BILINGUAL ASR MODEL WITH LANGUAGE IDENTIFICATION FOR BRAZILIAN PORTUGUESE AND SOUTH-AMERICAN SPANISH

Felipe Farias,Wilmer Lobato, William Castañeda, Marcellus Amadeus

Alana AI Research, Brazil
{felipe.farias,wilmer.lobato,william.cruz,marcellus}@alana.ai

ABSTRACT

This paper documents the development of a special case of multilingual Automatic Speech Recognition model, specifically tailored to attend two languages spoken by the majority of Latin America, Portuguese and Spanish. The bilingual model combines Language Identification and Speech Recognition developed with the Wav2Vec2.0 architecture and trained on several open and private speech datasets. In this model, the feature encoder is trained jointly for all tasks and different context encoders are trained for each task. The model is evaluated separately on two tasks: language identification and speech recognition. The results indicate that this model achieves good performance on speech recognition and average performance on language identification, training on a low quantity of speech material. The average accuracy of the language identification module on the MLS dataset is 66.75%. The average Word Error Rate in the same scenario is 13.89%, which is better than average 22.58% achieved by the commercial speech recognizer developed by Google.

***Index Terms*—** speech recognition, Automatic Speech Recognition, language identification, wav2vec2, multilingual

1. INTRODUCTION

The widespread use of mobile phones and automatic assistants all over the world is one of the many products of the research on multilingual speech recognition, a subtask of Automatic Speech Recognition (ASR) that has seen large improvements in the last years [1, 2]. Despite the existence of products, this is still considered an active area of research. One of the challenges is to improve the performance of these applications in languages with less speech resources, such as datasets and phonetic dictionaries, that are necessary to train models in a large continous vocabulary [3]. These resources are not equally available in all languages [4,5] and current solutions to speech tasks, such as classification, language identification and ASR require a large amount of data for training.

Portuguese and Spanish are highly popular languages with few speech technologies available [6], so they can be considered in-between high and low-resource languages. They are among the top 10 most spoken languages in the world [7], with huge online presence [8], however they lack the same amount of resources of Mandarin or English. These languages share other similarities, such as belonging to the same linguistic family [9], being originated on the Iberian peninsula, and being the current official languages of most countries in Latin America [10]. The positive effect of the development of speech applications for related languages simultaneously is well documented. It achieves good results in speech synthesis [11, 12].

In this work we investigate the development of a bilingual ASR model for Brazilian Portuguese and Latin-American Spanish combining the Wav2Vec2 architecture for two different activities: LID and ASR trained in a self-supervised manner. The LID model is trained on languages taht are closely related to the target languages and the ASR models are trained separately in a monolingual matter.

The remainder of this research paper is organized as follows. Section 2 presents related work on multilingual ASR. Section 3 describes the proposed bilingual ASR model. Section 4 details the experiments developed to evaluate each part of this model. Section 5 shows the results and Section 6 concludes the work.

2. RELATED WORKS

The latest developments on multilingual ASR has two general directions: the development of multilingual systems, in which one model is trained on a multilingual dataset, and the development of systems that combine multiple monolingual models. In the second case, the use of ASR monolingual models in a multilingual setting usually involves the use of Language Identification (LID) to select the adequate ASR model for each utterance.

Considering multilingual solutions, many studies propose end-to-end transcription models using the Seq2seq architecture in a multilingual dataset to good results. The model in [13] is trained on 10 languages and then ported to another 4 languages using transfer learning, while the study in [14] is trained on 51 languages and achieves improvement over monolingual models.

Considering solutions with multiple monolingual models, the RNN-T architecture is used to train jointly ASR and LID models in many streaming applications. The work developed

