# Improving Luxembourgish Speech Recognition with Cross-Lingual Speech Representations

Nguyen, Le Minh; Nayak, Shekhar; Coler, Matt

*Published in:*
2022 IEEE Spoken Language Technology Workshop, SLT 2022 - Proceedings

*DOI:*
10.1109/SLT54892.2023.10022706



*Publication date:*
2023

# IMPROVING LUXEMBOURGISH SPEECH RECOGNITION WITH CROSS-LINGUAL SPEECH REPRESENTATIONS

*Le Minh Nguyen, Shekhar Nayak, Matt Coler*

University of Groningen
leminhnguyen@outlook.com, {s.nayak, m.coler}@rug.nl

2022 IEEE Spoken Language Technology Workshop (SLT) | 979-8-3503-9690-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/SLT54892.2023.10022706

## ABSTRACT

*Luxembourgish* is a West Germanic language spoken by roughly 390,000 people, mainly in Luxembourg. It is one of Europe's under-described and under-resourced languages, not extensively investigated in the context of speech recognition. We explore the self-supervised multilingual learning of Luxembourgish speech representations for the speech recognition downstream task. We show that learning cross-lingual representations is essential for low-resourced languages such as Luxembourgish. Learning cross-lingual representations and rescoring the output transcriptions with language modelling while using only 4 hours of labelled speech achieves a word error rate of 15.1% and improves our Transfer Learning baseline model relatively by 33.1% and absolutely by 7.5%. Increasing the amount of labelled speech to 14 hours yields a significant performance gain resulting in a *9.3%* word error rate.[1]

***Index Terms—*** Luxembourgish, multilingual speech recognition, language modelling, wav2vec 2.0 XLSR-53, under-resourced language

## 1. INTRODUCTION

*Lëtzebuergesch* (Luxembourgish) is a West Germanic language spoken by roughly 390,000 people, mainly in Luxembourg. It remains one of Europe's under-described and under-resourced languages, not extensively investigated in the context of speech recognition. An initial attempt towards dealing with the low resource problem for Luxembourgish speech recognition was made by building unsupervised context-dependent models using 1200 hours of untranscribed Luxembourgish data [1]. DNN-based acoustic modelling was improved using target domain untranscribed data and out-of-domain transcribed data for low resource Luxembourgish speech recognition [2]. However, the performance of these approaches is not close to that of any high-resource language speech recognition systems.

A recent study on five different low resource languages demonstrated that different supervised deep learning based acoustic models work well for different languages and there is not a particular architecture which can work well for all languages [3]. A recent work on low resource speech recognition on Bemba language with 17.5 hours of labelled data using a supervised end to end model (Deep Speech [4]) reported a WER of 54.78% using a 5-gram language model. This seems to be far off from any rich resource speech recognition system performance and motivates towards making use of unlabelled resources for under-resourced languages.

More recently, self-supervised learning has become a paradigm for determining general data representations from unlabelled examples for downstream tasks such as speech recognition. This paradigm has shown the feasibility of speech recognition based on limited labelled data with the wav2vec 2.0 model [5]. The XLSR model is a multilingual speech recognition model based on wav2vec 2.0. It pre-trains on cross-lingual speech representations. Experiments have shown that cross-lingual pre-training significantly outperforms monolingual pre-training [6]. These self-supervised pre-trained models have been used to improve low resource speech recognition for different languages [7, 8, 9].

This paper explores this self-supervised paradigm to enable speech recognition for Luxembourgish by creating the first Luxembourgish wav2vec 2.0 model. We create a Luxembourgish speech recognition corpus of upto 14 hours from radio broadcast domain audio which can be readily found for low resource languages [10]. As our baseline, we conduct two experiments using this corpus. First, we train a monolingual wav2vec 2.0 model from scratch on Luxembourgish speech. We also apply transfer learning by fine-tuning Luxembourgish speech representations on a wav2vec 2.0 model pre-trained on the LibriSpeech corpus. To improve our baseline, we investigate multilingual learning by pre-training Luxembourgish speech representations on top of the XLSR-53 checkpoint pre-trained on 53 languages.

