ASRLUX: AUTOMATIC SPEECH RECOGNITION FOR THE LOW-RESOURCE LANGUAGE LUXEMBOURGISH

Peter Gilles, Léopold Hillah, Nina Hosseini-Kivanani

University of Luxembourg
peter.gilles@uni.lu, leopold.hillah@sproochtek.lu, nina.hosseinikivanani@uni.lu

ABSTRACT

We have developed an automatic speech recognition (ASR) system tailored to Luxembourgish, a low-resource language that poses distinct challenges for conventional ASR approaches due to the limited availability of training data and inherent multilingual nature. By employing transfer learning, we meticulously fine-tuned an array of models derived from pre-trained wav2vec 2.0 and Whisper checkpoints. These models have been trained on an extensive corpus of various languages and several hundred thousand hours of audio data, utilizing unsupervised and weak supervised methodologies, respectively. This includes linguistically related languages such as German, Dutch, and French, which expedite the cross-lingual training process for Luxembourgish-specific models.

Fine-tuning was executed utilizing 67 hours of annotated Luxembourgish speech data sourced from a diverse range of speakers. The optimal word error rate (WER) achieved for wav2vec 2.0 and Whisper models were 9.5 and 12.1, respectively. The remarkably low WERs obtained serve to substantiate the efficacy of transfer learning in the context of ASR for low-resource languages.

Keywords: Automatic speech recognition, wav2vec 2.0, Whisper, Luxembourgish

1. INTRODUCTION

Until recently, developing powerful Automatic Speech Recognition (ASR) systems for small and low-resource languages has been quite a challenging task. These languages often lack sufficient data for training and evaluating ASR systems, which makes it challenging to achieve high accuracy in recognition. It is the purpose of this study to explore and demonstrate how an ASR system for Luxembourgish can be developed, taking into account the specific constraints of a low-resource language, i.e., scarcity of training material, and the specific situation of multilingualism. In the course of the 20th century, the small language Luxembourgish evolved to the national language of the Grand-Duchy of Luxembourg [1], [2]. 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 the digitization of everyday life. This has led to the need for tailored Natural Language Processing (NLP) and voice tools (especially Speech-to-Text (STT) and Text-to-Speech (TTS)). Luxembourgish is typologically quite close to German and has in fact evolved out of a German dialect, which has been standardized recently to a certain degree with regard to spelling and lexicon. In addition to that, Luxembourgish features a high degree of multilingualism: Words and phrases coming mainly from French and German are part of the regular lexicon and occur quite frequently. And it goes without saying that these multilingual insertions pose challenges for traditional approaches to ASR, which were based on monolingual systems. In recent years, several approaches to address multilingual ASR for Luxembourgish have been presented, mainly by re-utilizing training resources from related languages (see [3], [4], [5]). However, recent advances in machine learning (ML) and NLP have made it possible to build more accurate ASR systems for small and low-resource languages. In particular, these ML advances have led to the improvement of techniques (e.g., unsupervised and semi-supervised learning methods) for transferring knowledge from large, well-resourced languages to smaller languages with fewer resources. Based on recent advances in ASR research, the present study will be based on two of the most promising approaches: wav2vec 2.0 [6], developed by Meta, and Whisper [7], developed by OpenAI. For both approaches, large pre-trained model checkpoints are provided, which can be fruitfully used for adapting to a low-resource language like Luxembourgish.

2. RECENT ADVANCES IN ASR

In recent years, there have been significant advancements in the field of ASR due to the use of deep learning techniques. These techniques have proven to be particularly effective for ASR, leading to rapid improvements in the accuracy and performance of ASR systems. Currently, self-supervised transformer-based models for audio processing have significantly improved the performance of ASR systems. These models are able to handle unlabeled data effectively, which is a significant advantage over traditional ASR systems (such as HMM-based approaches like Kaldi [1]) that require a large amount of labeled data and a pronunciation dictionary for training. With the availability of pre-trained multilingual base models, it is possible to make significant progress in developing ASR systems for low-resource languages even when training data is scarce.