doi:10.20944/preprints202210.0480.v1

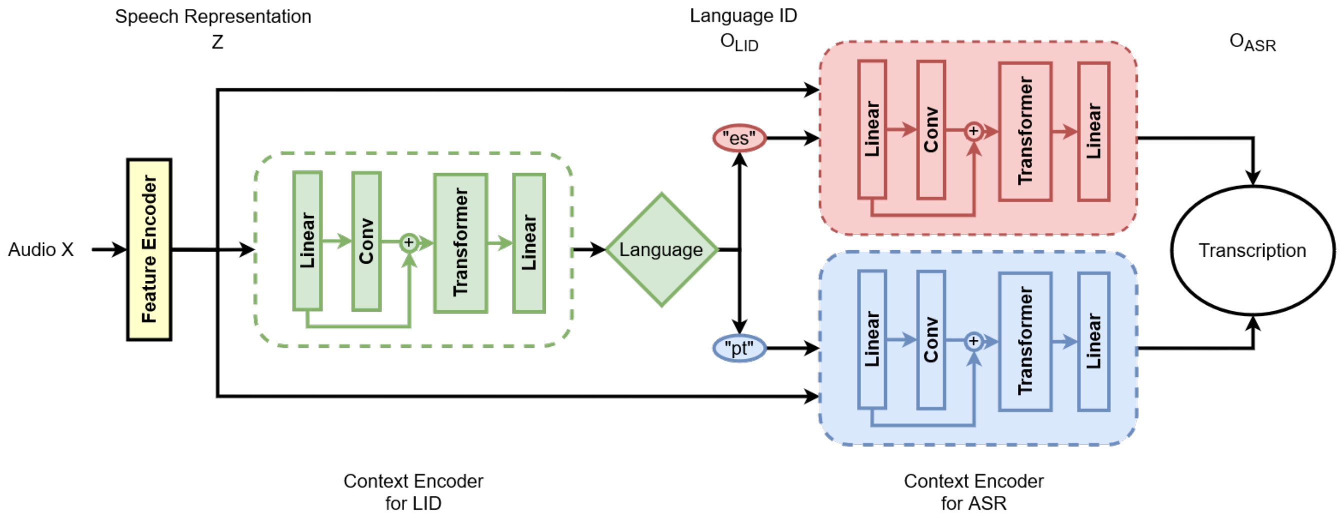

Fig. 1: Schematics of the bilingual ASR model. The audio signal X is fed to a feature encoder, that transforms the audio in speech representations Z. These representations are fed to context encoders. The output of the LID context encoder O_{LID} is the language of the signal and the output of the ASR context encoders O_{ASR} is the transctiption.

in [15] develops an LID model that uses acoustic and text embeddings to choose the correct ASR model in 4 different languages. The work in [16] uses pre-trained LID embeddings to choose between ASR models in English-Spanish and English-Hindi pairs.

Most of the approaches are trained on labeled data, which is difficult to obtain and is limited on all but high-resource languages.

3. THE BILINGUAL ASR MODEL

Figure 1 shows the structure of the bilingual ASR model proposed in this work. The model uses building blocks of the the Wav2Vec2.0 architecture [17], the feature encoder (represented in yellow), followed by context encoders fine-tuned for LID (represented in green), and for monolingual ASR in each of the target languages, Portuguese (represented in blue) and Spanish (represented in red). The following subsections present details on the structure of the Wav2Vec2.0, the training regimes and objectives and the adaptations made to our case.

3.1. Wav2Vec2 Architecture

The architecture of the LID and the ASR models is based on the Wav2vec2.0 [17]. This architecture can be broadly divided in two parts, the feature encoder and the context encoder. The feature encoder maps the audio into a set of speech representations and the context encoder maps the speech representations into context representations, that encode also the context and relative position of the representations. These context representations can be mapped to several downstream tasks in the fine-tuning step. For this, a classifier is added to the context encoder with the task outputs as targets.

3.2. Feature Encoder

As illustrated in Figure 1, the feature encoder maps a chunk of the raw audio X to a dense set of speech representations to be processed by the remainder of the network. This set $Z = z_1, ..., z_T$ represents T time steps. The speech representations are also quantized to a representation $\mathbf{q_t}$ that is considered the target in the self-supervised training.

3.3. Context Encoder

The speech representations created in the feature encoder are fed to a Transformer network that yields context representations $C = c_1, ..., c_T$. These representations do not use absolute position encoding, but relative position, thus being more robust. The context representations are linked to the output of a downstream task by a classification network on the top of the Transformers. The output O of the classifiers depend on the task at hand. For the LID model, the outputs O_{LID} are strings representing the language of the utterance. The output O_{ASR} of the ASR models consist in the letters of the target languages, added by tokens for space and padding.

