"¿Te vienes? Sure!" Joint Fine-tuning of Language Detection and Transcription Improves Automatic Recognition of Code-Switching Speech

Léopold Hillah
leopold.hillah.001@student.uni.lu
University of Luxembourg
Luxembourg

Mateusz Dubiel
mateusz.dubiel@uni.lu
University of Luxembourg
Luxembourg

Luis A. Leiva
name.surname@uni.lu
University of Luxembourg
Luxembourg

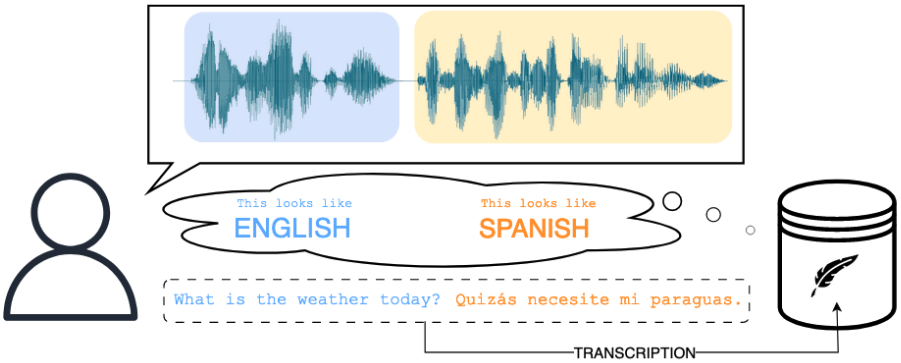


Figure 1: Incorporating language detection end-to-end improves automatic transcription of code-switching speech. This can make Conversational Agents more accurate and efficient in understanding the user's needs.

ABSTRACT

Human communication in multilingual communities often leads to code-switching, where individuals seamlessly alternate between two or more languages in their daily interactions. While this phenomenon has been increasingly prevalent thanks to linguistic globalization, it presents challenges for Automatic Speech Recognition (ASR) systems since they are designed with the assumption of transcribing a single language at a time. In this work, we propose a simple yet unexplored approach to tackle this challenge by fine-tuning the Whisper pre-trained model jointly on language identification (LID) and transcription tasks through the introduction of an auxiliary LID loss term. Our results show significant improvements in transcription errors, ranging between 14 and 36 percentage points of difference. Ultimately, our work opens a new direction for research on code-switching speech, offering an opportunity to enhance current capabilities of conversational agents.

CUI '24, July 8–10, 2024, Luxembourg, Luxembourg

ACM ISBN 979-8-4007-0511-3/24/07
https://doi.org/10.1145/3640794.3665579

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Accessibility technologies**; • **Computing methodologies** → **Speech recognition**.

KEYWORDS

Code Switching; Multilingual Conversations; Language Identification; Automatic Speech Recognition; Whisper; Speech

ACM Reference Format:
Léopold Hillah, Mateusz Dubiel, and Luis A. Leiva. 2024. "¿Te vienes? Sure!" Joint Fine-tuning of Language Detection and Transcription Improves Automatic Recognition of Code-Switching Speech. In *ACM Conversational User Interfaces 2024 (CUI '24), July 8–10, 2024, Luxembourg, Luxembourg*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3640794.3665579

1 INTRODUCTION

The ability to speak more than one language fluently is a fairly frequent phenomenon, even the norm in many countries [3, 20], in large part thanks to the ongoing linguistic globalization [6]. While the exact number of multilingual speakers is difficult to determine, plausible estimations of bilingualism range from 50% to 70% of the global population [20]. Alternating between different languages, typically in spoken form, is referred to as *code-switching*. This phenomenon has been widely studied in sociolinguistics, psycholinguistics, and cognitive science [18].

Research on text-based chatbots has demonstrated feasibility of recognising code-switched natural language, in turn leading to

higher user satisfaction [2, 39]. However, due to poor performance of Automatic Speech Recognition (ASR) technology, robust implementation of code-switching in voice-based interfaces remains impractical [11]. Indeed, ASR for code-switching represents an important and challenging research area, as most of the existing ASR systems, including multilingual models, assume that the audio to transcribe is in a single target language [4, 33, 41, 42]. As a result, such models tend to perform poorly when applied to code-switching scenarios.

