CROSS-LINGUAL TRANSFER FOR SPEECH PROCESSING USING ACOUSTIC LANGUAGE SIMILARITY

Peter Wu, Jiatong Shi, Yifan Zhong, Shinji Watanabe, Alan W Black

Carnegie Mellon University, PA, USA
peterw1@cs.cmu.edu

ABSTRACT

Speech processing systems currently do not support the vast majority of languages, in part due to the lack of data in low-resource languages. Cross-lingual transfer offers a compelling way to help bridge this digital divide by incorporating high-resource data into low-resource systems. Current cross-lingual algorithms have shown success in text-based tasks and speech-related tasks over some low-resource languages. However, scaling up speech systems to support hundreds of low-resource languages remains unsolved. To help bridge this gap, we propose a language similarity approach that can efficiently identify acoustic cross-lingual transfer pairs across hundreds of languages. We demonstrate the effectiveness of our approach in language family classification, speech recognition, and speech synthesis tasks.

Index Terms— cross-lingual, zero-shot, ASR, TTS

1. INTRODUCTION

As speech assistants and other speech-based technologies become more prevalent, it is important to make them accessible to everyone around the world. While there are thousands of languages in the world, these technologies currently only support a small subset of these languages. Since state-of-the-art speech processing algorithms typically require large training corpora, it is difficult, if not infeasible, to currently build them for low resource languages[1, 2].

Cross-lingual transfer is a promising direction that aims to bridge this gap by incorporating high-resource data into low-resource systems [3, 4]. For such approaches, choosing the data to transfer from is an important step [5, 6, 7, 8]. While many solutions for this step exist for text-based tasks, generalizable solutions still do not exist for speech-based ones [9]. In this work, we help bridge this gap by proposing language similarity metrics that pair target languages with source languages suitable for acoustic cross-lingual transfer. We find our approach effective in language family classification, speech recognition, and speech synthesis. Thus, our contributions are the following:

1. we introduce an acoustic language similarity approach,

2. we compare it with non-acoustic ones in downstream cross-lingual speech recognition and speech synthesis tasks, and

3. we demonstrate the usefulness of our acoustic approach for both language family classification and acoustic cross-lingual transfer.

We proceed by discussing cross-lingual transfer techniques in Section 2. In Section 3, we discuss current non-acoustic language similarity approaches. Then, we describe our proposed acoustic language similarity approaches and their distinction from non-acoustic ones in Section 4. Sections 5, 6, and 7 contain our language family classification, speech recognition, and speech synthesis experiments, respectively. Finally, we summarize our results and propose future directions in Section 8. All our code and other supplementary material can be found at https://github.com/peter-yh-wu/cross-lingual.

2. CROSS-LINGUAL TRANSFER

2.1. Problem Statement

Given the set of all languages S, cross-lingual transfer generally refers utilizing n source languages $\{s_1, \ldots, s_n\} \subset S$ to improve task performance in target language $t \in S$, where t is typically low-resource [9]. Assuming that we are using a data-driven approach, data D_i would need to be chosen for each source language s_i. This definition reveals three challenging tasks: (1) choosing source languages $\{s_i\}$ from S; (2) choosing data D_i for each s_i; (3) transferring the information in $\{D_1, \ldots, D_n\}$ to the task in target language t. We focus on the **first** challenge in this work, and discuss its relation to other tasks below.

2.2. Cross-lingual Techniques

Many speech- and text-based works have approached the aforementioned third challenge of improving performance in target language t using source data $\{D_i\}$ [3, 10, 11, 12, 4, 7, 13, 5, 14]. To handle the second challenge of choosing the D_i's, many works select all the source data from the same corpus [5, 13] or an arbitrary set [15, 12, 4]. Multiple works have approached the first challenge of choosing source languages $\{s_i\}$ in a similar manner [5, 13, 4, 15, 12]. These algorithms that utilize vast amounts of source language data are generally pre-training ones, and thus can be used in conjunction with our approach [3, 11, 12, 4, 16].

Many works have shown that certain subsets of source data are more suitable than others for cross-lingual transfer [5, 8]. At the dataset level, Li et al. [6] propose an approach to select suitable corpora by iteratively sampling from each dataset and gradually attending to less data using similarity scores. Since this work assumes a given initial set of candidate languages and datasets, it addresses the second challenge of selecting the data $\{D_i\}$. The output of our approach, which selects languages $\{s_i\}$, can thus be used as an input to theirs.

Within the text modality, Lin et al. explore using phylogenetic and typological features to select source languages [9]. Since speech data contains a wealth of additional information, we propose acoustic-based features and compare them with their features in speech recognition and speech synthesis. As discussed in Sections 5, 6, and 7, we observe unique advantages of our features in speech

processing tasks. We describe all of these features and how we utilize them to measure language similarity below.

2.3. Language Similarity

In this work, we explore eight different language similarity approaches to select cross-lingual transfer languages for speech processing tasks. Namely, we focus on addressing the first challenge posed in Section 2.1 of choosing source languages $\{s_i\}$. All eight approaches are dataset-independent. They can be used in any cross-lingual transfer task as a compact look-up table without requiring additional computation. We proceed by detailing non-acoustic similarity approaches in Section 3 and our proposed acoustic ones in Section 4.

