Bilingual ASR model with language identification for Brazilian Portuguese and South-American Spanish

Marcellus Amadeus, William Castañeda, Felipe Farias, Wilmer Lobato

***Abstract*— Creating accurate and reliable low-resource automatic speech recognition (ASR) models remains challenging due to limited curated data. This work proposes a bilingual ASR model for Brazilian Portuguese and Latin-American Spanish implemented with the Wav2Vec2.0 architecture and trained on multiple speech datasets. It combines Language Identification and Speech Recognition, employing a joint feature encoder and task-specific context encoders. Evaluation in the Multilingual Librispeech dataset demonstrates promising results, with an average accuracy of 75.98% for language identification and a competitive Word Error Rate of 30.45% in a bilingual setting, comparable to the Whisper model.**

***Keywords*— speech recognition, Automatic Speech Recognition, language identification, Wav2Vec2, multilingual**

I. Introduction

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

Portuguese and Spanish are broad languages with few speech technologies available [6], so they are considered between high and low-resource languages. Nevertheless, they are among the top 10 most spoken languages in the world [7], with a substantial online presence [8], but lack the same amount of resources as Mandarin or English.

Low-resource languages like Portuguese and Spanish share other similarities, such as belonging to the same linguistic family [9], originated on the Iberian peninsula, and are the current official languages of most countries in Latin America [10]. The impact of ASR models with the same quality and granularity as those in high-resources languages is felt by large communities still insufficiently attended. The positive effect of simultaneously developing speech applications for both languages is well documented. For example, it achieves good results in speech synthesis [11], [12]. Therefore, this study can contribute to the knowledge of training models in related languages in ASR.

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

Thus, this work proposes an ASR for Brazilian Portuguese and Latin-American Spanish. The main contribution is to establish a bilingual ASR model combining the Wav2Vec2 architecture for Language Identification (LID) and training in a self-supervised manner. LID model training is on languages closely related to the target languages, and the ASR training implements a monolingual approach.

The remainder of this paper is organized as follows: Section II presents related work on multilingual ASR. Section III describes the proposed bilingual ASR. Section IV details the experiments to evaluate each part of the bilingual ASR. Section V shows the results, and Section VI concludes the work.

II. Related Works

The latest developments in multilingual ASR have two general directions: multilingual systems in which one model is trained on a multilingual dataset and combine multiple monolingual models. [1], [2], [13], and the combined use of multiple monolingual models. In the second case, using ASR monolingual models in a multilingual setting usually involves Language Identification (LID) to select the adequate ASR model for each utterance. Considering multilingual solutions, studies propose end-to-end transcription models using the Seq2seq architecture in a multilingual dataset getting good results. For example, the model in [13] is trained on ten languages and then ported to another four languages using transfer learning. The study of [14] trained on 51 languages achieves improvement over monolingual models. Examinating solutions with multiple monolingual models, the RNN-T architecture trains ASR and LID models together in streaming applications. The work of [15] develops a 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.

Part of development approaches train labeled data, which is difficult to obtain and is limited to all languages that are not rich in acoustic resources. The emergence of techniques that create a good representation of speech in an unsupervised manner [17], [18] enabled research to use more abundant unlabeled data. These techniques perform a series of speech-related tasks, such as language identification [19], speaker recognition [20], emotion recognition [21], and ASR. However, do not

