LUX-ASR: BUILDING AN ASR SYSTEM FOR THE LUXEMBOURGISH LANGUAGE

Peter Gilles, Nina Hosseini-Kivanani, Léopold Hillah

University of Luxembourg
Department of Humanities & Department of Computer Science
L-4365 Esch/Belval, Luxembourg

ABSTRACT

We present a first system for automatic speech recognition (ASR) for the low-resource language Luxembourgish. By applying transfer-learning, we were able to fine-tune Meta's wav2vec2-xls-r-300m checkpoint with 35 hours of labeled Luxembourgish speech data. The best word error rate received lies at 14.47.

***Index Terms*—** ASR, low-resource language, Luxembourgish, wav2vec 2.0

1. INTRODUCTION

In the course of the 20th century, the small language Luxembourgish evolved to the national language of the Grand-Duchy of Luxembourg [1], [2]. The language is typologically quite close to German and has in fact evolved out of a German dialect, which has been standardized recently to certain degree with regard to spelling and lexicon. It is, however, characterised by a high degree of multilingualism: Words and phrases coming mainly from French and German are part of the regular lexicon and occur quite frequently. Luxembourgish is used by approximately 300,000 speakers, mainly as a spoken language of everyday use, but also as the only spoken language in parliament. As an identity symbol this language is used more and more in speaking and writing due to social media and digitization of everyday life. This has led to the need of tailored NLP and voice tools (especially STT and TTS) for this very specific context (first studies are [3], [4], [5]). Studies dedicated to ASR are [6], [7] and quite recently [8]. In this study, we will explore and demonstrate how a system for Automatic Speech Recognition (ASR) can be developed for Luxembourgish, taking into account the specific constraints of a low-resource language, i.e. scarcity of training material and multilingualism.

2. STATE-OF-THE-ART

In recent years, with the rise of highly efficient machine learning algorithms (deep learning), the field of speech recognition has seen significant advancements. ASR is one of these domains that has grown significantly, and among them, ASR systems for low-resource languages have shown a proliferation. Despite recent interest in ASR, approximately 128 languages have ASR systems and have experienced this progress. The lack of ASR systems for other languages is due to the fact that building a reliable ASR system requires a large amount of annotated data. The standard procedure for creating a speech recognizer requires 1) thousands of hours of annotated speech for an acoustic model, 2) a phonetic pronunciation dictionary, and 3) a large amount of text for creating a language model. Recently, ASR for under-sourced languages has also shown growing interest. However, ASR systems encounter difficulties with speech data from low-source languages.

Currently, the significant improvements over available approaches for ASR systems happened with self-supervised transformer-based models for audio processing (i.e., wav2vec 2.0) which helps ASR systems handle unlabelled data pretty well. Due to the fact that with traditional ASR systems (HMM-based approaches like Kaldi [1], etc.), a high amount of exactly transcribed audio data and a pronunciation dictionary was required for training, which is costly and not available for many low-resource languages. But with the advent of the huge pre-trained multilingual base models and wav2vec algorithms [9], considerable progress in developing ASR systems can be made for low-resource languages when training data is scarce. Furthermore, the performance of ASR systems is not related to the amount of labeled data for this scenario.

3. MAIN PURPOSE

We build an ASR system for recognizing Luxembourgish speech: We place particular emphasis on testing systems for a genre-specific task, i.e., transliteration of speeches and debates of the Luxembourgish parliament ('Chambre de Députes'), who also provided a substantial part of the training data. Further fields of applications for an ASR system for Luxembourgish are assistance tools for transcribing interviews or historic media archives and for human-machine interaction in general.

[1] Kaldi: `http://kaldi-asr.org/doc/`

4. MATERIAL AND METHOD

The training material consists of pairs of audio samples with their matching orthographic transcriptions, and it originates from the following sources in table 1.

1.	parliamentary speeches	25 hours
2.	crowd-sourced sentences [10]	9 hours
3.	MaryLux sentences [2]	1 hour

Table 1: Overview of the sources for the training data.

For all training data, highly reliable written transcripts are available, which also follow the orthographic standard of Luxembourgish. Numbers and special characters like "%, &, +" have been replaced by the corresponding words. The data is split into smaller chunks of audio files with a duration of between 1 second and 20 seconds. While sources 2. and 3. are already available in the correct format, the larger audio files of the parliamentary speeches have been split into smaller chunks by forced-alignment based on the written transcript. For this step, the 'Munich Automatic Unit Segmentation' tool (MAUS) has been used, for which an implementation for the Luxembourgish language is available [11]. For the training itself, we are building upon XLS-R, the large-scale model for cross-lingual speech representation learning based on wav2vec 2.0, provided by Meta [12]. For this fine-tuning with pytorch, the wav2vec2-xls-r-300m checkpoint with its 300 million parameters has been chosen. Early stopping has been applied to avoid overfitting. While Luxembourgish is not part of the base model, the fine-tuning will benefit from structurally close languages being included in xls-r, e.g., German, Dutch, and English.