In a survey of code-switching from linguistics and social sciences, Doğruöz et al. [16] noted the growing interest in code-switching and pointed out challenges for languages technologies that arise mainly due to the disconnect between the improvements attributed to computational methods and relevant linguistics and social aspects of code-switching. With the advent of Large Language Models (LLMs), there have been a few attempts to improve code-switching performance [26, 50, 53] although they have focused mainly on textual data. For example, Yong et al. [50] proposed generating code-switching synthetic text with LLMs, leading the authors to recommend human supervision to ensure high-quality results. Similarly, Zhang et al. [53] tested LLMs on four different tasks within the context of code-switching texts, concluding that LLMs are "not (yet) code-switchers", and thus calling for more research in this direction to help improve their performance.

In this paper, we propose a simple approach that allows for accurate transcription of code-switching speech, building upon Whisper [42], a transformer-based ASR model. We introduce a new loss term that accounts for predicted language tokens on a specific language, and combine it with the regular transcription loss. We show that ASR performance improves substantially, significantly exceeding between 14 and 36 percentage points in terms of Word Error Rate (WER), with target language identification accuracy surpassing 99% in most cases. Taken together, our paper makes the following contributions:

- We study code-switching speech and the role of language identification in ASR systems.
- We offer a simple method to fine-tune a pre-trained Whisper model on multiple languages simultaneously.
- We provide guidance to develop next-generation ASR systems that can be used to develop more sophisticated conversational interfaces.

The choice of both language identification and transcription tasks during fine-tuning, follows directly from the focus of our work on code-switching ASR in multilingual settings.

2 RELATED WORK

In recent years, code-switching research has gradually gained in popularity thanks to continued improvements of end-to-end and pre-trained ASR models [47]. However, code-switching speech research is still lagging compared to research on its text-related counterpart. In this section, we first present relevant studies on code-switching ASR and then discuss the potential of speech-based Conversational Agents (CAs) with code-switching capabilities to support multilingual users.[1]

[1]In this paper we use 'multilingual users' to refer to individuals who speak at least two languages.

2.1 Progress on Multilingual ASR

Barman et al. [5] investigated the challenge of language identification in code-switching text on social media for Indian languages. They used supervised machine learning methods with character-based *n*-gram features to solve word-level language classification with and without context. The best result was achieved with a Conditional Random Fields (CRF) classifier, yielding 95.76% accuracy. In another study, Ylmaz et al. [49] explored the use of semi-supervised learning for both acoustic and language models to improve the performance of English-isiZulu code-switching ASR. They found that the use of acoustic models based on factored time-delay neural nets (TDNN-F) helped achieve an absolute mixed-language WER reduction of 3.4% and 2.2% on first and second iterations, respectively. However, the proposed language model did not help in improving ASR performance.

More recently, Liu et al. [34] worked on enhancing code-switching ASR with a language alignment loss for frame-level language detection based on pseudo-language labels derived from an ASR decoder and a hybrid CTC/Attention model. They used LLMs and a linguistic hint to guide their prompting and achieved a WER ranging between 14.1% and 5.5% in two Mandarin-English datasets. The main challenges encountered were about balancing training for dominant languages in bilingual data and in generalizing to accented speech, which affected performance on secondary languages.

Dhawan et al. [14] developed a method for generating synthetic code-switching datasets for ASR from monolingual sources. They also proposed a novel concatenated tokenizer, which enables monolingual ASR models to generate a language ID with each emitted token. This was proved highly effective in spoken language identification task, achieving 98% performance on the Google Fleurs monolingual dataset [12]. They achieved a WER of 50% and 53% using aggregated and concatenated tokenizers, respectively, on the Bangor Miami dataset [13], which is one of the most popular benchmarks for research on code-switching speech.

2.2 Potential of Multilingual CAs

Cihan et al. [11] argued that implementation of code-switching recognition in voice-only CAs is crucial for increasing their accessibility for multilingual users. However, previous research has pointed out several challenges that these speakers experience when it comes to *speech production*. For example, Wu et al. [48] noted that, due to limited language coverage, many users need to use their second language (L2) when interacting with CAs, which results in higher cognitive workload. In comparison with monolingual speakers, multilingual speakers possess a smaller receptive vocabulary (i.e., the words that they can understand) in each of their languages [7]. Moreover, multilingual speakers may be disadvantaged when it comes to their first language (L1) production compared to monolingual speakers [27]. Consequently, this leads to higher cognitive effort when producing language due to less frequent retrieval of words from individual's lexicon [11, 22]. Therefore, by improving ASR capabilities for code-switched speech, we can reduce the resulting cognitive workload for multilingual users.