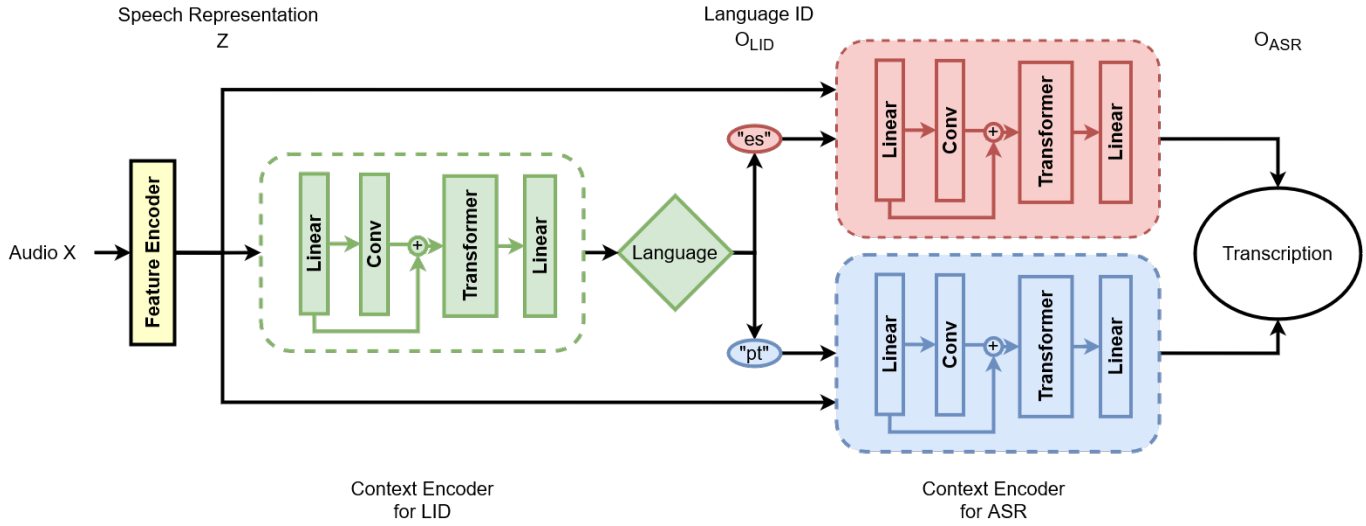


Fig. 1: Schematic drawing of the Bilingual ASR. The audio signal X is the input to a feature encoder that transforms it into speech representations Z. These representations enter into the LID and ASR context encoders. The LID output O_{LID} is the language, and the ASR output O_{ASR} is the transcription.

focus on low-resource languages, such as the ones targeted in this study.

III. Bilingual ASR Model

In this Section, we detail the architecture of our approach. Figure 1 shows the structure of the proposed bilingual ASR model. The model adopts building blocks of the Wav2Vec2.0 architecture [18], the feature encoder (represented in yellow), followed by context encoders fine-tuned for LID (illustrated in green), and monolingual ASR in each of the target languages, Brazilian Portuguese (depicted in blue) and Latin-American Spanish (represented in red). Following subsections present details on the structure of the Wav2Vec2.0, the training regimes and objectives, and the adaptations made to our case.

A. Wav2Vec2 Architecture

LID and ASR models have Wav2vec2.0 architecture as a base [18] divided into two parts: the feature 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 within relative positions. Thus, the context representations serve several downstream tasks in the fine-tuning step. For this, a classifier is part of the context encoder with the task outputs as targets.

B. 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 quantized to $\mathbf{q}_t$ considered the target in the self-supervised training.

C. Context Encoder

The speech representations created in the feature encoder are the input 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 connect to the output of a downstream task by a classification network on the top of the Transformers. Classifier output O depends on the task at hand. LID model outputs O_{LID} are strings representing the language of the utterance. The output O_{ASR} of the ASR models consists of the letters of the target languages, added by tokens for space and padding.

D. Training

1) Pre-training: The pre-training of the model achieves using only unlabeled data. After the training feature encoder part of the speech representation is masked. The Transformer network training allows learning the context representations c_t that identify the quantized speech representation $\mathbf{q}_t$ from a set $K + 1$, the correct $\mathbf{q}_t$ plus K distractors uniformly sampled from the masked representations from the same utterance.

2) Fine-tuning: The fine-tuning of the model is the process by which the context representations c_t map to output classes, and the class number and form depend on the model-tuned task. Thus, fine-tuning a pre-trained model happens when adding a classifier to the context network with O classes representing the possible outputs. The classifier training uses Connectionist Temporal Classification (CTC) loss function [27]. 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.

Dataset	Language	Size (h)	Size (utterances)	Speakers (m/f)
CETUC [22]	Portuguese	99	101k	100 (50/50)
Common Voice [4]	2[1]	54	45k	-
LapsBM [23]	Portuguese	0.9	1k	-
Latin American Spanish Corpora [24]	Spanish	37h	24k	174
MLS [5]	2[2]	1000	2800k	128 (62/66)
Voxforge [25]	Portuguese	4	4k	-
Voxlingua107 [26]	4[3]	196	64k	-
Alana Chatbot	Portuguese	1.3	1k	5 (3/2)

TABLE I: 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", "Í", "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", "ą", "á", "é", "í", "ñ", "ó", "ú", "ü", "|", "[UNK]", "[PAD]"]

TABLE II: Output of the LID and ASR classifiers.

IV. Experimental Setup

Performed experiments are with a cloud instance containing a V100 Tesla GPU with 16GB RAM, Linux operating system, using Python version 3.7 and the Huggingface[4] training framework.

