Where are you from? Geolocating Speech and Applications to Language Identification

Anonymous ACL submission

Abstract

In this paper we explore training models to answer the question, Where are you from? at a global scale. In other words, we are training models to geolocate speech based on language, accent and dialect. By leveraging radio broadcasts with known geographic locations, we train interpretable models for geolocation from audio and demonstrate that solving this task also provides a simple, but novel method for language identification (LID). We show that our method can outperform standard self-supervised models.

1 Introduction

LID is a critical component in many modern multilingual speech technologies (Barrault et al., 2023). In order to make more accurate predictions or generate more fluent outputs, speech technologies often condition their predictions on class labels describing the language, dialect, or accent of the input speech. As a result, tasks aimed at producing these class labels have been extensively explored (Zissman, 1996; Chen et al., 2023; Watanabe et al., 2017; Alumäe et al.). State-of-the-art systems perform remarkably well on common benchmarks for these tasks, including on the FLEURS (Conneau et al., 2023) and VoxLingua (Valk and Alumäe, 2021) corpora.

However, many speech phenomena are problematic for LID systems. Code-switching, receptive bilingualism, and symmetric and asymmetric mutual intelligibility of languages, challenges LID systems, but also the notion of using categorical labels for a phenomenon that occurs on a continuum (Haugen, 1966). Is a Hindi speaker who says a few words in English really switching to English? or are those words effectively part of the Hindi vernacular? and who makes that decision?

Geolocation of speech may be preferable to LID in many such circumstances. For instance, a code-switched Hindi-English utterance may cause problems for an LID system, but it is still likely to have occurred somewhere in India; a conversation between receptive English-Spanish bilinguals in the United States is still likely to have occurred in the United States regardless of which language was spoken. Furthermore, while some standard corpora for LID cover about 100 languages and dialects, evaluating LID on every accent and dialect is intractable.

Audio, however, often comes geolocated for free. For instance, the current location of a cell phone or home assistant can be passed as metadata along with any audio recordings on that device. IP addresses also correlate fairly well with geolocation (Li et al., 2012). These data can serve as soft labels for language, dialect, and accent. In this work, however, we use audio from radio stations (primarily FM broadcasts) that are simultaneous streamed on the web. Because FM radio generally travels only up to about 70 km (FCC), it is reasonable that speech heard on FM stations is at least understood by most people within a 70 km radius of the station, and possibly even representative of the local vernacular.

Our contributions are:

1. We demonstrate that geolocation associated with data collected from radio stations, can be used to train models that predict where people speak particular languages / dialects, or with specific accents.

2. We propose an interpretable model for speech geolocation whose predictions appear informed primarily by phonetics and accent.

