Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper

Mark Simmons
University of California San Diego
mjsimmons@ucsd.edu

Abstract

This paper explores finetuning Whisper for transcribing audio from linguistic elicitation of Tira, a Heiban language of Sudan. Audio originates from linguistic fieldwork and is bilingual in English and Tira. We finetune Whisper large-v3 using hand-labeled Tira audio and evaluate the resulting model on bilingual audio. We show that Whisper exhibits catastrophic forgetting of English after only a small amount of training, but that including automatically annotated English spans of audio in the training data dramatically reduces catastrophic forgetting of English while largely preserving ASR performance on monolingual Tira audio. This work is relevant to the study of automatic speech recognition for under-resourced languages and for contexts of bilingualism in a high and low-resourced language.

1 Introduction

Automatic speech recognition (ASR) tools convert speech into text, enabling rapid transcription or captioning of audio. Recent ASR models have reached or exceeded human performance at transcription on high-resource languages such as English (Radford et al., 2022), however performance lags in under-resourced languages and in contexts of code-switching (where multiple languages are used in a single conversation). While research on expanding Whisper's performance on low-resource languages exists (e.g. Lord and Newman, 2024; Liu et al., 2024; Williams et al., 2023; Qian et al., 2024), less work has been done on improving performance in code-switched scenarios. Code-switching is an under-addressed topic in ASR and in NLP in general, and research there often focuses on a few high-resource language pairs, such as Spanish-English, Mandarin-English or Hindi-English (Winata et al., 2023). Peng et al. (2023) evalute Whisper on Mandarin-English code-switched audio and Kulkarni et al. (2023), on Mandarin-English, Arabic-English and Hindi-English, for example.

The majority of languages in the world can be classified as 'ultra low-resource' in terms of the amount of NLP research and tools available for them (Liu et al., 2022). While ASR research for such languages exists (e.g. Prud'hommeaux et al., 2021; Adams et al., 2018; Amith et al., 2021; Mitra et al., 2016), the only work we are aware of that addresses ASR with an ultra low-resource language paired with a high resource language is San et al. (2022), which uses a corpus of single-speaker audio in English and Muruwari, though they only use ASR for English in their corpus. Thus, we are not aware of any work that directly addresses code-switched ASR involving at least one ultra low-resource language.

In this paper, we evaluate Whisper on bilingual audio in English and Tira, an ultra low-resource language of the Heiban family spoken in the Nuba mountains region of Sudan, before and after finetuning on monolingual audio in Tira. Audio comes from linguistic elicitation on Tira conducted by the authors and other colleagues in the Tira language project in collaboration with native Tira speaker Himidan Hassen. Linguistic elicitation refers to the process of studying the grammar of a language by "asking questions" from native speakers (Mosel, 2008). This often involves use of a metalanguage, a language spoken in common between the linguist and language speaker, in this case English, to ask for translations of words, paradigms, or sentences into the target language, or to elicit morphological paradigms for a given word in the target language. Audio from the Tira language project, then, contains speech both in Tira and English. While elicitation is different than classical code-switching, where interlocutors use multiple languages to communicate (often within the same utterance), the challenges faced in ASR for bilingual elicitation are largely the same as those faced in ASR for code-switched audio, thus, we use the term "bilingual

audio" to refer to either.

The contributions of this paper are as follows. We describe our process for using fieldwork data from linguistic elicitation of Tira, an out-of-domain language for Whisper, to create an ASR dataset. We then finetune Whisper on this dataset, and evaluate on bilingual audio in Tira and English. We also compare this with fine-tuning Whisper on Tira and English simultaneously by using existing hand-labeled annotations for Tira and automatically generated labels for English.

2 Dataset

We first created a Tira audio corpus using existing fieldwork recordings. Tira is a tonal language, meaning that pitch can distinguish words and morphemes. Tone has historically been difficult for ASR, as it is realized suprasegmentally, that is, simultaneous with the production of phonological segments such as consonants and vowels (Adams et al., 2018; Mortensen et al., 2016).