3. NON-ACOUSTIC LANGUAGE SIMILARITY

3.1. Non-Acoustic Approaches

For our non-acoustic language similarity approaches, we use the six distance metrics from Lin et al. [9, 17]: genetic, inventory, syntactic, phonological, featural, and geographic. Sections 3.2 and 3.3 describe the background and details of the first five approaches. Since we observed that their geographic distance method would sometimes map languages to the same location, we utilize an additional geographic measure, as discussed in Section 3.4.

3.2. Language Family Trees

Multiple sources have devised ways to categorize languages into families, using genealogical, typological, and other linguistic information [18, 19, 20]. Several works categorize languages under a hierarchy of language families, which we can visualize as a directed acyclic graph where family nodes have languages and subfamilies as children. Since categorization decisions differ between linguists, these language family trees can vary noticeably between sources, as shown in Figure 1. The large number of different linguistic attributes thus offers us many different ways of defining language similarity, which we discuss below.
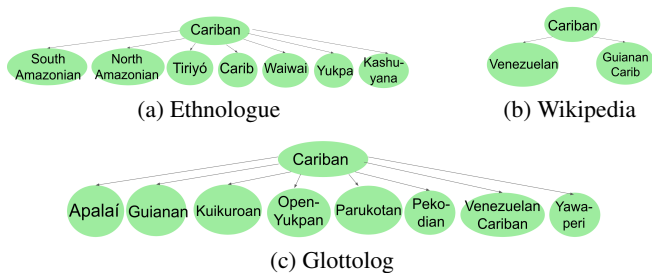

Fig. 1. Top two levels of the Cariban language family tree based on three different linguistics data sources [18, 21, 19].

3.3. Phylogenetic and Typological Distances

As in Lin et al. and the URIEL typological database [9, 17], our five language distance metrics that incorporate phylogenetic and typological information are:

1. **Genetic Distance:** The distance between languages in Glottolog's world language family tree [19]. For reference, a subset of this tree is depicted in Figure 1 above.

2. **Inventory Distance:** The cosine distance between binary vectors extracted from the PHOIBLE database [22]. Each vector dimension has phonetic information about the presence of sounds [17].

3. **Syntactic Distance:** The cosine distance between syntax feature vectors utilizing information from the World Atlas of Language Structures (WALS) and Syntactic Structures of World Languages (SSWL) [20, 23].

4. **Phonological Distance:** The cosine distance between vectors containing phonological information from Ethnologue and WALS [18, 20].

5. **Featural Distance:** The cosine distance between vectors incorporating features from the above four [17].

For these five language similarity approaches, we use the pre-computed lang2vec distances provided in URIEL [17]. The genetic distance between any two languages is a real value in $[0, 1]$, and all the aforementioned vectors are element-wise non-negative with norm greater than 0. Thus, since cosine distance is given by

$$d_{cos}(x, y) = 1 - \frac{x \cdot y}{\|x\| \, \|y\|}, \tag{1}$$

the other four distance functions have range $(0, 1]$.

3.4. Geographic Distance

The sixth non-acoustic language similarity approach that we use is geographic distance, which we measure in two ways. Our first way is via the orthodromic distance approach described in Lin et al. and URIEL, using pre-computed lang2vec distances [9, 17]. We observed that languages may be mapped to the same location, and thus utilize an additional geographic distance measure. Namely, we obtain language coordinates using Wilderness metadata and compute the geodesic distance with the GeoPy library [24, 25].[1] Unless stated otherwise, we report results in this work using the first approach.

4. ACOUSTIC LANGUAGE SIMILARITY

4.1. Acoustic Approaches

We propose two acoustic-based approaches for measuring language similarity. In contrast to the non-acoustic approaches in Section 3, these methods are entirely data driven and thus do not rely on expert linguistic knowledge. Additionally, they leverage information directly from speech data, offering opportunities to learn language properties suitable for downstream speech tasks. Our acoustic-based approaches are motivated by i-vector and x-vector language recognition works [26, 27]. While these works study tens of languages, we show that ours can successfully support hundreds. As far as we are aware, this work is also the first to compare speech vectors with other language similarity approaches and systematically study their uses in downstream speech tasks.

Generally, both involve training a neural network to classify languages given multilingual speech data. We generate a vector embedding for a language by averaging the outputs of a designated embedding layer across speech samples from that language. Then, we measure language distance as the distance between the language embeddings. The primary difference between our approaches is that one aligns its speech representations with another space, as detailed below.

---

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

Formally, our approaches are each defined by a model M and a size-N dataset $D = \{(x_n, y_n)\}_{n=1}^N$, where the x's are model inputs and the y's are language labels. x is a speech representation in our first approach and a (speech, text) pair in our second. Let $M_e(x)$ be the output of the embedding layer, and $M_c(M_e(x))$ be the output of the language classification layer. In our first approach, we train M on a subset $D_{train} \subset D$ using a classification loss $f_c(M_c(M_e(x)), y)$. Then, we generate an embedding e_l for language l using samples $D_l \subset \{x_n | y_n = l \wedge (x_n, y_n) \in D\}$. We compute e_l as the mean of the embedding layer outputs, or $\frac{1}{|D_l|}\sum_{x_n \in D_l} M_e(x_n)$. In this work, we measure language similarity using cosine distance, given in Equation 1. We refer to this approach as the **speech-based approach**.

