Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4684–4691
Marseille, 20-25 June 2022

Conversational Speech Recognition Needs Data? Experiments with Austrian German

Julian Linke[1], **Philip N. Garner**[2], **Gernot Kubin**[1], **Barbara Schuppler**[1]
[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology
[2]Idiap Research Institute
[1]Inffeldgasse 16c, A-8010 Graz, Austria
[2]Rue Marconi 19, 1920 Martigny, Switzerland
[1]{linke, kubin, b.schuppler}@tugraz.at, [2]phil.garner@idiap.ch

Abstract

Conversational speech represents one of the most complex of automatic speech recognition (ASR) tasks owing to the high inter-speaker variation in both pronunciation and conversational dynamics. Such complexity is particularly sensitive to low-resourced (LR) scenarios. Recent developments in self-supervision have allowed such scenarios to take advantage of large amounts of otherwise unrelated data. In this study, we characterise an (LR) Austrian German conversational task. We begin with a non-pre-trained baseline and show that fine-tuning of a model pre-trained using self-supervision leads to improvements consistent with those in the literature; this extends to cases where a lexicon and language model are included. We also show that the advantage of pre-training indeed arises from the larger database rather than the self-supervision. Further, by use of a leave-one-conversation out technique, we demonstrate that robustness problems remain with respect to inter-speaker and inter-conversation variation. This serves to guide where future research might best be focused in light of the current state-of-the-art.

Keywords: Speech Recognition, Conversational Speech, Austrian German, Low-Resource, Wav2vec2.0, Kaldi

1. Introduction

Solving automatic speech recognition (ASR) tasks for conversational speech is crucial especially for social robots interacting with humans or automatic transcriptions of multimedia meetings (Popescu-Belis et al., 2012). Two humans who interact spontaneously with each other introduce complex inter- and intra-speaker variation depending on for instance the speaker's attitude towards the listener and the speaking task (Wright, 2006). Especially casual face-to-face conversations are characterized by a large amount of speaker-dependent pronunciation variation, by disfluencies, and by broken words or incomplete utterance structures. The resulting high degree of variation on all linguistic levels affects the acoustic model, the lexicon and language model of an ASR system.

Given the high variation in spontaneous conversations, the amount of annotated training data needed for ASR experiments to enable generalization for an unseen test set can sometimes be misleading in the sense that avoiding the data sparseness problem appears not to be possible, especially in case of spontaneous speech (Furui et al., 2005; Furui, 2009). Such studies give insights into the relationship between data size for acoustic model training and WER in case of Japanese spontaneous speech recognition: Utilizing $1/8$ of available data ($63.75\mathrm{h}$) for acoustic model training results in a WER of approx. 27% whereby training with the entire data ($510\mathrm{h}$) gives an improvement of approx. 2%, but still no convergence.

In this paper, we deal with conversational speech from the "Graz corpus of Read And Spontaneous Speech" (GRASS) (Schuppler et al., 2014a), which contains about $19\mathrm{h}$ (or 19 conversations) of Austrian German conversations introducing a considerable complexity in light of both inter-speaker and inter-conversation variation (i.e., from conversation to conversation, the amount of laughter, overlapping speech and disfluencies varies (Schuppler et al., 2017)). Despite German being a well resourced language, for the Austrian variety there are few resources available. For conversational speech, GRASS is the only resource currently available. For less spontaneous and less casual speaking styles, using German German[1] data for training an ASR system still delivers reasonably good results for recognizing Austrian German (Adda-Decker et al., 2013), this is, however, not the case for casual conversations where speakers show a higher degree of dialectal pronunciations. Hence, with respect to this variation, ASR experiments may require larger amounts of annotated conversational speech data than for less spontaneous speaking styles and thus may be viewed as a case of low-resourced (LR) language processing.

With wav2vec2.0 (Baevski et al., 2020b), a framework for self-supervised learning of speech representations, powerful ASR models can be built also with small amounts of annotated data by finetuning pre-trained models. With the help of this modern architecture it is even possible to come close to state-of-the-art results with only $10\mathrm{min}$ of labeled training data in the case of Librispeech (Baevski et al., 2020a; Conneau et al., 2021; Hsu et al., 2021; Zhang et al., 2021; Panayotov et al., 2015). Hence, we hypothesize this innovative framework also to be effective in solving a LR speech

[1] With German German we refer to German as spoken by German speakers.

recognition task for Austrian conversational speech.