We will focus on two of the most effective algorithms, i.e., wav2vec 2.0 and Whisper, which are especially useful for the Luxembourgish case study because they provide pre-trained checkpoints that have been trained on multiple languages. The idea behind wav2vec 2.0 is to learn by predicting the context around a given audio segment. Wav2vec 2.0 uses a technique called "time-depth separable convolution," which involves applying a series of convolutional neural networks (CNNs) to the audio waveform data. CNNs are trained to extract features from the audio data at multiple scales.

For learning the contextual representations of speech data, both CNNs and recurrent neural networks (RNNs), which can be trained on large amounts of unannotated data, are being used. This is beneficial to use this model in low-resource languages where annotated data is often scarce. The Whisper [7] approach uses a transformer-based sequence-to-sequence (seq2seq) architecture for learning the contextual representation of speech data. This approach is robust to noise and other distortions, which is important for the data when the quality of the audio files is poor. In addition to the different architectures used in the two models, wav2vec 2.0 is designed to be trained on large amounts of unannotated audio files, while Whisper can be used on both annotated and unannotated data ('weakly supervised', see [8]). Both approaches have the potential to be used for low-resource languages such as Luxembourgish and achieve good performance with relatively little annotated data. Recently, a wav2vec 2.0 model for Luxembourgish has been presented by [9] [2], with results similar to the present study.

3. MATERIAL & METHOD

At the core of the ASRLux system under development at the University of Luxembourg is the training data, which consists of pairs of audio samples with their matching high-quality orthographic transcriptions. The 67 hours of training material come from several hundred speakers, thus representing a wide range of speaker characteristics and accents. Manual checks have been conducted to ensure that the audio and text match as precisely as possible. Table 1 provides information about the sources used.

	Datasets	Hours
1.	parliamentary speeches	32
2.	news commentaries	24
3.	crowd-sourced sentences [10]	9
4.	MaryLux sentences [11]	1
	Total duration	***67***

Table 1: Composition of the training data.

The data has been divided into smaller chunks of audio files with a duration ranging from 1 second to 20 seconds, with an average duration of 6 seconds (39724 chunks). Sources 2 and 3 are already available in the correct format, but larger audio files have been split into smaller chunks using a tool called MAUS (Munich Automatic Unit Segmentation), for which a reliable implementation for the Luxembourgish language is available [12]. The smaller chunks were created based on written transcripts using forced-alignment. Depending on the type of model, text normalization was applied to numbers and special characters like "%, &, +", which have been replaced by the corresponding words. Especially the parliamentary speeches and the news commentaries contain a lot of French and German words and phrases, which are then consequently included in the training.

As for the training itself, we have chosen recipes developed by Hugging Face, whose transformers framework ([13]) offers state-of-the-art scripts for the fine-tuning of wav2vec 2.0[3] and Whisper.[4] The dataset was split up into a training set of 75% and a test set of 15% and early stopping has been applied to circumvent overfitting. The fine-tuning of these highly parameterized models is still quite demanding regarding computing power. While the smaller wav2vec 2.0 checkpoint could be trained on four V100 GPUs on the University's HPC in approximately 30 hours, the large wav2vec 2.0 and

the Whisper-large checkpoints could be trained only on A4 GPUs with 46 GB of VRAM in a reasonable time.

To further improve the accuracy of the wav2vec 2.0 model, a language model has been compiled. The input data for the language model consists of text data from various sources (parliamentary debates and speeches, radio news articles and commentaries, Wikipedia, etc.) and sums up to approximately 130 million word tokens. By applying KenLM [14], the language model is constructed of n-grams with up to five words. The pyctcdecode library, a beam search decoder for CTC speech recognition, is used to attach the language model to the acoustic model. The weights of the n-grams then are used in the decoding pipeline to select certain words in cases when the probabilities of the acoustic recognition are not decisive. Using a language model improves the error rate by 2 to 5%. Note that Whisper does not currently support the attachment of a language model.

