A Family Affair: Counterintuitive transfer effects in ASR for endangered languages

Emily Prud'hommeaux[1], **Hunter Scheppat**[1], **Nicholas Hernandez**[1], **Éric Le Ferrand**[2]
[1]Boston College, [2]University at Buffalo
{prudhome,scheppat,hernantx}@bc.edu, ericlefe@buffalo.edu

Abstract

Recent work on ASR for endangered languages has focused on fine-tuning a pretrained multilingual model to a small amount of data in the target language, but even with large and diverse pretrained models, error rates can be stubbornly high. Including additional data from the target language can yield improvements, but acquiring this kind of data for endangered languages is challenging. One commonly proposed alternative is to leverage data in a more highly resourced language from the same language family, but there has been remarkably little work evaluating the real-world utility of this practice. In this paper, we explore the relevance of phylogenetic relatedness of transfer data for improving ASR for endangered languages with minimal training resources. We find, unexpectedly, that augmenting with unrelated donor language data improves word error rate (WER) more than related language data. In addition, the performance improvement observed via transfer of unrelated data is larger than that observed when nearly doubling the quantity of in-language training data. These surprising results hold even when controlling for factors such as acoustic similarity, audio quality, and subword overlap. Our findings point the way to practical new approaches for improving ASR for endangered languages.

1 Introduction

Most contemporary approaches to automatic speech recognition (ASR) for under-resourced languages involve fine-tuning a pretrained multilingual model to a small amount of transcribed audio data in the target under-resourced language. With the small quantity of transcribed audio available for a typical endangered language, this approach tends to yield stronger results than end-to-end monolingual training (Jimerson et al., 2023), but error rates often remain high. One possible remedy is to augment the training set with available data from a related language. There is relatively little work, however, comparing the utility of this practice with other alternatives, such as gathering more in-language data or including data from an unrelated language.

In this paper, we explore the impact of incorporating related versus unrelated donor language data in the ASR training pipeline for several under-resourced endangered languages, and we compare these results to those derived by including additional target language data. Using a two-stage fine-tuning approach from a multilingual pretrained model, we investigate whether first fine-tuning on data from a donor language and then fine-tuning to the target language can improve ASR accuracy for that target language. We then compare the impact of fine-tuning from a model trained on a closely related language with that of fine-tuning from a model trained on a similar quantity of data from a distantly related or entirely unrelated donor language. We also compare those results with training on additional target language data. Finally we perform a regression analysis to reveal the factors that predict positive transfer, including not only phylogenetic relatedness but also acoustic similarity and subword overlap between donor and target, and a general measure of donor audio quality.

We find that two-stage fine-tuning nearly always improves word error rate (WER), sometimes dramatically, regardless of whether the donor language belongs to the same family. Surprisingly, unrelated languages tend to yield larger improvements than closely related languages. In addition, two-stage fine-tuning with the most effective donor language yields larger WER improvements than simply including additional target language data. Factors including audio quality, acoustic similarity, and subword overlap appear to contribute to reductions in WER, as well.

While gathering additional in-language data or including more related language data in training

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 28389–28397
October 24-29, 2026 ©2026 Association for Computational Linguistics

Language	Role	Phones	Chars	Hours	Tokens	Types	NISQA
English	donor	222,310	308,656	5.728	72,046	6,731	3.621
Hungarian	donor	222,352	234,433	5.086	36,823	10,652	2.904
Swahili	donor	222,304	239,274	4.757	44,185	8,569	3.349
Tagalog	donor	222,319	222,925	7.423	44,442	6,312	2.234
Rukai	donor/large	222,286	239,340	11.453	38,445	11,816	3.313
	small	97,149	104,106	5.003	13,517	5,675	
Puyuma	donor/large	222,289	235,020	11.822	45,103	6,047	3.507
	small	94,541	99,642	5.002	15,445	3,386	
Paiwan	donor/large	222,337	228,086	10.421	42,908	7,064	3.629
	small	107,082	109,788	5.000	16,613	3,928	
Bunun	donor/large	222,280	232,630	11.312	44,671	7,647	3.742
	small	98,188	102,844	5.001	16,197	4,247	
Amis	large	226,090	226,226	10.839	43,491	6,263	3.219
	small	104,508	104,390	5.001	17,549	3,756	
Seediq	large	162,944	151,464	8.820	31,371	3,675	3.076
	small	92,418	86,156	5.001	16,196	2,697	
Atayal	large	222,286	223,493	11.684	46,251	6,686	3.313
	small	95,043	96,200	5.001	17,240	3,859	

