# Exploring Transliteration-Based Zero-Shot Transfer for Amharic ASR

**Hellina Hailu Nigatu**
hellina_nigatu@berkeley.edu
UC Berkeley
USA, CA

**Hanan Aldarmaki**
hanan.aldarmaki@mbzuai.ac.ae
MBZUAI
UAE, Abu Dhabi

## Abstract

The performance of Automatic Speech Recognition (ASR) depends on the availability of transcribed speech datasets—often scarce or non-existent for many of the world's languages. This study investigates alternative strategies to bridge the data gap using zero-shot cross-lingual transfer, leveraging transliteration as a method to utilize data from other languages. We experiment with transliteration from various source languages and demonstrate ASR performance in a low-resourced language, Amharic. We find that source data that align with the character distribution of the test data achieve the best performance, regardless of language family. We also experiment with fine-tuning with minimal transcribed data in the target language. Our findings demonstrate that transliteration, particularly when combined with a strategic choice of source languages, is a viable approach for improving ASR in zero-shot and low-resourced settings.

## 1 Introduction

Automatic Speech Recognition (ASR) is an essential technology used in digital accessibility, video captioning, and virtual assistants. The performance of ASR models depends on the availability of large transcribed speech data for supervised training; yet, such data is lacking for the majority of the world's languages. Attempts to address this data resource gap include data augmentation techniques via self-training and speech synthesis (Bartelds et al., 2023; Kahn et al.), transfer learning by multilingual pre-training alongside high-resourced languages (Radford et al., 2022), or zero-shot transfer (Żelasko et al., 2020; Feng et al., 2021). Zero-shot approaches are particularly appealing in low-resourced settings as they eliminate the requirement of aligned data in the target language.

Languages use different writing scripts; hence, direct zero-shot transfer to the target language writing system may not be possible. Prior works address this challenge by relying on phonetic transcriptions, namely the International Phonetic Alphabet (IPA), as a universal system that can be applied zero-shot to unseen languages (e.g. Feng et al., 2021). While IPA representations are beneficial for building text-free speech recognition systems for unwritten languages, they are not suitable for use cases where users interact directly with the ASR output, such as automatic dictation or video captioning, as most people cannot decode IPA. Another challenge in zero-shot ASR is that languages have different phonetic distributions. In such cross-lingual settings, a deeper investigation of the choice of transfer languages can improve performance (Do et al., 2022; Khare et al., 2021).

We explore how to best utilize transliteration as a mechanism for zero-shot ASR transfer. We focus on a single low-resourced language, Amharic, as a target language, and experiment with Arabic, Xhosa, French, and Spanish as our transfer languages. We selected Arabic and Xhosa based on language family and shared phonetic distribution(§3.1), and French and Spanish as high-resourced but unrelated languages. We automatically transliterate the transcriptions of the transfer languages to our target language script and experiment with zero-shot transfer with wav2vec2 XLS-R and GMM-HMM models (§3.3). While zero-shot speech recognition generally has high error rates (Gao et al., 2021), our approach demonstrates improvements over prior work and gives insights into best practices for cross-lingual transfer.[1]

**Contributions** Our results demonstrate how transliteration can be used for effective zero-shot transfer even when the source language does not fully cover the phonemes of the target language (§5.2). With only 22 hours of data from transfer

[1]Data, code and models will be aviable at https://github.com/hhnigatu/ASR-via-Translitration



*Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 64–73

languages–which is just 4% of the data size used in prior work–zero-shot transfer through transliteration results in performance gains over existing baselines (§5.3). In addition, with 10 minutes up to one hour of target language data, we find that transliteration offers an effective means for data augmentation, resulting in up to ∼30% absolute reduction in CER compared to augmentation with source language scripts (§5.4).

## 2 Related Work

In this section, we describe prior work on zero-shot ASR, the use of transliteration for ASR transfer, Amharic ASR, and the impact of transfer language selection in cross-lingual ASR.

**Zero-Shot ASR:** Prior work has explored zero-shot cross-lingual ASR, mainly relying on IPA-based transcriptions and measuring Phoneme Error Rates (PER) (Xu et al., 2022) or Phonetic Token Error Rates (PTER) (Żelasko et al., 2020). Cross-lingual settings involve shared acoustic models trained on single or multiple languages and tested on an unseen language(s). However, performance in this zero-shot setting has high error rates, in the 70-90% range (Gao et al., 2021). Prior work has relied on linguistic knowledge to improve zero-shot ASR under these constraints: Xu et al. (2022) mapped phonemes across transfer and target languages based on edit distance between articulatory features to capture Out-Of-Vocabulary (OOV) phonemes in the target language. Gao et al. (2021) improve zero-shot ASR by adding language embeddings to capture "phylogenetic similarity and phone inventory" of the target language, in addition to masking phonetic tokens that do not exist in the target language. However, IPA-based cross-lingual ASR requires mapping back to the original orthography of the target languages when used in user-facing applications. Additionally, PER and PTER do not reflect the performance at the word level, which is the basic unit for many languages.