Another issue highlighted by Kann [30] is that current capabilities of CAs are based on the 'one-language-at-a-time' paradigm [43]

with the ASR module committing exclusively to one of two monolingual transcription candidates as early as possible while processing an utterance. Consequently, this lack of flexibility hampers effective code-switching and creates communication obstacles for multilingual users [10, 30].

While enabling CAs with code-switching speech production capabilities is also important for making human-CA interactions more natural and successful, in this paper we will focus on improving ASR transcription as the first essential step to achieve this goal.

3 METHODOLOGY

We used Whisper [42], the latest state-of-the-art end-to-end ASR model. It has proved highly robust across languages, various accents, dialects and speaking styles, and in noisy environments.

Whisper is a transformer-based sequence-to-sequence model from OpenAI, trained on 680,000 hours of labeled audio data collected from the Internet, out of which 117,000 hours were multilingual. Whisper can be fine-tuned on various languages and tasks (e.g., transcription or voice activity detection), and its weights are available in several versions, from 39M parameters (tiny) to 1.5B parameters (large). In this work, we use the Whisper-Large-V2 model.

Whisper's encoder takes as input 80-bins log-Mel filterbanks with a window length of 25 ms and a hop length of 10 ms, generated from 30-second chunks of raw audio. Whisper's decoder generates text tokens as well as special tokens corresponding to different languages and tasks, transcription and translation being the two main tasks included with the pre-trained model [42].

3.1 Datasets

In this work we focus on bilingual English-Spanish speakers, for simplicity. We used one dataset for language identification, two different datasets for model fine-tuning, and two other different datasets for model evaluation.

3.1.1 Language Identification Dataset. We used the Voxlingua107 dataset [46] to build a language identification model for six languages (English, Spanish, French, German, Luxembourgish, and Portuguese) based on the Whisper-Large-V2 pre-trained encoder, to which a classification head was added. The choice of additional languages apart from English and Spanish is to ensure that the model's accuracy is not artificially inflated by the limited choice that a binary classifier would provide. The model yielded an F1 score of 0.99 on the validation split and both an F1 score and an accuracy score of 0.999 on the test split. We used this model to generate the reference labels for the final model evaluation.

3.1.2 Fine-Tuning Datasets. We first fine-tuned Whisper on the Spanish language, using several datasets of Spanish speakers from Latin America, namely Mexico [24], Puerto Rico, Argentina, and Colombia [21]; all are accessible at OpenSLR.org [1]. For the sake of conciseness, we will refer to this combination of Spanish datasets as 'SpaSLR'. Then we fined-tuned Whisper on both Spanish and English using a subset of the MLCommons People's Speech English dataset [17], which will refer to 'MLComm' for short. Table 1 summarizes these datasets.

Table 1: Datasets used for model fine-tuning.

Dataset	Accent	Lines	Running words	Duration
SpaSLR	Argentina	5739	3845	8.03 h
SpaSLR	Colombia	4903	4070	7.58 h
SpaSLR	Mexico	11243	19730	24.49 h
SpaSLR	Puerto Rico	617	1708	1.00 h
MLComm	American English	22500	17616	85.54 h

3.1.3 Evaluation Datasets. We use the Bangor Miami dataset [13], which consists of recordings of daily interactions among four selected Spanish-English bilingual speakers (35.66 h of raw audio, $M = 8.91$ h/speaker). This dataset was initially created for linguistic analysis and recorded in the CHAT annotation format [37]. Because of the spontaneous nature of the recordings, there is usually some background noise, varying volume levels, and a large amount of silence within the audio files. The data therefore needs some additional processing to be suitable for ASR training and evaluation:

(1) Splitting wave files and transcription files based on timestamp information in the CHAT files into segments close to and not exceeding 20 s of duration, following current standards [19, 31].
(2) Applying Voice Activity Detection (VAD) to each audio segment using Silero VAD [45] to remove unneeded silence and merging back those segments with a silence gap of 20 ms or less.
(3) Applying audio normalization to each wave file: first resampling to 16KB and then equalizing gain levels.

The processed dataset comprises 23.14 h of audio ($M = 5.78$ h/speaker).

We also use the Google Fleurs dataset [12] for model evaluation, which is already processed for ASR tasks and represents monolingual out-of-domain data. By using this additional dataset we can investigate the performance of models fine-tuned for code-switching scenarios on monolingual speech, to see if they can generalize, which also facilitate comparisons against previous work.