Our results demonstrate that pre-training cross-lingual speech representations are essential for low-resourced languages such as Luxembourgish. Learning cross-lingual representations and rescoring the output transcriptions with language modelling yield better results than the monolingual model or applying transfer learning from LibriSpeech. When using only 4 hours of labelled Luxembourgish speech, our

[1]Models and datasets are available at `https://huggingface.co/lemswasabi`

978-1-6654-7189-3/22/$31.00 ©2023 IEEE

multilingual XLSR-53 wav2vec 2.0 model with Language Model (LM) rescoring achieves a WER of 15.1% and improves the baseline result relatively by 33.1% and absolutely by 7.5%. Increasing the transcribed speech dataset to 14 hours, our model sets the new best WER for Luxembourgish Speech Recognition of 9.3%. Our results achieve a 58.8% relative and 13.3% absolute improvement over our baseline.

## 2. LUXEMBOURGISH CORPUS CREATION

We use labelled radio broadcast audio from *Radio Telé Lëtzebuerg* (RTL.lu)[2]. As unlabelled data, we collected 842 hours of unlabelled speech from the same domain. Further, we normalize Luxembourgish speech labels by verbalizing Non-Standard Words (NSWs). After normalizing the transcription texts, we segment the transcribed audio into smaller utterances that can be used to fine-tune the acoustic model. We apply forced alignment to match speaker utterances from an audio file to their candidate labels extracted from the audio transcription. Forced alignment, used from the *aeneas* library, generates segmented utterances by first chunking the transcription text into text fragments. Then each text fragment is synthesized with a Text-to-Speech (TTS) engine. MFCCs are extracted from both signals, and the Sakoe-Chiba Band DTW algorithm is applied to align the synthesized text fragment with the reference signal. This procedure generates timestamps of the segment boundaries. However, forced alignment in aeneas does not estimate the confidence level within a region of similarity as in the approaches of [11, 12]. Without a confidence metric evaluating the correctness of a segment label, we need language experts to validate the labelled utterances. Therefore, we implement the idea of the human in the loop similar to the data processing pipeline from [13].

To those ends, we design a tool where a user can validate aligned audio-transcription pairs and correct the transcription if necessary. A detailed view of a segmented audio item is pictured in Figure 1. First, it visualizes the waveform of the current audio to the user and highlights the currently selected segment area within the waveform. Underneath the waveform component, we list different information about the audio to the user. The information box displays the audio file duration and the source from where it was collected. Besides this information, we show the user the total number of segments. Additionally, a progress bar tells the user how many segments are already validated. The third visual component is the editor area, where the user can correct the current transcription, start and end timestamps. In this editor area, we provide the user with the functionalities to play back or validate the current segment. Furthermore, the user can create and delete segments.

This segmentation process yielded 6874 correctly transcribed utterances with an average duration of 7 seconds. To divide the 842 hours of unlabelled audio files into 15 to 20 seconds long segments, we used the same procedure from [12].