This study presents ASR experiments for Austrian German conversational speech from two ASR frameworks, the Kaldi speech recognition toolkit (Povey et al., 2011) and the wav2vec2.0 implementation of fairseq (Ott et al., 2019). In case of wav2vec2.0, we finetune a cross-lingual speech representation (XLSR) pretrained model (Conneau et al., 2021) with different training data splits by testing each of the 19 GRASS conversations individually. Referring back to the problem of conversational speech complexity, we compare the XLSR experiments with an LR approach by pretraining and finetuning only with available GRASS conversational speech data. Ultimately, this paper aims at investigating three hypotheses to gain more insight about the role of data for conversational speech:

1. Performing cross-validation by testing each conversation individually points out conversational speech complexity and reinforces a LR language processing assumption.
2. Finetuning a data-driven pre-trained cross-lingual speech representation model is effective for Austrian conversational speech.
3. Finetuning a LR speech representation model pretrained only on Austrian conversational speech is not effective for Austrian conversational speech.

These hypotheses are investigated by the experiments presented in section 4. After answering our hypotheses, the corollary section 5 discusses further findings which result from comparing the results from our ASR experiments.

2. Related Work

Training acoustic models for conversational speech typically needs large amounts of training data, because generalization to pronunciation variation introduced by this speaking style is difficult.

For comparison with a *traditional* approach to train acoustic models for speech recognition, we use the Kaldi speech recognition toolkit. Kaldi is a speech recognition system based on finite-state transducers, where the core library supports acoustic modeling with standard Gaussian mixture models (GMM). Resulting alignments from basic GMM-HMM models can be used for further training with, e.g., time-delay neural networks (TDNN) (Peddinti et al., 2015).

In case of LR scenarios, where the amount of annotated data is limited, unsupervised and representation learning techniques have gained a lot of attention recently (Glass, 2012; Chung and Glass, 2018; Schneider et al., 2019; Synnaeve et al., 2019). *Self-supervised* learning from unlabeled speech data in end-to-end ASR systems creates general and powerful speech representations which can be used as a basis for finetuning on small amounts of labeled data yielding both the possibility of high capacity model training and the prevention of over-fitting (Baevski et al., 2020a; Baevski et al., 2020b). It turned out that exploiting a XLSR model is an effective strategy in building state-of-the-art speech recognition systems given only few labeled speech data, because shared discrete acoustic units across languages can be adapted to a specific task by finetuning a classifier representing self-chosen target units plus word boundary tokens. We can solve this task, e.g., with a Connectionist Temporal Classification (CTC) loss when training a speech recognition system (Conneau et al., 2021; Graves et al., 2006).

One system to exploit an XLSR model is wav2vec2.0 (Baevski et al., 2020b), a framework for self-supervised learning of contextual representations obtained from the raw waveform of speech. The architecture consists of a transformer network (Vaswani et al., 2017) which is fitted by encoded speech audio coming from a multi-layer convolutional neural network. Primarily, the transformer network tries to learn contextualized representations by solving a contrastive task.

3. Materials

3.1. GRASS

The Graz Corpus of Read and Spontaneous Speech (GRASS corpus) (Schuppler et al., 2014a; Schuppler et al., 2017) contains about 19h of Austrian conversational speech collected from 38 Austrian speakers (19f/19m). As language use in conversational speech varies strongly with educational level, social background and dialect region, GRASS contains only speakers who were born in the same broad dialect region (Eastern Austria), have been living in an urban area for years and have a higher education degree. For the conversational speech component, 19 pairs of speakers who had been knowing each other for several years were recorded for one hour each without interruption in order to encourage a fluent, spontaneous conversation. There was no restriction in terms of chosen topic or speaking behavior leading to the use of authentic, partly dialectal pronunciation with typical characteristics such as frequently occurring overlapping speech, laughter, or the use of swear words (Schuppler et al., 2017). No other person was present in the recording studio during the conversation. Despite the speakers' awareness of being recorded, they appeared to completely forget about the studio recording situation after a period of five to ten minutes, resulting in completely casual conversations.

3.2. Lexicon

All words from the GRASS corpus remaining after preprocessing are included in a lexicon file. For all phone-based experiments, we used the G2P online tool (Reichel and Kisler, 2014) for German German to create canonical German pronunciations, as a similar resource is not available for Austrian German.