3.2 Data Preparation

We randomly split the fine-tuning datasets into 80% training and 20% evaluation. The data was further preprocessed by extracting features from the raw audio input data and the reference transcriptions were tokenized with a language-specific option for the shared tokenizer created from the pre-trained Whisper model. This allowed us to set the right language token for each target label. This step is crucial when fine-tuning for multiple languages simultaneously. We further filtered the data on audio length having at most 30 s to stay in line with the original processing window of Whisper, and on label sequence length, with a maximum of 448 tokens.

3.3 Fine-tuning Loss

Whisper allows fine-tuning to specific languages and datasets. However, Whisper has proved not to perform well on language identification in multilingual settings, even after fine-tuning on well-equipped languages [40]. Since this is a critical step, we propose to

add an auxiliary language-specific loss $\mathcal{L}_{\text{LID}}$ and combine it with the original ASR prediction loss $\mathcal{L}_{\text{ASR}}$ of Whisper.

To compute the LID loss, we first extract all tokens available in the vocabulary of each language considered, and compute the categorical cross-entropy over the corresponding logits (predicted tokens). Let C be the number of languages we want to work with, $p_c(x)$ the true distribution of the actual language tokens, and $q_c(x)$ the distribution of the predicted tokens:

$$\mathcal{L}_{\text{LID}} = -\sum_{c=1}^{C} p_c(x) \log(q_c(x)). \tag{1}$$

We then introduce a parameter $\alpha \in [0, 1)$ which helps adjust the importance of the language-specific loss and the transcription loss in the final model loss, so that:

$$\mathcal{L} = \alpha \mathcal{L}_{\text{LID}} + (1 - \alpha)\mathcal{L}_{\text{ASR}}. \tag{2}$$

Based on grid-search experiments, we have observed that $\alpha = 0.2$ is a good trade-off value, since it leads to over 99% language detection accuracy, while keeping good transcription results.

The rationale behind our approach is the fact that in code-switching speech there is always a dominant language [9], often referred to as the "matrix language", and a non-dominant language, referred to as the "embedded language" [38]. Therefore, if the ASR model correctly identifies the matrix language, then it would be easier for it to also detect the embedded language.

3.4 Fine-tuning Procedure

We used the Adam optimizer [32] with momentums $\beta_1 = 0.9$ and $\beta_2 = 0.98$. For all fine-tuning experiments, a cosine scheduler was used to decay the learning rate, starting with an initial warmup phase of 150 steps and a weight decay of 0.001 afterwards. We used batch sizes of 64 utterances for training and 32 for validation, and set gradient accumulation steps to 10.

To speed up computations, we used half-precision floating point arithmetic format (FP16). Also, when using the LID loss, we decreased the training batch size to 48, to accommodate for available GPU memory. The maximum number of training epochs was set to 30, with an early stopping callback implemented with a patience of 3 epochs and Word Error Rate (WER) as the monitoring metric. At the end of this training process, we also compute the Character Error Rate (CER), for completeness.

We also investigated the use of Parameter-Efficient Fine-tuning (PEFT) [23, 25, 52], giving preference to its most recent variant of Weight-Decomposed Low-Rank Adaptation (DoRA) [35] to reduce the model's required trainable parameters. In this case, the validation loss was used as the monitoring metric. A training batch size of 80 was used while training with PEFT.

4 RESULTS

We conducted ablation experiments to assess the contribution of our proposed modified loss and the fine-tuning approach. The results shown in Table 2 confirmed that the contribution of the LID loss term is substantial, improving both language detection accuracy and error rates; cf. model IDs M3-4 and M7-8. Moreover, the results from fine-tuning on both English and Spanish simultaneously indicated that, without the LID loss, the model fails at language identification and transcription. Further, Figure 2 shows a negative correlation between language identification accuracy and transcription errors. Based on these results, we can conclude that the LID loss is key for successfully fine-tuning Whisper on multiple languages simultaneously.

In zero-shot settings, both language identification and audio transcription for high-resource languages are somewhat competitive with the pre-trained Whisper model, without fine-tuning. After fine-tuning, without the LID loss ASR performance becomes worse due to the well-known "catastrophic forgetting" phenomenon [36, 44], limiting the model's ability to be extended to new speakers or dialects, for example. To address this issue, our proposed approach has yielded a strong performance in language identification, reaching almost perfect accuracy and, as a result, ensuring that transcription performance does not degrade with fine-tuning. In fact, on the Miami corpus, we have noticed a reduction in WER of 14 points when models were fine-tuned on Spanish and a reduction of 25 points when they were fine-tuned for both English and Spanish. Similarly, the reductions in WER ranged between 21 and 36 points on the monolingual Google Fleurs dataset.