Table 1: Counts of characters, Allosaurus-generated phones, hours of audio (for train+test), and word tokens and word types (for train only) for each dataset. The test utterances, which were the same for all experiments, were included in training the donor models but not the large baseline models. The NISQA audio quality metric is the median calculated over a random sample of 200 utterances per dataset.

are both commonly proposed ways to try to improve performance in endangered language ASR, as much or more benefit may be derived via transfer from unrelated language data. Creating data for endangered languages is challenging, and many such languages belong to families that are themselves under-resourced. Our findings offer practical alternatives to researchers and communities working on endangered language ASR.

2 Related work

Including related language data is an established practice for low-resource ASR and was often found to be effective in reducing word error rates (WER) over monolingual training in pre-transformer architectures (Nouza et al., 2016; Billa, 2021; Novitasari et al., 2020). While some prior work in statistical ASR found advantages in training with related vs. unrelated languages (Van Heerden et al., 2010; Juan et al., 2015), other studies reported that training with any additional data is helpful, regardless of its relatedness (Vu and Schultz, 2013; Tachbelie et al., 2020; Abate et al., 2021).

Related language data is less well explored when fine-tuning from a large pretrained model within transformer architectures, which is currently the most common way to train ASR models in extreme low-resource settings. A few studies have shown that pretraining a feature extractor on monolingual related language data rather than multilingual data can provide more robust modeling, including for Guadeloupean and French (Macaire et al., 2022), Pomak and Slavic languages (Tsoukala et al., 2023), and Indic languages (Javed et al., 2022). These studies, however, have limitations. Most notably, the related languages used in these cases have abundant resources, a situation that is rare among endangered languages. Including related languages during fine-tuning from a multilingual model yields mixed results with some showing improvements in WER (Coto-Solano et al., 2024) and others degradations (Abdou Mohamed et al., 2024; Li et al., 2024), with other factors being equally important, including phonetic overlap (Nigatu and Aldarmaki, 2025) or geographic proximity (Adams et al., 2019).

Given the variable utility of relatedness, some work has explored other similarity features for facilitating multilingual and cross-lingual NLP. One

approach is the lang2vec metric with URIEL (Littell et al., 2017; Khan et al., 2025), a database of linguistic information. Other metrics of similarity are derived from the data itself, including Token Distribution Similarity (Gogoulou et al., 2024), Angular Similarity of Phoneme Frequencies (ASPF) (Do et al., 2022), and Acoustic Token Distribution Similarity (ATDS) (San et al., 2024).

Our work is distinct in a number of ways. First, we experiment with multiple closely related languages, one distantly related language, and three unrelated languages. Second, we control carefully for dataset size, which we measure in terms of phonetic content rather than raw duration. Third, we compare our transfer results to training simply with additional in-language data. Fourth, we carry out a statistical analysis to determine whether factors other than relatedness, such as acoustic or textual similarity, contribute to positive transfer.

Although this paper focuses on ASR, we note that the impact of phylogenetic relatedness in cross-lingual transfer has also been explored in multilingual language modeling (Pires et al., 2019; Wu and Dredze, 2019; Adams et al., 2017) and tasks like machine translation, dependency parsing, and morphological and POS tagging (Lin et al., 2019; Cotterell and Heigold, 2017; Johnson et al., 2017).