Only for the Kaldi experiments, we derived additional pronunciation variants from the canonical pronunciations with 29 phonological rules based on findings from

(Schuppler et al., 2014b). Some of the rules were assimilation and deletion rules relevant for conversational speech of all German varieties, whereas other rules cover pronunciations typical for the Austrian German variety. We added manually created pronunciation variants in order to capture specific pronunciations that cannot be generated in an automated way.

For wav2vec2.0 models, we create simplified lexicons where each word maps only to one pronunciation. In case of the character-based models, words are directly mapped to character sequences and in case of the phone-based models, words are directly mapped to canonical pronunciations.

4. Experiments

To investigate our three hypotheses, we first present experiments with Kaldi (section 4.1) and then experiments with fairseq (section 4.2).

4.1. Experiments with Kaldi

This section describes our experiments with Kaldi, which serve as a baseline for the main investigation.

4.1.1. Methods

When preprocessing GRASS transcriptions files for Kaldi, we excluded chunks involving artefacts, laughter and noise, resulting in a deletion of approx. 3.3h of all available chunks ($\approx 17.5\text{h}$ of all chunks from GRASS contain lexical items). In the end approx. 14h of the data were used in the experiments.

We performed leave-p-out cross-validation (with $p = 2$ speakers of the same conversation) resulting in approx. 0.75h of test data and 13.5h of training data per split. Hence, we trained 19 baseline models (Kaldi-LR) where each training split involves 18 conversations.

We reduced the initial phone set from 65 to 38 target phones by performing phone set minimization rules based on phonetic studies on Austrian German (Moosmüller, 2007): First, a replacement rule (silibant devoicing of the alveolar fricative $/z/$, as usual in Austrian German); second, a rule which split all diphthongs into two separate phones; third, a rule which united short vowels and long vowels.

We extracted 13-dimensional MFCCs and performed cepstral mean and variance normalization (CMVN). For acoustic models (AM), initial GMM-HMM-models comprised basic monophone and triphone training. On top of the triphone GMM, a speaker independent GMM model with linear discriminative analysis (LDA) and maximum likelihood linear transform (MLLT) (Gopinath, 1998) was trained. This model was the new basis for training with constrained maximum likelihood linear regression (fMLLR) (Gales, 1998). Finally, latter triphone alignments were used to train a TDNN with 13 layers and hidden dimensions of 512 utilizing only existing MFCC features.

For Kaldi experiments, the language models (LM) were built using the SRILM toolkit with a Witten-Bell discounting for N-grams of different orders (Stolcke, 2002). LMs were trained on data coming from one training split. The experiments with 3-grams and 4-grams indicated a 4-gram model to be superior.

4.1.2. Results

With this Kaldi experiment, we aimed at investigating the hypothesis that testing each conversation individually points out conversational speech complexity and reinforces a LR language processing assumption. Table 1 shows the WERs achieved with our baseline Kaldi-LR system. They range between 43.89% and 65.12%, where the resulting mean WER lies at approx. 56%, with a standard deviation of 5.4%. Hence, we observe a lack of performance and also high variation between the conversations with respect to the WERs.

The problem with conversational speech in LR scenarios is well-known: Results from (Laurent et al., 2016) give WERs of $\approx 40\%$ in case of conversational-like data. (Sriranjani et al., 2015), for instance, showed that based on very limited LR Indian language data ($\leq 3\text{h}$) recorded in a rural environment WERs ranged from $\approx 10\%$ to $\approx 34.5\%$. Furthermore, WERs from baseline experiments described in (Yi et al., 2020) range from 33.77%...51.54% in case of LR multilingual telephone conversation data.

We find that performing cross-validation by testing each conversation individually points out conversational speech complexity and indicates a data sparsity problem. At this stage we conclude that our first hypothesis cannot be rejected.

4.2. Experiments with Fairseq

This section describes our experiments with fairseq in order to further investigate our hypotheses.

4.2.1. Methods

In comparison to our Kaldi experiments, the preprocessing of GRASS transcriptions files was slightly different: We additionally had to exclude chunks involving foreign words and dialect lexemes, resulting in a total deletion of approx. 4h of all available chunks (i.e., approx. 0.7h more than for the Kaldi experiments). Other chunks which can involve breathings, speaker noise, singing, smacking, laughed speech, coughing, sighing, broken words or multi-word expressions were maintained. When parsing the transcriptions, we automatically corrected inconsistent orthography of fillers (e.g., `hm` and `hmm`), as these tokens can cause a high number of substitution errors. In the end approx. 13.5h of the data remained for our experiments.