5 DISCUSSION

To improve performance for code-switching ASR systems we should leverage the potential of the LID loss term, to determine the language in which utterances are spoken in order to ensure a better transcription. As indicated by our experiments results, the inclusion of this loss term is essential.

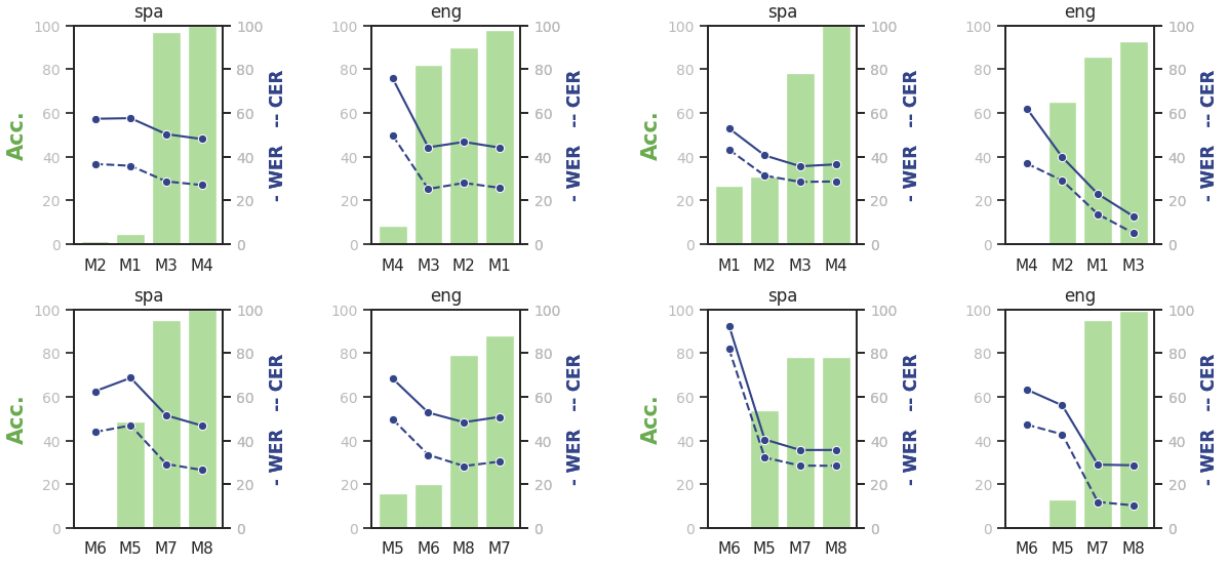
Model 6 is of particular interest, in that it is the perfect counter-example of our approach. In fact, the model was fine-tuned on both English and Spanish using a PEFT Dora approach without any special consideration for languages. As a result, the model could transcribe the audio but was unsuccessful at identifying the languages, on both evaluation datasets and on both English and Spanish. On the other hand, Model 4 was fine-tuned only on Spanish data for both language identification and transcription using a PEFT Dora approach. As a result, the model scored 100% on Spanish LID and very poorly on English LID. The poor performance of both M4 and M6 models in the absence of the adequate LID fine-tuning, was probably amplified by the PEFT approach, which has been recently shown to "learn less and forget less" than full fine-tuning [8].

Our work has potential to improve CA code-switching interactions in several ways. Firstly, it can reduce cognitive load of users by promoting flexibility and enabling them to use vocabulary from both L1 and L2 — effectively making their lexicon retrieval easier. Secondly, as highlighted by Kann [30], it has potential to make structural and spontaneous language learning more accessible by empowering users and preventing communication breakdowns. Thirdly, in the context of task-based interactions, it has potential to improve users' performance by reducing the number of speech recognition errors and minimizing the likelihood of communication breakdowns — consequently leading to improved service satisfaction.

Our work is also of particular relevance to low-resource languages, where communities tend to code-switch more often between low-resource and high-resource languages [15, 51]. Fine-tuning for such low-resource languages, including new languages

Table 2: Results on code-switching (Bangor Miami) and monolingual (Google Fleurs) datasets. The best results for each dominant language and dataset is highlighted in bold. The ID column indicates each of the studied model variants (i.e., with/without LID loss and with/without PEFT).