**Amharic ASR:** Prior works have investigated both zero-shot and supervised ASR systems for Amharic. Tachbelie et al. (2014) found that using morphemes in lexical and language modeling led to improved performance gain for Amharic with GMM-HMM models. In multilingual settings, Whisper (Radford et al., 2022) which contains 32 hours of Amharic speech with translated English corpus reports a 140% WER. MMS (Pratap et al., 2023) which contains Amharic speech data achieved 52.9% WER with CTC decoding and 30.1% WER with an external language model for Amharic. Previous work (Feng et al., 2021) included Amharic in a cross-lingual setting and found that, when using a monolingual 3-gram language model for decoding, the PER for Amharic was 74.8% on the Babel (Gales et al., 2014) data. Żelasko et al. (2020) got a similar performance for Amharic in zero-shot cross-lingual transfer with a PTER of 75.2% on the Babel dataset.

**Transfer Language Selection:** Prior work showed that phonetic similarity of transfer and target languages improves performance (Khare et al., 2021; Tachbelie et al., 2020a). Phonemes that are not shared between transfer and target languages suffer in cross-lingual ASR (Li et al., 2022; Khare et al., 2021). Do et al. (2022) found that languages that had higher Angular Similarity of Phoneme Frequencies (ASPF) scores were better transfer languages for cross-lingual Text-To-Speech (TTS) as compared to selecting a transfer language based on language family. Tachbelie et al. (2020b) used phonetic overlap to select a transfer language for training an acoustic model and test ASR performance on the target language. However, Tachbelie et al. (2020b) used a phonetic dictionary and language model in the target language.

**Transliteration:** When the transfer language orthography is different from the target language, one potential solution is to use transliteration. By transliterating all transcripts in a multilingual setting to a single writing system, models can benefit from cross-lingual transfer more effectively (Datta et al.). Transliteration has also been used as a data augmentation strategy: Khare et al. (2021) found that further pre-training a model on transliterated English data before finetuning on target language data improved performance for all languages in their experiments except Amharic. To the best of our knowledge, zero-shot transfer with transliteration has not been explored.

## 3 Transliteration-Based Zero-Shot ASR for Amharic

As described in the previous section, most previous works on zero-shot ASR are based on phonetic transcriptions, which limits the usability of the resulting ASR system. In addition, previous works show relatively poor performance in zero-

shot Amharic ASR, even as measured in phoneme error rates, compared to other languages. We utilize transliteration as a means to achieve zero-shot ASR directly in the target language orthography. Additionally, we experiment with four transfer languages, looking at phonetic coverage and approximation through transliteration. We experiment with fine-tuning a XLS-R model for zero-shot ASR. Additionally, we experiment with GMM-HMM models with a Language Model (LM) trained in the target language data. We report performance in terms of Word Error Rate (WER), Character Error Rate (CER), and Phone Token Error Rate (PTER).

### 3.1 Source & Target Languages

There are several strategies for selecting transfer languages in cross-lingual speech systems, such as using similarity in unigram phonetic distribution for ASR (Khare et al., 2021), or Angular Similarity of Phoneme Frequencies (ASPF) (Do et al., 2022). Mismatch in phonetic inventories between source and target languages presents a challenge for cross-lingual zero-shot ASR, which degrades performance (§2). We experiment with transfer language (1) from the same language family (Arabic) (2) maximum unigram phonetic coverage (Xhosa), and (3) unrelated higher resourced languages (Spanish and French).

**Target Language: Amharic** is an Afro-Semitic language spoken in Ethiopia. It is written using the Ge'ez script (Adugna, 2023) and has an Abugida[2] writing system, which consists of consonant-vowel sequences written as a unit. Amharic has 38 phonemes (31 consonants and 7 vowels) (Leslau, 2000). It includes glottalized sounds or ejectives[3] that are not found in many higher-resourced languages (Tachbelie et al., 2014).

**Source Language: Arabic** is an Afro-Semitic language, which is the same language family as Amharic. The Arabic language has only three vowels with long and short versions (ara, 2023) and short vowels are not always marked in writing as they are in the form of diacritics (Contributors to Wikimedia projects, 2023).

**Source Language: Xhosa** is a Niger-Congo language spoken in Southern Africa. It uses the Latin script and is known for having a heavy load of click sounds[4]. Xhosa has 30 common phonemes with Amharic, the highest coverage from all of our other transfer languages. Specifically, Xhosa covers the 5 ejective phonemes (k', p',t', /t͡s'/,/t͡ʃ'/ ) in Amharic that are not found in any of the other three transfer languages.

**Source Language: Spanish** is an Indo-European language that uses the Latin script. It has 5 vowels and fewer than 20 consonants (Hualde, 2005); it only covers 21 out of the 38 phonemes in Amharic. Spanish is considered a high-resourced language based on the availability of data, the number of speakers, and the availability of language technologies.