A. Datasets

The characteristics of the speech datasets used in experimental procedures are detailed in Table I. The training dataset for ASR in Brazilian Portuguese consists of 157k utterances taken from the CETUC [22], Common Voice [4], LapsBM [23], Multilingual Librispeech (MLS) [5], Voxforge [25] and Alana Chatbot datasets. The Latin-American Spanish ASR training consists of 20k utterances from the Common Voice [4] dataset and the Argentinian partition from the Latin-American Spanish Corpora [24].

The LID training dataset consists of Portuguese, Spanish, Catalan, and Galician partitions of the Voxlingua107 [26] to train the model. Selected languages are due to their proximity to the target languages. The evaluation dataset consists of 4k utterances, half taken from the Portuguese and half from Spanish partitions of the MLS [5] dataset. The datasets used in the pre-training step are discussed in 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 audio to 16kHz and mixing it so each audio is monaural. As well as labeled audio with at least the transcription of the utterance and a token indicating the spoken language.

B. Metrics

The ASR evaluation uses two metrics: Word Error Rate (WER) and Word Information Lost (WIL). The WER of a sentence or set of sentences is the rate between the number of substitutions (S), insertions (I), and deletions (D) over the total number of words (N). Is calculated as [28]:

$$WER = \frac{D + I + S}{N} \tag{1}$$

WIL measure, based on Mutual Information, is a statistical dependence between two pairs of sentences. For example, considering S as the number of substitutions in a sentence, D as the number of deletions, I as the number of insertions, and H as the number of hits or correct words, the WIL of a sentence or set of sentences is calculated as [28]:

$$WIL = 1 - \frac{H^2}{(H + S + D)(H + S + I)} \tag{2}$$

Accuracy is the models' evaluation metric, in terms of LID, calculated as the rate between the correct number and the total number of classifications in the test.

C. LID Setup

In the LID fine-tuning step, the wav2vec pre-trained encoder XLR-S [2] append a pooling layer and a linear layer with the output dimension of $L = 3$. Table II shows the model's output consisting of tokens representing two possibilities: audio languages, Portuguese or Spanish, and an unknown language if the audio is not one of the intended ones. Models training implements Adam optimizer with learning rate $lr = 3e^{-}3$ and linear decay learning schedule. For a maximum duration of 1 second during training speech signal is truncated.

D. ASR Setup

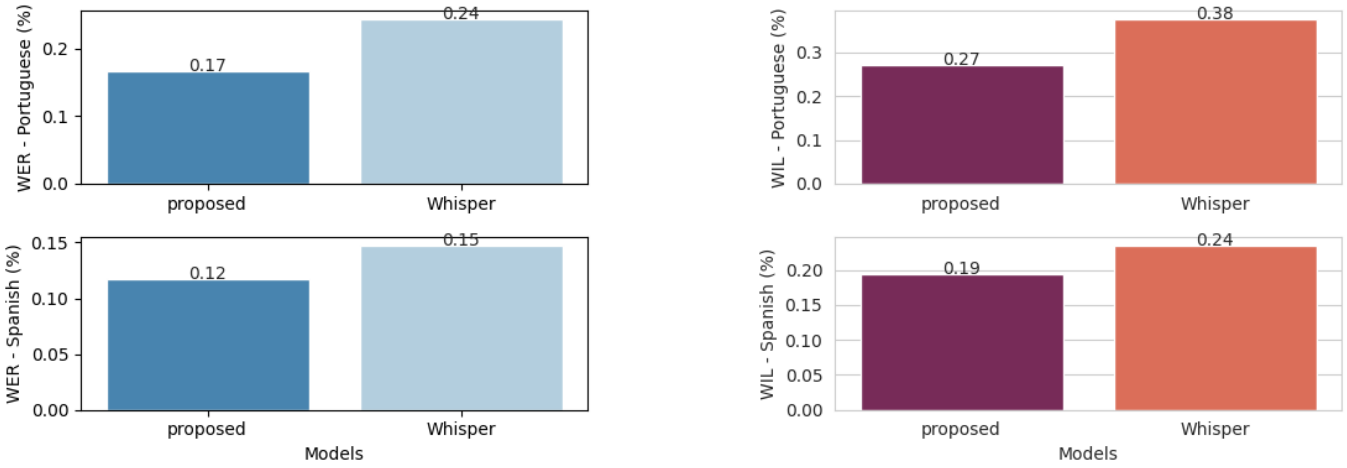
In the ASR fine-tuning, the wav2vec pre-trained encoder XLR-S [2] appended a pooling layer and a linear layer with L output dimension, being L the vocabulary size for each language. In Portuguese, $L = 44$, and in Spanish $L = 37$. Table II

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

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

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

[4]https://huggingface.co/



(a) WER results.