The quality of ASR systems is usually evaluated using the word error rate (WER), which measures the number of errors in the transcription, including substitutions (S), deletions (D), and insertions (I). The WER is calculated by summing up the three types of errors and dividing by the number of words (N) in the reference transcription. A lower WER indicates a higher-quality ASR system.

- $\text{WER} = \frac{\text{S+D+I}}{\text{N}} * 100$

WER can be used to compare the performance of different models and identify potential areas for improvement. In this project, we use n-gram stochastic language models with n equal to 5, considering phones as units.

4. RESULTS

We will report the results for four fine-tuned models here. Two models are based on the wav2vec 2.0 XLS-R checkpoints with 300 million and 1 billion parameters, respectively. For Whisper, we first experimented with the smaller Whisper checkpoints (small and medium), but after obtaining unsatisfactory results, we used the checkpoint Whisper-large with 1.5 billion parameters. Table 2 lists the WERs obtained for our four models and the original Whisper model. Based on the test dataset, the WERs all lie below 20 and range among rates for state-of-the-art systems for big languages. The 1 billion model for wav2vec 2.0 proves to be clearly superior to the 300 million parameter version. The additional use of a language model lowered the WER by a few points more, with a WER of 9.5 being the lowest for all models. The rates for Whisper are similar for the uncased version. For the cased version, the WER is considerably higher, which is most likely due to punctuation and capitalization issues.

Model	WER w/o language model	WER w/ language model
wav2vec 2.0 (300m)	19.1	14.5
wav2vec 2.0 (1b)	12	**9.5**
Whisper-large uncased	12.1	n.a.
Whisper-large-v2 cased	18.6	n.a.

Table 2: WERs for Luxembourgish ASR models trained with and without a language model.

To demonstrate the quality of the models, two longer examples are presented in table 3 along with their ground truth transcription. With incorrect recognized words in bold, it is obvious that both models make only a few errors, which are mostly related to rare words (e.g. *wellechen* 'some', *Ustiechlechkeet* 'contagiousness'), unstressed function words (e.g., confusion of *just* with *net*) or word endings (e.g., *nimmt* for *nimm* 'names', *bann* for *band* 'volume').

Contrary to wav2vec 2.0, the Whisper model works relatively well with multilingual speech. While the recognition of isolated French, German, or English sometimes leads to errors, longer passages in these languages (phrases, sentences) are rendered more or less correctly. The reason for this lies in Whisper's cross-lingual pre-training, which uses audio and text data and is able to manage multilingual input much better than wav2vec 2.0 Although the latter is trained on numerous languages, its multilingual capacities are linked only to the acoustic and not to the textual representation. The quality of the output thus depends on the training material used in fine-tuning. As a final example, an extract of a speech by the Grand-Duc of Luxembourg shows how well a longer stretch of French (in bold), inserted in his Luxembourgish speech, is recognized well by our Whisper model.

> *... Mir sollen houfreg sinn op d'Diversitéit an den Zesummenhalt an eise Gesellschaft.* ***A cet endroit, je voudrais remercier les non luxembourgeois qui resident ou qui travaillent à notre pays pour leur contribution précieuse à notre société. cohésion économique, mais aussi la cohésion sociale de notre pays sont des atouts qui nous appartiennent de défendre à tout prix. Ils sont au coeur de notre projet et de notre réussite. C' est notre bien commun à tous.*** *Haut, op dësem chrëschtel Wënd, wëll ech meng Unerkennung awer net nëmmen op de politesche Plang begrenzen. ...*

Wav2vec 2.0, on the other hand, is running into serious difficulties with this sample and is producing numerous errors.

One of the further advantages of Whisper's

Table 3: Examples for recognition: Ground truth and the results of wav2vec 2.0 and Whisper; recognition errors in bold face.