**Source Language: French** is an Indo-European language that also uses the Latin script. It is also considered a high-resourced language. French covers 23 of the 38 phonemes of Amharic.

### 3.2 Transliteration

We transliterate the transfer language transcriptions to the target language script. None of the languages fully cover the phonemes in the target language (see Table 2). There are also phonemes in the source languages that do not exist in Amharic. In both cases, the transliteration process approximates the phonemes to the target language in a way that maximizes coverage; as an example, the Arabic غ /ɣ/ character is transliterated into the Ge'ez 'ግ'/g/. For Xhosa, French, and Spanish we used the `google-transliteration-api`[5] and for Arabic, we built a rule-based transliterator.

| Language | Original Word | Lexicon Entry | Pronunciation |
|---|---|---|---|
| Arabic | رَحِمَ /rahima/ | ራሒማ rahima | ር ኣ ሕ ኢ ም ኣ r a h i m a |
| Xhosa | waguqa /waguk!a/ | ዋጉቃ /waguk'a/ | ው ኣ ግ ኡ ቅ ኣ w a gu k' a |
| Amharic | ጠቀሜታ t'ək'əmeta | ጠቀሜታ t'ək' əmeta | ጥ ኧ ቅ ኧ ም ኤ ት ኣ t'ək' əmeta |

Table 1: **Sample lexicon entries for training (Arabic and Xhosa) and testing (Amharic).** We show both original and IPA transcriptions for readability.

### 3.3 Models

**GMM-HMM** are traditional ASR models, in which the distribution of acoustic features at each time step is modeled as Gaussian Mixture Models (GMMs), and the transitions between phones (or sub-phones) are modeled using HMMs. For inference, a word-level grammar transducer G, a pronunciation lexicon L, context dependency graph C, and learned HMM states H are used to create a WFST graph for decoding. To use this architecture

[2] https://www.omniglot.com/writing/ethiopic.htm
[3] https://wals.info/chapter/7
[4] https://www.omniglot.com/writing/xhosa.htm
[5] pypi.org/project/google-transliteration-api

in zero-shot transfer, we create a training lexicon using the transliterated words from our transfer languages. Each entry in the lexicon consists of a transliterated source language word, along with the sequence of Ge'ez characters which we use in place of phonemes. For the pronunciation lexicon, we split the consonant-vowel sequences of the Ge'ez script so each resulting character represents a single phoneme[6]. Table 1 presents sample lexicon entries. For decoding, we use an Amharic lexicon and language model. Hence, the L and G graphs at test time include words in the target language, which are combined with the H and C graphs trained on the transliterated transfer language data to create our decoding graph. This way, the model is equipped with knowledge of the target language without the need for aligned speech data.

**XLS-R-53** is a self-supervised end-to-end neural acoustic model pre-trained on 56k hours of 53 languages (Conneau et al., 2021). The model can be fine-tuned for speech recognition by adding a linear projection layer and optimizing it using the CTC loss (Conneau et al., 2021). For our proposed transliteration-based zero-shot ASR, we use the audio and transliterated transcripts from the source languages to fine-tune the XLS-R model. Hence, the model is trained to directly predict the graphemes of our target language.

## 4 Experimental Settings

In this section, we describe the datasets we used for our experiments, the training settings for our models and the language combinations we tried.

### 4.1 Datasets

Table 2 shows the datasets we used for each of our transfer languages. For Arabic, the majority of speech data sets do not contain diacritics (Aldarmaki and Ghannam, 2023), which is a shortcoming that may negatively impact the effectiveness of transliteration[7]. Hence, we used the ClArTTS dataset (Kulkarni et al., 2023), which consists of read speech by a single male speaker in Classical Arabic and is transcribed with complete diacritics. To control for the effect of data size on performance, we downsample all train sets to match the size of the smallest set, which is around 12 hours.

For Amharic, we use two publicly available datasets: FLEURS (Conneau et al., 2022) and ALFFA (Tachbelie et al., 2014). FLEURS is a 102-way parallel read corpus of sentences translated from English Wikipedia with about 12 hours of speech per language. The Amharic test set of FLEURS includes 516 utterances. In all our experiments with the FLEURS test set, we ran both the hypothesis and predictions through Whipser's normalizer[8]. ALFFA contains about 20 hours of Amharic speech from the news domain. The transcriptions for ALFFA have been segmented using Morfessor (Creutz and Lagus, 2005) to obtain morphemes; we manually reconstructed the test set transcriptions, which has 359 utterances, as we are interested in word-level performance. We also used the Babel dataset (Gales et al., 2014), which contains scripted phone conversation data, to compare to prior work. We resampled all data to 16 kHz.