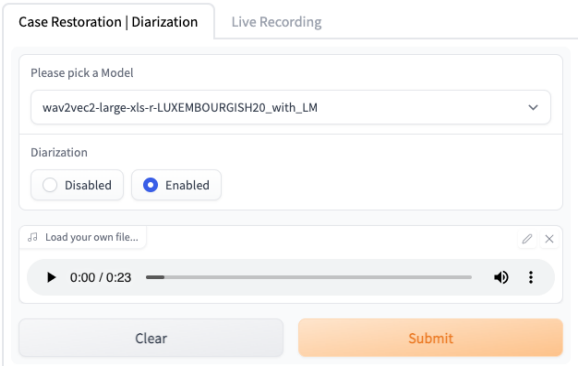
The specific genre of parliamentary speeches is highly multilingual and contains numerous French and German words and phrases. Our training data does also contain sections in these languages, and the ASR model then allows us to decode these French and German passages as well. In contrast to the system proposed in [6], where separate ASR models are developed for the different languages and then combined into one, the advantage of the current approach is to train only one multilingual model.

Besides the acoustic wav2vec2 2.0 model, a language model has been compiled to improve the Word Error Rate (WER). The language model comprises of texts from various sources (parliamentary debates and speeches, radio news articles and commentaries, Wikipedia etc.) and consists of approximately 130 million word tokens. Using KenLM [13] the language model is made up of n-grams with up to five words. By drawing upon pyctcdecode module in the transformer framework [14], the language model was attached to the acoustic model. After evaluation with the test set, the following promising word error rates were obtained: WER 19.12, CER: 6.65 (without language model), WER 14.47, CER: 5.93 (with language model). Recently, a further ASR model for Luxembourgish has been developed by applying a quite similar approach, obtaining a slightly better WER [8].

[2] https://github.com/mbarnig/Marylux-648-TTS-Corpus

5. DEMONSTRATION

During the demonstration, we will showcase several aspects of our ASR system using unseen data. Fig. 1 gives an overview of the interface. As additional features, the capitalization of certain words (Luxembourgish follows the same system of capitalization of nouns as German), the restoration of punctuation, and the diarization of speakers have been implemented.



(a)

(b)

Fig. 1: Demo page of the Lux-ASR tool, showing the interface (a) and the result window with the recognized text (b) in plain format (top) and with case and punctuation restored (bottom).

6. CONCLUSION

The presented Lux-ASR system will be further optimized in the near future by enlarging both the training data and the language model.

7. REFERENCES

[1] Peter Gilles, "Luxembourgish," in *Oxford Encyclopedia of Germanic Linguistics*, Sebastian Kürschner and Antje Dammel, Eds. Oxford University Press, (in press).

[2] Peter Gilles and Jürgen Trouvain, "Illustrations of the ipa: Luxembourgish," *Journal of the International Phonetic Association*, vol. 43, no. 1, pp. 67–74, 2013.

[3] Martine Adda-Decker, Thomas Pellegrini, Eric Bilinski, and Gilles Adda, "Developments of "lëtzebuergesch" resources for automatic speech processing and linguistic studies," in *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, 2008.

[4] Martine Adda-Decker, Lori Lamel, and Natalie D Snoeren, "Initializing acoustic phone models of under-resourced languages: A case-study of luxembourgish," in *Spoken Languages Technologies for Under-Resourced Languages*, 2010.

[5] Martine Adda-Decker, Lori Lamel, and Natalie D Snoeren, "Studying luxembourgish phonetics via multilingual forced alignments.," in *ICPhS*, 2011, vol. 11, pp. 196–199.

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

[7] 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.

[8] Le Minh Nguyen, "Improving luxembourgish speech recognition with cross-lingual speech representations," M.S. thesis, Master thesis, Voice Technology (VT), University of Groningen, 2022.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[10] Nathalie Entringer, Peter Gilles, Sara Martin, and Christoph Purschke, "Schnëssen. surveying language dynamics in luxembourgish with a mobile research app," *Linguistics Vanguard*, vol. 7, no. s1, 2021.

[11] Thomas Kisler, Uwe Reichel, and Florian Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[12] 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.

[13] Kenneth Heafield, "Kenlm: Faster and smaller language model queries," in *Proceedings of the sixth workshop on statistical machine translation*, 2011, pp. 187–197.

[14] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, 2020, pp. 38–45.