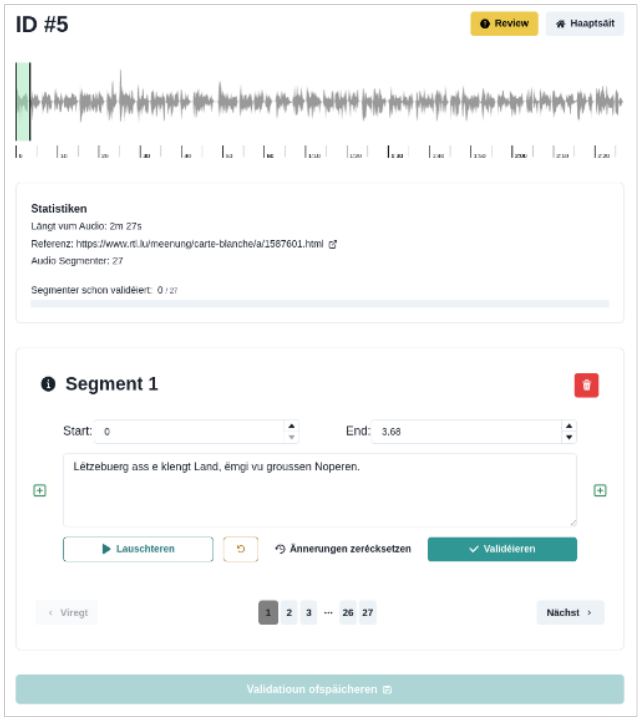


**Fig. 1**. Detail page of the audio segments validation tool.

After successfully segmenting the collection of large audio files into smaller utterances, we split the utterances into different sets for training, validating and testing our speech recognition model. To create the speech corpus, we shuffle the utterances first and allocate 95% to the training set. The rest 5% is evenly split between the validation and testing sets. Furthermore, for our experiments, we created four training data configurations to study the effect of the amount of labelled data on the performance of wav2vec 2.0. These configurations represent training sets of different sizes that are 4, 8, 11 and 14 hours, respectively. The validation and testing sets are kept the same for each configuration. For the scope of this study, we are not ensuring speaker generalization. Thus, we shuffle the speech samples before splitting them into the different sets without ensuring that each speaker's utterances appear only in one data split.

To catalyze further contributions to Luxembourgish ASR, we release a subset of the Luxembourgish dataset situated in the public domain under the copyright of RTL.lu.

[2]https://www.rtl.lu/

## 3. EXPERIMENTS AND RESULTS

For our baseline, we pre-train a monolingual wav2vec 2.0 model from scratch in one experiment and use transfer learning in another to fine-tune Luxembourgish speech representations on a wav2vec 2.0 checkpoint pre-trained on English speech representations from LibriSpeech. Furthermore, we investigate if multilingual wav2vec 2.0 models improve the performance of our baseline models. Thus, we apply the wav2vec 2.0 XLSR model that pre-trains cross-lingual speech representations of 53 different languages from BABEL, Common Voice and MLS [6]. We pre-train Luxembourgish speech representations on top of this checkpoint. Finally, we fine-tune each model configuration on labelled speech for the speech recognition downstream task.

### 3.1. ASR models

As our baseline, we train the first Luxembourgish base wav2vec 2.0 models. In the first experiment, we pre-trained a model from scratch only on the 842 hours of unlabelled speech. In the second experiment, we fine-tuned the same unlabelled speech on a wav2vec 2.0 model pre-trained on 960 hours of English speech from the LibriSpeech corpus. Both models were introduced with a linear layer to fine-tune them on 4 hours of labelled Luxembourgish speech from broadcast news using the CTC loss. The two models achieved a test WER of 23.09% and 22.57%, respectively. Table 1 shows the validation and testing evaluation results. This represents the baseline to benchmark our multilingual models with LM rescoring against it.

To improve our baseline models, we investigate the self-supervised learning of cross-lingual speech representations to verify the claim of [6, 14] that learning cross-lingual speech representations significantly outperform monolingual wav2vec 2.0 models pre-trained only on a single language. For this evaluation, we fine-tuned a wav2vec 2.0 XLSR-53 checkpoint [6] on 842 hours of unlabelled speech from the RTL.lu domain. Pre-training Luxembourgish speech representations in addition to the 53 languages took three days on four Nvidia V100 GPUs with 32 GB VRAM each. We fine-tuned the pre-trained acoustic model on different labelled data setups of 4, 8, 11, and 14 hours. All our fine-tuning experiments were executed on a mobile Nvidia 1070 GPU with 8 GB of VRAM, and the duration lasted between 12 hours to 1.5 days. We used the default hyper-parameters for the pre-training and fine-tuning from Fairseq[3] and Hugging Face[4] implementations respectively.