For our second approach, we add a term to our loss function and assume each x_n is a pair (v_n, w_n) of speech and text data. Our new loss is given by $f(x, y) = f_c(M_c(M_e(v)), y) + f_a(M, x, y)$, where f_a is an alignment loss. In this work, we train a text encoder T jointly with M and define f_a as

$$f_a(M, x, y) = f_c(M_c(T(w)), y) + \alpha d_{cos}(M_e(v), T(w)), \quad (2)$$

where d_{cos} is the cosine distance function in Equation 1 and α is a hyperparameter. We create language embeddings in a similar manner as the first approach, feeding v into M_e instead of x. We refer to this approach as the **multimodal-based approach** and discuss specific architectures used for both approaches in Section 4.3.

4.2. Acoustic Dataset

We use the Wilderness dataset to train the neural models in our acoustic similarity approaches from Section 4.1 [24]. This dataset is suitable for our acoustic approaches since it contains parallel speech and text data for over 700 languages. In this work, we focus on using data that are considered as being good or very good quality in Wilderness paper. Namely, we choose languages with data that could train a random forest Clustergen speech synthesizer with Mel-cepstral distortion (MCD) under a threshold c [28, 29, 30, 31]. We let $c = 5.5$ since this is the middle value of the good quality range used in the Wilderness work [24]. As discussed in that paper, MCD serves as an objective measure of synthesis quality, which in turn has been shown to correlate with alignment and dataset quality [24, 32, 33]. Among this subset of languages, we choose the low-resource ones, which we define as those with limited amounts of public data outside of Wilderness. This yields 195 languages, with 22.8 ± 5.9 hours of speech per language. Figure 2 visualizes the location of these languages.

Fig. 2. Chosen Wilderness languages, colored by Ethnologue language family [24, 18].

4.3. Models for Acoustic Approaches

For both acoustic approaches proposed in Section 4.1, we use a convolutional network based on VGGVox as our speech encoder M_e, with an output dimension of 512 [34] and two-dimensional adaptive max pooling. Each input to the speech encoder is an 80-channel normalized Mel-scaled spectrogram with an FFT window length of 800,

hop size of 200, window size of 800, and sampling rate of 16000 Hz. For all experiments, we use normalized embeddings as in Khosla et al. [35], Adam optimization with learning rate 10^{-3}, and batch size 128. We explore two classification losses for our speech-based approach: cross-entropy and SupCon [35], referred to as CE and SC in Tables 3, 4, and 5. Classifier M_c is a ReLU activation followed by a fully connected (FC) layer with output dimension 195 in the former, and an FC layer with 128 outputs in the latter. We replace the original SupCon augmentations with SpecAugment [36].

For the multimodal-based approach described in Section 4.1, we use an LSTM-based model as our text encoder T. Specifically, we feed each input into an embedding matrix with embedding dimension of 256, followed by an LSTM with hidden dimension 128. We then apply max pooling across the sequence dimension and feed the resulting vector into an FC layer with output dimension 128. We chose max pooling instead of a more complex approach like attention, since the latter did not improve performance in the language family classification task in Section 5. To standardize text inputs across languages, we romanize them using UniTran, as described in the Wilderness paper [24, 37]. For all experiments with the multimodal-based approach, we set the alignment hyperparameter α to $3 * 10^{-2}$. Each model takes a few GPU days to train and a few GPU hours to extract all the language embeddings on an RTX 2080 Ti.

To check whether our language embeddings are interpretable, we compare their k-means clustering assignments with their geographic locations. Here, we train the classifier for the speech-based approach described in Section 4.1 on a random 80%-10%-10% train-val-test split of our 195-language Wilderness subset using cross-entropy. For each language, we make an embedding using all of the test data in that language. Figure 3 colors the locations of these languages based on their k-means clustering assignments. Since assignments to the same cluster generally seem close on the map, it appears that our embeddings contain some degree of geographical information. Additionally, Figure 2 and Figure 3 share many boundaries where language colorings change, suggesting that our embeddings capture some language family information as well.

Fig. 3. k-means assignments of language embeddings extracted from speech data, where $k = 5$. More details are in Section 4.3.

5. LANGUAGE FAMILY CLASSIFICATION

In addition to computing language similarity, our acoustic approaches can also be used for classifying language families from speech. We explore this through a zero-shot task using our 195-language Wilderness subset from Section 4.2 and our two neural classifiers from Section 4.3. More specifically, we separate the languages into a random 80%-10%-10% train-validation-test split, yielding 156, 19, and 20 languages. Then, we train our models to classify the 156 training languages. During evaluation, we map the predicted language to its Ethnologue language family and compare whether the family matches that of the ground truth language. Validation and test accuracy are calculated over the 24 Ethnologue language families containing our 195 languages [18], and we select the best models based on validation accuracy. We also evaluate an

Table 1. Language family classification from speech on unseen languages. More details are in Section 5.

