy used as a measure of mutual distance. Russian **ROG** *and* **UST** *form a tight group, as do* **SPS** *dialect and Rusyn* **LEM**, *both spoken in Poland. Rusyn* **TRA** *spoken in Ukraine shows greater similarity to* **ROG-UST** *group.*

A visualization of dialect/variety clustering based on the cross-entropy in the form of a phylogenetic tree (generated by SplitTree4 [2]).



Note the primary split along the majority language (East-West Slavic)

SplitTree clustering of models for individual speakers. Russian dialects occupy mostly the left hand side, while the varieties spoken in Poland are mostly in the right hand side. The models of speakers of the **TRA** *variety cluster in the middle.*



Discussion

- The presented methodology was able to "correctly" recognize the Russian dialects as more closely related mutually than with other Slavic languages (despite the vast geographical distance).
- Majority language (group) shows a strong influence on prosodic patterns of Rusyn language spoken in different countries.
- Next steps:
  1. Add more languages/varieties.
  2. Investigate the sources of the differences by finding the prosodic patterns that contribute most to the cross-entropy.

References

1. J. Šimko, A. Suni, K. Hiovain, and M. Vainio (2017) "Comparing languages using hierarchical prosodic analysis," in *Proceedings of Interspeech 2017*, Stockholm.
2. D. H. Huson and D. Bryant, Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.*, 23(2):254-267, 2006