Model Details				Dominant Language	Bangor Miami			Google Fleurs		
ID	Training data	$\mathcal{L}_{\text{LID}}$	PEFT		Acc.	WER	CER	Acc.	WER	CER
M1	SpaSLR	✗	✗	eng	**98.30**	**44.17**	25.75	86.0	22.91	13.68
M2	SpaSLR	✗	✓	eng	89.90	46.77	27.97	65.20	39.76	29.05
M3	SpaSLR	✓	✗	eng	82.20	44.27	**25.22**	**93.10**	**12.67**	**5.01**
M4	SpaSLR	✓	✓	eng	8.60	75.81	49.68	0.50	62.15	36.90
M1	SpaSLR	✗	✗	spa	4.59	57.64	35.84	26.80	52.95	43.23
M2	SpaSLR	✗	✓	spa	1.33	57.37	36.67	31.00	40.60	31.34
M3	SpaSLR	✓	✗	spa	97.04	50.33	28.68	78.20	**35.61**	**28.45**
M4	SpaSLR	✓	✓	spa	**100.00**	**48.08**	**27.02**	**100.00**	36.49	28.65
M5	SpaSLR + MLComm	✗	✗	eng	15.90	68.32	49.71	13.00	56.08	42.82
M6	SpaSLR + MLComm	✗	✓	eng	20.10	52.78	33.45	0.10	63.34	47.30
M7	SpaSLR + MLComm	✓	✗	eng	**88.10**	50.81	30.35	95.40	28.94	11.87
M8	SpaSLR + MLComm	✓	✓	eng	79.40	**48.38**	**28.30**	**99.50**	**28.67**	**10.27**
M5	SpaSLR + MLComm	✗	✗	spa	48.67	68.69	46.90	53.90	40.50	32.24
M6	SpaSLR + MLComm	✗	✓	spa	0.00	62.58	43.90	0.00	92.58	81.88
M7	SpaSLR + MLComm	✓	✗	spa	95.12	51.52	29.24	**78.20**	**35.64**	28.49
M8	SpaSLR + MLComm	✓	✓	spa	**99.85**	**46.83**	**26.51**	**78.20**	35.68	**28.44**



(a) Bangor Miami dataset. **(b) Google Fleurs dataset.**

Figure 2: Comparison of language identification accuracy (leftmost Y axis) and transcription errors (WER and CER, rightmost Y axis) on code-switching (Bangor Miami) and monolingual (Google Fleurs) datasets. The horizontal axis depicts the Model IDs indicated in Table 2, sorted by language detection accuracy. As can be observed, higher detection accuracy leads to lower transcription errors.

that were not considered by pre-trained ASR models, may improve with the help of the LID loss. To support new languages, once a new language token is added to the vocabulary, or an existing one is reused as a language token, the LID loss will help ensure that these languages are properly detected by the ASR model, thus improving the quality of transcriptions. Ultimately, this will help bridge the technology gap in terms of speech applications for low-resource languages.

5.1 Limitations and Future Work

We should note that the transcriptions in the Miami corpus are not always verbatim, as they were not built for ASR research in the first place but rather for linguistic analysis. Moreover, the names of speakers in the audio files were systematically replaced with other names in the reference transcriptions, and thus they do not correspond to their actual speech. While we did not consider speaker identification in our research, these discrepancies might have slightly affected negatively our transcription results.

We have focused on code-switching scenarios involving two languages. In the future, we plan to apply our approach to multilingual communities that use several languages on a regular basis, as is the case for example in Luxembourg, where natives speak the three official languages of the country (Luxembourgish, French, German) plus English. We also plan to explore alternative model architectures to improve code-switching language segmentation and transcription through a Mixture of Experts architecture (MoE) [28, 29]. A MoE consists of two main elements: sparse MoE layers and a gate network or router. The router could use language identification to send tokens to a language-specific Expert, which would generate a more accurate transcription.

6 CONCLUSION

We have proposed a simple yet powerful approach to fine-tune the pre-trained ASR Whisper model on multiple languages simultaneously, by combining the original ASR loss with an auxiliary LID loss for language identification. Our results suggest that this approach leads to substantial improvements on speech transcription metrics, on both the code-switching Bangor Miami dataset and the out-of-domain monolingual Google Fleurs dataset. To the best of our knowledge, this is the first attempt to address jointly language identification and speech transcription of multiple languages simultaneously. Ultimately, this paper can serve as the basis for further research related to code-switching ASR in conversational interfaces.