We rely on n-gram LMs to reestimate our speech decodings. Even though language modelling based on neural network architectures outperform their n-gram counterpart, neural-based LMs are still very computationally expensive [15, 16, 17]. Additionally, we decided on n-gram language modelling since it is well integrated into the Hugging Face wav2vec 2.0 implementation. We have access to two text corpora from a similar domain as the audio data for this task. The Zenter fir d'Lëtzebuerger Sprooch (ZLS) provided access to the parliamentary debate transcriptions from 2020 to 2022, representing 4 million tokens. The second corpus was given to us by RTL.lu and represents 16 million tokens from the same news domain. In total, the text corpus corresponds to 20 million tokens. Our experiments explore which corpus domain performs better when rescoring the model outputs.

### 3.2. Results

In our experiments, the evaluation of our model experiments on the test set shows that learning cross-lingual representations are essential for low-resourced languages such as Luxembourgish. Learning cross-lingual representations and rescoring the output transcriptions with language modelling while using the same amount of labelled speech achieves a WER of 15.1% and improves the baseline result by 33.1% and absolutely by 7.5%. Increasing the amount of labelled speech yields a significant performance gain. Our best multilingual wav2vec 2.0 model fine-tuned on 14 hours of speech reaches a 9.3% WER. The detailed results of our fine-tuning experiments with different training data sizes are reported in Table 1.

We visualize the findings of our multilingual fine-tuning experiments in Figure 2 which showcases better the data from Table 1. The larger the transcribed dataset increases, the more the WER trend of the speech decoding decreases. We observe the same trend when decoding with language modelling. Overall the 5-gram LM decreases the WER between one and two per cent.

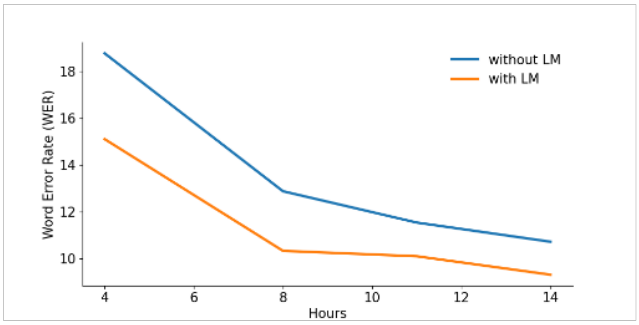


**Fig. 2**. Word Error Rate respective to labelled dataset size. With and without Language Modelling (LM).

Additionally, in our multilingual experiments, we study the impact of the text corpus size on recognition performance. In language modelling, the perplexity metric measures how well an LM predicts a sequence of words. A low perplexity

[3] https://github.com/facebookresearch/fairseq/tree/main/examples/wav2vec/config/pretraining

[4] https://huggingface.co/hf-test/xls-r-300m-sv/blob/main/run.sh

| Model | Unlabelled data | Language Model | WER dev | WER test | CER dev | CER test |
|---|---|---|---|---|---|---|
| **Baseline:** | | | | | | |
| **4h labelled** | | | | | | |
| Base wav2vec 2.0 | LB-842 | None | 23.95 | 23.09 | 7.97 | 7.63 |
| Base wav2vec 2.0 | LS-960 + LB-842 | None | 23.39 | 22.57 | 8.15 | 7.60 |
| **Multilingual Models:** | | | | | | |
| **4h labelled** | | | | | | |
| XLSR-53 wav2vec 2.0 | LB-842 | None | 19.44 | 18.77 | 7.16 | 6.43 |
| | | 5-gram | 16.11 | 15.10 | 6.63 | 5.79 |
| **8h labelled** | | | | | | |
| XLSR-53 wav2vec 2.0 | LB-842 | None | 13.86 | 12.87 | 3.11 | 2.91 |
| | | 5-gram | 10.94 | 10.32 | 2.48 | 2.39 |
| **11h labelled** | | | | | | |
| XLSR-53 wav2vec 2.0 | LB-842 | None | 12.68 | 11.53 | 2.76 | 2.55 |
| | | 5-gram | 9.98 | 10.09 | 2.27 | 2.22 |
| **14h labelled** | | | | | | |
| XLSR-53 wav2vec 2.0 | LB-842 | None | 11.68 | 10.71 | 2.64 | 2.31 |
| | | 5-gram | **9.50** | **9.30** | **2.17** | **2.08** |