### 4.2 GMM-HMM Model

**Training** We trained triphone GMM-HMM models using the Kaldi[9] toolkit on the transliterated Arabic and Xhosa data[10]. As described in §3, we used the transliterated words in the two languages to create the training lexicon. For decoding in Amharic, we created a lexicon using text data in the Amharic language from (Azime and Mohammed, 2021). For experiments using a single source language for training, we used the same training lexicon with transliterated words from both languages to avoid OOV characters in decoding. Hence, phonemes that are not in Arabic but are in Xhosa, for example, would be initialized but not trained.

**Monolingual vs Multilingual Transfer** For our GMM-HMM experiments, we selected the languages with the highest and lowest phonetic coverage with our target language: Arabic and Xhosa. We trained monolingual models using data from each language and multilingual transfer models with data combined from the two languages.

### 4.3 wav2vec2 XLS-R

**Training** As described in §3, we fine-tune XLS-R using transliterated data from our transfer languages. We used the XLS-R 53 model (Conneau

[6]For instance, ራ /ra/ is split into ር /r/ and ኣ /a/ characters.

[7]We performed preliminary experiments without diacritic marks and obtained poor performance.

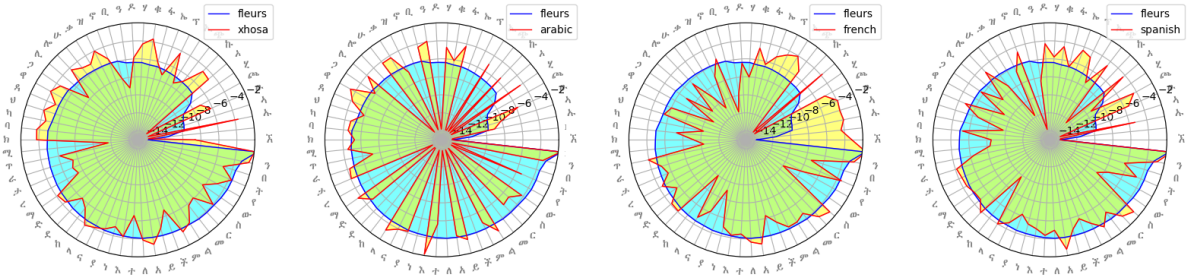
[8]https://pypi.org/project/whisper-normalizer

[9]https://kaldi-asr.org

[10]We experimented with speaker adaptive training (SAT), but found that speaker-independent triphone models perform better. This is in line with prior work (Rouhe et al., 2022) with low-resourced languages using GMM-HMM.

| Language | Language Family | No. of Common Phonemes | Source Dataset | Domain | Hours |
|---|---|---|---|---|---|
| Arabic | Afro-Semitic | 19 | ClArTTS (Kulkarni et al., 2023) | Religious | 12 |
| French | Indo-European | 23 | VoxPopuli (Wang et al., 2021) | Parliament | 211 |
| Spanish | Indo-European | 21 | Common Voice 9.0 (Ardila et al., 2020) | Diverse | 408* |
| Xhosa | Niger-Congo | 30 | NCHLT isiXhosa Speech Corpus (de Vries et al., 2014) | Diverse | 56 |

Table 2: **Comparison of phoneme overlap and datasets used for transfer languages.** Xhosa has the highest number of common phonemes with Amharic, while Spanish has the lowest. The datasets vary by domain and duration, with the Spanish dataset showing the number of validated hours. All datasets were down-sampled to match the size of the Arabic dataset for uniformity.




(a) Xhosa (b) Arabic (c) French (d) Spanish

Figure 1: **Log frequency of characters in the train sets (yellow) compared with the FLEURS test set (blue).** Only characters that have a minimum relative frequency of 0.01 in all sets are included in the visualization.

| | | ALFFA | | FLEURS | |
|---|---|---|---|---|---|
| | | WER | CER | WER | CER |
| GMM-HMM | Arabic | 97.23 | 84.07 | 97.93 | 87.31 |
| | Xhosa | 92.94 | 75.74 | 93.17 | 76.24 |
| | Combined | **92.23** | 75.12 | **93.16** | 77.40 |
| XLS-R | Arabic | 100.33 | 86.67 | 100.10 | 82.72 |
| | Xhosa | 99.91 | 78.70 | 99.91 | 77.88 |
| | Combined | 99.85 | **73.46** | 99.81 | **73.72** |
| XLS-R + LM | Combined | 99.14 | 77.98 | 99.17 | 78.57 |

Table 3: **Zero-shot performance on test sets for Amharic using GMM-HMM and XLS-R models** We report performance on training with Arabic only, Xhosa only, or both (combined) data.

et al., 2021) which has 317M parameters. The model was trained on a total of 56K hours of data from 53 languages, which includes Arabic, French, and Spanish but not Amharic or Xhosa[11]. We experimented with different learning rates [3e-5, 1e-4, 1e-6, 3e-4] and used a linear learning rate scheduler with 500 steps as warmup. We trained for a maximum of 18.5K steps, with early stopping based on the performance on the validation set. All our experiments were conducted on two 24GB Titan RTX GPUs with CUDA Version 11.2.