Ground truth	wav2vec 2.0 (1B)	Whisper-large-v2
Villmools merci, Här President. Den Avis vun den Experten huet kloer gewisen, datt d'Covid-Kris nach net eriwwer ass, dass de Risk nach ëmmer do ass an datt d'Expektative fir September, zumindest wat d'Experten ugeet, déi sinn, datt mer eventuell virun enger neier Well kënne stoen. Op wellechem Datum, wellech Variant, mat wellecher Virulenz, welleger Ustiechlegkeet, dat wësse mer an dësem Moment selbstverständlech net. Déi Zuelen, déi mer haut kennen, soen dat selwecht. De Staatsminister huet gëschter zitéiert: 1200 Infektiounen den Dag. Dat ass eppes, muss ech soen, wat eis virun enger Rei Joer, virun enger Rei Méint jo weesentlech méi erschreckt huet, wéi et eis haut erschreckt, well d'Situatioun	villmools merci här president den avis vun den experten huet kloer gewisen datt d'covidkris nach net eriwwer ass dass de risk nach ëmmer do ass an datt d'expektative fir september zumindest wat d'experten ugeet déi sinn datt mer eventuell virun enger neier well kënne stoen op wellechem datum **well ech** variant mat wellecher virulenz wellecher **ustieche keet** dat wësse mer an dësem moment selbstverständlech net déi zuelen déi mer haut kennen soen dat selwecht de staatsminister huet gëschter zitéiert den **auszweehonnert** infektiounen den dag dat ass eppes muss ech soen wat eis virun enger rei **t eis** haut erschreckt **war** d'situatioun	Villmools merci, Här President. Den Avis vun den Experten huet kloer gewisen, datt d'Covidkris nach net eriwwer ass, dass de Risk nach ëmmer do ass an datt d'Expektative fir September zumindest wat d'Experten ugeet, déi sinn, datt mer eventuell virun enger neier Well kënne stoen. Op **wellegem** Datum, **welleg** Variant, mat **welleger** Virulenz, welleger Ustiech**tegkeet**, dat wësse mer an dësem Moment selbstverständlech net. Déi Zuelen, déi mer haut kennen, soen dat selwecht. De Staatsminister huet gëschter zitéiert: 1200 Infektiounen den Dag. Dat ass eppes, muss ech soen, wat eis virun enger Rei Joer, virun enger Rei Méint jo weesentlech méi erschreckt huet, wéi et eis haut erschreckt, well d'Situatioun
Se si gesond, gesi schéin aus, schmaache gutt a schéin Nimm hunn se och nach! Ech schwätze vun den Uebst- a Geméiszorten, vun de Kraider, de Gewierzer an den Nëss. An dësem Buch huet den Zenter fir d'Lëtzebuerger Sprooch ronn 300 Nimm fir déi geleefegst Produite vun der Natur gesammelt. Dat ass natierlech just eng Selektioun, mee déi meescht Zorten, déi kann een iessen an déi et hei am Gaart oder am Buttek gëtt, fënnt een an dësem drëtte Band vun der beléifter Serie Lëtzebuerger Wuertschatz.	se si gesond gesi schéin a mech maachen a se nach ech schwätze vun den **uebstzorten deezer** an den nëss an dësem buch huet den zenter fir d'lëtzebuerger sprooch ronn dräihonnert **nimmt** fir déi **geleeft** produite vun der natur gesammelt dat ass natierlech just eng selektioun mee déi meescht zorten déi kann een iessen an déi **déi** et hei am **gaard uerder** am buttek gëtt fënnt een an dësem drëtte **bann** vun der beléifter **seng** lëtzebuerger **wirtschaft**	se si gesond, gesi schéin aus, schmaache gutt a schéin Nimm hu se och nach ech schwätze vun den Ueb**es** a Geméiszorte vun de Kr**ä**ider, de Gewierzer an den Nëss**en** an dësem Buch huet de **Centre** fir d'Lëtzebuerger Sprooch ronn dräihonnert Nimm fir déi geleeftegst Produite vun der Natur gesammelt dat ass natierlech **net** eng Selektioun **wéi** déi meescht Zorten, déi kann een iessen an déi **déi** et hei am Gaart oder am Buttek gëtt, fënnt een an dësem drëtte Band vun der beléifter Serie Lëtzebuerger Wuertschatz.

pretraining on text can be seen in the integration of capitalization of nouns (Luxembourgish is here following the spelling rules of German) and of punctuation. In the Whisper examples above, the capitalization is achieved quite well, whereas the placement of period and comma is partially not yet correct. Wav2vec 2.0, on the other hand, capitalization, and punctuation have to be restored in an additional step, e.g. from the representation of a language model.