3 Data

Throughout we refer to the language used for the initial fine-tuning stage as the *donor language*, and the language to which that model is adapted and on which it is evaluated as the *target language*. Our target languages are 7 of the 14 Indigenous languages of Taiwan, known as the Formosan language subfamily of the larger Austronesian language family: Amis, Atayal, Bunun, Paiwan, Puyuma, Rukai, and Seediq. The audio data and accompanying Latin alphabet transcripts are extracted from the ePark corpus (Indigenous Language Research and Development Foundation, 2023) of studio-recorded scripted speech made available in FormosanBank.[1] We use four Formosan languages (Bunun, Paiwan, Puyuma, Rukai) as related donor languages, along with Tagalog (Fleurs (Conneau et al., 2023)), a more distantly related Austronesian language. For unrelated donor languages, we use English (TEDlium v3 (Hernandez et al., 2018)) Swahili (ALFFA (Gelas et al., 2012)), and Hungarian (VoxPopuli (Wang et al., 2021)).

[1]https://github.com/FormosanBank

To ensure that data quantity is not responsible for any WER gains we might observe, we must control for dataset size, but the usual metric, duration, does not account for the large variation in speaking rate observed across our corpora. The ePark Formosan corpus was created as an educational resource, with most utterances spoken at an unusually slow rate. In contrast, the non-Formosan data is delivered at the typical speaking rate for presentation-style or scripted speech. While character count is roughly equivalent to phone count in many languages, there is a weak association between characters and speech sounds in English. To more accurately measure speech content, we use phone counts derived from the output of Allosaurus (Li et al., 2020), a pretrained universal phone recognizer. While the identities of the predicted phones themselves were sometimes inaccurate in our data since Allosaurus tries to select the most precise phoneme from the hundreds of possible IPA phonemes, we found that the *number* of phones proposed corresponded strongly to the actual number of phones observed in an utterance, with minimal deletions or insertions.

To create our donor datasets, we first identified four Formosan languages whose ePark data exceeds 10 hours. Of those, the language with the least data was Bunun. We used Allosaurus to generate a phone-level transcription of the full Bunun dataset, and we then set the number of phones for our donor datasets to be roughly the same as the number of phones in Bunun, approximately 222K. For each other donor language (English, Hungarian, Swahili, Tagalog, Rukai, Puyuma, and Paiwan), we randomly selected utterances from its associated corpus until we had roughly 222K phones for that language. We used the same process to create the large baseline datasets for the target languages.

Since we are focusing on endangered languages, only the Formosan languages serve as targets. To create our small target Formosan datasets, we selected 5 hours of the data from the corresponding donor version of that language dataset or, in the case of target-only languages, from the available ePark data. By using "10 hour" donor datasets and "5 hour" target datasets, we simulate the linguistic scenario typical of many endangered and oral languages, where closely related languages are often under-resourced as well. All small target datasets were partitioned randomly into 80/20 train/test splits. These test sets were used to evaluate both the small and the large target baselines.

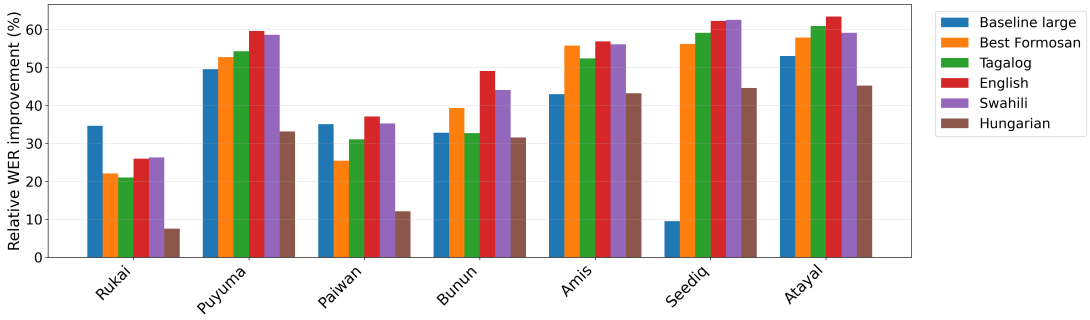


Figure 1: Relative improvements in WER over the small baseline for the seven target Formosan languages. Best Formosan is the best of the four Formosan donors. Full WER results are presented in Table 2.