3. We demonstrate simple methods by which the model can be repurposed for LID.

2 Related Work

Van Leeuwen and Orr (2016) first proposed the

* equal contribution

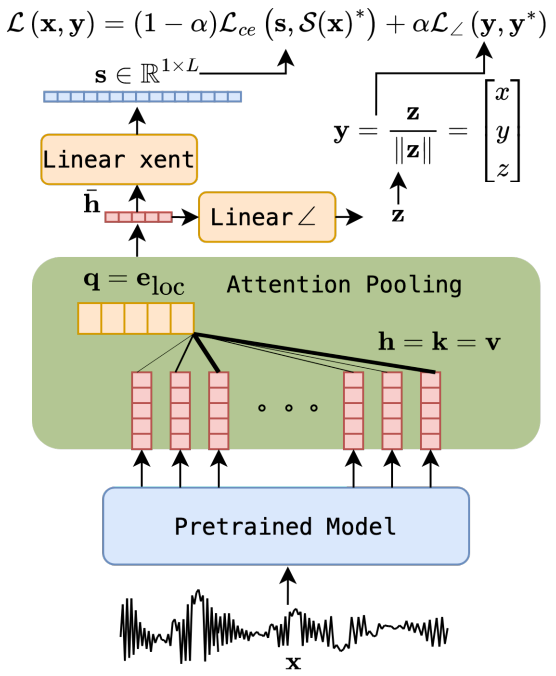

Figure 1: The geolocation model described in this paper. See Section 3 for more details. $\mathbf{x}$ is the input utterance. $\mathbf{e}_{\text{loc}}$ is a trained embedding representing the geolocation task. $\mathbf{h}$ is the sequence of extracted embeddings. $\bar{\mathbf{h}}$ is a vector representation of audio produced via cross-attention with $\mathbf{e}_{\text{loc}}$. In the attention block $\mathbf{e}_{\text{loc}}$ is the query, $\mathbf{q}$, and outputs of the pretrained model are the keys, $\mathbf{k}$, and values, $\mathbf{v}$. `Linear`$\angle$, transforms $\bar{\mathbf{h}}$ into Cartesian coordinates, $\mathbf{y}$, used to compute the angular distance, $\mathcal{L}_{\angle}(\mathbf{y},\mathbf{y}^*)$, between the predicted and ground-truth locations. `Linear xent`, produces scores, $\mathbf{s}$, for L possible locations, from which either the cross-entropy or binary cross-entropy loss, $\mathcal{L}_{ce}(\mathbf{s},\mathbf{y}_{idx})$, over the set of possible locations is computed. $\mathcal{S}(\mathbf{x})^*$ is the set of ground truth locations for input $\mathbf{x}$.

task of accent location in the context of identifying Dutch accents in the *Sprekend Nederland* corpus and presented various formulations for describing a person's linguistic origins. Lohfink (2017) used regression and classification-based approaches to locate accent from i-vectors (Dehak et al., 2010). To our knowledge this is the only other work on geolocation of audio from the linguistic context of speech, i.e., not the background noise, or channel, which, on the other hand, *have* been previously explored (Kumar et al., 2016).

Prior work (Ye et al., 2016) has shown that geolocation can be used to improve ASR systems, as geolocation tends to be correlated with accent and also device preference. A similar line of work (Xiao et al., 2018) described how geolocation can be used in language modeling to bias ASR predictions towards locally relevant lexical items, including points-of-interest.

A key challenge we address is how to train neural networks to produce points on a sphere. This problem has been previously addressed in the literature on audio source localization. Perotin et al. (2019), for instance, examined whether regression based, or classification approaches were best suited for localization.

There has been a significant amount of work on language, dialect, and accent ID. Perhaps the most similar effort to ours from this body of work was Pratap et al. (2023), which attempted to scale speech technologies to thousands of languages by relying primarily on recordings of religious texts. The primary challenge of that effort was to see whether models trained on clean, single speaker recordings with *known* language labels would generalize to out-of-domain scenarios. In contrast, we are examining whether models can be trained on *heterogenous* data collected from radio with *soft* language labels in the form of geolocations. To our knowledge this is the first attempt at language, dialect, or accent localization on a global scale and the first to apply it to language ID.

3 Method

3.1 The Task of Speech Geolocation

Van Leeuwen and Orr (2016) proposed a probabilistic formulation of the speech geolocation task. Let $\mathbf{x}$ be an input audio sample spoken by a single speaker. Let $\mathbf{z}$ be a point estimate of that speaker's origin and $\mathbf{x}$ be an input speech utterance. Then the task of speech geolocation is to estimate the distribution,

$$p(\mathbf{y}|\mathbf{x}). \tag{1}$$

Given a model, $p(\mathbf{y}|\mathbf{x})$, and the ground truth distribution over locations, $p(\mathbf{y})$, one can use point estimates,

$$\mathbf{y}^* = \mathbb{E}_{p(\mathbf{y})}[\mathbf{y}] \tag{2}$$

$$\mathbf{y} = \mathbb{E}_{p(\mathbf{y}|\mathbf{x})}[\mathbf{y}], \tag{3}$$

of the ground-truth and prediction locations respectively to evaluate model quality. The angular distance between points can be used to this end*). Note that we have not specified a coordinate

*The spherical law of cosines formulation can suffer from loss of precision at small distances (~1 km). The Haversine formulation does not. Both versions seemed to work equally well, nor is this level of precision needed for our application

system for $\mathbf{y}$. As a convention in this paper, $\mathbf{y}$, represents Cartesian coordinates of point using a spherical approximation of the Earth with radius, $\rho = 6378.1$ km. Approximating the shape of the earth as a sphere incurs minor errors in location (< 30 km) which we consider negligible for our purposes. Let $\Theta = (\phi, \lambda)$ be the corresponding point on a sphere specified by latitude, ϕ, and longitude, λ. Then the angular distance between two points is

$$d_\theta(\Theta, \Theta^*) = \arccos\{\sin\phi \sin\phi^* + \cos\phi\cos\phi^*\cos(\lambda - \lambda^*)\}. \quad (4)$$

The corpus error, $D(\cdot,\cdot)$, between a set of N paired predictions, Θ_1^N, and targets, ${\Theta^*}_1^N$, can be evaluated by the average angular distance,

$$D\left(\Theta_1^N, {\Theta^*}_1^N\right) = \frac{1}{N}\sum_{i=1}^{N} d_\theta(\Theta_i, {\Theta^*}_i). \quad (5)$$

Given the near impossibility of labeling speech with all perceptible origins of influence, this is likely the only realistic evaluation metric for this task. However, in some circumstances, modeling a speaker's origins with a single point is insufficient: the speech of bilingual speakers will likely reflect two disparate origins; an audio sample may contain more than one speaker; a person's speech is likely influence by two parents.