Approach\Split	Val	Test
Random	0.00	0.00
Multimodal	0.29	0.22
Speech	0.25	0.20

untrained version of our Section 4.3 speech-based classifier, called the random baseline in Table 1.

The two trained models have noticeably higher test set accuracies than the random baseline, indicating that the models learned to perform zero-shot language family classification. Also, the multimodal-based approach outperforms the speech-based one, suggesting that the former encoded more linguistic content into its embedding space by aligning with textual representations. While speech-based classification has been widely explored [38, 39], we believe that this is the first work to indicate its potential for classifying language families.

6. SPEECH RECOGNITION

6.1. ASR Task

We also analyze the performance of our language similarity approaches in a cross-lingual speech recognition (ASR) task. Namely, we train monolingual ASR models on a set of source languages and evaluate them on a set of target languages. Thus, experiments using source languages different from target ones are zero-shot. For all ASR experiments, we romanize text using Unidecode[2] and evaluate performance using character error rate (CER). We note that the transliteration helps to make zero-shot ASR possible. To measure the suitability of each language similarity approach for cross-lingual ASR, we calculate the Spearman correlation between the language similarities and the CERs. Details on the specific sets of source-target language pairs used are described below. As mentioned in Section 2.1, we focus here on how to choose the source languages rather than how to improve the subsequent steps in the cross-lingual transfer procedure.

Unless mentioned otherwise, we use an ESPNet Transformer ASR model comprised of a 12-block Transformer encoder, a 6-block Transformer decoder, and a CTC module [40, 41]. Additionally, we use a byte pair encoding vocabulary size of 1000, decoding beam size of 10, and an optimization procedure similar to that of Vaswani et al. [42]. Further details for reproducibility are provided in the accompanying code.

In addition to data from Wilderness [24], we also conduct ASR experiments using other public datasets. Namely, we use Javanese and Sundanese data from Kjartansson et al. [43], Iban data from Juan et al. [44], and Indonesian, Hindi, Vietnamese, and Hakha Chin data from Common Voice Corpus 6.1 [45]. We chose these datasets since they are all public, their languages are relatively near each other, and most of these languages are low-resource. Train-validation-test splits follow those defined in the respective works [43, 44, 45]. We train our cross-lingual models using the source language training and validation sets and evaluate them on the target language test sets. Further details on how we use these datasets are described below.

---

[2]https://pypi.org/project/Unidecode

6.2. Motivation for Cross-lingual ASR

Our first ASR experiment compares cross-lingual models with monolingual and multilingual ones, all evaluated on the Indonesian and Iban test sets described in Section 6.1. We train each monolingual model using data from the same dataset as the respective test set. Each cross-lingual model trains on a monolingual dataset in a language different from the test ones. Here, these monolingual datasets are the Javanese dataset by Kjartansson et al. [43], LibriSpeech [46], and GigaSpeech [47]. Our two multilingual models are trained on the 8-language MLS [48] and the 52-language version of Hou et al. [4, 40]. The training data for all models are the same as those defined by the original works. The monolingual and Javanese models are described in Section 6.1, and the rest are Conformers and ESPnet Transformers detailed in the ESPNet Model Zoo [49, 40, 41].[3] Table 2 contains the CERs of all models on the two test sets.

By leveraging more training data, multiple models outperform the monolingual ones, which train on the training sets in the target datasets. Also, despite using less training data than multilingual and other cross-lingual approaches, Javanese performs the best on both test sets. Since Javanese is closer to the target languages than the other sources, these results suggest the usefulness of leveraging data from similar languages to build ASR systems for low-resource languages. The language similarity approaches in Sections 3 and 4 offer ways to automatically identify such similar languages. For example, Javanese is closer to Indonesian than English for both our acoustic measures and the majority of the non-acoustic ones. Thus, we proceed to study the correlation between language similarity and performance in downstream speech tasks.

Table 2. ASR performance (CER) of monolingual, cross-lingual, and multilingual models evaluated on Indonesian and Iban test sets [45, 44]. Row 1 contains the monolingual ASR CERs. For reference, we provide the amount of training data for each model. More details are in Section 6.2.

Training Data	Hrs.	Indo.	Iban
Original Set [45, 44]	< 10	54.7	75.7
Javanese [43]	300	**35.4**	**50.4**
LibriSpeech [46]	1,000	74.2	56.3
GigaSpeech [47]	10,000	55.3	59.8
MLS [48]	50,000	70.5	62.0
OpenLI52 [4]	5,000	41.8	50.5

6.3. Wilderness Zero-Shot ASR

To compare the language similarity methods in Sections 3 and 4, we first perform zero-shot ASR using Javanese and Sundanese as our source languages and Wilderness languages as our targets. Here, we randomly select ten languages from our 195-language Wilderness subset and evaluate each language using all of its data.[4] Table 3 contains the Spearman correlations between ASR performance (CER) and different similarity measures, where each number uses a single source language and all the targets. Our multimodal-based similarity approach performs the best overall, followed by the genetic, speech-based, and geographic measures. Also, our speech-based approaches perform about the same. This suggests that cross-entropy

---

[3]https://github.com/espnet/espnet_model_zoo
[4]Wilderness ASR 6-letter codes are ACCIBS, AGUNVS, BOATBL, CSOTBL, FRDWBT, JACWBT, NODWBT, POHPOC, STNBSP, and VUNBST.

