Making *Sense* of What *Sounds* Similar: Modelling Mutual Intelligibility among Related Languages

Badr M. Abdullah

Department of Language Science and Technology (LST)
Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
babdullah@lsv.uni-saarland.de

Abstract

Closely-related languages, or language varieties, usually form a temporal and/or spatial continuum, whereby speakers of different varieties can usually communicate with each other efficiently using their own mother tongue. The degrees of similarity at different levels of the linguistic structural organization can be seen as preconditions, as well as predictors, of successful oral intercomprehension. For closely-related languages, similarities at the pre-lexical, that is the acoustic-phonetic and phonological, level have been found to be better predictors of cross-lingual speech intelligibility than lexical similarities.

In my PhD research, the goal is to take inspiration from the speech technology field and build computational models to quantify mutual intelligibility and oral intercomprehension between related languages.

Index Terms: mutual intelligibility, speech recognition, related languages

1. Introduction

One of the goals of human language is to communicate intentions from speaker to listener. When they wish to (verbally) communicate an intent to the listener, the speaker encodes their intent in a linguistic expression realized as a sequence of acoustic events. Upon observing the acoustic realization of the expression, the listener attempts to decode the communicative intent using their linguistic competence, knowledge of the communicative situation, and assumptions about the speaker's intentions. If the speaker's intent was expressed in linguistic codes the listener can decode and comprehend, the communication would be successful. On the other hand, if the listener cannot decode the observed expression due to structural differences between the speaker's language and listener's language, the communication would be less optimal and might fail.[1]

Human languages share structural details at different levels of linguistic organization. Phylogenetically related languages exhibit a high degree of similarity along several dimensions, including: (1) phonological (e.g., sound inventory and phonotactics), (2) lexical (e.g., the relation between word forms and meaning), (3) morpho-syntactic (e.g., word formation and word order). If two languages (A and B) are closely-related and share linguistic structures at different levels, there is a high chance that language A and language B are mutually intelligible. That is, the listener can use their linguistic competence of language B to successfully decode expressions of language A, and vice versa. This linguistic phenomenon can be observed between many pairs of closely-related languages. In many cases, the phenomenon of mutual intelligibility is not symmetric. For example, it has been observed that Portuguese speakers can understand Spanish much better than the other way around.

2. Research Objectives

Mutual intelligibility has been extensively investigated from a linguistic point of view in several studies (cf. [1] and [2]). Within this thesis, the aim is to address the problem from an information-theoretic point of view and validate/complement the assumptions/findings of the literature by means of computational modelling of the cross-lingual comprehension process. The Slavic language family is taken as a case study of mutual intelligibility among closely-related languages. I aim to investigate the acoustic-phonetic and phonological factors that contribute to mutual intelligibility and oral intercomprehension. Concretely, my objectives are three-fold:

(1) to (computationally) model the process of cross-lingual language comprehension by taking inspiration from recent advances in the field of automatic speech recognition (ASR) and natural language processing (NLP),

(2) to develop signal-driven metrics that quantify the degree of acoustic similarity between a group of languages,

(3) to gain further insights into the linguistic factors that contribute to oral intercomprehension.

3. Progress Report

3.1. Quantified Measure of Acoustic Similarity

I started my PhD endeavor by asking the question: how can we quantify the acoustic similarity between a group of closely-related languages or spoken language varieties? To answer this question, I surveyed the recent advances in speech technology and came to the conclusion that models of spoken language identification (LID)[3] are suitable candidates for the task of quantifying acoustic similarity. LID models take as an input an acoustic realization of a linguistic expression (usually a few seconds of a spoken utterance) and produce as an output a probability distribution over the candidate languages. I hypothesised that the confusion patterns of LID models would be a good starting points to analyze the acoustic similarity between languages. I implemented an LID model based on deep neural networks (DNNs), which are state-of-the-art techniques and have shown to be effective for many tasks in spoken language recognition.

3.2. First Results

The first challenge I encountered was that DNN-based LID models perform remarkably well when predicting the language of a 3-second spoken utterance, even when the candidate languages are closely-related as in the case of the Slavic language

[1] Although a simplistic view of communication with human languages, this level of abstraction is sufficient for this abstract.