Therefore, we extend the formulation in (van Leeuwen and Orr, 2016) to include the possibility of multiple points of origin. We achieve this by specifying a closed set of locations, $\mathcal{S}$, e.g., a list of cities with population $> p$, or in our case, the set of locations broadcasting radio stations. The problem is then to estimate the subset, $\mathcal{S}(\mathbf{x}) \subseteq \mathcal{S}$, of locations associated with speech, $\mathbf{x}$, i.e., we want to estimate the distribution,

$$p(\mathcal{S}(\mathbf{x}) \mid \mathbf{x}), \quad (6)$$

over these locations. Unfortunately, to our knowledge, there exist no data in sufficient quantities and annotated in any consistent way with this information, so evaluating such models requires defaulting to Eq. 5. Point estimates from $p(\mathcal{S}(\mathbf{x}) \mid \mathbf{x})$ can be estimated by averaging over the locations $s \in \mathcal{S}(\mathbf{x})$. Note that we want the spherical mean,

$$\mathbf{y} = \frac{\sum_{s\in\mathcal{S}(\mathbf{x})} s}{\left\|\sum_{s\in\mathcal{S}(\mathbf{x})} s\right\|}, \quad (7)$$

i.e., the MLE estimate of the Von Mises distribution mean parameter.

3.2 Model

Our model is depicted in Figure 1. We describe the depicted components below.

3.2.1 Speech representations

The only prior work on geolocation from audio (Lohfink, 2017), relied on i-vectors to contain all necessary information for predicting geographic location. More recently, self-supervised representations have become the state-of-the-art representation used in speaker identification. For this reason we build our geolocation models from various pretrained models. We limited ourselves to various versions of the Wav2Vec2 (Baevski et al., 2020) architecture as many multilingually pretrained models exist, including XLSR-53 (Conneau et al., 2020), XLS-R (Babu et al., 2021), and MMS (Pratap et al., 2023) models.

3.2.2 Interpretable Pooling

Once a sequence of embeddings, $\mathbf{h}$, has been extracted from a pretrained model, those representations need to be pooled to produce a single class label. While average pooling is commonplace, we take inspiration from (Girdhar and Ramanan, 2017), and use an attention based pooling mechanism instead to let the model learn which embeddings are relevant for the task of geolocation. The advantage of this approach is its interpretability – we can inspect the attention weights to see which sequence positions contributed most to the location prediction.

This is important as we are explicitly aiming to classify the audio based on linguistic features and not channel artifacts. If the model places high attention weights on regions of silence, the model is likely cuing on channel artifacts, whereas high weights on specific recurring phonemes, or phoneme sequences indicates the model has learned to associate those phonemes, or phoneme sequences with a particular location.

To this end, we train a task-specific embedding vector, $\mathbf{e}_{\text{loc}}$ that encodes the task of geolocating audio. This vector is treated as a query, $\mathbf{q}$, against which keys, $\mathbf{k}$, are compared. For this task, we use a single-headed, scaled-dot-product attention (Vaswani et al., 2017). The resulting attention weights are used to select which embeddings, i.e., the values, $\mathbf{v}$, to pool for subsequent prediction of location. We denote the pooled representation as $\bar{\mathbf{h}}$.

3.2.3 Regression-based Prediction

As discussed in Section 3.1, a practical evaluation metric is the average angular distance. We therefore also explore training models to produce single point estimates for the origin of $\mathbf{x}$. We use a simple classifier, `Linear`$\angle$, responsible for converting the pooled representation $\bar{\mathbf{h}}$ into Cartesian coordinates.

Since we are restricted to produce points on the surface of the Earth, we project $\mathbf{z}$ onto the unit-sphere representing the Earth and denote this quantity

$$\mathbf{y} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}.$$

We also experimented with directly producing latitude and longitude. However, the results were not substantially different, except for the the training was less stable and appeared more sensitive to the learning rate. We train using the angular distance as the objective function, $\mathcal{L}_{\angle}(\mathbf{y}, \mathbf{y}^*)$, where convert Cartesian coordinates $\mathbf{y}$ into spherical coordinates Θ as

$$\Theta = \begin{bmatrix} \phi \\ \lambda \end{bmatrix} = \begin{bmatrix} \arctan \frac{y}{x} \\ \arcsin z \end{bmatrix}. \quad (8)$$