Just as in the Kaldi experiments, we perform cross-validation resulting in 19 training splits where each split results from leaving out 1 conversation. Subsequently, we receive approx. 0.75h of test data and 12.75h of training data per split. Finally, we randomly choose 10% of resulting training splits as validation sets (approx. 1.25h) to adjust the LM weights in the decoder.

Table 1: Summary of best and worst conversation-dependent WERs (test set abbreviations include speaker IDs plus sex): Character-based (CHR) and phone-based (PHN) models with wav2vec2.0 are finetuned on LR pre-trained models (only GRASS) or XLSR. Kaldi-LR models are also phone-based and incorporate additional pronunciations in the lexicon. We present results coming from 3 decoding strategies: Decoding without a lexicon (**Lexfree**), decoding with a lexicon (**Lex**) and decoding with a lexicon and LM (**4-gram**).

	Phone-based						
Kaldi-LR	**Lexfree**	**Lex**	**4-gram**				
009M010M	-	-	65.12				
021F022F	-	-	43.89				
μ/σ	-	-	56.19/5.4			**Character-based**	
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

Training a speech recognition system with wav2vec2.0 involves two steps: 1) self-supervised learning from unlabeled speech data (pre-training) and 2) finetuning an obtained pre-trained model with labeled speech data. For all speech representation models, we used the same architecture with 315M parameters containing 24 transformer blocks with model dimensions 1024, inner dimension 4096 and 16 attention heads.

When finetuning wav2vec2.0 models, we compared two basic target sets: 1) a character-based (CHR) set resulting in 31 characters as targets and 2) a phone-based (PHN) set resulting in 65 phonetic units as targets. Both target sets included a white space unit which models silence parts. Similar to our lexicon creation, in case of the character-based models, the orthographic transcriptions given by GRASS were directly mapped to character sequences. In case of the phone-based models, the orthographic transcriptions were mapped to canonical phonetic sequences.

The available pre-trained XLSR model was trained with 56000h of multilingual speech data built on top of wav2vec2.0. The training data of XLSR contains CommonVoice (36 languages, 3600h) (Ardila et al., 2020), BABEL (17 languages, 1700h) (Gales et al., 2014) and MLS (8 languages, 50000h) (Pratap et al., 2020). We finetuned XLSR with our labeled speech data with a CTC loss (Graves et al., 2006) after introducing a classification layer representing our targets. Here, we present results coming from 19 phone-based models (PHN-XLSR) and 19 character-based models (CHR-XLSR), as there are 19 conversations in GRASS.

For our experiments with LR wav2vec2.0 models, we pre-trained merely with in-domain GRASS data followed by finetuning the pre-trained GRASS models given the labels from our training splits. Also these models were trained with a Connectionist Temporal Classification (CTC) loss after introducing a classification layer representing the two target types. Thus, we view the resulting models as LR approaches, because we used exactly the same training data (11.5h) for both, pre-training and finetuning. Also for this LR experiment with fairseq, we compare WERs coming from 19 phone-based (PHN-LR) and 19 character-based (CHR-LR) models.

For both XLSR and LR experiments with fairseq, we used a greedy decoder (**Lexfree**) in case of CHR models and a beam-search decoder without language model weighting (**Lex**) or with language model weighting (**4-gram**) in case of CHR and PHN models. The greedy decoder searches the greedy best path by using only acoustic model predictions. The AM search space of the beam-search decoder is restricted by a lexicon and we incorporated an LM by providing an LM weight. Here, when incorporating an LM, we trained LMs of order 4 with modified Kneser-Ney smoothing and default pruning, which removes singletons of order 3 or higher by utilizing the KenLM toolkit (Heafield, 2011). LMs were trained merely on data coming from one training split and we choose an LM weight from the set of LM weights $\{1, 2, 3\}$ with respect to best WERs coming from the additional validation data. In case of beam-search decoding, we chose a beam size of 100. For the phone-based models, we do not provide results of the greedy decoder, because reasonable results could only be produced with the help of a lexicon introducing a target set which allows for word disambiguations.

4.2.2. Results from XLSR Pre-Training and GRASS Fine-Tuning

This experiment investigates the hypothesis that finetuning a data-driven pre-trained cross-lingual speech representation model is effective for Austrian conversational speech.

