Conversational Speech Recognition Needs Data: Experiments with Austrian German

Julian Linke, Philip N. Garner, Gernot Kubin, Barbara Schuppler

Signal Processing and Speech Communication Laboratory / Idiap Research Institute

Motivation

- Low-resourced (LR) conversational speech recognition is challenging
- More data means better performance?

Materials

- GRASS: The *Graz Corpus of Read and Spontaneous Speech* contains about 19h of Austrian conversational speech collected from 38 Austrian speakers

ASR

- *Traditional* approach with Kaldi: TDNN, 4-gram and pronunciation lexicon
- *Self-supervised* approach with wav2vec: LR and XLSR

Hypotheses

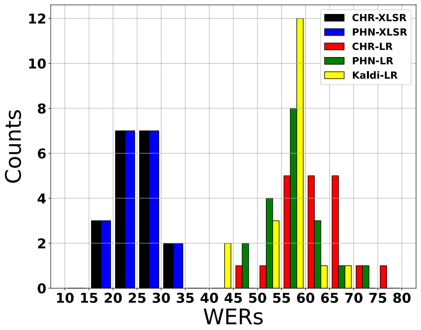
1. Low-resourced Kaldi not effective
2. Data-driven wav2vec effective
3. Low-resourced wav2vec not effective

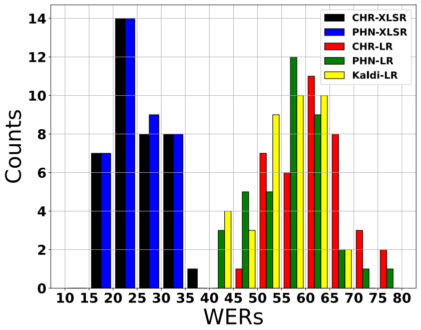
Performances

- Hypotheses demonstrate importance of data-driven approach
- Additional insights from conversation-dependent decodings

	Phone-based				Character-based		
Kaldi-LR	**Lexfree**	**Lex**	**4-gram**				
009M010M	-	-	65.12				
021F022F	-	-	43.89				
μ/σ	-	-	56.19/5.4				
PHN-XLSR	**Lexfree**	**Lex**	**4-gram**	**CHR-XLSR**	**Lexfree**	**Lex**	**4-gram**
006M007M	-	42.03	32.71	006M007M	41.5	38.95	34.49
038F039F	-	26.63	17.44	038F039F	22.37	19.88	17.36
μ/σ	-	33.15/4.32	24.69/4.10	μ/σ	31.23/4.86	28.06/4.92	25.06/4.42
PHN-LR	**Lexfree**	**Lex**	**4-gram**	**CHR-LR**	**Lexfree**	**Lex**	**4-gram**
016M018M	-	90.44	73.45	016M018M	95.32	98.11	76.98
021F022F	-	64.93	45.14	038F039F	75.61	72.32	48.52
μ/σ	-	75.14/5.86	57.28/6.46	μ/σ	85.5/4.63	84.75/6.36	62.54/6.36

Distributions of WERs





Conclusions

- Effectivness of data-driven wav2vec
- Importance of linguistic knowledge
- Complexity results in lack of robustness