Finally, for audio recordings involving several speakers, the diarization method from pyannote.audio has been implemented [15].

5. OUTLOOK

From our evaluation and testing with various audio samples, we found that both systems produced high-quality results. The WER for all models lays clearly below 20, indicating a highly promising perspective for the ASR of Luxembourgish. Not only because of the restoration of capitalization and punctuation but mainly due to its generally better word recognition, it could be shown that Whisper and it is 'weakly supervised pretraining' outperform wav2vec 2.0.

In the next steps, it is foreseen to improve the quality and to enlarge the amount of training data, particularly by including more informal and conversational genres. Methods of audio data augmentation will be applied to make the system more robust for different levels of audio quality [16].

Regarding the applications of LuxASR, it is planned to develop an assistant system for transcribing the speeches and debates of the Luxembourgish parliament. LuxASR will also be used to extract text data from audio archives to assist in research projects in the Humanities and to develop tools for pronunciation assessment.

LuxASR can be tested for inference online on Hugging Face.[5]

6. REFERENCES

[1] P. Gilles, "Luxembourgish," in *Oxford Encyclopedia of Germanic Linguistics*, S. Kürschner and A. Dammel, Eds. Oxford: Oxford University Press, 2023.

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

[3] M. Adda-Decker, L. Lamel, and N. D. Snoeren, "Studying luxembourgish phonetics via multilingual forced alignments." in *ICPhS*, vol. 11, 2011, pp. 196–199.

[4] M. Adda-Decker, L. Lamel, and G. Adda,

"Speech alignment and recognition experiments for luxembourgish," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[5] K. Veselỳ, C. Segura, I. Szöke, J. Luque, and J. Cernockỳ, "Lightly supervised vs. semi-supervised training of acoustic model on luxembourgish for low-resource automatic speech recognition." in *INTERSPEECH*, 2018, pp. 2883–2887.

[6] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[8] J. C. Vásquez-Correa and A. Álvarez, "Novel Speech Recognition Systems Applied to Forensics within Child Exploitation: Wav2vec2.0 vs. Whisper," ENGINEERING, Preprint, Dec. 2022.

[9] L. M. Nguyen, "Improving luxembourgish speech recognition with cross-lingual speech representations," Master's thesis, Master thesis, Voice Technology (VT), University of Groningen, 2022.

[10] N. Entringer, P. Gilles, S. Martin, and C. Purschke, "Schnëssen. surveying language dynamics in luxembourgish with a mobile research app," *Linguistics Vanguard*, vol. 7, no. s1, 2021.

[11] M. Barnig, "Marylux-648-TTS-Corpus," 2022.

[12] T. Kisler, U. Reichel, and F. Schiel, "Multilingual processing of speech via web services," *Computer Speech & Language*, vol. 45, pp. 326–347, 2017.

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. 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.

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

[15] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "pyannote.audio: neural building blocks for speaker diarization," in *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2020.

[16] I. Jordal, A. Tamazian, E. T. Chourdakis, C. Angonin, Askskro, N. Karpov, T. Dhyani, O. Sarioglu, Kvilouras, E. B. Çoban, F. Mirus, Jeong-Yoon Lee, K. Choi, MarvinLvn, SolomidHero, and T. Alumäe, "Audiomentations," Zenodo, Aug. 2022.

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

[2] https://schreifmaschinn.lu

[3] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

[4] https://huggingface.co/blog/fine-tune-whisper

[5] https://huggingface.co/unilux