Table 3. Spearman correlations between language similarity measures and cross-lingual ASR performance (CER), using Javanese and Sundanese as source languages. More details about Wilderness (Wild.) and non-Wilderness (Non-W.) experiments are in Sections 6.3 and 6.4.

| Similarity | Javanese | | Sundanese | |
	Wild.	Non-W.	Wild.	Non-W.
Syntactic	−0.70	0.52	0.12	0.14
Geographic	0.65	0.68	0.29	0.72
Phonological	0.00	0.04	−0.08	0.31
Genetic	0.70	**0.81**	**0.52**	0.45
Inventory	−0.25	0.43	0.04	0.39
Featural	−0.35	0.54	0.12	0.22
Multimodal	**0.85**	0.43	0.44	0.29
Speech (SC)	0.71	0.36	0.25	0.64
Speech (CE)	0.73	0.64	0.27	**0.79**
Ensemble	**0.93**	0.79	0.50	0.75

Table 4. Spearman correlations between language similarity measures and fine-tuned cross-lingual ASR performance (CER), using Javanese and Sundanese as source languages. More details are in Section 6.5.

| Similarity | Javanese | | Sundanese | |
	Wild.	Non-W.	Wild.	Non-W.
Syntactic	−0.45	0.48	0.26	−0.04
Geographic	0.77	0.42	0.65	0.72
Phonological	−0.53	0.34	−0.17	0.11
Genetic	0.66	0.58	0.65	0.34
Inventory	−0.60	0.21	−0.07	0.29
Featural	−0.67	0.42	−0.32	0.04
Multimodal	**0.93**	**0.64**	**0.77**	0.18
Speech (SC)	0.84	0.50	0.54	0.61
Speech (CE)	0.84	0.43	0.43	**0.75**
Ensemble	0.90	**0.64**	**0.77**	0.71

may be sufficient to train our acoustic measure, potentially since its training dataset is large and consists of non-noisy labels. Another benefit of cross-entropy is its low computational costs during training relative to objectives like contrastive loss. To explore the effects of ensembling, we combined the best linguistic language similarity measure for each source language with our best acoustic one by rescaling their language similarities to [0, 1] and then averaging the rescaled values. This combination performs the best overall, suggesting the usefulness of leveraging both non-acoustic and acoustic information. We note that both our speech-based approach and the target languages here use the 195-language Wilderness subset described in Section 4.2. Thus, in the next section, we study how well our similarity measure generalizes to cases where both the source and target language data are not from Wilderness.

6.4. Non-Wilderness Zero-Shot ASR

To study the generalizability of our speech-based similarity measure, we perform cross-lingual ASR using source and target language data that are both not from Wilderness. Namely, our source languages are Javanese and Sundanese, and our target languages are these two plus Iban, Indonesian, Hindi, Vietnamese, and Hakha Chin [43, 44, 45]. Section 6.1 provides more information about these data. Table 3 contains the Spearman correlations between ASR performance (CER) and different language similarity measures, where each number uses a single source language and all the targets. The cross-entropy speech-based approach performs the best overall, followed by the geographic, genetic, and SupCon speech-based measures. These approaches also performed well in the Wilderness ASR task discussed in Section 6.3. Moreover, ensembling in the same manner as that task also yields the highest average correlation, suggesting that these trends may hold even in non-Wilderness data settings.

6.5. ASR with Fine-tuning

We also explore the performance of our similarity measures with fine-tuned ASR models. For each source-target pair in our zero-shot experiments above, we fine-tuned the ASR model on a random ten-minute subset of the target language dataset that is disjoint from the evaluation data. We again observe that the speech-based, multimodal-based, geographic, and genetic measures perform the best, as well as the ensemble having the highest mean correlation.

7. SPEECH SYNTHESIS

7.1. TTS Task

We also analyze our similarity measures in cross-lingual text-to-speech (TTS) tasks. Namely, we train monolingual TTS models on a set on source languages and evaluate them on a set of target languages. Thus, experiments using source languages different from target ones are zero-shot. Section 7.2 describes our experiments using a statistical model with Wilderness data, and Section 7.3 describes our experiments using a neural model with non-Wilderness data. Similarly to our ASR experiments in Section 6, we romanize text using Unidecode, which helps make zero-shot TTS possible. Details on evaluation metrics and the specific sets of source-target language pairs used are described below. As mentioned in Section 2.1, we focus here on how to choose the source languages rather than how to improve the subsequent steps in the cross-lingual transfer procedure.