3.2.4 Classification-based Prediction

Rather than directly producing a point estimate, we can try to predict the posterior $p(\mathbf{y}|\mathbf{x})$. We achieve this by training a classifier, `Linear xent`, to produce a vector of scores, $\mathbf{s}$, for each location in our set $\mathcal{S}$. For $\mathcal{S}$, we use the set of all locations in the training data. We use as a loss function, $\mathbf{L}_{ce}(\mathbf{s}, \mathbf{s}^*)$, which in this case is the cross-entropy,

$$\mathbf{L}_{ce}(\mathbf{s}, s^*) = H(\text{Sofmax}(\mathbf{s}), \mathbf{s}^*), \quad (9)$$

between predicted locations and the one-hot ground-truth location $\mathbf{s}^*$.

3.3 Multi-label Binary Classification

In the event that multiple ground truth locations, $\mathcal{S}(\mathbf{x})^*$, exist, we can train using the binary cross-entropy between $\mathbf{s}$ and $\mathcal{S}(\mathbf{x})$. In other words, we assume that the prediction for each location is made independently. In practice, no available data are labeled with multiple ground-truth locations. However, we none-the-less experiment with the multi-label loss. To induce multiple ground-truth locations from our labels, we pick the top-k closest locations to the ground-truth and include those as additional ground-truth locations. This may serve to regularize the model slightly as in densely populated regions, i.e., where there are many radio stations, the top-k locations will cover a narrow region, whereas areas where radio stations are sparse, and where the model should not over-fit to specific locations, the top-k location will cover a broad area.

Whether the model is trained to produce one or more labels, point estimates for the distributions can be computed by Eq. 7, possibly restricting the summation to only the top-k most probably locations. Finally, as it may be advantageous to combine both objective functions, we experiment by interpolating both of them via a constant parameter, α, controlling the extent to which each function is used.

4 Data

4.1 Training

As previously mentioned, we rely on data collected from radio broadcasts to train our networks. Streams were recorded using an API to the radio.garden aggregator which provides user-submitted geographic coordinates for radio stations. Two separate sets of data were used. The first set consists of ∼400 hours of speech collected August 9, 2023 to August 13, 2023. The second set consists of ∼4000 hours of speech collected between September 27, 2023 to October 1, 2023.

Stations were randomly sampled from the aggregator and recorded for 30 seconds. The first set of recordings were sampled uniformly at random among all possible station locations. During the second collection, data were sampled proportional to the linguistic diversity since the primary application of this method is to support LID tasks.

We evenly distribute k points on a sphere each corresponding to the center of a region from which we sample radio stations to record. Specifying evenly spaced points on a sphere has no analytical solution for all k, but can be efficiently approximated by mapping the Fibonacci lattice to points on the sphere. Each possible radio station is mapped to the closest such point according to the angular distance.

When recording radio stations, each point on the Fibonacci lattice is sampled proportionally to the language density of that region. We used the set of languages and their coordinates list in the `Phoible` database to this end (Moran and McCloy,

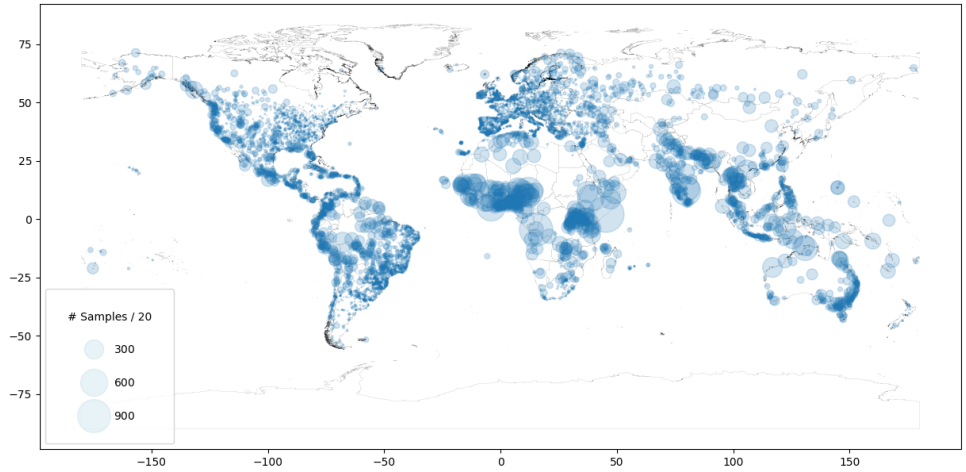

Figure 2: The distribution of the radio training data. Each circle represents a location with at least one training sample. The size of the circle is proportional to the number of utterances from a particular location.

