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

Patrick Foley*, **Matthew Wiesner***
Bismarck Bamfo Odoom, Leibny Paola Garcia, Kenton Murray, Philipp Koehn
Johns Hopkins University, Hackerman 22 3400 North Charles Street,
Baltimore, MD 21218-2680
pfoley2006@gmail.com
{wiesner,bodoom1,lgarci27,kenton,phi}@jhu.edu

Abstract

We train models to answer the question, Where are you from? and show how such models can be repurposed for language identification (LID). To our knowledge, this paper is the first to introduce data sources, methods and models to tackle the task of geolocation of speech at a global scale, and the first to explore using geolocation as a proxy-task for LID. Specifically, we explore whether radio broadcasts with known origin can be used to train regression and classification-based models for geolocating speech. We build models on top of self-supervised pretrained models, using attention pooling to qualitatively verify that the model geolocates the speech itself, and not other channel artifacts. The best geolocation models localize speaker origin to around 650km. We confirm the value of speech geolocation as a proxy task by using speech geolocation models for zero-shot LID. Finally, we show that fine-tuning geolocation models for LID outperforms fine-tuning pretrained Wav2Vec2.0 models, and achieves state-of-the-art performance on the FLEURS benchmark.

1 Introduction

Language identification (LID) is a critical component in many modern multilingual speech technologies (Barrault et al., 2023). 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.) and state-of-the-art systems perform remarkably well on common benchmarks, including on the FLEURS (Conneau et al., 2023) and VoxLingua (Valk and Alumäe, 2021) corpora.

While these corpora cover ∼100 *languages*, there are orders of magnitudes more *accents* and *dialects* and annotating all of them is intractable. To address this problem Pratap et al. (2023), proposed to scale speech technologies to thousands of languages and dialects 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 data. In contrast, we demonstrate that models can be trained on widely available *heterogenous* data collected from radio with *soft* language labels in the form of geolocations. To our knowledge, we are the first to demonstrate geolocation of speech on a global scale and the first to apply it to LID.

Geolocation of speech may also be preferable to LID in many circumstances. For instance, a code-switched Hindi-English utterance or a conversation between receptive English-Spanish bilinguals in the United States may cause problems for LID systems, but they still likely occurred, respectively, in India and in the United States. These phenomena* not only challenge LID systems, but also the notion of using categorical labels for a phenomenon that occurs on a continuum (Haugen, 1966).

Furthermore, audio can be approximately geolocated for free: the current location of a cell phone can be passed as metadata along with any audio recordings on that device; IP addresses can be resolved to 100-200 kilometers (Li et al., 2012). We investigate whether these data can serve as soft labels for language, dialect, and accent. In this work, we use audio from FM radio broadcasts, that are simultaneous streamed on the web. The *key assumption*† used in this work is that because FM radio travels only up to about 70 km (FCC), on average, the 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.

* equal contribution

*receptive bilingualism, symmetric and asymmetric mutual intelligibility of languages, accent and dialect

†Rebroadcasts, speakers of various origins, and foreign language broadcasting break this assumption.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 5114–5126

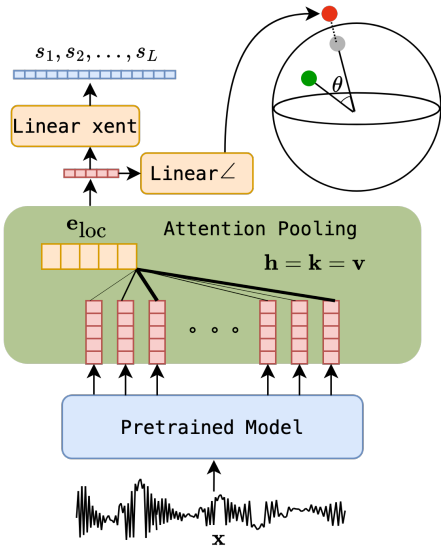


Figure 1: The geolocation model described in this paper. See Section 2 for more details. $\mathbf{x}$ is the input utterance. $\mathbf{e}_{\text{loc}}$ is a trained embedding representing the geolocation task. $\mathbf{h}$ is the sequence of extracted embeddings. In the attention block, $\mathbf{e}_{\text{loc}}$ is the query, and outputs of the pretrained model are the keys, $\mathbf{k}$, and values, $\mathbf{v}$. `Linear`$\angle$, transforms pooled representations into Cartesian coordinates (in red), which are projected onto the surface of the earth (gray). Error from the ground-truth (green), is measured by θ. *Alternatively*, `Linear xent`, produces scores, $\{s_1, \ldots, s_L\}$, for L locations.

1.1 Related Work

Van Leeuwen and Orr (2016) first proposed the 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 the speech signal, i.e., not the background noise, channel (Kumar et al., 2016), or extracted text (Bell et al., 2015), which, on the other hand, *has* been previously explored. Plchot et al. (2009) and (Sikasote et al., 2023) mined radio broadcasts to support language identification and ASR, respectively, but did not explore speech geolocation.

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.

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

Our contributions are: (i) 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. (ii) We propose classification and regression-based approaches built on top of self-supervised models for speech geolocation. Our use of an interpretable attention pooling mechanism indicates that predictions appear informed primarily by phonetics and accent and not spurious channel artifacts. (iii) We demonstrate that geolocation models can be repurposed for zero-shot LID. (iv) We show that fine-tuning speech geolocation models for LID out-performs fine-tuning the original pretrained models on the FLEURS benchmark.