Dataset statistics are found in Table 1. The table also illustrates the weak correspondence between characters and phones in English and the slow speaking rate in the Formosan datasets.

4 Method

We fine-tune our models with XLSR-53 (Conneau et al., 2021), a multilingual speech model based on wav2vec2 2.0 that is widely used in low-resource settings. Following a popular tutorial[2], we train a CTC layer for 30 epochs[3] without freezing the speech features, using a uniform character vocabulary across all languages. We decode using `pyctcdecode` with a trigram word-level LM built on the training transcripts for the target language using KenLM (Heafield, 2011). English, Swahili, and Tagalog appear in the XLSR-53 training data, though not the particular corpora we use here; Hungarian and the Formosan languages do not. (See the Limitations section for some discussion.)

We first fine-tune and evaluate two baseline models for each Formosan language: one using the large size of "10 hours" (∼210K phones: large baseline) and one using the small target size of "5 hours" (∼100K phones: small baseline). Five to ten hours of data is quite typical for endangered languages, which, by definition, have few speakers, and in practice have even fewer speakers who have the skills and time to transcribe audio data.

We then carry out our two-stage fine-tuning training using our donor corpora for the three unrelated languages, plus the distantly related Tagalog, and the closely related Formosan languages, Rukai, Puyuma, Paiwan, and Bunun, as follows. We first fine-tune XLSR-53 to each donor language to create a trained model for each donor language. Each of these 8 models is then fine-tuned in a second stage to one of the 7 Formosan target languages. That two-stage-tuned model is then tested on the test data for that target language.

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

[3]All models converged before 30 epochs.

5 Results and Analysis

Figure 1 shows the relative improvement in WER for each target language over the small baseline for each of the training conditions: training with additional in-language data (large baseline), two-stage fine-tuning with the most helpful Formosan donor, and two-stage tuning with Tagalog, English, Swahili, and Hungarian. Table 2 presents the raw WER results for both baselines and each donor-target pair. As expected, we see that training with double the amount of in-language data (large baseline) yields substantial improvements in WER over the small baseline model. We also observe that two-stage training consistently improves WER for the small baseline, regardless of the donor language.

Surprisingly, however, we see that two-stage training with an *unrelated* language always yields larger WER reductions than two-stage training with the best related language. A mixed-effects regression with random intercept per target language confirms this surprising observation. Formosan donors yield significantly less WER improvement than unrelated donors ($\beta = -0.06$, 95% CI $[-0.11, -0.01]$, $p = 0.010$). Tagalog (also Austronesian, but not Formosan) did not differ significantly from unrelated donors in this model ($\beta = +0.01$, $p = .78$).

In order to further explore these interesting results, we consider a few factors other than phylogenetic relatedness that might contribute to positive transfer. One of these is **subword overlap** between target and donor. To obtain subwords, we tokenize the training data for each language with the BPE-

Lang.	Small	Large	Rukai	Puyuma	Paiwan	Bunun	Tagalog	Eng.	Hung.	Swa.
Rukai	54.94	**35.92**	–	48.30	42.83	44.85	43.40	40.67	50.80	40.49
Puyuma	32.36	16.33	22.20	–	15.30	16.71	14.81	**13.08**	21.64	13.40
Paiwan	44.36	28.82	38.78	36.78	–	33.09	30.59	**27.91**	38.98	28.72
Bunun	48.85	32.83	33.32	33.05	29.67	–	32.89	**24.89**	33.45	27.33
Amis	37.48	21.38	23.05	21.90	16.59	18.20	17.85	**16.17**	21.29	16.46
Seediq	35.44	32.07	21.99	20.81	15.70	15.53	14.50	13.38	19.64	**13.28**
Atayal	44.70	21.01	25.11	24.46	18.96	18.83	17.49	**16.36**	24.50	18.28