[11]The model includes Arabic data from Common Voice. While Xhosa is not included, XLS-R training data include Zulu, a related and mutually intelligible language to Xhosa (Spiegler et al., 2010).

**Monolingual vs Multilingual Transfer** We experiment with monolingual transfer where we train on an equal amount of transliterated data from each of the transfer languages separately. This results in four models trained on transliterated transcripts and speech data in each of the transfer languages. Then, we trained on pairs of the four languages resulting in 6 unique pairs for multi-lingual transfer.

**Comparison with GMM-HMM Models** To compare with the GMM-HMM models, we used data from (Azime and Mohammed, 2021) to train a trigram language model for decoding using the SIRLM[12] toolkit. The shallow fusion with this external LM is used only in comparison with GMM-HMM performance, ensuring our results in other settings are fully zero-shot.

## 5 Results

In this section, we report the results of the various experimental settings described above.

### 5.1 Transfer with GMM-HMM

As Table 3 shows, we find that the GMM-HMM models outperform the XLS-R models in zero-shot settings in terms of WER. The XLS-R model, on

[12]http://www.speech.sri.com/projects/srilm

| | | ALFFA | | FLEURS | |
|---|---|---|---|---|---|
| | | WER | CER | WER | CER |
| Monolingual | Arabic | 100.33 | 86.67 | 100.10 | 82.72 |
| | Xhosa | **99.91** | 78.70 | **99.91** | **77.88** |
| | Spanish | 101.67 | **75.25** | 121.72 | 81.46 |
| | French | 116.95 | 87.29 | 140.81 | 84.09 |
| Multilingual | French-Arabic | 98.79 | 85.57 | 100.11 | 87.57 |
| | French-Xhosa | **99.19** | 82.10 | 99.95 | 87.29 |
| | French-Spanish | 99.51 | 73.14 | 105.19 | 74.08 |
| | Spanish-Arabic | 99.87 | 69.70 | 115.06 | 72.71 |
| | Spanish- Xhosa | 99.63 | **69.61** | 103.24 | **70.39** |
| | Arabic-Xhosa | 99.85 | 73.46 | **99.81** | 73.72 |

Table 4: **Performance of models trained on monolingual and multilingual settings.** Models trained on Spanish and Xhosa data significantly outperform the models trained on Arabic and French. Pairing the least-performing transfer languages with the better-performing ones improves performance.

| Training Set | ALFFA | FLEURS |
|---|---|---|
| Arabic | 7.65% | 1.78% |
| French | 11.23% | 4.02% |
| Spanish | 2.05% | 0.08% |
| Xhosa | 3.12% | 0.03% |

Table 5: **Percentage of characters that are not found in the test sets but are found in the training set of the transliterated data.** Each percentage quantifies how much percent of the total number of characters in the total training set the unique characters account for.

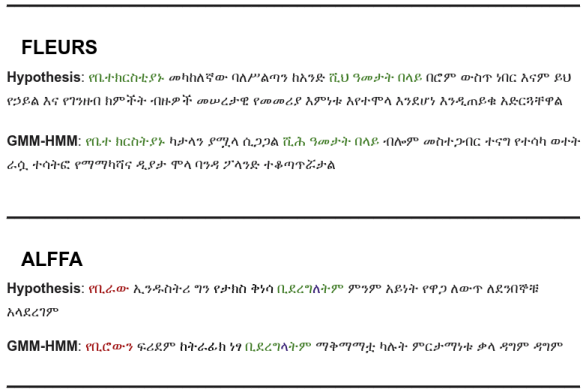



Figure 2: **Samples showing the predictions of the GMM-HMM model trained on Arabic-Xhosa data.** While the full sentences of the predictions do not make sense, highlighted in green are words and characters that the model correctly predicted.

the other hand, achieved similar or slightly better CER, but much higher WER, as the model was unable to predict correct words in the target language. In the GMM-HMM set-up, we enforce the language structure during test time through an Amharic lexicon and an Amharic LM. Adding an external LM for decoding with the XLS-R model improved the WER but only slightly, which shows the advantage of the HMM model where target language structure can be incorporated at decoding time. We see these patterns in examples for both test sets using the Arabic-Xhosa combined model in Figure 2. While the full sentences do not make sense, we see highlighted in green full words that were captured by the GMM-HMM model. In the ALFFA example in Figure 2, we see the first word highlighted in red having a similar sound with the word in the hypothesis but a completely unrelated meaning: the word in the hypothesis says "ye biraw" meaning "The beer" while the word in the prediction says "ye birow" meaning "The bureau." Table 3 also presents results using Arabic-only and Xhosa-only data for training; we find that the Xhosa-only models outperform the Arabic-only models. This is likely due to the higher coverage of Amharic phonemes in Xhosa compared to Arabic. The best performance is achieved when both languages are combined.

