Julian Linke Thesis

Evaluation report

Louis ten Bosch

Feb 20, 2025

General

This thesis concerns the automatic recognition of read speech and conversational speech, with a focus on the performance of an array of HMM-based and end-2-end ASR systems in different configurations. The thesis text – based on published papers – is written in a transparent way and rich in detail. In fact, this thesis presents the first transformer-based prominence aware ASR system.

It is known that conversational speech poses unique challenges compared to read or prepared speech – primarily due to e.g. turn-taking, ungrammatical/incomplete utterances, disfluencies, restarts, informality in register, and a high degree of pronunciation variation. Prosodic annotation tools and ASR systems perform significantly worse on conversational speech, compared to prepared speech. This thesis aims 1) to analyze acoustic representations for conversational speech using explainable AI methods, and 2) to improve the performance of prosodic prominence classification and ASR systems, as measured with standard performance measures.

The interesting and challenging aspect in this work is the combination of ASR systems and prosodic information as encoded through fine-tuning of self-supervised speech representations. The results clearly show a strong interaction with ASR WER measures and prosodic features of various kinds.

The ASR tasks as described were challenging in the sense that data came from (under-resourced) conversational Austrian German, so with high variation from a (low-resourced) regional variety of a (well-resourced) language in addition to the high inter-speaker and inter-dyad variation.

Assessment (“very good (1)”, “good (2)”, “satisfactory (3)”, “acceptable (4)”, “fail (5)”)

topic	**comment**	**grade**
Research questions	Pro: the RQs are well formulated and very clear. Interestingly they relate local phenomena (e.g. spectral details, local duration information) with higher level phenomena (e.g. pitch, sentence duration). Contra: a few RQs are really big. Utterance duration might be confounded with language model and so indirectly with content of context.	1-2
Embedding of the work in the literature	Pro: The literature is very well covered and integrated in a senseful way. The number of references to papers from 2024 is much lower compared to references to 2021-2-3 but this is of course logical given the time frame for aa PhD. Contra: none	1-2
Originality	Pro: The thesis is original in combining old (and newer) questions with modern techniques. Contra: what could be strengthened is the future view of how to proceed. Now, the text about future prospects is somewhat flat (section 5.2):	2

	“First, **future work should explore all possible combinations of system configurations.** For example, researchers could investigate how wav2vec2 fine-tuned on a corpus of canonical German read speech performs in the conversational speech setting. **Furthermore, given the peculiar performance of Whisper, an analysis of Whisper fine-tuned on GRASS is recommended.** This would provide insights into the performance of a transformer-based ASR system without explicit linguistic knowledge that has been fine-tuned on data from the target language and style. Second**, future experiments should extend beyond one corpus to enhance the generalizability of the findings**” But we know it is quite hard to come up with good recommendations – especially given the very fast progress at the computational (model) side.	
Coherence	The thesis is a coherent piece of work, nicely covered by two overarching big themes. The chapters are clearly closely related.	1
Presentation and language use	The thesis is a nice read and presents the topics in a thorough manner.	1
Overall		1

Based on all considerations, I recommend to accept this dissertation and propose an overall grade of “sehr gut (1)”.

Dr. Louis ten Bosch

CLS, Radboud University, Nijmegen, NL

My summary:

The research has two primary aims: (1) analysis of acoustic representations and models for conversational speech with explainable machine learning methods, and (2) evaluation of acoustic representations and models for conversational speech with standard performance measurements

For the **first aim** of this thesis, several research questions were addressed related to the analysis of acoustic representations and models for conversational speech with explainable machine learning methods.

[a] analyzing the main acoustic cues of prosodic prominence for conversational speech (RQ1)

b] whether WERs of conversational speech are affected by utterance-level features (RQ2)

[c] what shared discrete speech representations encode with respect to language varieties, speaking styles, and speakers (RQ3)

[d] whether the fine-tuning of self-supervised speech representations implicitly encodes prosody (RQ4)

The first research objective involves analyzing acoustic cues of prosodic prominence. Using explainable machine learning models, the study confirms that durational features play a critical role in prominence classification. Novel entropy-based features, which encode F0 and RMS variations, provide an alternative to complex durational calculations while maintaining classification performance.

The second research focus examines the effects of utterance-level features on ASR WERs. The study compares ASR architectures (Kaldi, wav2vec2, and Whisper) and finds that utterance length and articulation rate significantly impact performance. Whisper performs worse on short utterances but improves with longer ones, while other ASR models show different trends. Systems using linguistic information, such as a lexicon or language model, are more robust against high articulation rates.