**Table 1**. Results of our Luxembourgish wav2vec 2.0 fine-tuning experiments. CER and WER on the Luxembourgish dev/test sets when training on the labelled data setups of 4h, 8h, 11h and 14h. The models were pre-trained using the audio from the Luxembourgish dataset (LB-842) as unlabelled data. In the transfer learning setup, the model was pre-trained on audio from LibriSpeech (LS-960). The best WERs and CERs on dev and test sets are highlighted in bold.

metric indicates a good model. The perplexity of an LM is related to the sparseness of a reference text corpus used to train the model. To reduce the perplexity, we generally increase the corpus size or apply word decomposition algorithms to reduce the out-of-vocabulary rate. Therefore, we investigate the impact of two 5-gram LMs on speech recognition decoding. The first LM is trained on 4 million tokens from the parliament transcription corpus. The second model is trained on the combination of the parliament and RTL.lu text resources, totalling 20 million tokens. In Figure 3, we plot the WERs for each LM decoder used in our different-sized acoustic models.

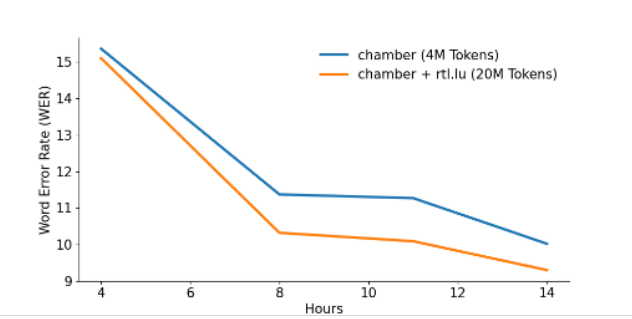


**Fig. 3**. Word Error Rate respective to labelled dataset size while comparing the impact of LM sizes.

In the first configuration, in which the acoustic model was fine-tuned on only 4 hours of labelled speech, the increase in text data did not improve the WER considerably. However, when the size of the labelled dataset increases, the larger text corpus significantly decreases the WER for the acoustic models fine-tuned with 8 and 11 hours, respectively. This observation could imply that using more text data for language modelling is less effective in low resource than in high resource acoustic models.

To analyze the errors emitted by the speech decoder, we inspect the model inference on the test dataset and compare it to the ground truth labels. We can summarize the common decoding errors into the following categories:

1. Words composed of multiple words are decoded as separate words. E.g. *schlësselroll* is decoded by the model as *schlëssel roll*
2. The speaker pronounces loan words from contact languages how they would sound in Luxembourgish. E.g. *caféepicerie* which is a loan word from French, is decoded by the model as *kaffi episserie*.
3. The speaker pronounces a correct variation of a word that is not used in the ground truth. E.g. *interessi* is pronounced instead of *interesse*.
4. The n-rule is often misused while speaking. The letter *n* at the end of words is not dropped when speaking even though it is dropped in the written form.
5. On some occasions, the article *d'* is missing in the output transcription.

6. The speaker pronounces silent vowels and consonants.

7. The model decodes the wrong vowel. E.g. *d'schwemm* instead of *d'schwämm*.

## 4. CONCLUSION

We presented our work on creating a Luxembourgish speech corpus to enable Luxembourgish speech recognition. We segmented the transcribed audio into utterances by applying forced alignment combined with validation from language experts. We also trained the first Luxembourgish wav2vec 2.0 models in a monolingual and transfer learning setting for our baseline. To improve our baseline, we investigated the self-supervised multilingual learning of Luxembourgish speech representations for the downstream speech recognition task. Therefore, we fine-tuned a wav2vec 2.0 XLSR-53 checkpoint on 842 hours of unlabelled Luxembourgish speech.

Our experiments validate that learning cross-lingual representations and LM rescoring are essential for low-resourced languages such as Luxembourgish. Learning cross-lingual representations and rescoring the output transcriptions with language modelling while using only 4 hours of labelled speech achieves a word error rate of 15.1% and improves the baseline result for Luxembourgish speech recognition relatively by 33.1% and absolutely by 7.5%. Increasing the amount of labelled speech to 14 hours yields a significant performance gain resulting in a *9.3%* word error rate.