Table 2: Word error rates (WER) as percentages by target language (rows) and donor language model (columns). Best model per language in **bold**. Recall that Amis, Seediq, and Atayal served only as targets and not as donors.

based pretrained Whisper (Radford et al., 2023) tokenizer and compute the overlap of unique subwords between each donor-target pair.[4] A second feature is **acoustic token distribution similarity (ATDS)** San et al. (2024). ATDS measures the acoustic similarity between two languages based on the frequency of recurring acoustic-phonetic sequences extracted from a pretrained wav2vec 2.0 model (here XLSR-53). These two measures are reported in the heatmaps in Appendix Figure 2. While our Formosan corpora were all recorded under the same conditions, the donor corpora were not. For this reason, we also consider **NISQA** (Non-Intrusive Speech Quality Assessment) (Mittag et al., 2021), a popular measure of audio quality intended to capture features related to recording setting and equipment, which we calculate over a random sample of 200 utterances per dataset using the VERSA toolkit (Shi et al., 2025). NISQA scores are reported in the final column of Table 1.

When subword overlap, acoustic similarity (ATDS), and audio quality (NISQA) were added as covariates, both un-relatedness effects strengthened substantially. Formosan donors showed a -0.20 penalty ($p < 0.001$) and Tagalog a $+0.35$ advantage ($p < .001$). Subword overlap ($\beta = +0.04$, $p = 0.001$) and NISQA ($\beta = +0.16$, $p < .001$) both contributed positively to WER improvement in the joint model, aligning with the "common sense" expectation that transfer from acoustically and textually similar donors should yield improvements. The strengthening of this effect when accounting for these features suggests that the disadvantage of Formosan donors is not explained by these features. In other words, Formosan donors seem like they ought to help more than unrelated languages given their greater similarity to Formosan targets, but they actually help less.

Finally, we note that in all but one language, transfer from an unrelated donor outperforms training with additional in-language data. A paired Wilcoxon signed-rank test confirms that best-donor WERs were significantly lower than large baseline WERs ($W = 25$, $p < 0.05$). This could be due to the overall increased data volume involved in two-stage training, but the result raises a practical question: when only 5 hours of target-language data is available, is it more worthwhile to invest in collecting and transcribing additional in-language data or to leverage existing donor-language resources? Experimentation across a more diverse and larger set of languages is needed to provide an answer.

6 Conclusions

In this short paper, we explore the utility of related vs. unrelated language data augmentation for improving ASR performance for a family of endangered languages. We find that two-stage transfer learning yields substantial improvements, with donors outside the language family showing the strongest results, even over training on additional in-language data. While it is common to propose to use related language data in transfer learning for low-resource ASR, it may actually be just as effective to use data from unrelated languages. Our findings offer practical alternatives to collecting additional in-language data when training ASR models for endangered language documentation. Future work will focus on expanding this line of experimentation to other endangered language families, including features like speaker diversity, and probing the impact of the presence of potential donor languages in the pretraining data.

[4]We use the Whisper tokenizer because it purports to be language-agnostic and trained on large amounts of multilingual transcripts of audio data rather than text data.