ACKNOWLEDGMENTS

Research supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding (grant CHIST-ERA-20-BCI-001) and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

REFERENCES

[1] [n. d.]. OpenSLR Open Speech and Language Resources. https://www.openslr.org/. Accessed: 2024-02-11.
[2] Emily Ahn, Cecilia Jimenez, Yulia Tsvetkov, and Alan Black. 2020. What code-switching strategies are effective in dialogue systems? *Society for Computation in Linguistics* 3, 1 (2020).
[3] Larissa Aronin and David Singleton. 2012. *Multilingualism*. Vol. 30. John Benjamins Publishing.
[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.
[5] Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code-Mixing: A Challenge for Language Identification in the Language of Social Media. In *Proceedings of the First Workshop on Computational Approaches to Code-Switching*. 13–23.
[6] Aisha Bhatti, Sarimah Shamsudin, and Seriaznita Hj Mat Said. 2018. Code-Switching: A Useful Foreign Language Teaching Tool in EFL Classrooms. *English Language Teaching* 11 (2018), 93–101.
[7] Ellen Bialystok and Gigi Luk. 2012. Receptive vocabulary differences in monolingual and bilingual adults. *Bilingualism: Language and Cognition* 15, 2 (2012), 397–401.
[8] Dan Biderman, Jose Gonzalez Ortiz, Jacob Portes, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John P. Cunningham. 2024. LoRA Learns Less and Forgets Less. arXiv:2405.09673 [cs.LG]
[9] Barbara Bullock, Wally Guzmán, Jacqueline Serigos, Vivek Sharath, and Almeida Jacqueline Toribio. 2018. Predicting the presence of a Matrix Language in code-switching. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg (Eds.). Association for Computational Linguistics, Melbourne, Australia, 68–75. https://doi.org/10.18653/v1/W18-3208
[10] Yunjae J Choi, Minha Lee, and Sangsu Lee. 2023. Toward a Multilingual Conversational Agent: Challenges and Expectations of Code-mixing Multilingual Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.
[11] Helin Cihan, Yunhan Wu, Paola Peña, Justin Edwards, and Benjamin Cowan. 2022. Bilingual by default: Voice Assistants and the role of code-switching in creating a bilingual user experience. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–4.
[12] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. FLEURS: Few-shot Learning Evaluation of Universal Representations of Speech. arXiv:2205.12446 [cs.CL]
[13] Margaret Deuchar, Peredur Davies, Jon Russell Herring, M. Carmen Parafita Couto, and Diana Carter. 2014. *Building Bilingual Corpora*. Multilingual Matters, Bristol, Blue Ridge Summit, 93–110. https://doi.org/doi:10.21832/9781783091713-008
[14] Kunal Dhawan, KDimating Rekesh, and Boris Ginsburg. 2023. Unified Model for Code-Switching Speech Recognition and Language Identification Based on Concatenated Tokenizer. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 74–82. https://doi.org/10.18653/v1/2023.calcs-1.7
[15] Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. MUCS 2021: Multilingual and Code-Switching ASR Challenges for Low Resource Indian Languages. In *Interspeech 2021*. ISCA. https://doi.org/10.21437/interspeech.2021-1339
[16] A Seza Doğruöz, Sunayana Sitaram, Barbara E Bullock, and Almeida Jacqueline Toribio. 2021. A Survey of Code-switching: Linguistic and Social Perspectives for Language Technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 1654–1666. https://doi.org/10.18653/v1/2021.acl-long.131
[17] Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The People's Speech: A Large-Scale Diverse English Speech Recognition Dataset for Commercial Usage. *CoRR* abs/2111.09344 (2021). arXiv:2111.09344 https://arxiv.org/abs/2111.09344
[18] Penelope Gardner-Chloros. 2009. *Code-Switching*. Cambridge university press.
[19] Abhinav Goyal and Nikesh Garera. 2023. Building Accurate Low Latency ASR for Streaming Voice Search in E-commerce. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 276–283.
[20] François Grosjean. 2021. *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press.
[21] Adriana Guevara-Rukoz, Isin Demirsahin, Fei He, Shan-Hui Cathy Chu, Supheakmungkol Sarin, Knot Pipatsrisawat, Alexander Gutkin, Alena Butryna, and Oddur Kjartansson. 2020. Crowdsourcing Latin American Spanish for Low-Resource

Text-to-Speech. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*. European Language Resources Association (ELRA), Marseille, France, 6504–6513. https://www.aclweb.org/anthology/2020.lrec-1.801

[22] John J Gumperz. 1982. *Discourse strategies*. Number 1. Cambridge University Press.

[23] Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2021. Towards a unified view of parameter-efficient transfer learning. *arXiv preprint arXiv:2110.04366* (2021).

[24] Carlos D. Hernandez-Mena. 2019. TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license. Web Download.

[25] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*. PMLR, 2790–2799.