Table 5. Spearman correlations between language similarity methods and TTS performance, measured with MCD for the Wilderness experiment and MOS for the non-Wilderness one. A more positive correlation with MCD is better, and a more negative correlation with MOS is better. More details are in Sections 7.2 and 7.3.

| Similarity | Sim. vs. MCD ↑ | Sim. vs. MOS ↓ | |
		Hindi	Telugu
Syntactic	0.32	−0.76	−0.30
Geographic	0.27	−0.82	−0.50
Phonological	−0.03	−0.65	−0.95
Genetic	0.40	−0.80	−0.53
Inventory	0.12	−0.67	−0.70
Featural	0.33	−0.65	−0.71
Multimodal	0.50	**−0.87**	**−1.0**
Speech (SC)	**0.53**	**−0.87**	**−1.0**
Speech (CE)	0.47	**−0.87**	−0.90
Ensemble	0.52	**−0.97**	**−1.0**

7.2. Wilderness Statistical Speech Synthesis

We first check the suitability of our similarity measures for cross-lingual TTS with Wilderness data [24]. Namely, we randomly select 15 languages from our 195-language Wilderness subset described

Table 6. MOS ratings of cross-lingual Indic TTS models, where Hindi and Telugu are our target languages. We also provide the distances calculated by our top six language similarity measures from Table 5. More details are in Section 7.3.

| Source | Hindi | | | | | | | Telugu | | | | | |
	MOS	SC	Mul.	CE	Pho.	Inv.	Fea.	MOS	SC	Mul.	CE	Pho.	Inv.	Fea.
Hindi	**4.8 ± 0.4**	0.00	0.00	0.00	0.00	0.00	0.00	3.4 ± 0.9	0.43	0.57	0.33	0.30	0.31	0.40
Kannada	2.8 ± 1.1	0.05	0.10	0.07	0.30	0.44	0.40	4.2 ± 0.7	0.42	0.56	0.28	0.00	0.36	0.40
Marathi	2.8 ± 1.5	0.12	0.11	0.14	0.59	0.40	0.50	2.9 ± 0.9	0.55	0.65	0.33	0.64	0.36	0.40
Tamil	1.5 ± 0.5	0.15	0.24	0.23	0.59	0.47	0.50	1.9 ± 0.3	0.58	0.83	0.41	0.64	0.43	0.40
Telugu	2.1 ± 1.2	0.43	0.57	0.33	0.30	0.31	0.40	**4.8 ± 0.4**	0.00	0.00	0.00	0.00	0.00	0.00

in Section 4.2 and use all 15 as both sources and targets, yielding 225 source-target language pairs.[5] We found that the data in each language was insufficient to build a neural TTS model [1], and thus instead used the random forest Clustergen speech synthesizer as our TTS model [28, 29]. Our training steps are the same as those in the Wilderness paper, and we evaluate models on the first chapter texts in each target language [24]. As discussed in Section 4.2, we measure TTS performance objectively using Mel-cepstral distortion (MCD).

Table 5 shows the Spearman correlations between TTS performance (MCD) and the language similarity measures. For each language similarity approach, we calculate the mean of 15 Spearman correlations, each using one source and all the targets as in Section 6. Our acoustic approaches described in Section 4.1 perform the best, followed by the genetic distance measure. We note that the trends observed here are similar to those from our ASR tasks in Section 6.

7.3. Non-Wilderness Neural Speech Synthesis

Non-Wilderness TTS Data: In Section 7.2 above, both our speech-based approach and the TTS data use our 195-language Wilderness subset described in Section 4.2. Thus, we also study how well our similarity measure generalizes to cases where the TTS data are not from Wilderness. Namely, we train five monolingual neural TTS models on different Indic source languages and evaluate them on two Indic target languages. Our source languages are Hindi, Kannada, Marathi, Tamil, and Telugu, and our target languages are Hindi and Telugu, all from the CMU INDIC corpus [50]. For each language, we sort the utterances by file name and let the last 100 be the test set, the 100 before that be the validation set, and the remaining be the training set.

Non-Wilderness TTS Model: Our models all use the same Transformer-TTS architecture, initialized using pre-trained weights from an English TTS model [51, 52]. Namely, we use 4 attention heads, 6 layers and 1536 units in both our encoder and decoder, and an optimization procedure similar to that of Vaswani et al. [42]. We train models using the train and validation sets of source languages and evaluate synthesized texts in the target languages' test sets. TTS samples and further details for reproducibility are provided in the accompanying GitHub.

Non-Wilderness TTS Evaluation: For each target language, we measure TTS performance using a MOS naturalness test performed by four listeners who are native speakers of that language. Namely, we ask evaluators to rate from 1 to 5 how much each sample sounds like natural speech in the target language, where 5 is the highest. Our MOS evaluation data contains two samples from each source-target pair, randomly chosen from the synthesized test set samples. Listen-

ers rate all the samples that have their target languages, yielding 40 total MOS ratings per target.

Table 6 summarizes these MOS test results, including the distances calculated by our top six language similarity measures from Table 3. Evaluators rated TTS samples with the same source and target the highest, as expected. Our proposed speech-based language similarity approach correctly identified the top two similar but different source languages for both targets. We compare this approach with others below.