Limitations

There are a number of popular architectures that could have been used in our experiments. We chose to focus on XLSR-53 for several reasons. First, because XLSR-53 trains quickly, uses relatively few resources, and creates small models, it is very widely used for endangered language ASR, our specific use case. Second, in our preliminary work, we found that other popular models yielded weak baseline results, failed to generate improvements with two-stage training, or both. MMS (Pratap et al., 2024) showed few benefits in our donor/target two-stage fine-tuning setup. HuBERT (Hsu et al., 2021) yielded baseline WERs consistently and substantially higher than XLSR-53. Prior work has shown that multi-stage training in Whisper (Radford et al., 2023) with out-of-domain data from within the same language produces much weaker results than single-stage in-domain-only training (Tapo et al., 2024), making Whisper unsuitably fragile for our experiments. We also note that the data used to train the Whisper multilingual models has not been disclosed, opening the possibility of data contamination with our train and test sets. Finally, we note the impact on computing resources required to fine-tune dozens of additional models, particularly when prior work suggests that no one architecture consistently outperforms any other in extreme low-resource settings (Jimerson et al., 2023).

While none of the data used here was used to pretrain XLSR-53, tens of thousands of hours of English were used in pretraining, along with 100 hours or so each of Swahili and Tagalog. Our experiments were not designed to probe this particular aspect of XLSR-53, but we note that Hungarian, which does not appear in the XLSR-53 training data, is an effective donor, outperforming three of the four Formosan donors. A definitive test would require donor languages balanced across both dimensions, which we leave for future work.

Finally, we recognize that there is one important factor we do not consider in our statistical analysis: speaker diversity. Prior work has shown that increasing the number of speakers in the training data can yield substantial improvements in WER (Berrebbi et al., 2023). Unfortunately, only the Hungarian data is specifically labeled for speaker identity. We attempted to estimate speaker counts in the other datasets via normalized maximum eigengap spectral clustering (NME-SC) (Park et al., 2019) over ECAPA (Guo et al., 2024) speaker embeddings. While the resulting counts seemed reasonable for Tagalog, English, and Swahili, they did not align well with what the ePark corpus creators have told us about their data collection process.

Ethical Considerations

All data used for this project were extracted from publicly available datasets constructed for research purposes. Consent was obtained from all communities and speakers at the time the resources were created.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant #2319296. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2021. End-to-end multilingual automatic speech recognition for less-resourced languages: The case of four Ethiopian languages. In *Proceedings of ICASSP*, pages 7013–7017.

Naira Abdou Mohamed, Anass Allak, Kamel Gaanoun, Imade Benelallam, Zakarya Erraji, and Abdessalam Bahafid. 2024. Multilingual speech recognition initiative for African languages. *International Journal of Data Science and Analytics*, pages 1–16.

Oliver Adams, Adam Makarucha, Graham Neubig, Steven Bird, and Trevor Cohn. 2017. Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of EACL*, pages 937–947.

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of NAACL-HLT*, pages 96–108.

Dan Berrebbi, Ronan Collobert, Navdeep Jaitly, and Tatiana Likhomanenko. 2023. More speaking or more speakers? In *Proceedings of ICASSP*, pages 1–5.

Jayadev Billa. 2021. Leveraging non-target language resources to improve ASR performance in a target language. In *Proceedings of Interspeech*, pages 2581–2585.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech*, pages 2426–2430.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Rolando Coto-Solano, Tai Wan Kim, Alexander Jones, and Sharid Loáiciga. 2024. Multilingual models for ASR in chibchan languages. In *Proceedings of NAACL-HLT*, pages 8521–8535.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual character-level neural morphological tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759.

Phat Do, Matt Coler, Jelske Dijkstra, and Esther Klabbers. 2022. Text-to-speech for under-resourced languages: Phoneme mapping and source language selection in transfer learning. In *Proceedings of SIGUL 2022*, pages 16–22.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *Proceedings of SLTU*, pages 94–101.

Evangelia Gogoulou, Timothée Lesort, Magnus Boman, and Joakim Nivre. 2024. Continual learning under language shift. In *International Conference on Text, Speech, and Dialogue*, pages 71–84.