Audio labels for Tira come from pre-existing annotations recorded in ELAN (Sloetjes and Wittenburg, 2008), a software for annotation of multimedia recordings. The annotations relevant to this work are narrow IPA transcription and free translation into English. IPA transcriptions along with timestamps were extracted using the Python `pympi-ling` package[1]. A total of 28k annotated utterances were found from across 202 elicitation session recordings, totalling to 16 hours of audio. As these annotations were made to be used as reference for the purposes of linguistic documentation, certain noise is present in the labels relative to ASR training data. For example, 2123 records that did not have tone marked were excluded from the dataset. Sometimes Himidan's metacommentary in English is included alongside a Tira utterance in a single annotated label. We used the `pyenchant`[2] library to look for any sentences containing English words in the transcription and discarded these sentences from the monolingual Tira dataset. Sometimes, Himidan hums or whistles a Tira sentence for purposes of hearing the tones. Many records were explicitly labeled as such either in the transcription or translation tier, i.e. [kəvəlɛðéŋí únɛrɛ] "Whistling: I pulled him here yesterday.", but some whistled or hummed speech is included with no overt indication. To account for this, we used PyAnnote voice activity detection[3] (VAD) (Bredin and Laurent, 2021; Bredin, 2017) to determine the percentage of total duration for each record that was detected as speech. We found that the majority of records contained $\geq 60\%$ speech, so we excluded all records beneath this threshold, 825 records in total. Manual inspection showed that records beneath the 60% threshold were often completely silent, contained humming, whistling, excessive noise, or static.

Another metric we use for assessing audio quality is cosine similarity of text and audio embeddings using CLAP-IPA (Zhu et al., 2024). CLAP-IPA consists of an audio encoder, which takes audio input in any language and returns an acoustic embedding $\mathbf{s}$, and a phoneme encoder, which takes a sequence of IPA characters as input and returns a phoneme embedding $\mathbf{p}$ such that the speech embedding for a given word should have a small cosine distance to the phone embedding for its respective IPA sequence. CLAP-IPA was intended for keyword spotting (the task of identifying a given word, or in this case phoneme sequence, in a stream of speech) and forced alignment (the task of mapping each unit in a given word or phoneme sequence to its timestamps in the audio). However, we adopt it here as a metric for summarizing transcription noise with the assumption that audio which is clearer and is free of noise, cross-talk or other artefacts will have a high cosine similarity to its respective transcription. We calculated the cosine similarity of the embedding for each audio record with the phoneme embedding for its respective transcription. We found that most records in the dataset were above or equal to the threshold $\mathrm{sim}(\mathbf{s}, \mathbf{p}) = 0.6$, so we excluded any record whose cosine similarity fell beneath this value, 2156 records in total. Manual inspection of excluded records indicated that they generally contained significant noise or echo, or included commentary in English run along with the Tira utterance where only the Tira utterance had been transcribed in the label.

Annotations were made in a narrow phonetic transcription rather than in an established orthography, which can introduce variation as transcribers are required to make subjective decisions of how to represent phonetic variation (cf. Michaud et al. 2018). We compensated for this by normalizing the set of IPA symbols used in the dataset. For

[1] https://pypi.org/project/pympi-ling/
[2] https://pypi.org/project/pyenchant/
[3] https://huggingface.co/pyannote/voice-activity-detection

example, the phoneme /ɟ/ might be transcribed [ɟ, ʝ, dʒ, dʑ, dʐ]. Each of these symbols were replaced with [ɟ], and similarly for other phonemes. We also used NFKD normalization from the Python `unicodedata`[4] package.

Data splits should be chosen so as to minimize overlap between partitions. For fieldwork audio datasets, splits may be decided on speaker identity or grouped by narrative. For the Tira dataset, only one speaker is present, and different recordings may have significant overlap in their content. For example, across several elicitation sessions focusing on syntactic structure utterances may begin with [ùɾnɔ̀ kəlɛ̀ŋìt̪ɔ́ àpɾí. . .] 'grandfather knows that the boy. . . '. To maximize the difference across data partitions, we calculated the phone embedding for each transcription using CLAP-IPA and sampled records so as to maximize the cosine distance of embeddings between the train, validation and test splits. Statistics for the size of this and other datasets are given in Table 1. We refer to this dataset as the "hand-labeled monolingual" or "hand-labeled" dataset.

To evaluate the model's generalization to bilingual audio, we hand annotated labels from two elicitation recordings containing both Tira and English speech. We picked one recording that supplied Tira labels used for training (the "in-domain" bilingual set) and one recording that was not used in training (the "out-of-domain" bilingual set). Note that English audio for both recordings will be unseen for the model. Each label was taken from up to 30 seconds of speech, always segmented to end at the end of a speech turn.