## 5.2 Transfer with XLS-R

**Monolingual Transfer** We find that the monolingual XLS-R model trained on Xhosa outperforms all the other models for both dataset, except for CER on the ALFFA dataset where the Spanish-trained model outperforms (see Table 4). The French model is the least-performing model in both settings across both metrics. Interestingly, French performs worse than Spanish despite having a higher overlap with Amharic phonemes. We hypothesize the good performance of the Xhosa model can be explained by the coverage of 30 out of 38 of the Amharic phonemes by Xhosa. Additionally, since we did soft approximation through transliteration (§3.2), we hypothesize that even if the phoneme is not present in the language, the transliteration might still approximate the character

| Method | LM | No. Languages | Train Data Hours | PER/PTER |
|---|---|---|---|---|
| *Prior Work* | | | | |
| Feng et al. (2021) Cross Mono-tg | 3-gram | 12 | 554.40 | 74.80 |
| Żelasko et al. (2020) Cross | None | 12 | 554.40 | 75.20 |
| *Ours* | | | | |
| Xhosa-Arabic | None | 2 | 22 | 76.32 |
| Spanish-Xhosa | None | 2 | 22 | **73.54** |

Table 6: **Comparison of our top two best performing models with prior work reported performance.** With just 4% of training data size and two transfer languages, our best performing model outperforms the reported PTER in zero-shot ASR for Amharic.

representing the phoneme.

To understand the performance gap further, we looked at the distribution of the characters in the test sets and the transliterated training data of each of the languages. Figure 1 shows radar plots of each distribution in terms of log frequencies (the log is used to enable interpretable visualization of the power distribution of characters). Due to the large number of composite characters in the Amharic script, we only show the characters that have a minimum relative normalized frequency of 0.01 in each set. The plots show a clear pattern: both Xhosa and Spanish train sets have better coverage of the frequent characters in the test set. Arabic is missing many of the frequent characters, and French includes a high relative frequency of characters that are infrequent in the test set. As Table 5 shows, both French and Arabic have characters that are not found in the test set that account for a higher percentage of their total number of characters. For example, characters that are in the training set of transliterated French but not in the ALFFA test set account for 11.23% of the total. On the other hand, for both Spanish and Xhosa training sets, the characters that exist in the training set but do not exist in the test set account for less than 4% of the total. This analysis suggests that character distribution plays a larger role than phoneme coverage in zero-shot performance.

**Multilingual Transfer** In testing with models trained by combining two languages, we find that the combination of Spanish and Xhosa gives the best performance, which is expected since the two languages had the top two best performances in the single-language setting. The combination of the least-performing models resulted in an improvement over performance in either of the languages independently for ALFFA ( 86.67% with Arabic only and 87.29% in French only to 85.57% in French-Arabic combined) However, for FLEURS, the performance degraded, with a 3% absolute increase of CER from the French-only model and 5% increase in the CER from the Arabic-only model. We also find that pairing the least performing transfer languages with the better performing languages improves performance on the single-language models: pairing Arabic and Xhosa data reduced CER from the Arabic-only model by a 10% absolute drop for both test sets.

### 5.3 Comparison with Baselines

We compare with two prior works that experiment with Amharic in zero-shot: Feng et al. (2021) trained hybrid DNN-HMM models with training data from 12 phonetically diverse languages and tested cross-lingually on Amharic. Żelasko et al. (2020) trained an end-to-end ASR model with CTC loss on 12 languages and tested on Amharic. Both works train models with IPA transcriptions and report Phone Error Rate (PER) and Phone Token Error Rate (PTER) respectively on the Babel dataset. In Table 6, we show the reported results for our two best models on Babel and compare them to the baselines. Since our models are trained to predict graphemes of the target language, we use LanguageNet grapheme-to-phone (g2p)[13] converter, which is also used in (Żelasko et al., 2020), to covert our model predictions and Babel hypothesis to IPA. We then calculate PTER on the IPA transcriptions. With just 4% of training data size compared to prior work, our best-performing model trained with only two languages outperforms the baselines.

### 5.4 Few-Shot Fine-Tuning

As noted in Section 2, performance in zero-shot cross-lingual ASR typically has high error rates, even at the phoneme level. Hence, we investigate the performance of further fine-tuning of the ASR models in low-resource settings, where we only use 10 minutes to 1 hour of target language data. In this setting, the transliterated data serve as a