3.4. Training

3.4.1. Pre-training

The pre-training of the model can be done using only unlabeled data. After training the feature encoder, part of the speech representations is masked. The Transformer network is trained to learn the context representations c_t that correctly identify the quantized speech representation $\mathbf{q}_t$ from a set made of $K + 1$ representations, the correct $\mathbf{q}_t$ plus K distractors uniformly sampled from the masked representations from the same utterance.

doi:10.20944/preprints202210.0480.v1

Dataset	Language	Size (h)	Size (utterances)	Speakers (m/f)
CETUC [18]	Portuguese	99	101k	100 (50/50)
Common Voice [4]	2[1]	54	45k	-
LapsBM [19]	Portuguese	0.9	1k	-
Latin American Spanish Corpora [20]	Spanish	37h	24k	174
MLS [5]	2[2]	1000	2800k	128 (62/66)
Voxforge [21]	Portuguese	4	4k	-
Voxlingua107 [22]	4[3]	196	64k	-
Alana Chatbot	Portuguese	1.3	1k	5 (3/2)

Table 1: Characteristics of the datasets used in the ASR and LID experiments.

Context Encoder	Output
LID	["pt", "es","unk"]
ASR Portuguese	["", "<pad>", "</s>", "<unk>", "|", "A", "E", "O", "S", "R", "I", "N", "D", "M", "T", "U", "C", "L", "P", "V", "G", "F", "H", "Q", "B", "Ã", "Ç", "É", "Á", "Z", "J", "X", "I", "Ó", "Ê", "-", "Õ", "À", "Ú", "Ô", "Â", "Y", "K", "W"]
ASR Spanish	["a", "b", "c", "d", "e", "f", "g", "h", "i", "j", "k", "l", "m", "n", "o", "p", "q", "r", "s", "t", "u", "v", "w", "x", "y", "z", "¡", "á", "é", "í", "ñ", "ó", "ú", "ü", "|", "[UNK]", "[PAD]"]

Table 2: Output of the LID and ASR classifiers.

3.4.2. Fine-tuning

The fine-tuning of the model is the process by which the context representations c_t are mapped to output classes. The number and form of the classes depends on the task for which the model is being tuned. A pre-trained model is fine-tuned by addding a classifier on top of the context network with O classes representing the possible outputs. This classifier is trained using Connectionist Temporal Classification (CTC) loss function [23]. In the case of ASR, the output classes O_{ASR} are the characters that form the output text. These characters may be letters, accents, space and even punctuation. In the case of LID, the output classes O_{LID} are the languages of each utterance.

4. EXPERIMENTAL SETUP

The experiments described on this paper were performed on a cloud instance containing a V100 Tesla GPU with a 16GB RAM, Linux operating system, using Python version 3.7 and the Huggingface[4] training framework.

4.1. Datasets

The characteristics of the speech datasets used in experimental procedures are detailed in Table 1. The training of ASR models use Portuguese and Spanish datasets, such as CETUC [18] for Portuguese, the Latin American Spanish Corpora [20] for Spanish and partitions of multilingual datasets such as Common Voice [4], Voxforge [21] and Multilingual Librispeech [5]. The Alana Chatbot dataset consists in recordings of sentences used by customers in interactions between with chatbots in 11 different scenarios.

For LID the Portuguese and Spanish partitions are joined by partitions in Catalan and Galician of the Voxlingua107 [22] in order to train the model. These languages are selected due to their proximity with the target languages. The datasets used in the pre-training step are discussed on detail on [1] for the Portuguese ASR model and on [2] for the Spanish ASR model and LID model.

The pre-processing of the datasets consists in resampling the audios to 16kHz and mixing it so each audio is monaural. Each audio is labeled with at least the transcription of the utterance and a token indicating the language in which it is spoken.

4.2. LID Setup

In the LID fine-tuning step, the wav2vec pre-trained encoder XLR-S [2] is appended by a pooling layer and a linear layer with the output dimension of $L = 3$. The output of the model is presented in Table 2 and consists in a token representing the two possible languages of the audio, Portuguese and Spanish, and a token in case of the language of the audio is not one of the intended ones. The models are trained using Adam optimizer with learning rate $lr = 3e^{-}3$ and linear decay learning

[1] Only the Portuguese and Spanish partitions of the Common Voice dataset are used in these experiments. The details presented in this table refer to those partitions.