The middle row of table 1 shows the WERs that achieved with the XLSR models. When decoding CHR-XLSR without a lexicon, WERs ranged between 22.37% and 41.5%, resulting in a mean value of 31.23% and standard deviation of approx. 5%. In case of PHN-XLSR, the **Lex** WERs are more similar to CHR-XLSR **Lexfree** results. **Lex** results of CHR-XLSR, on the other hand, are approx. 5% better with respect to mean value. We note that no big differences between **4-gram** PHN-XLSR and CHR-XLSR models can be observed, i.e., mean values and standard deviations are very similar. We observe that the powerful XLSR models give satisfactory results considering the high difficulty level of given face-to-face conversational data. As a matter of fact, all WERs of XLSR models are much lower than those from the Kaldi experiment, regardless of LM incorporation, and in case of CHR-XLSR even without utilizing a lexicon.

wav2vec2.0 models pre-trained on 50000h of English data were tested on various languages showing their effectiveness in LR scenarios (Yi et al., 2020): Results with German telephone speech (approx. 13h of training data) demonstrated an absolute improvement of approx. 20% compared to a baseline system. In general, they achieve more than 20% relative improvements in case of all six tested LR languages. Overall, their WERs are in the same range as ours.

We conclude that solving ASR tasks for GRASS conversational speech by finetuning speech representation models pre-trained on a high amount of out-of-domain data is effective. Thus, our second hypothesis cannot be rejected.

4.2.3. Results from GRASS Pre-Training and GRASS Fine-Tuning

This experiment investigates the last hypothesis that finetuning an LR speech representation model trained only on Austrian conversational speech is not effective for Austrian conversational speech.

The final row of table 1 shows WERs achieved with the models PHN-LR and CHR-LR. The PHN-LR **4-gram** results were slightly worse than the results from the LR Kaldi approach, both with respect to mean WERs and to the worst conversation (i.e., by a difference of 8.33%). All PHN-LR models performed better than CHR-LR models, resulting in mean WER differences of 8.26% (**4-gram**) and 9.6% (**Lex**). Interestingly, **Lexfree** and **Lex** results were similarly bad in case of CHR-LR, with mean WERs of approx. 85%.

We refer back to section 4.1.2 which presents WERs from the literature in case of LR conversational speech recognition, because the results from this experiment again demonstrate problems with respect to both, performance and robustness in case of LR scenarios.

From this experiment, we conclude that finetuning a LR speech representation model which is pre-trained merely on Austrian conversational speech is not effective. Also our Kaldi-LR results demonstrate similar performance issues. Consequently, for neither of the two LR ASR approaches presented in this paper, where models were trained merely on GRASS conversational speech, resulted in state-of-the-art WERs for conversational speech. Hence, our third hypothesis is true, and we show that training on approx. 11.5h hours of conversational speech emphasizes the data sparsity problem. Additionally, these results are also reinforcing our first hypothesis, i.e., that performing cross-validation by testing each conversation individually points out conversational speech complexity and certifies the LR language processing assumption.

5. Corollary

After answering our hypotheses, this section discusses further findings which result from our experiments: we discuss the role of linguistic knowledge, the role of targets and the role of inter-speaker vs. inter-conversation variation.

5.1. Role of Linguistic Knowledge

We made several observations when looking at the influence of incorporating knowledge given by a lexicon or LM in case of wav2vec2.0 models.

Both, lexicon-based PHN/CHR-LR and PHN/CHR-XLSR models benefit from LM probabilites whereby higher differences in WERs can be observed in case of LR models ($\approx$ 20% with respect to mean values). When comparing PHN-models with CHR-models those improvements are similar in the LR cases, but they differ more strongly in the XLSR cases, despite the overall WERs being similar. Hence, we notice that incorporating a LM has an higher impact on lexicon-based PHN models compared to lexicon-based CHR models in the XLSR case. At the same time, however, lexicon-free CHR-XLSR solutions are similar to lexicon-based PHN-XLSR solutions.

The experiments presented in (Conneau et al., 2021) showed WER improvements of $\approx$ $2 \ldots 4\%$ due to LM incorporation when finetuning a smaller CHR-XLSR model. Another study showed improvements by adding LM probabilities and more advanced lexicons via dialect variation modeling (Khosravani et al., 2021). To the best of our knowledge, comparisons between PHN/CHR-XLSR models showing varying impacts of LM probabilities, have not yet been reported.