(b) WIL results.

Fig. 2: Performance of the proposed models in the Portuguese and Spanish partitions of the MLS dataset.

Model	Accuracy
Proposed model	75.98%
Whisper	99.82%

TABLE III: Accuracy of the LID model.

shows detailed vocabulary with the letters, space, and special tokens. These tokens are used instead of unknown characters, space between letters, and padding, which is important for CTC decoding. Models training implements Adam optimizer with learning rate $lr = 3 \times 10^3$ and linear decay learning schedule.

E. Comparative Model

Comparing the proposed model with the Whisper model [29], a multilingual ASR model trained in over 680k hours of speech covering 96 languages. Model selection is due to its remarkable performance in ASR and other speech-related tasks.

V. Results

A. Monolingual ASR

The results are illustrated in Figure 2. Experiments show that the proposed model achieves high performance compared to the SOTA Whisper model on the studied languages in WER (14.23% against 19.56%) and WIL (23.26% against 30.53%). Despite the little training dataset, the result corroborates previous studies in this testing scenario (see Table 9 in [2]). Moreover, the results in Portuguese are even better than those achieved in the XLR-S paper.

B. LID

The test accuracy of the Language ID in the target languages, Portuguese and Spanish, is shown in Table III. The average accuracy is 75.98%, which is worse than the performance of the comparative model, Whisper (99.82%). The experimental results show that the MLS dataset provides a challenging scenario for the LID model proposed in this work.

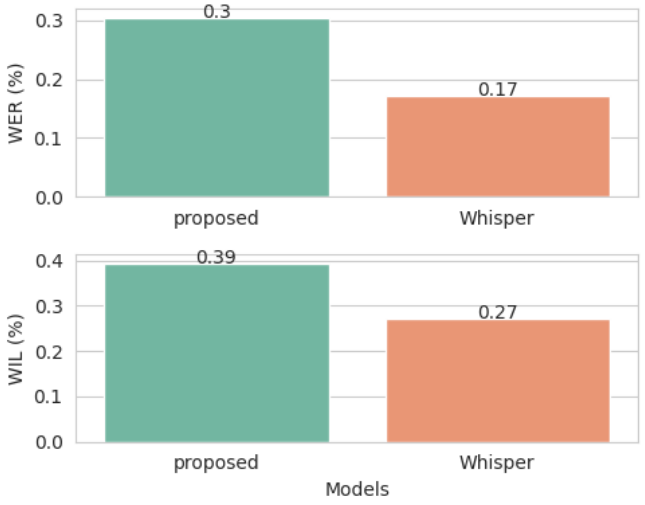


Fig. 3: Performance of the proposed model in a bilingual dataset formed by the Portuguese and Spanish partitions of the MLS dataset.

One possibility is that the training with only four different languages is insufficient to provide good separation between the two languages. An alternative is that training the model with historically and linguistically close languages such as Portuguese, Spanish, Catalan, and Galician does not provide enough information to separate the language correctly.

C. Bilingual ASR

The result of the final model, combining LID and ASR, in a bilingual dataset formed by the test datasets from the Portuguese and Spanish partitions of the MLS dataset, is shown in Figure 3. The results indicate that the final model can be competitive when compared with the current SOTA, and the low accuracy of the LID module restricts the model performance in a bilingual setting.

VI. Conclusion

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

VII. Acknowledgements

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

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 under-resourced 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, Michael Auli, and Abdelrahman Mohamed, "Effectiveness of self-supervised pre-training for speech recognition," *arXiv preprint arXiv:1911.03912*, 2019.

[18] 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.

[19] Zhiyun Fan, Meng Li, Shiyu Zhou, and Bo Xu, "Exploring wav2vec 2.0 on speaker verification and language identification," *arXiv preprint arXiv:2012.06185*, 2020.

[20] Sergey Novoselov, Galina Lavrentyeva, Anastasia Avdeeva, Vladimir Volokhov, and Aleksei Gusev, "Robust speaker recognition with transformers using wav2vec 2.0," *arXiv preprint arXiv:2203.15095*, 2022.

[21] Leonardo Pepino, Pablo Riera, and Luciana Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[22] 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.

[23] 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.

[24] 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," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020, pp. 6504–6513.

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

[26] 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.

[27] 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.

[28] Andrew Cameron Morris, Viktoria Maier, and Phil Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in *Eighth International Conference on Spoken Language Processing*, 2004.

[29] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.