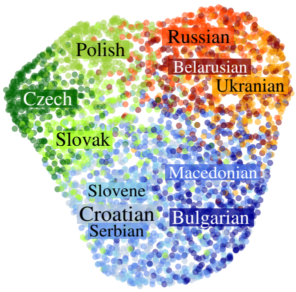

Figure 1: *UMAP projection of representations.*

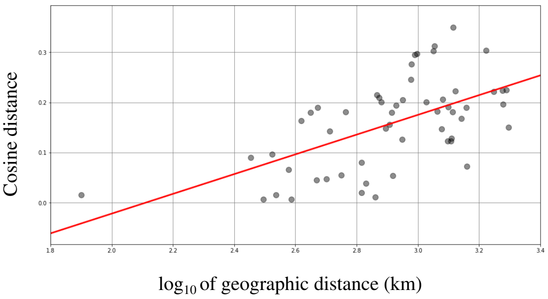

Figure 2: *Correlation of cosine vs. geographic distance.*

family. As a result, the observed confusion patterns were not as obvious nor interesting as I have initially anticipated. Moreover, it turned out that the remarkable performance of my LID models (and many models in the LID literature as a matter of fact) could be attributed to the striking ability of deep neural networks to pick up spurious correlations in the dataset they have been trained and evaluated on. I observed that LID models are very likely to exploit channel-related characteristics in the audio recordings when trained on a single dataset. To assess the generalization ability of LID models on the learning task, I conducted a cross-dataset evaluation of my LID models and observed a significant drop in accuracy when evaluating on a dataset with different acoustic/channel conditions. To improve LID generalization across different datasets, I explored unsupervised domain adaptation techniques from the field of applied machine learning. In one of my experiments, the domain-adaptive LID model has shown to improve the cross-dataset accuracy from 50.94% to 90.56% with relative accuracy improvement of 77.7% [4].

3.3. Analysis of Emerging Representations

The reasonable performance of the adaptive-LID models reported across datasets is a better estimate of model ability to extract the language-related features from audio recordings with little impact of speaker/channel characteristics on the learning task. LID models that are based on multi-layer DNNs learn to predict the identity of the spoken language by transforming a low-level spectrotemporal representation of the speech signal into a high-level feature vector where language classes are linearly separable. This transformation is realized as a series of non-linear transformations within the network layers where each layer re-represents the output of its preceding layer to optimize for the learning objective. I analyzed the representations of the final layer in the LID models to get insights into what kind of similarities are encoded in these high-level representations. In the first analysis, I used the UMAP [5] dimensionality reduction technique to visualize the emerging space in the model. The outcome of this visualization is shown in Figure 1. The UMAP algorithm attempts to preserve the global structure of the space thus the proximity in the depicted 2-dimensional space reflects the distance in the higher-dimensional space. It can be observed in Figure 1(a) that the sub-spaces emerged in the model correspond to the Slavic branches that are widely established by historical linguistics and Slavic studies (i.e., the 3-way division between East-, West-, and East-Slavic). In the second analysis, I obtained language prototype vectors by averaging the representations of the evaluation utterances and then correlated the distance between prototype vectors (measured by cosine distance) with the geographic proximity of the languages (measured by the $\log_{10}$ of geographic distance in km). The outcome of this analysis is shown in Figure 2. It can be observed that the distance in the LID representation space highly correlates with the geographic distance with a correlation coefficient of 0.58. These two analyses confirm my hypotheses that LID models capture acoustic similarity in their representations.

4. Future Work

4.1. Cross-lingual Spoken Word Recognition

Spoken word recognition (SWR) is a research topic at the intersection of psycholinguistics, speech perception, and cognitive modeling. SWR research aims to develop theories and (computational) models to build a better understanding of how humans access lexical knowledge upon observing an acoustic realization of a word form. There is a rich literature on this topic that has addressed word recognition from a monolingual point of view using connectionist models. I plan to explore computational modelling of SWR within the context of my PhD thesis in three dimensions: (1) extending the SWR connectionist models in the literature by incorporating relevant approaches from the field of representation learning to build up better abstractions of pre-lexical and lexical representations than what has been explored in the literature, (2) addressing the problem of SWR as lexical meaning retrieval instead of word form classification, and (3) investigating the cross-lingual aspects of SWR to simulate word-level intercomprehension.

4.2. Beyond Word Recognition

SWR research has addressed the problem of accessing the lexical meaning of word forms out of context. Nevertheless, the linguistic context in which the word has occurred in plays a very important role in comprehension. Therefore, I plan to extend the idea of SWR modelling to account for linguistic expressions beyond the word level. This goal can be realized by developing an “L1 listener” model that takes as input an auditory stimulus and induces a meaning representation as output (schematized in Figure (2)). Modelling the “L1 listener” poses three questions:

- **RQ1** What are the internal building blocks of the “L1 listener” model that maps from the auditory stimuli (input) to the meaning representation (output)?
- **RQ2** How to represent the auditory stimuli to the model?
- **RQ3** How to represent the meaning of a linguistic expression realized as an auditory stimulus?

5. References

[1] C. Gooskens, “The contribution of linguistic factors to the intelligibility of closely related languages,” *Journal of Multilingual and multicultural development*, vol. 28, no. 6, pp. 445–467, 2007.

[2] V. J. Van Heuven, “Making sense of strange sounds:(mutual) intelligibility of related language varieties. a review,” *International journal of humanities and arts computing*, vol. 2, no. 1-2, pp. 39–62, 2008.

[3] H. Li, B. Ma, and K. A. Lee, “Spoken language recognition: from fundamentals to practice,” *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[4] B. M. Abdullah, T. Avgustinova, B. Möbius, and D. Klakow, “Cross-domain adaptation of spoken language identification for related languages: The curious case of slavic languages,” 2020.

[5] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv preprint arXiv:1802.03426*, 2018.