[13] https://github.com/uiuc-sst/g2ps

| | | WER/CER without LM | | | | WER/CER with LM | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | ALFFA | | FLEURS | | ALFFA | | FLEURS | |
| 10 minutes | Amharic Only | **101.08** | 79.77 | **101.16** | 78.32 | 99.32 | 84.74 | 98.91 | 81.74 |
| | Source script + Amharic | 101.94 | 78.38 | 104.06 | 76.91 | 99.12 | 81.32 | 99.12 | 78.52 |
| | Transliterated + Amharic | 102.87 | **71.42** | 102.98 | **70.72** | **98.80** | **66.99** | **97.72** | **70.21** |
| 20 minutes | Amharic Only | 100.52 | 80.38 | 101.35 | 80.37 | 99.57 | 83.02 | 99.44 | 80.67 |
| | Source script + Amharic | 101.61 | 69.54 | 100.72 | 68.17 | 99.02 | 71.19 | 97.47 | 68.13 |
| | Transliterated + Amharic | **95.32** | **42.22** | **92.41** | **40.34** | **83.41** | **38.10** | **80.24** | **36.42** |
| 30 minutes | Amharic Only | 101.29 | 74.68 | 100.55 | 73.96 | 98.89 | 79.23 | 98.64 | 78.08 |
| | Source script + Amharic | 99.54 | 51.86 | 98.84 | 49.37 | 91.37 | 50.16 | 89.09 | 46.73 |
| | Transliterated + Amharic | **91.46** | **36.75** | **88.25** | **33.88** | **76.00** | **32.04** | **70.01** | **28.46** |
| 1 hour | Amharic Only | 83.55 | 30.34 | **74.77** | **26.29** | **64.30** | 26.76 | **55.95** | 23.19 |
| | Source script + Amharic | 99.54 | 51.01 | 99.41 | 47.48 | 77.10 | 35.31 | 71.85 | 30.90 |
| | Transliterated + Amharic | **82.57** | **30.19** | 75.51 | 26.54 | 66.67 | **25.77** | 57.91 | **21.47** |

Table 7: **Performance of XLS-R further fine-tuned with small amounts of Amharic data, from 10 minutes to 1 hour.** We compared direct fine-tuning on Amharic data, vs. fine-tuning first with transfer language data, original script, or transliterated script.

form of data augmentation, where the model is first fine-tuned on the source languages, then further optimized on the target language. For these experiments, we used a linear learning rate scheduler with 100 steps as warmup and we trained models with smaller steps depending on data size to avoid overfitting. For comparison, we (1) directly fine-tune XLS-R on the target Amharic data and (2) fine-tune XLS-R with transfer language data without transliteration then further fine-tune on Amharic data. The results are shown in Table 7.

We note how further fine-tuning with small amounts of supervised data in the target language results in significant performance improvements. With 20 minutes or more, we observe large reductions in error rates. We observe that the model trained on the transliterated data outperforms both the model trained on original source transcripts (up to 30% absolute reduction in CER) as well as the model directly fine-tuned on Amharic data alone. The performance gap between the three setups is most pronounced as the data is smaller, indicating the benefits of using transliteration with carefully selected transfer languages for low-resource ASR. Compared to zero-shot, we observe roughly 40% and 10% absolute reduction in CER and WER, respectively with 30 minutes of Amharic data.

## 6 Discussion

Our experiments show how to use transliteration for zero-shot transfer in low-resourced settings. With just a fraction of the training data size compared to prior work, our best-performing model outperforms the reported performance on Amharic in a zero-shot setting. Additionally, by training on transliterated data, we predict directly in the target language orthography.

Error rates in zero-shot ASR are generally high for direct use of the systems (Gao et al., 2021). However, zero-shot approaches give us insights to how to best select transfer languages when we have limited data available. In line with prior work, we find that languages that have high unigram phonetic coverage with the target language are better transfer languages. Further, we find that through soft approximation via transliteration, even languages that do not have high phonetic coverage can be good transfer languages. Our analysis reveals that transfer languages with the least post-transliteration rate of Out-Of-Vocabulary (OOV) characters in the target test set perform best as transfer languages, regardless of their language family or degree of inherent phonetic coverage.

In zero-shot settings, GMM-HMM models result in significantly lower WER, which is ascribed to the fact that the models incorporate the target language lexicon in decoding, unlike the end-to-end models that lack such linguistic knowledge without supervised training. However, CER is much lower using the XLS-R model. In low-resource settings, with 10 minutes to 1 hour of training data in the target language, transliteration results in improved performance compared to direct fine-tuning on the target language or using the transfer languages without transliteration.

## 7 Conclusion

In this study, we explored the use of transliteration for zero-shot and low-resource cross-lingual ASR transfer. We find that, with careful selection of source languages, using $\sim$22 hours of source data, we can build zero-shot ASR systems that can transcribe words directly in the target language orthography. With small amounts of transcribed data

in the target language, large reductions in error rates can be achieved through using transliteration for data augmentation.

## Limitations

While our results show promising results for zero-shot transfer for Amharic, there are several avenues for improvement. First, the Arabic and French data are domain-limited. The Arabic data is further constrained by having a single speaker. As discussed in Section 4.1, we could not find multi-speaker diverse domain data with diacritic markers for Arabic. While this is a limitation, it is also reflective of the real state of building language technologies for low-resourced languages. Our current work explores how far we can go with data and tools that are currently available to us in a low-resourced setting. For future work, we will explore using automated methods for adding diacritic markers to existing Arabic datasets. Additionally, we were limited to trying multilingual transfers with just two transfer languages due to compute resource constraints. However, our results still demonstrate that our transliteration-based approach outperforms the previously reported performance for zero-shot ASR for Amharic. Future work can explore adding more languages and trying more combinations of languages in the multi-lingual setting. Additionally, our work focused on just one target language; future work could explore our approach on more languages.