[2] Only the Portuguese and Spanish partitions of the MLS dataset are used in these experiments. The details presented in this table refer to those partitions.

[3] Only the Portuguese, Spanish, Catalan and Galician partitions of the Voxlingua are used in these experiments. The details presented in this table refer to those partitions.

[4] https://huggingface.co/

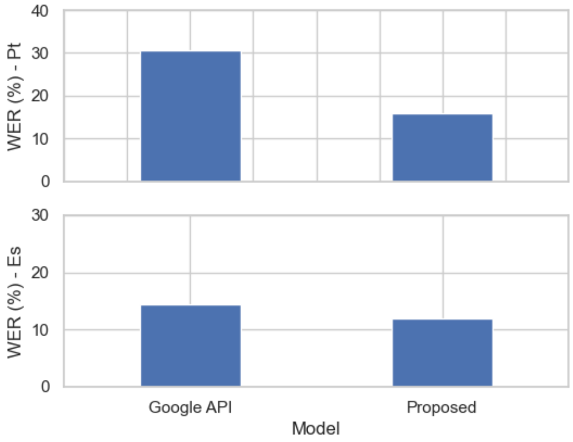

Fig. 2: Automatic Speech Recognition results.

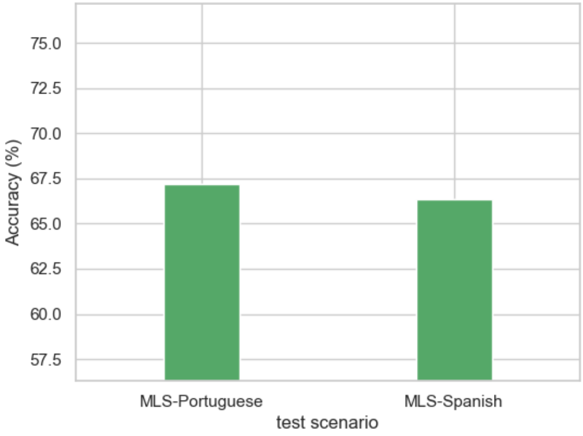

Fig. 3: Accuracy of the LID model.

schedule. The speech signal is truncated for a maximum duration of 1 second during training.

4.3. ASR Setup

In the ASR fine-tuning, the wav2vec pre-trained encoder XLR-S [2] is appended by a pooling layer and a linear layer with L output dimension, being L the vocabulary size for each language. In Portuguese $L = 44$ and in Spanish $L = 37$. Table 2 shows detailed vocabulary, which consists of the letters, space and special tokens. These tokens are used instead of unknown characters, space between letters and padding, which is important for the CTC decoding. The models are trained using Adam optimizer with learning rate $lr = 3e^-3$ and linear decay learning schedule.

4.4. Comparative Model

The evaluation of the proposed in the ASR is compared to the Google Speech transcription service, which can be accessed by the SpeechRecognition python package [24] and is available in Portuguese and Spanish among other languages.

5. RESULTS

5.1. ASR

The evaluation metric for the ASR experiments is the Word Error Rate (WER) and the results are illustrated in Figure 2. Experimental results show that the monolingual ASR models achieve an average of 13.89% in this scenario, with a small advantage for the Spanish partition of the dataset. This result is actually in tune with previous studies in this testing scenario (see Table 9 in [2]), despite the little training dataset. The results in Portuguese are even better than the achieved in the XLR-S paper.

5.2. LID

The test accuracy of the Language ID in the target languages, Portuguese and Spanish, is showns in Figure 3. The average accuracy is 66.75%, considered low in comparison to other LID models that use the pre-trained XLR-S feature encoder [25]. The experimental results show that the MLS dataset provides a challenging scenario for LID model proposed in this work. One possibility is that the training with only 4 different languages is insufficient to provide good separation between two languages. Other possibility is the fact that training the model with languages so historically and linguistically close such as Portuguese, Spanish, Catalan and Galician do not provide the model with enough information to separate the languages properly.