As for future research, we expect performance improvements by collecting more Luxembourgish text resources for language modelling. In addition, to collecting more text data, we can now use our ASR model to augment the transcribed speech corpus by applying the Smith-Waterman alignment algorithm. This approach inexpensively increases the Luxembourgish speech corpus since there are publicly available speech-transcription pairs, and we do not need to rely on language experts to supervise audio segmentation and correct speech labels.

wav2vec 2.0 enables speech recognition for Luxembourgish with little labelled data. Additionally, collaborating with the Zenter fir d'Lëtzebuerger Sprooch (Center for the Luxembourgish Speech) and *Radio Telé Lëtzebuerg* (RTL.lu), we expect to implement this solution in production to make Luxembourgish speech recognition accessible to end-users and improve the digital inclusion in Luxembourg. We hope to catalyze research in natural language processing for Luxembourgish by generating more text resources from broadcast news with our speech recognition system.

## 5. ACKNOWLEDGMENTS

We thank the Faculty Board of the University of Groningen for their funding to cover expenses for attending the SLT22 conference. Additionally, we thank the Zenter fir d'Lëtzebuerger Sprooch for their help normalizing the transcriptions of the segmented audio.

## 6. REFERENCES

[1] Martine Adda-Decker, Lori Lamel, and Gilles Adda, "Speech alignment and recognition experiments for Luxembourgish," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[2] Karel Veselỳ, Carlos Segura, Igor Szöke, Jordi Luque, and Jan Cernockỳ, "Lightly supervised vs. semi-supervised training of acoustic model on Luxembourgish for low-resource automatic speech recognition.," in *INTERSPEECH*, 2018, pp. 2883–2887.

[3] Ethan Morris, Robbie Jimerson, and Emily Prud'hommeaux, "One size does not fit all in resource-constrained ASR," *Proceedings of INTERSPEECH 2021*, 2021.

[4] Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, et al., "Deep speech: Scaling up end-to-end speech recognition," *arXiv preprint arXiv:1412.5567*, 2014.

[5] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[6] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *CoRR*, vol. abs/2006.13979, 2020.

[7] Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu, "Applying wav2vec2. 0 to speech recognition in various low-resource languages," *arXiv preprint arXiv:2012.12121*, 2020.

[8] Mitchell DeHaven and Jayadev Billa, "Improving low-resource speech recognition with pretrained speech models: Continued pretraining vs. semi-supervised training," *arXiv preprint arXiv:2207.00659*, 2022.

[9] Jing Zhao and Wei-Qiang Zhang, "Improving automatic speech recognition performance for low-resource languages with self-supervised models," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[10] Moussa Doumbouya, Lisa Einstein, and Chris Piech, "Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, pp. 14757–14765.

[11] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015*, pp. 5206–5210.

[12] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "Mls: A large-scale multilingual dataset for speech research," *arXiv preprint arXiv:2012.03411*, 2020.

[13] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[14] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al., "Xls-r: Self-supervised cross-lingual speech representation learning at scale," *arXiv preprint arXiv:2111.09296*, 2021.

[15] Hainan Xu, Tongfei Chen, Dongji Gao, Yiming Wang, Ke Li, Nagendra Goel, Yishay Carmiel, Daniel Povey, and Sanjeev Khudanpur, "A pruned rnnlm lattice-rescoring algorithm for automatic speech recognition," in *IEEE international conference on acoustics, speech and signal processing (ICASSP) 2018*, pp. 5929–5933.

[16] Shankar Kumar, Michael Nirschl, Daniel Holtmann-Rice, Hank Liao, Ananda Theertha Suresh, and Felix Yu, "Lattice rescoring strategies for long short term memory language models in speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 165–172.

[17] Ebru Arisoy, Abhinav Sethy, Bhuvana Ramabhadran, and Stanley Chen, "Bidirectional recurrent neural network language models for automatic speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)2015*, pp. 5421–5425.