Looking at lexicon-based beam-search decoding results from PHN-XLSR and knowing that the AM search space is entirely restricted by the lexicon, one might argue that word mapping ambiguities lead to some substitution errors due to homophones. However, in case of our small canonical lexicon, only $\approx$ 1.8% of all words are ambiguous introducing those unpredictable errors[2] and we believe that those errors are small in comparison to errors which arise from missing Austrian German pronunciations. Hence, we conclude

[2] We hypothesize that, when introducing ambiguous pronunciations in the lexicon, words are randomly selected during beam-search decoding.

that the canonical pronunciations in the lexicon, which introduce 65 target phones, lead to higher amounts of training errors in comparison to errors occurring from 31 character targets due to more noisy labels in case of Austrian German. We hypothesize this error to be lower in case of character-based systems, because they have only 31 character targets. Nevertheless, for both phone-based and character-based systems, incorporating LM probabilities resulted to help to reduce the impact of ambiguities.

5.2. Role of Targets

Here, we discuss the role of target labels by comparing our phone-based and character-based systems. If we look at performances from the LR models, character-based systems performed worse than phone-based systems in case of both, **Lex** and **4-gram**, whereby Kaldi-LR WERs were more similar to PHN-LR **4-gram** WERs. XLSR models showed similar results when decoding with a LM, but in case of only lexicon-based decoding, CHR-XLSR models achieved better performances.

The systematic comparisons between character-based and phone-based ASR systems by (Basson and Davel, 2012) showed that increasing training data leads to similar performances in character-based and phone-based ASR systems. (Zeineldeen et al., 2020) compared results for attention-based encoder-decoder models and found similar performances for character-based and phone-based systems regardless of lexicon or LM incorporation with more training data in general. Additionally, they also achieved similar results of 18.2% (PHN) and 18.6% (CHR) with a simplified decoder without LM nor lexicon by inserting word-disambiguate and end-of-word symbols in case of their phone-based models.

Our results are in line with results reported in the literature and suggest that character-based systems give similar performances as phone-based systems if enough data is available. However, our differences between phone-based and character-based models in case of lexicon-based decoding results indicate the relevance of knowledge, and that, for instance, the incorporation of more advanced lexicons might lead to further improvements.

5.3. Inter-Conversation vs. Inter-Speaker Variation

Our results indicate that variation in WERs with respect to each conversation and with respect to each speaker differs when comparing LR and XLSR models.

Table 1 shows that in case of beam-search decoding standard deviations of WERs are always higher in the LR scenario with wav2vec2.0 models than the WERs of the XLSR scenario. Figure 1 clarifies this variation by comparing speaker-dependent WERs of **4-gram** models. In general, histograms over bins with 5%-width show that overall WERs and the range of speaker-dependent WERs are lower in case of XLSR models compared to LR models. Furthermore, when comparing WER ranges normalized by mean value we measure values of 0.79 (PHN/CHR-XLSR), 0.6 (PHN/CHR-LR) and 0.46 (Kaldi-LR). Figure 2 clearly demonstrates that range of conversation-dependent WERs is lower in case of Kaldi models (21.23%) compared to wav2vec2.0 LR models (31.84%). In case of normalized conversation-dependent WER ranges we measure values of 0.69 (PHN/CHR-XLSR), 0.53 (PHN/CHR-LR) and 0.38 (Kaldi-LR). Corresponding entropy measurements which address directly to the shape of the distributions are 0.83 (PHN/CHR-XLSR), 0.96 (PHN/CHR-LR) and 0.79 (Kaldi-LR). Even if absolute WER ranges of XLSR models are lowest, our measurements demonstrate that Kaldi distributions appear to have the least unexplained variability, especially in case of conversation-dependent WERs.

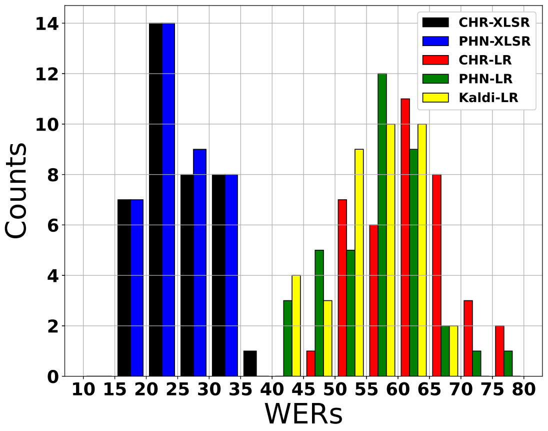

Figure 1: Histogram showing speaker-dependent WERs (**4-gram**). WERs and range of WERs are lower in case of finetuned XLSR models ($35.71\% \ldots 16.09\% = 19.62\%$) in comparison to finetuned LR models ($79.37\% \ldots 43.36\% = 36.01\%$). Kaldi-LR model WERs range from $69.19\% \ldots 43.42\% = 25.77\%$.

