Graz University of Technology

Signal Processing and Speech Communication Laboratory

Univ.-Prof. DI Dr. Gernot Kubin
Head of Laboratory

Inffeldgasse 16c
8010 Graz, Austria

Tel: +43 316 873 4430
Fax: +43 316 873 104430

gernot.kubin@tugraz.at
www.spsc.tugraz.at

DVR: 008 1833 UID: ATU 574 77 929

To the Dean of Studies
Faculty of Electrical and Information Engineering
Univ.-Prof. Mag. Dr. Alexander Bergmann
Graz University of Technology
Inffeldgasse 18
8010 Graz, Austria

via Email

Graz, Feb 19, 2025

Re: Evaluation of the Doctoral Thesis "What's so complex about conversational speech? Prosodic prominence and speech recognition challenges" by Dipl.-Ing. Julian Linke, BSc

Dipl.-Ing. Julian Linke presents in his dissertation *What's so complex about conversational speech? Prosodic prominence and speech recognition challenges* two fundamentally new insights into conversational speech: first, how can we analyze and use prosodic concepts like "prominence" in conversational speech and, second, what we can learn from the difficulties of state-of-the-art technologies for automatic speech recognition with conversational speech data.

A large part of the thesis is performed on a unique database of conversational speech (the GRASS corpus) that exhibits a set of properties and conversational freedom which makes it a one-of-a-kind resource: the conversations are *unscripted, open topic, casual, face-to-face, and dyadic with a close relationship of the two conversation partners*. The only similar database found by the doctoral candidate is a corpus from Japan and consists of everyday speech usage recorded with a portable device, with lower audio quality and less annotations than for the GRASS corpus. Note that, in general, casual conversations do not use a language variety close to an official standard; thus the GRASS corpus reflects the German variety spoken in Austria which should not distract from the fact that the main objective was to capture casual conversations as close as possible to real life and for a German language variety. This database with its rich annotations served as the *strong basis for the experiments* designed, performed, and evaluated in this thesis.

Secondly, the extensive experiments required a set of state-of-the-art methods and software implementations for model building based on machine learning methods and for analyzing the statistical influence of certain variables on the model outcomes. These tools span *several generations of technology*, e.g., from Hidden-Markov-Model based automatic speech recognition like Kaldi to transformer-based architectures like Wav2vec and Whisper. While the doctoral candidate had access to most of these highly complex tools as open resources on the internet, he was the first to implement them on the SPSC lab's compute infrastructure, to get all of them up and running, and to validate their performance after extensive tuning. Only such *tuned and validated models* are close enough to the state-of-the-art to produce scientifically useful evidence from and confidence in experiments that often included sophisticated fine-tuning with small datasets. This impressive contribution in terms of engineering skills was key to setting the scene for the work on the six research questions of the thesis and will have a lasting impact on the future research at the SPSC lab.

After the introductory chapter (showcasing the research objectives and main contributions of the thesis), chapter 2 actually presents the GRASS corpus and two other corpora of German varieties which contain conversational speech for comparison purposes, however with much less freedom and less annotations in the recorded conversations. *Initial experiments* with automatic speech recognition using the 'oldest' technology (Kaldi) help to set the scene.

The following chapter addresses *prosodic prominence* and its automatic classification in read and conversational speech, building on the unique GRASS feature that it contains both types of speech material from the same speakers. In the quest for *interpretable classification results*, the focus is on decision trees and random forests. A large set of standard features is extended by original features based on entropy measures. One of the results is that durational features have the highest explanatory power for the detected prominence level (which is somehow controversial in the literature) and, surprisingly, the *new entropy-based features* capture this aspect without explicit segmentation.

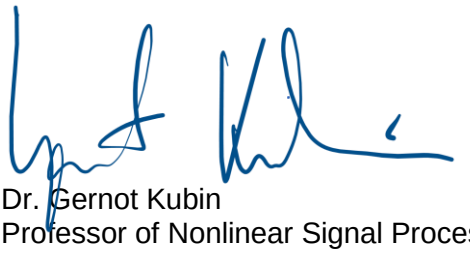
Chapter 4 is the heart of the thesis where the first two sections have been published in one of the leading journals of the field (*Computer Speech and Language*), the others in leading conferences. The main question is what we can learn from experiments with state-of-the art automatic speech recognition technologies. The main thrust is not to use the learnings to advance these technologies themselves but, as all these systems have been trained on huge amounts of speech data (up to 680,000 hours or more than 77 years of 24/7 non-stop talking), they implicitly represent a lot of knowledge about speech that can be extracted from the recognition results in carefully crafted experiments. So instead of searching (manually or automatically) through a database of 680,00 hours of speech, *speech science can produce answers* to its research questions by analyzing the behavior of the huge models available for these data sets. This may be likened to the study of physical phenomena in digital twins built by numerical simulation or similar techniques. In a further enrichment of this methodological approach, the comparison of speech recognition systems from different technology generations helps to complement the *insights for the speech scientist with insights for the speech technologist.* This program is carried through in all sections of this chapter, using again the best available statistical analysis tools to isolate the most important influence variables. One striking result is that the variability patterns of speech recognition errors found across different pairs of conversation partners persist across all studied generations of technology, corroborating that, while new technologies certainly have led to a dramatic overall reduction in recognition errors, the speakers' idiosyncracies still constitute *impenetrable barriers for machines*. Further experiments show that the latent representations of state-of-the-art models capture information, e.g., about speaking style (independent of language variety) or prominence. Such analyses can serve further downstream processing even if it can hardly assist the automatic speech recognition output itself which is computed with the same models.

The last chapter wraps up with a good general discussion of the knowledge gained through the experiments and how it can be connected with a broader perspective to the fields of speech science and technology. After a rather short outlook to future work it concludes with the *four most important take-home messages* of the thesis, condensed in a nice concise way.

A short appendix provides additional material on the new entropy-based features and a very comprehensive bibliography rounds off the thesis.

In conclusion, Dipl.-Ing. Julian Linke has built up a large toolbox of methods related to automatic speech recognition, has designed a substantial number of experiments applying these tools on highly specialized conversational speech datasets, and has elaborated the answers to six research questions with their implications for speech science and speech technology. The thesis is based on *6 high-quality international publications with peer review* and he has authored or co-authored *10 more such publications* on a variety of topics showing his research interest and expertise beyond the core of the doctoral dissertation. The language use is clear throughout the thesis and all arguments are carefully supported by evidence and discussed in

an adequate manner. Some of the new tools have already found interest from other research groups in Austria and beyond. Based on all considerations, I recommend to accept this dissertation and propose an overall grade of “sehr gut (1)”.



Dr. Gernot Kubin
Professor of Nonlinear Signal Processing