Jingxiang Guo, Jinxuan Zhu, Sixu Lin, and Feng Shi. 2024. ECAPA-TDNN embeddings for speaker recognition. In *2024 5th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT)*, pages 1488–1491.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Indigenous Language Research and Development Foundation. 2023. Yuanzhumin Yuyan Leyuan (ePark). https://web.klokah.tw/.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of AAAI*, pages 10813–10821.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of ACL*, pages 1008–1016.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop ASR for a very under-resourced language: A case study for Iban. In *Proceedings of Interspeech*, page 1270–1274.

Aditya Khan, Mason Shipton, David Anuragha, Kaiyao Duan, Phuong H Hoang, Eric Khiu, A Seza Doğruöz, and En-Shiun Annie Lee. 2025. Uriel+: Enhancing linguistic inclusion and usability in a typological and multilingual knowledge base. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6937–6957.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *Proceedings of ICASSP*, pages 8249–8253.

Zhaolin Li, Monika Rind-Pawlowski, and Jan Niehues. 2024. Speech recognition corpus of the khinalug language for documenting endangered languages. In *Proceedings of LREC-COLING 2024*, pages 15171–15180.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Patrick Littell, David R Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of EACL*, pages 8–14.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics*, pages 2512–2520.

Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. NISQA: A deep cnn-self-attention model for multidimensional speech quality

prediction with crowdsourced datasets. *Proceedings of Interspeech*, pages 2127–2131.

Hellina Hailu Nigatu and Hanan Aldarmaki. 2025. Exploring transliteration-based zero-shot transfer for Amharic ASR. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP)*, pages 64–73.

Jan Nouza, Radek Safarik, and Petr Cerva. 2016. ASR for South Slavic languages developed in almost automated way. In *Proceedings of Interspeech*, pages 3868–3872.

Sashi Novitasari, Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Cross-lingual machine speech chain for Javanese, Sundanese, Balinese, and Bataks speech recognition and synthesis. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 131–138.

Tae Jin Park, Kyu J Han, Manoj Kumar, and Shrikanth Narayanan. 2019. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters*, 27:381–385.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of ACL*, pages 4996–5001.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

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 International Conference on Machine Learning*, pages 28492–28518.

Nay San, Georgios Paraskevopoulos, Aryaman Arora, Xiluo He, Prabhjot Kaur, Oliver Adams, and Dan Jurafsky. 2024. Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 100–112.

Jiatong Shi, Hye-jin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, Dareen Safar Alharthi, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, and Shinji Watanabe. 2025. VERSA: A versatile evaluation toolkit for speech, audio, and music. In *Proceedings of NAACL-HTL (System Demonstrations)*, pages 191–209.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020. Development of multilingual ASR using globalphone for less-resourced languages: The case of Ethiopian languages. In *Proceedings of Interspeech*, pages 1032–1036.

Allahsera Tapo, Éric Le Ferrand, Zoey Liu, Christopher Homan, and Emily Prud'hommeaux. 2024. Leveraging speech data diversity to document indigenous heritage and culture. In *Proceedings of Interspeech*, pages 5088–5092.

Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. ASR pipeline for low-resourced languages: A case study on Pomak. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45.

Charl Van Heerden, Neil Kleynhans, Etienne Barnard, and Marelie Davel. 2010. Pooling ASR data for closely related languages. In *Proceedings of the Workshop on Spoken Languages Technologies for Under-Resourced Languages*, pages 17–23.

Ngoc Thang Vu and Tanja Schultz. 2013. Multilingual multilayer perceptron for rapid language adaptation between and across language families. In *Proceedings of Interspeech*, pages 515–519.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of ACL*, pages 993–1003.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of EMNLP-IJCNLP*, pages 833–844.

A Appendix

The data splits and code used here are available at https://github.com/FormosanBank/A-Family-Affair-ASR.