Non-Wilderness TTS Correlations: Table 5 contains the Spearman correlations between the MOS values and different similarity measures, where each number uses a single target language and all the sources. Since the first geographic similarity measure from Section 3.4 considers these languages to be the same location, we use the second one instead. We observe that our acoustic approaches yield the strongest correlations, as with our Wilderness TTS experiment in Section 7.2. Compared to our previous tasks, the ensemble again performs the best overall, but the genetic and geographic measures did not rank as high. This further suggests that language similarity measures grounded in speech are suitable for acoustic cross-lingual transfer.

8. CONCLUSION AND FUTURE DIRECTIONS

In this work, we study how to choose source languages for acoustic cross-lingual transfer tasks through exploring eight different language similarity measures. Our proposed acoustic approaches often outperform other techniques in downstream tasks like speech recognition and text-to-speech tasks. This same approach can also be used to classify the families of unseen languages. Moving forward, we plan to train our acoustic approaches on all of the Wilderness data in order to strengthen their performance on languages outside our 195-language subset [24]. Our work can also be extended to a range of new acoustic cross-lingual transfer directions. These include leveraging language similarity in tasks like speech translation [53] and voice conversion [54], as well as integrating cross-domain [6, 55] and multimodal learning [56, 57] insights into our methods.

9. ACKNOWLEDGEMENTS

---

[5]Wilderness TTS 6-letter codes are APRWBT, JACWBT, KEKSBG, KJBSBG, MAKLAI, MOPWBT, NASPNG, NPLWYI, POHPOC, QEJLLB, QULSBB, QWHLLB, TZBSBM, TZCSBM, and VUNBST.

10. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark, and Rif A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017.

[2] SaiKrishna Rallabandi, Peter Wu, and Alan W Black, "Submission from CMU for blizzard challenge 2019," *Proceedings of Blizzard Challenge*, vol. 2019, 2019.

[3] Siddharth Dalmia, Ramon Sanabria, Florian Metze, and Alan W Black, "Sequence-based multi-lingual low resource speech recognition," in *ICASSP*. IEEE, 2018, pp. 4909–4913.

[4] Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki, "Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning," in *Interspeech*, 2020.

[5] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli, "Unsupervised cross-lingual representation learning for speech recognition," *Interspeech*, 2021.

[6] Xinjian Li, Siddharth Dalmia, A. Black, and Florian Metze, "Multilingual speech recognition with corpus relatedness sampling," in *Interspeech*, 2019.

[7] Xilun Chen, Ahmed Hassan Awadallah, Hany Hassan, Wei Wang, and Claire Cardie, "Multi-source cross-lingual model transfer: Learning what to share," in *ACL*, 2019.

[8] Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky, "Massively multilingual adversarial speech recognition," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics*. June 2019, pp. 96–108, Association for Computational Linguistics.

[9] Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, M. Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig, "Choosing transfer languages for cross-lingual learning," in *ACL*, 2019.

[10] Tom Sercu, George Saon, Jia Cui, Xiaodong Cui, Bhuvana Ramabhadran, Brian Kingsbury, and Abhinav Sethy, "Network architectures for multilingual speech representation learning," in *ICASSP*, 2017, pp. 5295–5299.

[11] Shubham Toshniwal, Tara N. Sainath, Ron J. Weiss, Bo Li, Pedro J. Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," *ICASSP*, pp. 4904–4908, 2018.

[12] Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Y. Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert, "Massively multilingual asr: 50 languages, 1 model, 1 billion parameters," *Interspeech*, pp. 4751–4755, 2020.

[13] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *ACL*, 2020.

[14] Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang, "Cross-lingual natural language generation via pre-training," in *AAAI*, 2020, vol. 34, pp. 7570–7577.

[15] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU*, 2017, pp. 265–271.

[16] Xinjian Li, Siddharth Dalmia, David Mortensen, Juncheng Li, Alan Black, and Florian Metze, "Towards zero-shot learning for automatic phonemic transcription," *AAAI*, vol. 34, pp. 8261–8268, 04 2020.

[17] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin, "URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017.

[18] *Ethnologue: Languages of the World*, SIL International, 23rd edition, 2020.

[19] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, *Glottolog 4.3*, Jena, 2020.

[20] Matthew S. Dryer and Martin Haspelmath, *The World Atlas of Language Structures Online*, Jena, 2013.

[21] Peter Wu, Yifan Zhong, and Alan W Black, "Automatically identifying language family from acoustic examples in low resource scenarios," *arXiv preprint arXiv:2012.00876*, 2020.

[22] Steven Moran and Daniel McCloy, Eds., *PHOIBLE 2.0*, Max Planck Institute for the Science of Human History, Jena, 2019.

[23] Chris Collins, "Syntactic structures of the world's languages (SSWL)," 2009.

[24] Alan W Black, "CMU wilderness multilingual speech dataset," in *ICASSP*. IEEE, 2019, pp. 5971–5975.

[25] Charles FF Karney, "Algorithms for geodesics," *Journal of Geodesy*, vol. 87, no. 1, pp. 43–55, 2013.

[26] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Interspeech*, 2011.