A broad study on domain shifts in self-supervised pre-training (Hsu et al., 2021) observe that adding more out-of-domain data during pre-training is beneficial and simultaneously pre-training on more domains improves robustness in general.

Our findings confirm the effectiveness of finetuning GRASS conversational speech with XLSR with respect to performance, but we still observe lack of robustness with respect to resulting WER distributions. Additionally, in case of Kaldi models variation per conversation appears to be better modeled than variation per speaker.

6. Conclusions

In this paper we presented ASR experiments for Austrian German conversational speech from two ASR frameworks, the Kaldi speech recognition toolkit and fairseq (i.e., wav2vec2.0). We investigated the impact

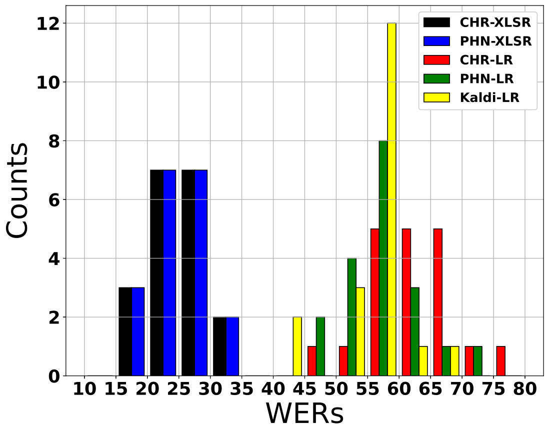

Figure 2: Histogram showing conversation-dependent WERs (**4-gram**). WERs and range of WERs are lower in case of finetuned XLSR models ($34.49\% \ldots 17.36\% = 17.13\%$) in comparison to finetuned LR models ($76.98\% \ldots 45.14\% = 31.84\%$). Kaldi-LR model WERs range from $65.12\% \ldots 43.89\% = 21.23\%$.

of data size, inter-speaker and inter-conversation variation, and structural knowledge for ASR performance, and compared phone-based and character-based ASR approaches.

Our results showed the effectiveness of finetuning a pre-trained cross-lingual speech representation model when solving LR ASR tasks with Austrian conversational speech. Even though performances were already satisfying with the data-driven approach, we still observed the importance of including structural linguistic knowledge via a lexicon or LM, as WERs decreased in case of both, LR and XLSR models. Furthermore, WERs varied strongly from speaker to speaker and from conversation to conversation, indicating the complexity of conversational speech, and also indicating the lack of robustness to speaker variation in case of all ASR approaches shown here.

In future, we will further investigate whether the impact of more advanced lexicons and LMs is larger for ASR of conversational speech in comparison to ASR of other less spontaneous and less casual speaking styles. Given our findings from this paper, we hypothesize that better performing systems do not necessarily result in systems which are also more robust to inter-speaker variation.

7. Acknowledgements

The work by B. Schuppler and the cooperation between SPSC Laboratory and Idiap was funded by grant V-638-N33 from the Austrian Science Fund (FWF). J. Linke received additional support from the Rudolf Chaudoire Foundation.

8. Bibliographical References

Adda-Decker, M., Schuppler, B., Lamel, L., Morales-Cordovilla, J. A., and Adda, G. (2013). What we can learn from ASR errors about low-resourced languages: A case-study of Luxembourgish and Austrian. In *ERRARE Workshop - Ermenonville, Paris, France*.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *LREC*, pages 4218–4222.

Baevski, A., Auli, M., and Mohamed, A. (2020a). Effectiveness of self-supervised pre-training for speech recognition. *ArXiv*, abs/1911.03912.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020b). wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, et al., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Basson, W. and Davel, M. (2012). Comparing grapheme-based and phoneme-based speech recognition for Afrikaans. In *Pattern Recognition Association of South Africa (PRASA)*, pages 144–148, 11.

Chung, Y.-A. and Glass, J. R. (2018). Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *ArXiv*, abs/1803.08976.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2021). Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Furui, S., Nakamura, M., Ichiba, T., and Iwano, K. (2005). Why is the recognition of spontaneous speech so hard? In Václav Matoušek, et al., editors, *Text, Speech and Dialogue*, pages 9–22.