A key finding is the role of F0 and RMS entropy-based features in ASR performance. Low pseudo-entropy values (indicating stable F0 contours) correlate with lower WERs, particularly for Whisper on short utterances. Additionally, the interaction between F0 entropy and articulation rate suggests that flatter F0 contours and slower speech improve ASR accuracy.

This study also investigates pronunciation variation, particularly in Austrian German. Higher pronunciation deviations from standard Austrian German lead to increased WERs, especially for Whisper, which lacks in-domain fine-tuning. ASR systems with linguistic resources (e.g., Kaldi, wav2vec2 with lexicon) show better robustness against pronunciation variation.

Lastly, the thesis examines the effect of utterance-level perplexity on ASR performance. While lower perplexity correlates with better WERs, this effect is weaker than other factors such as utterance length, articulation rate, F0 entropy, and pronunciation variation. The study recognizes the limitations of simple n-gram language models for capturing conversational speech nuances.

The **second aim** of this thesis addressed several research questions related to the evaluation of acoustic representations and models for conversational speech using standard performance measurements. This study began by investigating whether word-level prominence classification results with prosodic features or word level prominence detection results with fine-tuned speech representations align with inner-annotator agreements (RQ5). Subsequently, it explored how low-resourced HMM-based ASR systems compare to low-resourced or data driven transformer-based ASR systems in terms of their effectiveness for recognizing Austrian German conversational speech (RQ6)

The study investigated how prominence detection and classification can be automated and how ASR systems perform under conversational speech conditions. This study first focuses on developing prominence classification and detection tools for conversational speech. Using prosodic features and random forest models, the classification tool aligns closely with human inter-rater agreements. For three prominence levels, the model achieved a cross-validation accuracy of 63% ± 7%, while for a binary classification (prominent vs. non-prominent), it reached 88% ± 5%. These results suggest that prominence classification can be effectively automated and that well-designed feature sets can eliminate the need for complex durational calculations.

A more advanced prominence detection tool based on **wav2vec2** is also explored. Unlike traditional classification models, this tool operates directly on raw audio data, achieving results comparable to human annotations. The three-level prominence detector had an error rate of **36.54% ± 0.92%**, while the two-level detector performed better with an error rate of **24.83% ± 1.79%** and accuracy of **89.72% ± 3.26%**. Unlike previous methods that rely on forced alignments requiring transcriptions, this approach aligns speech with prominence levels directly. These results indicate that modern self-supervised speech representations can effectively model prosodic prominence without the need for manual text-based alignment.

Next, the study compares different ASR **architectures** to determine their effectiveness in transcribing Austrian German conversational speech. The study evaluates Kaldi, wav2vec2, and Whisper, considering various training strategies and data availability. One key question was how much training data is required for ASR to work effectively with conversational speech. Different training approaches yielded varying **word error rates (WERs)**:

- Kaldi trained on GRASS SC with **cross-entropy loss**: **51.87% ± 4.83%**
- Kaldi with **LF-MMI criterion**: **42.86% ± 4.78%**
- wav2vec2 trained only on Austrian German speech: **57.28% – 62.54%**
- wav2vec2 fine-tuned on multilingual data: **25.06% – 22.79%**
- Whisper (zero-shot, no fine-tuning): **41.78% ± 8.23%**

The results suggest that **fine-tuning on in-domain data significantly improves ASR performance**, reducing WERs by approximately **20%**. However, large standard deviations indicate that conversational speech remains difficult to model consistently.

The robustness of ASR systems is another critical factor. While all systems performed well on **read speech** (WERs between **1.01% – 11.8%**), they struggled with **conversational speech**:

- Whisper: **41.78%**
- Kaldi: **42.86%**
- wav2vec2: **29.81% (without lexicon), 22.79% (with lexicon)**

These findings indicate that **no current ASR system is fully robust for Austrian German conversational speech**, as all models showed significant performance variability across different conversations.

Modern ASR systems, even with extensive pretraining, still struggle with conversational speech variability. While fine-tuning pre-trained models on domain-specific data improves performance, large standard deviations in WERs highlight the complexity of conversational speech recognition. The results reinforce the assumption that Austrian German remains a low-resourced variety, necessitating further advancements in ASR adaptation and robustness.