2019). Specifically, the Gaussian kernel using the angular distance is used to compute scores for each language-lattice point pair, (l, f_i), where $l \in \mathcal{L}$, is a language in the set of languages, $\mathcal{L}$, from Phoible, and f_i is the i^{th} Fibonacci lattice point. The sum of scores across all languages determines the weight of that Fibonacci lattice point. Here, l is represented by the canonical longitude, latitude coordinates for that language.

$$w_i = \sum_{l \in \mathcal{L}} e^{-\frac{d(l,f_i)^2}{\sigma^2}} \quad (10)$$

This heavily biases samples toward south-east Asia, Africa, and North and South America. Unfortunately, many of the radio stations in Australia, North America, and South America, are not broadcasting the indigenous languages responsible for the high linguistic density in these regions. This likely leads to more English, Spanish, and Portuguese than desired.

These chunks were then segmented using the `inaSpeechSegmenter` (Doukhan et al., 2018) as in (Pratap et al., 2023). Segments are labeled as male, female, music, or NoEngery. Segments which were primarily labeled as male or female were kept, converted to the FLAC files and resampled to 16kHz. The other segments were discarded. On a subset of 1000 manually annotated samples the precision of this speech detection system was 95%. In total of 3748 hrs of audio remained after discarding music and keeping only the subsegments that the `inaSegmenter` labeled as speech. Figure 2 shows the global distribution of collected samples.

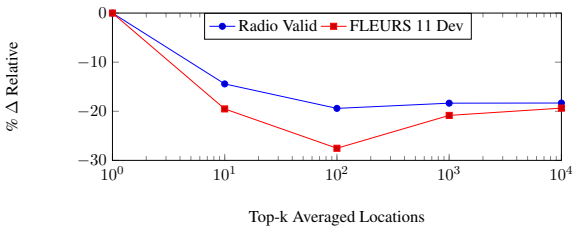

Figure 3: Averaging the Multi-label predictions

Pretrained Model	Radio Valid	FLEURS 11 Dev	FLEURS 11 Test
XLSR-53	3248 km	2345 km	2253 km
XLS-R-300m	2893 km	2576 km	2509 km
mms-300m	**2563 km**	**818 km**	**780 km**
mms-300 v2.0	2614 km	**774 km**	**741 km**

Table 1: Average prediction error (km) of models built starting from different pretrained models.

4.2 Evaluation

We use 3 different datasets for evaluation.

Radio Valid: We held-out all segments from 50 randomly selected radio stations among the collected data. Segments shorter than 2 seconds were discarded leaving 4.47 hrs of audio. Holding out broadcasts reduces the risk of speaker overlap between the train and test sets. These data were only used for evaluation of geolocation. This set was also used as a development set on which geolocation model parameters were tuned.

We note that while ground-truth locations of the radio stations are generally trust-worthy, i.e., the problem of rebroadcasts is relatively minor, a multitude of speakers of sometimes disparate origin speak during broadcasts. For instance, an American may regularly speak on an Australian news

	Radio Valid	FLEURS 11 Dev	FLEURS 11 Test
CE	3720	1982	1848
BCE (1)	3792	1959	1913
BCE (10)	3289	1483	1427
BCE (3)	3285	1286	1278
BCE (3) avg	3056	932	919
$\angle$ dist	**2614**	**774**	**741**

Table 2: Error of models trained with different objective functions. CE is cross-entropy, BCE (k) is binary cross-entropy, using the k nearest locations as targets. $\angle$ dist is regression using the angular distance. For the BCE (3) avg model, point estimates are the average of the top-100 predictions, rather than taking the single best prediction.

program. Furthermore, some expatriate and immigrant communities also have broadcasts in certain cities. These kinds of stations artificially deflate the reported accuracy of models.

FLEURS: We use the FLEURS corpus to evaluate both LID and geolocation. While the FLEURS utterances are not annotated with geolocation, or speaker demographic information to our knowledge, they are guaranteed to be labeled with the correct language, contrary to the radio data.

We therefore create a simulated geolocation evaluation set by assigning a point location to each language and using that as the ground-truth. We use the language locations from the Phoible (Moran and McCloy, 2019) database where applicable. For US English, Brazilian Portuguese, and Russian, we the population center of the country where the language was spoken. In the case of Latin American Spanish, a single point in Peru was chosen as an approximate geographic center of Latin American Spanish. We primarily focused on a subset of 11 FLEURS languages: US English, Latin American Spanish, Brazilian Portuguese, French, Polish, Macedonian, Russian, Malayalam, Hong Kong Yue, Filipino, and Japanese. We refer to this subset as **FLEURS 11**.