6. CONCLUSION

This paper demonstrates the use of the Wav2Vec2.0 architecture to build a bilingual ASR model in Portuguese and Spanish that combines sequentially a LID model and two monolingual ASR models. We train monolingual ASR models in the target languages and a bilingual LID model to choose the proper ASR model to an utterance from the identified language. Experimental results show that the monolingual ASR models achieve a performance comparable to the state of the art in the widely used MLS dataset, despite training on a small amount of data. The LID model trained with only 4 languages yields below average accuracy in the MLS test dataset, indicating that training with a different set or number of languages is more indicated for LID. This paves the way to different studies that combine speech recognition and audio classification using the same feature encoding, such as multi-dialect speech recognition, emotion recognition with ASR, speaker identification with language identification.

7. ACKNOWLEDGEMENTS

The authors would like to thank Alana AI for the financial and technical support to develop this work.

8. REFERENCES

[1] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *arXiv preprint arXiv:2006.13979*, 2020.

[2] 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.

[3] Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz, "Automatic speech recognition for underresourced languages: A survey," *Speech communication*, vol. 56, pp. 85–100, 2014.

[4] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[5] 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.

[6] Thales Aguiar de Lima and Márjory Da Costa-Abreu, "A survey on automatic speech recognition systems for portuguese language and its variations," *Computer Speech & Language*, vol. 62, pp. 101055, 2020.

[7] M. Paul Lewis, Ed., *Ethnologue: Languages of the World*, SIL International, Dallas, TX, USA, 2022.

[8] Ramesh Pandita, "Internet: A change agent an overview of internet penetration & growth across the world," *International Journal of Information Dissemination and Technology*, vol. 7, no. 2, pp. 83, 2017.

[9] Murray B Emeneau, "India as a lingustic area," *Language*, vol. 32, no. 1, pp. 3–16, 1956.

[10] Francesc Alías, Antonio Bonafonte, and António Teixeira, "Editorial for special issue "iberspeech2018: Speech and language technologies for iberian languages"," 2020.

[11] Pallavi Baljekar, Sai Krishna Rallabandi, and Alan W Black, "An investigation of convolution attention based models for multilingual speech synthesis of indian languages.," in *Interspeech*, 2018, pp. 2474–2478.

[12] Jaka Aris Eko Wibawa, Supheakmungkol Sarin, Chen Fang Li, Knot Pipatsrisawat, Keshan Sodimana, Oddur Kjartansson, Alexander Gutkin, Martin Jansche, and Linne Ha, "Building open javanese and sundanese corpora for multilingual text-to-speech," 2018.

[13] Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori, "Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 521–527.

[14] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *arXiv preprint arXiv:2007.03001*, 2020.

[15] Chander Chandak, Zeynab Raeesy, Ariya Rastrow, Yuzong Liu, Xiangyang Huang, Siyu Wang, Dong Kwon Joo, and Roland Maas, "Streaming language identification using combination of acoustic representations and asr hypotheses," *arXiv preprint arXiv:2006.00703*, 2020.

[16] Surabhi Punjabi, Harish Arsikere, Zeynab Raeesy, Chander Chandak, Nikhil Bhave, Ankish Bansal, Markus Müller, Sergio Murillo, Ariya Rastrow, Sri Garimella, et al., "Streaming end-to-end bilingual asr systems with joint language identification," *arXiv preprint arXiv:2007.03900*, 2020.

[17] 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.

[18] VFS Alencar and Abraham Alcaim, "Lsf and lpc-derived features for large vocabulary distributed continuous speech recognition in brazilian portuguese," in *2008 42nd Asilomar conference on signals, systems and computers*. IEEE, 2008, pp. 1237–1241.

[19] Nelson Neto, Carlos Patrick, Aldebaro Klautau, and Isabel Trancoso, "Free tools and resources for brazilian portuguese speech recognition," *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[20] Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson, "Crowdsourcing latin american spanish for low-resource text-to-speech," 2020.

[21] KEN Mclean, "Voxforge," https://voxforge.org/pt, Accessed: 2022-09-01.

[22] Jörgen Valk and Tanel Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[23] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[24] A. Zhang, "Speech recognition (version 2.1)," 2015.

[25] Andros Tjandra, Diptanu Gon Choudhury, Frank Zhang, Kritika Singh, Alexis Conneau, Alexei Baevski, Assaf Sela, Yatharth Saraf, and Michael Auli, "Improved language identification through cross-lingual self-supervised learning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6877–6881.