Table 3 presents the character error rate (CER) obtained for each target language: small in-language baseline; large in-language baseline; and after transfer from each donor. These results largely parallel the WER results: transfer from a non-related language always outperforms the small in-language model by a large margin, and English is the best donor for most languages except Rukai, where the large in-language model wins out, and Seediq, where Swahili is the best donor. The only notable difference is that the large in-language Paiwan model is slightly but probably insignificantly better than the English donor model (4.09 vs 4.13).

Figure 2 shows heatmaps for ATDS acoustic similarity and subword overlap between targets and

Lang.	Small	Large	Rukai	Puyuma	Paiwan	Bunun	Tagalog	Eng.	Hung.	Swa.
Rukai	11.87	**5.90**	–	8.24	7.53	7.52	21.50	6.79	9.27	6.84
Puyuma	8.77	4.53	4.95	–	3.91	5.18	10.94	**3.62**	4.99	3.71
Paiwan	10.87	**4.09**	5.65	5.53	–	6.02	5.35	4.13	6.27	4.56
Bunun	11.37	6.82	5.01	5.08	4.82	–	12.63	**3.81**	5.06	4.01
Amis	9.67	8.54	4.17	4.57	3.17	4.88	7.32	**2.98**	3.81	3.20
Seediq	30.06	13.94	7.54	7.36	6.38	6.28	8.18	6.06	6.98	**5.74**
Atayal	20.59	10.61	7.55	7.24	6.15	6.25	9.16	**5.65**	7.36	6.01

Table 3: Character Error Rate (CER) expressed as percentages, by target language (rows) and donor language model (columns). Best model per language is rendered in **bold**.

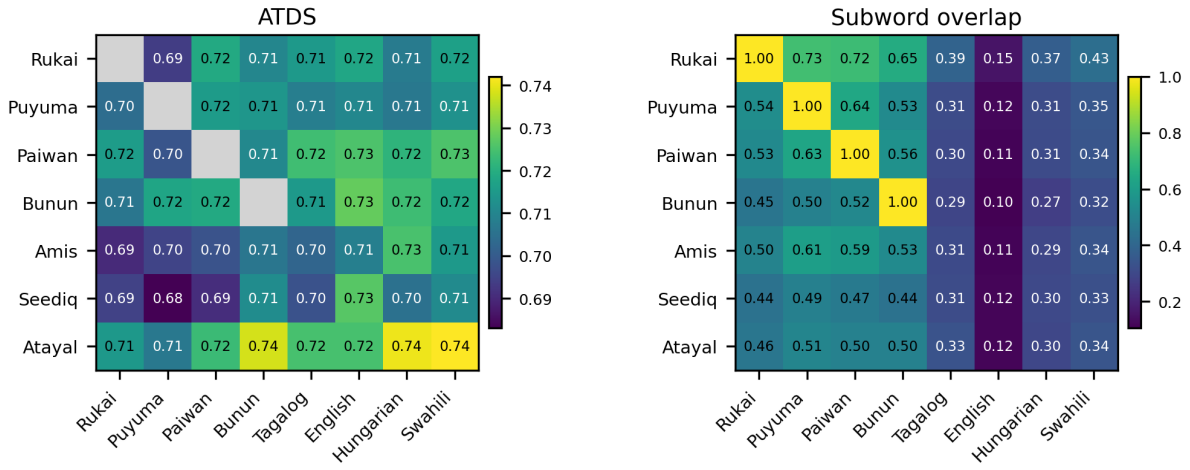


Figure 2: Left: Heatmap of ATDS acoustic similarity scores between targets (y-axis) and donors (x-axis). Right: Heatmap of subword overlap percentages between targets (y-axis) and donors (x-axis).

donors. The ATDS scores are slightly lower than the range reported in San et al. (2024), likely due to our using XLSR-53 representations rather than XLS-R, but the tight range is comparable. Recall that subword overlap was calculated using tokens produced by the language-agnostic Whisper tokenizer. The very low values for English are presumably related to the fact that English makes up the majority of the Whisper tokenizer training data.