We also created bilingual labels through data augmentation. For bilingual label creation, we used PyAnnote VAD to detect regions of speech from the longform elicitation recordings that were not included in the hand-labeled dataset. Since not all Tira utterances from the elicitation recordings were hand-labeled, several of these detected utterances contain speech in Tira. To distinguish between Tira and English audio, we trained a logistic regression model on hand-labeled Tira and English spans from the elicitation corpus to perform language identification (LID), using embeddings from the SpeechBrain ECAPA TDNN for language identification (Ravanelli et al., 2021), similar to the protocol outlined in (San et al., 2022). We trained LID on a dataset of 3637 Tira and 3637 English utterances, and it achieved 90% classification accuracy on a test dataset of 1818 Tira and 1818 English utterances. Tira utterances were all taken from Himidan, whereas English utterances were sampled from both Himidan and other speakers. Once utterances were segmented and labeled for language identity, English utterances were transcribed using Whisper large-v2 (which we found to perform better than Whisper large-v3 on English) and Tira utterances were transcribed using the fine-tune of Whisper large-v3 on monolingual Tira audio, as described in the following section. We then used these annotations to make a bilingual ASR dataset. For each hand-transcribed label from the monolingual dataset, we concatenated adjacent transcribed speech regions in the same elicitation recording to create a new label of up to 30 seconds. We excluded utterance transcriptions with excessive repetition (e.g. "Yeah. Yeah. Yeah. Yeah. . . "), a known failure mode of Whisper. For all training labels from the in-domain elicitation recording used for bilingual evaluation, we only included the hand-labeled Tira utterances to ensure both the model trained on the monolingual dataset and the bilingual dataset have seen the same set of data from the in-domain elicitation recording during training. We refer to this dataset as the "augmented bilingual" or "augmented" dataset.

Textual analysis of the labels revealed 14,017 words (5.3% of the whole dataset) were not identified as English (using `pyenchant` as above) or Tira (defined as any word containing only Tira IPA characters). Manual inspection of such words suggests that several are Tira words that were missed by the VAD+SLI pipeline and thus were transcribed by Whisper large-v2 rather than the checkpoint trained on Tira, e.g. "kukungapitito" instead of [kúkù ŋgápìt̪ìt̪ɔ́] 'Kuku hunted (in someone's place)', or "ngiyol" instead of [ŋìjɔ́l] 'eat'.

3 Experiment

We finetuned Whisper large-v3 using a learning rate of $3e-4$ with 500 warmup steps. We used the AdamW optimizer (Loshchilov and Hutter, 2019) with betas of 0.9 and 0.99, and trained with a batch size of 4 with 2 gradient accumulation steps for an effective batch size of 8. All models were trained with an Nvidia GeForce RTX 4090 with 24 gigabytes of VRAM. Due to GPU VRAM limitations, we were not able to finetune all of the weights

[4] https://docs.python.org/3/library/unicodedata.html

Dataset	Split	N records	Length (total)	Avg record len	%Tira	%Unk
Monolingual	train	16,384	9h29m	2.08s	100	
Monolingual	dev	2,048	1h8m	1.99s	100	
Bilingual	train	16,384	51h48m	11.38s	21.0	5.3
Bilingual in-domain	test	88	39m	26.25s	8.60	
Bilingual out-domain	test	65	29m	26.31s	2.86	

Table 1: Size of datasets used for training, validation (dev) and testing.

of Whisper large-v3, and had to rely on parameter efficient finetuning (Han et al., 2024; Houlsby et al., 2019). We used LoRA (Hu et al., 2021) applied to the query and value weights of the attention modules for parameter-efficient finetuning, following the example given in PEFT (Mangrulkar et al., 2022)[5], similar to Liu and Qu (2024). Models are evaluated in terms of word error rate (WER) and character error rate (CER). As Tira is out-of-domain for Whisper, labels were prefixed with a language ID for Yoruba (another tonal African language) for purposes of knowledge transfer (Qian et al., 2024), though we leave more thorough investigation of language ID choice for later research. We compared finetuning Whisper large-v3 using a LoRA, Whisper medium using a LoRA, and a full finetune of Whisper medium, by training each model size on the Tira dataset for 4 epochs. We found the best results came from Whisper large-v3 with LoRA, so we use this configuration for our experiment. We finetune one model for 10 epochs on each dataset respectively (hand-labeled monolingual and augmented bilingual).

4 Results

We evaluate Whisper large-v3 out of the box and compare it to a finetune using LoRA at each epoch of training using both monolingual hand-labeled data and augmented bilingual data, evaluating on both monolingual Tira data and bilingual Tira-English data. WER and CER on each evaluation set across training are given in Figure 1, where "epoch 0" corresponds to Whisper large-v3 with no finetuning.

In general, the model trained on augmented bilingual data outperforms the monolingual model when evaluated on bilingual data. When evaluated on monolingual data, both models perform similarly, with the monolingual model slightly outperforming the bilingual model.