[27] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken language recognition using x-vectors.," in *Odyssey*, 2018, pp. 105–111.

[28] Alan W Black, "Clustergen: A statistical parametric synthesizer using trajectory modeling," in *Ninth International Conference on Spoken Language Processing*, 2006.

[29] Alan W Black and Prasanna Kumar Muthukumar, "Random forests for statistical speech synthesis," in *Interspeech*, 2015.

[30] John Kominek, Tanja Schultz, and Alan W Black, "Voice building from insufficient data-classroom experiences with web-based language development tools.," in *SSW*, 2007, pp. 322–327.

[31] Robert Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*. IEEE, 1993, vol. 1, pp. 125–128.

[32] Alan W Black, Heiga Zen, and Keiichi Tokuda, "Statistical parametric speech synthesis," in *ICASSP*. IEEE, 2007, vol. 4, pp. IV–1229.

[33] W Voiers, "Diagnostic acceptability measure for speech communication systems," in *ICASSP*. IEEE, 1977, vol. 2, pp. 204–207.

[34] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 06 2017.

[35] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan, "Supervised contrastive learning," in *NeurIPS*, 2020.

[36] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech*, 2019.

[37] Ting Qian, Kristy Hollingshead, Su-youn Yoon, Kyoung-young Kim, and Richard Sproat, "A python toolkit for universal transliteration," in *LREC*, Valletta, Malta, May 2010, European Language Resources Association (ELRA).

[38] Peter Wu, Sai Krishna Rallabandi, Alan W Black, and Eric Nyberg, "Ordinal triplet loss: Investigating sleepiness detection from speech.," in *Interspeech*, 2019, pp. 2403–2407.

[39] Elizabeth Salesky, Badr M. Abdullah, Sabrina Mielke, Elena Klyachko, Oleg Serikov, Edoardo Maria Ponti, Ritesh Kumar, Ryan Cotterell, and Ekaterina Vylomova, "SIGTYP 2021 shared task: Robust spoken language identification," in *Proceedings of the Third Workshop on Computational Typology and Multilingual NLP*, Online, June 2021, pp. 122–129, Association for Computational Linguistics.

[40] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Yalta, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Interspeech*, 09 2018, pp. 2207–2211.

[41] Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Yalta, Ryuichi Yamamoto, Xiao fei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang, "A comparative study on transformer vs RNN in speech applications," *ASRU*, pp. 449–456, 2019.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, vol. 30.

[43] Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha, "Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali," in *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018, pp. 52–55.

[44] Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab, "Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban," in *Interspeech*, Dresden, Germany, September 2015.

[45] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *LREC*, Marseille, France, May 2020, European Language Resources Association.

[46] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *ICASSP*. IEEE, 2015.

[47] Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al., "Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio," *Interspeech*, 2021.

[48] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert, "MLS: A large-scale multilingual dataset for speech research," *Interspeech*, 2020.

[49] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *Interspeech*, 2020.

[50] Andrew Wilkinson, Alok Parlikar, Sunayana Sitaram, Tim White, Alan W Black, and Suresh Bazaj, "Open-source consumer-grade indic text to speech.," in *SSW*, 2016, pp. 190–195.

[51] Tomoki Hayashi, Ryuichi Yamamoto, Katsuki Inoue, Takenori Yoshimura, Shinji Watanabe, Tomoki Toda, Kazuya Takeda, Yu Zhang, and Xu Tan, "ESPnet-TTS: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit," in *ICASSP*. IEEE, 2020, pp. 7654–7658.

[52] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *AAAI*, 2019, vol. 33, pp. 6706–6713.

[53] Changhan Wang, J. Pino, and Jiatao Gu, "Improving cross-lingual transfer learning for end-to-end speech recognition with speech translation," in *Interspeech*, 2020.

[54] Peter Wu, Paul Pu Liang, Jiatong Shi, Ruslan Salakhutdinov, Shinji Watanabe, and Louis-Philippe Morency, "Towards client-side privacy for downstream speech tasks," in *APSIPA*, 2021.

[55] Akshat Gupta, Sai Krishna Rallabandi, and Alan W Black, "Task-specific pre-training and cross lingual transfer for sentiment analysis in Dravidian code-switched languages," in *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, 2021.

[56] Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, and Ruslan Salakhutdinov, "Cross-modal generalization: Learning in low resource modalities via meta-alignment," *ACM Multimedia*, 2021.

[57] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al., "Multibench: Multiscale benchmarks for multimodal representation learning," *NeurIPS*, 2021.

[58] John Towns, Timothy Cockerill, Maytal Dahan, Ian Foster, Kelly Gaither, Andrew Grimshaw, Victor Hazlewood, Scott Lathrop, Dave Lifka, Gregory D Peterson, et al., "Xsede: Accelerating scientific discovery computing in science & engineering, 16 (5): 62–74, sep 2014," *URL https://doi.org/10.1109/mcse*, vol. 128, 2014.

[59] Nicholas A Nystrom, Michael J Levine, Ralph Z Roskies, and J Ray Scott, "Bridges: a uniquely flexible hpc resource for new communities and data analytics," in *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, 2015, pp. 1–8.