5 Experiments

5.1 Geolocation

5.1.1 Pretrained Models

We first ran experiments to determine the sensitivity of the geolocation model to the underlying pretrained model. To this end we explored using 3 different 300M parameter Wav2Vec2.0 models (Baevski et al., 2020): XLSR-53 (Conneau et al., 2020), XLS-R(Babu et al., 2021), and the 300M parameter MMS model (Pratap et al., 2023). They are all of the same size and architecture, but trained on increasing amounts of multilingual data.

For these experiments we trained on 4, A100 GPUs using a batch size of 400s of audio. All radio segments longer than 10s were cut into 10s windows, and any chunks shorter than 2s were discarded. We use the **Radio Valid** set to determine when the model had converged. We trained using a learning rate of 3×10^{-6}. We update all parameters in the model except for the convolutional layers of the pretrained base, as in Baevski et al. (2020).

We also froze the entire pretrained model for the first 1000 steps, and trained the attention pooling module using a fixed learning rate of 1×10^{-5}. We use the `OneCycle` (Smith and Topin, 2019) learning-rate schedule, where the learning rate is warmed-up for the first 8% of of the steps. Models were trained for up to 800000 steps, but in practice they converged around 136000 steps, which is the checkpoint for which we reported results.

Subsequently, we trained a new model using the best pretrained model with a higher learning rate (3×10^{-05}, and faster warm-up (4% of iterations), and fewer total steps (140000), as it sped up convergence and slightly improved performance.

5.1.2 Pretrained Model Results

Table 1 shows the effect of the pretrained model on the geolocation performance. Bolded values are the best among the first set of models, while bolded and underlined values indicate the best overall scores. We see that the MMS model was responsible for all of the best results on the three test conditions, outperforming both the XLSR-53 and XLS-R models by a wide margin.

Somewhat surprisingly, the XLSR-53 model slightly outperformed the XLS-R model as seen comparing rows 1, and 2 of Table 1, despite being trained on significantly fewer data. One possibility is that the XLS-R model, which is trained primarily on European Parliamentary speech from the VoxPopuli (Wang et al., 2021) corpus, is better suited for European languages, or is biased towards that channel.

The MMS model, is very similar to the XLS-R model, except for it was additionally trained on 55k hrs of audio in 1,362 languages. This language coverage appears to play an important role in improving geolocation.

5.1.3 Objective Functions

We then explored training using different objective functions. We used the exact same training configuration as during the pretrained model experiments and swapped out objective functions. We trained using cross-entropy (CE) and a single ground-truth location, using binary cross-entropy (BCE) where either the 1, 3, or 10 closests neighbors where considered to be the ground truth. The classifier produce one or more of 9449 unique locations.

During inference we averaged the top-k most probable locations to create a point-estimate of the distribution used for model comparison. We swept this parameter on the **Radio Valid** and **FLEURS 11 dev** sets to determine the optimal parameter. Figure 3 shows the results of this experiment. Using the top-100 candidates gave the best results so that is what we used on the **FLEURS 11 test** set.

Table 2 shows the effect of training with various objective functions on model performance. The rows are ordered by performance. First, looking at the top row, we see that cross-entropy (CE), and binary cross-entropy using a single ground-truth location were worst performing models. We noticed that model training was somewhat unstable, and we hypothesized that this may be due to similar, often very nearby locations, which can cause model confusions. There is likely little signal in the audio that could differentiate between two such areas.

Training using multiple ground truth locations (10) and (3) appears to help. Finally producing a point estimate by averaging the most likely 100 locations improves over picking the single mostly likely location and comes close, but does not outperform the best regression-based approach (bottom row).

5.2 Language Identification

A potential downstream use-case for geolocation models is as a strong initialization for LID models. To this end, we ran several experiments in order to explore the use of geolocation in language identification. To first ascertain how well location on its own is predictive of language, we ran two simple experiments using the FLEURS 11 subset and a fine-tuned XLSR-53 model which we trained on the first data collection ($\sim$ 400 hr see Section 4).

Our first approach was to use the point estimates produced for each location as fixed parameters in a classifier. We refer to this approach as $\mathbf{y}^*$. We can improve slightly on this approach by calibrating our point location estimates on some small amount of data, in this case the FLEURS 11 dev set, and update point estimates according to our model's predictions on the entire development set. This enables us to correct of any consistent biases in location predictions. We use μ_{geo} to denote this approach since we are reëstimating the *geographic* mean locations from data.

Finally, languages may be better separated in our model's latent representations, $\bar{\mathbf{h}}$, since these ultimately have to be mapped down to the surface of a sphere, and many languages may map to similar locations. Therefore, we similarly can estimate language-specific mean embeddings. We use $\mu_{\bar{\mathbf{h}}}$ to denote this approach since in this case we are reëstimating *embedding* means from data. If languages are geographically localized and well-separated this simple approach should work well.