## References


2023. The Arabic Alphabet: Vowels. [Online; accessed 14. Dec. 2023].

Gabe Adugna. 2023. Research: Language Learning - Amharic: Home. [Online; accessed 14. Dec. 2023].

Hanan Aldarmaki and Ahmad Ghannam. 2023. Diacritic Recognition Performance in Arabic ASR. In *Proc. INTERSPEECH 2023*, pages 361–365.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Israel Abebe Azime and Nebil Mohammed. 2021. An amharic news text classification dataset. In *AfricaNLP*.

Martijn Bartelds, Nay San, Bradley McDonnell, Dan Jurafsky, and Martijn Wieling. 2023. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation. *ACL Anthology*, pages 715–729.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition.

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*.

Contributors to Wikimedia projects. 2023. Arabic diacritics - Wikipedia. [Online; accessed 14. Dec. 2023].

Mathias Creutz and K. Lagus. 2005. Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0.

Arindrima Datta, Bhuvana Ramabhadran, Jesse Emond, Anjuli Kannan, and Brian Roark. Language-Agnostic Multilingual Modeling. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 04–08. IEEE.

Nic J. de Vries, Marelie H. Davel, Jaco Badenhorst, Willem D. Basson, Febe de Wet, Etienne Barnard, and Alta de Waal. 2014. A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, 56:119–131.

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. *ACL Anthology*, pages 16–22.

Siyuan Feng, Piotr ´Zelasko, Laureano Moro-Vel´azquez, Ali Abavisani, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2021. How Phonotactics Affect Multilingual and Zero-Shot ASR Performance.

M. Gales, K. Knill, A. Ragni, and S. Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at CUED. *Workshop on Spoken Language Technologies for Under-resourced Languages*.

Heting Gao, Junrui Ni, Yang Zhang, Kaizhi Qian, Shiyu Chang, and Mark Hasegawa-Johnson. 2021. Zero-Shot Cross-Lingual Phonetic Recognition with External Language Embedding.

José Ignacio Hualde. 2005. The sounds of spanish.

Jacob Kahn, Ann Lee, and Awni Hannun. Self-Training for End-to-End Speech Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 04–08. IEEE.

Shreya Khare, Ashish R. Mittal, Anuj Diwan, Sunita Sarawagi, P. Jyothi, and Samarth Bharadwaj. 2021. Low Resource ASR: The Surprising Effectiveness of High Resource Transliteration. *Interspeech*.

Ajinkya Kulkarni, Atharva Kulkarni, Sara Abedalmon'em Mohammad Shatnawi, and Hanan Aldarmaki. 2023. ClArTTS: An Open-Source Classical Arabic Text-to-Speech Corpus. In *Proc. INTERSPEECH 2023*, pages 5511–5515.

Wolf Leslau. 2000. *Introductory grammar of Amharic*, volume 21. Otto Harrassowitz Verlag.

Xinjian Li, Florian Metze, David R. Mortensen, Alan W. Black, and Shinji Watanabe. 2022. ASR2K: Speech Recognition for Around 2000 Languages without Audio.

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. 2023. Scaling Speech Technology to 1,000+ Languages. *arXiv*.

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

Aku Rouhe, Anja Virkkunen, Juho Leinonen, and Mikko Kurimo. 2022. Low Resource Comparison of Attention-based and Hybrid ASR Exploiting wav2vec 2.0. In *Proc. Interspeech 2022*, pages 3543–3547.

Sebastian Spiegler, Andrew van der Spuy, and Peter A. Flach. 2010. Ukwabelana - An open-source morphological Zulu corpus. *ACL Anthology*, pages 1020–1028.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Laurent Besacier. 2014. Using different acoustic, lexical and language modeling units for ASR of an under-resourced language – Amharic. *Speech Communication*, 56:181–194.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020a. Analysis of GlobalPhone and Ethiopian languages speech corpora for multilingual ASR. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4152–4156, Marseille, France. European Language Resources Association.

Martha Yifiru Tachbelie, Solomon Teferra Abate, and Tanja Schultz. 2020b. DNN-based multilingual automatic speech recognition for Wolaytta using Oromo speech. 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 265–270, Marseille, France. European Language Resources association.

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 the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2022. Simple and Effective Zero-shot Cross-lingual Phoneme Recognition.

Piotr Żelasko, Laureano Moro-Velázquez, Mark Hasegawa-Johnson, Odette Scharenborg, and Najim Dehak. 2020. That Sounds Familiar: An Analysis of Phonetic Representations Transfer Across Languages.