For monolingual data, we see a precipitous drop in CER (0.86 to 0.15 for the monolingual model, 0.20 for the bilingual model) and WER (1.70 to 0.59 for the monolingual model, 0.72 for the bilingual model) in epoch 1, with much smaller improvements each subsequent epoch. For bilingual data, we see conflicting results with the model trained on monolingual Tira data. On the out-of-domain bilingual dataset, the monolingual model underperforms Whisper large-v3 at all epochs of training. For the in-domain bilingual dataset, there is a slight reduction in WER and CER by epoch 2, likely owing to the model's ability to transcribe Tira it has recognized in training, followed by a decline in performance in all subsequent epochs. Unlike the monolingual model, the augmented bilingual model's performance improves on both monolingual and bilingual datasets with training, achieving the best WER and CER at epoch 4 for the in-domain dataset and 10 for the out-of-domain dataset.

Dataset	Model	WER	CER	Epoch
Tira monoling	Tira only	**0.48**	**0.11**	8
	Augmented	0.53	0.13	10
In-domain biling	Tira only	0.83	0.57	2
	Augmented	**0.55**	**0.34**	4
Out-domain biling	Tira only	0.57	0.83	0
	Augmented	**0.49**	**0.34**	10

Table 2: Best WER and CER on validation sets

In Figure 2, we break down CER and WER by language. This plot confirms the trend suggested in Figure 1, namely that both the monolingual and augmented bilingual models perform similarly on Tira, but the augmented bilingual model significantly outperforms the monolingual model on English, likely owing to the inclusion of an English transcription task in training, even on synthetic labels.

Manual inspection gives further evidence that the worsening performance following epoch 2 for the monolingual model is due to catastrophic forgetting of English. For example, the span "on the computer" uttered by Himidan is transcribed correctly

[5] https://github.com/huggingface/peft/tree/main/examples/int8_training

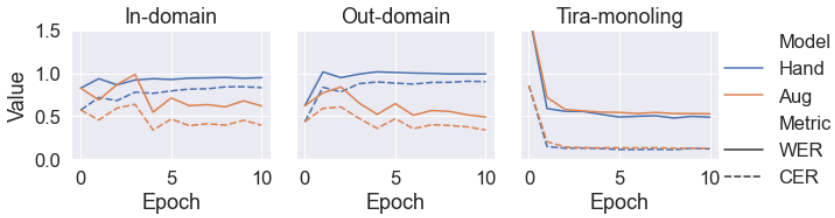

Figure 1: CER and WER on Tira validation sets. $\text{Epoch} = 0$ is equivalent to Whisper large-v3 with no finetuning.

in English by Whisper large-v3 and the model fine-tuned on monolingual hand labels after the first epoch, but in the second epoch is already transcribed as "aða kʊɔɛmpídɔ". This happens even to linguist's speech in English, particularly in proximity of Himidan giving a Tira production e.g. the span "[Himidan] ŋɔ́ɽɔ́n [linguist] Yeah I saw in the Stephen dictionary it was written as ŋicəlo" from the out-of-domain dataset was rendered "ŋɔ́ɽɔ́n, jà ìs ìn sn̪ìjən̪ dìkʃənɛ̀ ìwɛ̀s rɛ̀n̪ìŋ ìcəlò" after only one epoch of training. Manual inspection of the output of the augmented bilingual model shows that it is more common for Tira spans to be transcribed in a non-IPA pseudo-English orthography, even if the same span is correctly transcribed in the same proximity, e.g. "I'll pull them... okay **La lovela. lál ló válɛ̀ða nd̪ɔ̀bà**". This is likely due to the presence of similar spans in the augmented dataset, owing to the imperfect nature of the VAD>SLI pipeline, and could likely be ameliorated by improving the quality of the augmented dataset.

5 Conclusion

We describe the steps to create an ASR dataset from linguistic elicitation of an ultra low-resource language, Tira, including various strategies for data cleaning. We use the dataset to train Whisper large-v3, and evaluate on bilingual audio in Tira and English. We compare training on hand-labeled monolingual Tira audio with training on an augmented dataset where English (and additional Tira) audio is included with machine-generated labels. We show that the model exhibited catastrophic forgetting of English and overfitting after only one epoch of training, but this can be minimized by adding synthetic labels in English.