The results of these approaches are shown in Table 3. We also compute precision, P, and recall, R, for each language treating each language separate using a one-versus-all binary classifier.

First, for some languages (Spanish, Portuguese, Filipino, Malayalam, and Hong Kong Yue) these approaches work well. Second, calibrating the mean geographic location appears to improve recall in most cases. Over all, these methods give an accuracy of around 70% and in the case of $\mathbf{y}^*$, *no* training is required. In both experiments, precision and recall for Japanese was low, possibly due to

	$\mathbf{y}^*$	μ_{geo}	$\mu_{\bar{\mathbf{h}}}$
Lang	P, R	P, R	P, R
en_us	93, 57.7	93.7 , **58.7**	**96.0**, 48.3
es_419	**99.2**, 95	99.1, 97.0	92.7, **98.5**
pt_br	94.8, 84	**96.3**, 86.7	95.4, **88.4**
fr_fr	**68**, 75.6	52.2, **91.0**	51.9, 90.7
pl_pl	**56.2**, 31.7	45.0, **44.9**	37.9, 38.3
mk_mk	44.6, **45**	58.3, 30.3	**58.6**, 20.3
ru_ru	42, **93.9**	**74.1**, 77.3	62.3, 77.8
ml_in	77.7, **98.7**	**82.9**, 98.4	82.2, 98.5
yue_hk	78.7, 98	83.2, 99.2	**84.9**, **99.2**
fil_ph	**98.5** 80.3	97.9, 87.3	96.1, **90.9**
ja_jp	**52.5**, 3.2	23.2, 24.9	27.6, **27.1**
avg	**73.2**, 69.4	72.3, **72.3**	71.4, 70.7

Table 3: Precision (P) and recall (R) of Language ID on the subset of the FLEURS languages when using the geographical means (Geo-mean), and Calibrating Embeding

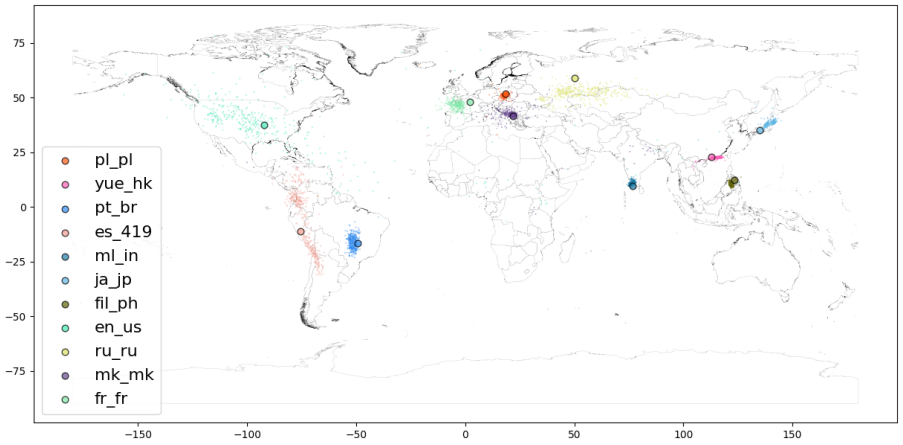

Figure 4: The geolocation model predictions on FLEURS 11 dev.

	FLEURS 11 Dev	FLEURS 11 TEST
MMS	89.4	89.6
MMS+Geoloc	**99.1**	**99.4**

Table 4: The language ID accuracy using geolocation based pretrained models. MMS is the MMS-300M model. Geoloc is our best performing geolocation model.

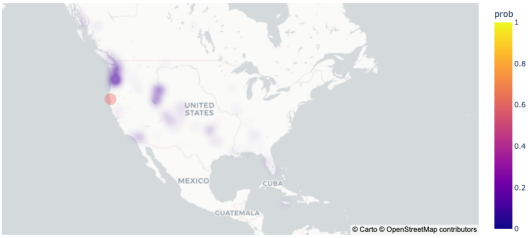

Figure 5: The binary-cross entropy geolocation model predictions on a radio station in the Radio Valid set. The red dot marks the broadcast location.

scarcity of Japanese radio in the training data. The low precision and recal for many of the European languages is likely explained by the close proximity of these test languages to each other which can cause false positive and negative detections. English in particular had a low recall, likely because it is spokenacross the globe and so the model produced less precise geolocation estimates for English.