[26] Ke Hu, Tara N. Sainath, Bo Li, Yu Zhang, Yong Cheng, Tao Wang, Yujing Zhang, and Frederick Liu. 2023. Improving Multilingual and Code-Switching ASR Using Large Language Model Generated Text. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 1–7. https://doi.org/10.1109/ASRU57964.2023.10389644

[27] Iva Ivanova and Albert Costa. 2008. Does bilingualism hamper lexical access in speech production? *Acta psychologica* 127, 2 (2008), 277–288.

[28] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. 1991. Adaptive mixtures of local experts. *Neural computation* 3, 1 (1991), 79–87.

[29] Michael I Jordan and Robert A Jacobs. 1994. Hierarchical mixtures of experts and the EM algorithm. *Neural computation* 6, 2 (1994), 181–214.

[30] Amanda Kann. 2022. Voice Assistants Have a Plurilingualism Problem. In *Proceedings of the 4th Conference on Conversational User Interfaces*. 1–5.

[31] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. 2021. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 855–859.

[32] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. (2017). arXiv:1412.6980 [cs.LG]

[33] Bo Li, Ruoming Pang, Tara N. Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W. Ronny Huang, Min Ma, and Junwen Bai. 2021. Scaling End-to-End Models for Large-Scale Multilingual ASR. (2021). arXiv:2104.14830 [cs.CL]

[34] Hexin Liu, Xiangyu Zhang, Leibny Paola Garcia, Andy W. H. Khong, Eng Siong Chng, and Shinji Watanabe. 2024. Aligning Speech to Languages to Enhance Code-switching Speech Recognition. https://api.semanticscholar.org/CorpusID:268351705

[35] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024. DoRA: Weight-Decomposed Low-Rank Adaptation. (2024). arXiv:2402.09353 [cs.CL]

[36] Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An Empirical Study of Catastrophic Forgetting in Large Language Models During Continual Fine-tuning. (2023). arXiv:2308.08747 [cs.CL]

[37] Brian MacWhinney and Catherine Snow. 1990. The Child Language Data Exchange System: an update. *Journal of Child Language* 17, 2 (1990), 457–472. https://doi.org/10.1017/S0305000900013866

[38] Carol Myers-Scotton. 1989. Codeswitching with English: types of switching, types of communities. *World Englishes* 8, 3 (1989), 333–346.

[39] Tanmay Parekh, Emily Ahn, Yulia Tsvetkov, and Alan W Black. 2020. Understanding linguistic accommodation in code-switched human-machine dialogues. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. 565–577.

[40] Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee weon Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing Whisper-Style Training Using an Open-Source Toolkit and Publicly Available Data. (2023). arXiv:2309.13876 [cs.CL]

[41] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. (2020). arXiv:2007.03001 [eess.AS]

[42] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 28492–28518. https://proceedings.mlr.press/v202/radford23a.html

[43] Johan Schalkwyk and Ignacio Lopez Moreno. 2018. Teaching the Google Assistant to be multilingual. *Google AI Blog* (2018).

[44] Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 6107–6122.

[45] Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[46] Jörgen Valk and Tanel Alumäe. 2020. VoxLingua107: a Dataset for Spoken Language Recognition. arXiv:2011.12998 [eess.AS]

[47] Genta Indra Winata, Alham Fikri Aji, Zheng-Xin Yong, and Thamar Solorio. 2022. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. (12 2022). https://arxiv.org/abs/2212.09660

[48] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R Doyle, Leigh Clark, and Benjamin R Cowan. 2020. See what I'm saying? Comparing intelligent personal assistant use for native and non-native language speakers. In *22nd international conference on human-computer interaction with mobile devices and services*. 1–9.

[49] Emre Ylmaz, Mitchell McLaren, Henk van den Heuvel, and David A van Leeuwen. 2018. Semi-supervised acoustic model training for speech with code-switching. *Speech Communication* 105 (2018), 12–22.

[50] Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 43–63. https://doi.org/10.18653/v1/2023.calcs-1.5

[51] Xianghu Yue, Grandee Lee, Emre Yılmaz, Fang Deng, and Haizhou Li. 2019. End-to-End Code-Switching ASR for Low-Resourced Language Pairs. (2019). arXiv:1909.12681 [cs.CL]

[52] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. 2021. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models. *arXiv preprint arXiv:2106.10199* (2021).

[53] Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual Large Language Models Are Not (Yet) Code-Switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 12567–12582. https://doi.org/10.18653/v1/2023.emnlp-main.774