Furui, S. (2009). Generalization problem in asr acoustic model training and adaptation. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 1–10.

Gales, M., Knill, K., Ragni, A., and Rath, S. (2014). Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. *Fourth International Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU-2014)*, pages 16–23.

Gales, M. (1998). Maximum likelihood linear transformations for hmm-based speech recognition. In *Computer Speech and Language*, volume 12, pages 75–98.

Glass, J. (2012). Towards unsupervised speech processing. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1–4.

Gopinath, R. A. (1998). Maximum likelihood modeling with Gaussian distributions for classification. In *Proc. of ICASSP*, volume 2, pages 661–664.

Graves, A., Fernández, S., Gomez, F., and Schmidhuber, J. (2006). Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning*, pages 369–376.

Heafield, K. (2011). KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

Hsu, W.-N., Sriram, A., Baevski, A., Likhomanenko, T., Xu, Q., Pratap, V., Kahn, J., Lee, A., Collobert, R., Synnaeve, G., and Auli, M. (2021). Robust wav2vec 2.0: Analyzing Domain Shift in Self-Supervised Pre-Training. In *Proc. Interspeech 2021*, pages 721–725.

Khosravani, A., Garner, P. N., and Lazaridis, A. (2021). Modeling Dialectal Variation for Swiss German Automatic Speech Recognition. In *Proc. Interspeech 2021*, pages 2896–2900.

Laurent, A., Fraga-Silva, T., Lamel, L., and Gauvain, J. (2016). Investigating techniques for low resource conversational speech recognition. In *Proc. of ICASSP*, pages 5975–5979.

Moosmüller, S. (2007). *Vowels in Standard Austrian German. An Acoustic-Phonetic and Phonological Analysis*. Habilitation Thesis, University of Vienna.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Proc. Interspeech 2015*, pages 3214–3218.

Popescu-Belis, A., Lalanne, D., and Bourlard, H. (2012). Finding information in multimedia meeting records. In *IEEE MultiMedia*, volume 19, pages 48–57.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. *Proc. of IEEE ASRU Workshop*.

Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Reichel, U. D. and Kisler, T. (2014). Language-independent grapheme-phoneme conversion and word stress assignment as a web service. *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2014*, pages 42–49.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Schuppler, B., Hagmüller, M., Morales-Cordovilla, J. A., and Pessentheiner, H. (2014a). GRASS: the Graz corpus of Read And Spontaneous Speech. In *LREC*, pages 1465–1470.

Schuppler, B., Adda-Decker, M., and Morales-Cordovilla, J. A. (2014b). Pronunciation variation in read and conversational austrian German. In *Proc. Interspeech 2014*, pages 1453–1457.

Schuppler, B., Hagmüller, M., and Zahrer, A. (2017). A corpus of read and conversational Austrian German. In *Speech Communication*, volume 94, pages 62–74.

Sriranjani, R., Karthick, B. M., and Umesh, S. (2015). Investigation of different acoustic modeling techniques for low resource Indian language data. In *2015 Twenty First National Conference on Communications (NCC)*, pages 1–5.

Stolcke, A. (2002). Srilm – an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)*, pages 901–904.

Synnaeve, G., Xu, Q., Kahn, J., Grave, E., Likhomanenko, T., Pratap, V., Sriram, A., Liptchinsky, V., and Collobert, R. (2019). End-to-end ASR: from supervised to semi-supervised learning with modern architectures. *ArXiv*, abs/21911.08460.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In I. Guyon, et al., editors, *Advances in Neural Information Processing Systems*, volume 30.

Wright, R. (2006). Intra-speaker variation and units in human speech perception and ASR. In *ITRW on Speech Recognition and Intrinsic Variation (SRIV 2006)*, pages 39–42.

Yi, C., Wang, J., Cheng, N., Zhou, S., and Xu, B. (2020). Applying wav2vec2.0 to speech recognition in various low-resource languages. *ArXiv*, abs/2012.12121.

Zeineldeen, M., Zeyer, A., Zhou, W., Ng, T., Schlüter, R., and Ney, H. (2020). A systematic comparison of grapheme-based vs. phoneme-based label units for encoder-decoder-attention models. *ArXiv*, abs/2005.09336.

Zhang, Y., Park, D. S., Han, W., Qin, J., Gulati, A., Shor, J., Jansen, A., Xu, Y., ..., and Wu, Y. (2021). Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *ArXiv*, abs/2109.13226.