5.2.1 Geolocation as Pretraining

Finally, we preliminarily explore using our best geolocation model as an initialization for an LID system. In these experiments, we train two LID models on the FLEURS 11 Training data, both with a learning rate of 1×10^{-5}, freezing pretrained model for the first 1000 steps, while using a fixed learning rate of 1×10^{-5}. In these experiments we also initialized $\mathbf{e}_{\text{loc}}$ with the value from the geolocation model. All segments 20 s or longer were discarded from training.

Table 4 shows the results of this experiment on the 11 FLEURS languages we trained on. When training from scratch (row 1), the model converges more slowly and gives about 90% accuracy on the FLEUR 11 dev and test sets. However, training when initializing from the geolocation model trained on radio data, the model converged more quickly to very close to 100% accuracy.

6 Interpreting Geolocation Predictions

While quantitative analysis of this task is difficult, it is very amenable to qualitative analysis. In Figures 4 and 5, we show example corpus level predictions for speech for the FLEURS 11 dev set, as well as a single heat-map produced from the multi-label prediction models.*

7 Conclusion

We have demonstrated that radio stations with geolocation can be harvested to train language, dialect, and accent models at a global scale. Furthermore, because geolocation and language are so correlated, training models using geolocations can be used to initialize language ID models. Future work should scale up these experiments and examine their application to accent recognition.

*Interactive examples can be found at geolocation-from-speech-demo.github.io

8 Ethical Considerations and Limitations

The primary limitation of our work is the availability of geolocated audio. We resorted to using radio stations for this purpose, but in general we cannot release the data collected from these stations to the public as it is almost certainly copyrighted. Furthermore, while our work covers a large portion of the world, we are ultimately limited by the availability of radio stations and what they choose to broadcast. We have no control over the content, which is often religious in nature, and the speakers tend to be male.

Furthermore, while identifying an individual's origins from speech is an interesting linguistic question, on its own, it could cause issues of data privacy. However, it could have broad applications in forensic analysis of speech, or biometric based security.

References

Fm broadcast station classes and service contours. https://www.fcc.gov/media/radio/fm-station-classes. Accessed: 2023-12-13.

Tanel Alumäe, Kunnar Kukk, Viet-Bac Le, Claude Barras, Abdel Messaoudi, and Waad Ben. Exploring the impact of pretrained models and web-scraped data for the 2022 nist language recognition evaluation.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

William Chen, Brian Yan, Jiatong Shi, Yifan Peng, Soumi Maiti, and Shinji Watanabe. 2023. Improving massively multilingual asr with auxiliary ctc objectives. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

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 *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. 2010. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798.

David Doukhan, Jean Carrive, Félicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *Acoustics Speech and Signal Processing (ICASSP), 2018 IEEE International Conference on*. IEEE.

Rohit Girdhar and Deva Ramanan. 2017. Attentional pooling for action recognition. *Advances in neural information processing systems*, 30.

Einar Haugen. 1966. Dialect, language, nation 1. *American anthropologist*, 68(4):922–935.

Anurag Kumar, Benjamin Elizalde, and Bhiksha Raj. 2016. Audio content based geotagging in multimedia. *arXiv preprint arXiv:1606.02816*.

Dan Li, Jiong Chen, Chuanxiong Guo, Yunxin Liu, Jinyu Zhang, Zhili Zhang, and Yongguang Zhang. 2012. Ip-geolocation mapping for moderately connected internet regions. *IEEE Transactions on Parallel and Distributed Systems*, 24(2):381–391.

Georg Lohfink. 2017. The" sprekend nederland" project applied to accent location. Master's thesis.

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

Lauréline Perotin, Alexandre Défossez, Emmanuel Vincent, Romain Serizel, and Alexandre Guérin. 2019. Regression versus classification for neural network based audio source localization. In *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 343–347. IEEE.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

David A van Leeuwen and Rosemary Orr. 2016. The" sprekend nederland" project and its application to accent location. *arXiv preprint arXiv:1602.02499*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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. *arXiv preprint arXiv:2101.00390*.

Shinji Watanabe, Takaaki Hori, and John R Hershey. 2017. Language independent end-to-end architecture for joint language identification and speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271. IEEE.

Xiaoqiang Xiao, Hong Chen, Mark Zylak, Daniela Sosa, Suma Desu, Mahesh Krishnamoorthy, Daben Liu, Matthias Paulik, and Yuchen Zhang. 2018. Geographic language models for automatic speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Guoli Ye, Chaojun Liu, and Yifan Gong. 2016. Geolocation dependent deep neural network acoustic model for speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5870–5874. IEEE.

Marc A Zissman. 1996. Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on speech and audio processing*, 4(1):31.