6 Limitations and future directions

Our training dataset comes from one speaker alone and is limited in its subject domain. As such the models produced in this work are overfit to his speech and to the domain of linguistic elicitation,

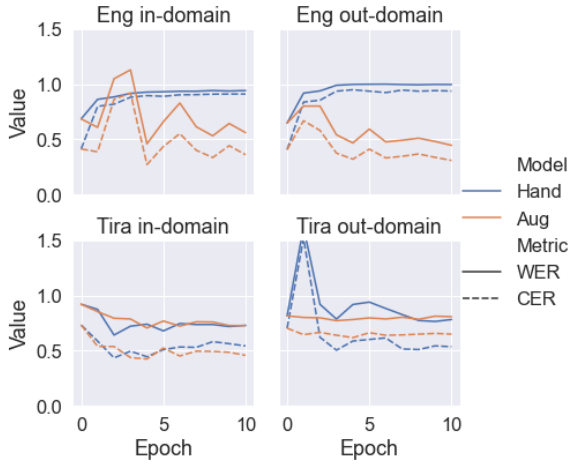

Figure 2: Language-specific WER and CER for bilingual datasets.

and would not generalize well to conversational speech in Tira or to other Tira speakers. However, our goal is a model suited to transcribing audio specifically from a context of linguistic fieldwork or pedagogy. We hope that our method can be extended to aid documentation and revitalization efforts on other low resource languages.

Future directions include comparing machine learning techniques for preventing catastrophic forgetting to training with artificial bilingual labels to see which causes the least degradation of English ASR performance, improving the quality of the augmented dataset, and reproducing these experiments with other datasets of bilingual audio from fieldwork corpora.

7 Ethical considerations

Data gathered on Tira were recorded with the consent of the speaker and the permission of UC San Diego's IRB (Protocol 805624). Annotations were produced by the authors and other academic colleagues. Data in English come from Himidan as well as the authors and other linguists present during elicitation sessions. No other data were used.

References

Oliver Adams, Trevor Cohn, Graham Neubig, Hilaria Cruz, Steven Bird, and Alexis Michaud. 2018. Evaluation Phonemic Transcription of Low-Resource Tonal Languages for Language Documentation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-End Automatic Speech Recognition: Its Impact on the Workflowin Documenting Yoloxóchitl Mixtec. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80, Online. Association for Computational Linguistics.

Hervé Bredin. 2017. Pyannote.metrics: A Toolkit for Reproducible Evaluation, Diagnostic, and Error Analysis of Speaker Diarization Systems. In *Proc. Interspeech 2017*, pages 3587–3591.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-Efficient Fine-Tuning for Large Models: A Comprehensive Survey. *Preprint*, arXiv:2403.14608.

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. *Preprint*, arXiv:1902.00751.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Atharva Kulkarni, Ajinkya Kulkarni, Miguel Couceiro, and Hanan Aldarmaki. 2023. Adapting the adapters for code-switching in multilingual ASR.

Yunpeng Liu and Dan Qu. 2024. Parameter-efficient fine-tuning of Whisper for low-resource speech recognition. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1522–1525.

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Zoey Liu, Crystal Richardson, Richard Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 3933–3944.

Laurel Lord and Mark Newman. 2024. Automatic Speech Recognition Variance: Consecutive Runs of Low-Resource Languages in Whisper. *International Journal of Machine Learning*, 14(2).

Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. *Preprint*, arXiv:1711.05101.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Alexis Michaud, Oliver Adams, Trevor Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Vikramjit Mitra, Andreas Kathol, Jonathan D. Amith, and Rey Castillo García. 2016. Automatic Speech Transcription for Low-Resource Languages — The Case of Yoloxóchitl Mixtec (Mexico). In *Interspeech 2016*, Interspeech_2016. ISCA.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A Resource for Mapping IPA Segments to Articulatory Feature Vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Ulrike Mosel. 2008. Chapter 3 Fieldwork and community language. In *Essentials of Language Documentation*, pages 67–86. De Gruyter Mouton.

Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. 2023. Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization. In *INTERSPEECH 2023*, pages 396–400. ISCA.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15:491–513.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate M. Knill, and Mark J. F. Gales. 2024. Learn and Don't Forget: Adding a New Language to ASR Foundation Models. *arXiv preprint*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv preprint*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba,

Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A General-Purpose Speech Toolkit.

Nay San, Martijn Bartelds, Tolulope Ogunremi, Alison Mount, Ruben Thompson, Michael Higgins, Roy Barker, Jane Helen Simpson, and Dan Jurafsky. 2022. Automated speech tools for helping communities process restricted-access corpora for language revival efforts. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by Category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Aiden Williams, Andrea Demarco, and Claudia Borg. 2023. The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43. ISCA.

Genta Winata, Alham Fikri Aji, Zheng Xin Yong, and Thamar Solorio. 2023. The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2936–2978, Toronto, Canada. Association for Computational Linguistics.

Jian Zhu, Changbing Yang, Farhan Samir, and Jahurul Islam. 2024. The taste of IPA: Towards open-vocabulary keyword spotting and forced alignment in any language. *Preprint*, arXiv:2311.08323.