2 Method

2.1 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 and let $\mathbf{y}$ be a point estimate of that speaker's origin. Then the task of speech geolocation is to estimate the distribution,

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

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

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

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

of the ground-truth and predicted 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 system for $\mathbf{y}$. As a convention in this paper, $\mathbf{y}$ represents Cartesian coordinates of a point using a spherical approximation of the Earth with radius, $\rho = 6378.1$ km. Approximating the shape of the

*The spherical law of cosines formulation suffers from loss of precision at small distances (~1 km). The Haversine formulation does not. Both versions seemed to work equally well. We do not currently need this level of precision.

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 great-cirle distance,

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

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 influenced 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 $\mathbf{y} \in \mathcal{S}(\mathbf{x})$. Note that we want the spherical mean,

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

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

2.2 Model

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

2.2.1 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. Therefore, we build our geolocation models from various pretrained models. We limited ourselves to the multilingually pretrained XLSR-53 (Conneau et al., 2020), XLS-R (Babu et al., 2021), and MMS (Pratap et al., 2023) versions of Wav2Vec 2.0 (Baevski et al., 2020).

2.2.2 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 to let the model learn which embeddings are relevant for geolocation. The advantage of this approach is its interpretability – high attention weights on regions of silence indicate that the model is cuing on channel artifacts, while high weights on frames corresponding to specific phonemes, or phoneme sequences indicate that the model is geolocating audio using lingustic information.

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 single-headed, scaled-dot-product attention (Vaswani et al., 2017). The attention weights are used to select the embeddings, i.e., the values, $\mathbf{v}$, to pool for prediction of location. We use $\bar{\mathbf{h}}$ to denote the pooled representation.

2.2.3 Regression-based Prediction

As discussed in Section 2.1, a practical evaluation metric is the average angular distance. We therefore explore training models to produce single point estimates for the origin of $\mathbf{x}$. We use a linear regressor, $\texttt{Linear}\angle$, to convert the pooled representation, $\bar{\mathbf{h}}$, into 3-dimensional Cartesian coordinates, $\mathbf{z} \in \mathbb{R}^{1\times 3}$. Since we are predicting points on the surface of the Earth, we then project (i.e., normalize) $\mathbf{z}$ onto the unit-sphere representing the Earth and denote this quantity as

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

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 Cartesian coordinates, $\mathbf{y}$, are converted into spherical coordinates Θ as

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

2.2.4 Classification-based Prediction

Rather than directly producing a point estimate, we can estimate the posterior distribution, $p(\mathbf{y}|\mathbf{x})$, 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,

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

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

2.3 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. We nonetheless 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 regularize the model as in densely populated regions, i.e., where there are many radio stations, the top-k locations will cover a narrow region, whereas the top-k locations will cover a broad area in regions where radio stations are sparse and the model should not over-fit to any specific location.

Whether the model is trained to produce one or more labels, point estimates of speaker origin can be computed by Eq. 7, possibly restricting the summation to the top-k most probable locations.

3 Data

3.1 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.

To sample radio stations, 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, 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 ith 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. Formally, weights are given by

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

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. The first set of recordings were sampled uniformly at random among all possible station locations. During the second collection, data were sampled from regions proportional to their estimated linguistic diversity since the primary application of this method is to support LID.

Recorded chunks were then segmented and labeled using the `inaSpeechSegmenter`

(Doukhan et al., 2018) as "male", "female", "music", or "NoEnergy" as done in (Pratap et al., 2023). Segments which were primarily labeled as "male" or "female" were kept, converted to the FLAC files and re-sampled 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, 3748 hrs of audio remained after discarding music and keeping only the subsegments that the `inaSegmenter` labeled as speech. Figure 3 in Appendix A shows the global distribution of collected samples.

3.2 Evaluation

We used 2 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 speech geolocation. This set was also used as a development set on which geolocation model parameters were tuned.

We note that treating these data as ground-truth technically corresponds to the slightly different, and more challenging task, of predicting the broadcast location of radio, rather than predicting a speaker's origin. However, we *assume* in this paper, that the speakers in the broadcast are indeed representative of speakers from the ground-truth location. In this sense the labels are noisy. While, we *do not have* ground-truth annotations for speakers, the ground-truth locations of the radio stations appear to generally be trust-worthy, i.e., the problem of rebroadcasts is relatively minor. In addition to rebroadcasts, a multitude of speakers of sometimes disparate origins can sometimes speak during broadcasts, e.g., an American may regularly speak on an Australian news program. Furthermore, some expatriate and immigrant communities also have broadcasts in certain cities.

From the perspective of the speech geolocation task, i.e., not the broadcast localization task, these situations result in erroneous labels of the speech segments and artificially inflates model error (in km) on this task. The error metric on this dataset is, however, representative of performance on the task of predicting broadcast location, but we have *not* explicitly attempted to address that task here.

FLEURS: We use the FLEURS corpus (Conneau et al., 2023) to evaluate both LID and geolocation. While, the FLEURS utterances are not annotated with geolocation, they *are* labeled with language and speakers are L1 speakers.

Since L1-language is itself a label on a speaker's origin, we create a simulated geolocation evaluation set by assigning a point location to each language and using that as ground-truth. We use the language locations from the Phoible (Moran and McCloy, 2019) database where applicable. For US English, Brazilian Portuguese, and Russian, which had no entries in the database, we use the approximate population centers of the respective countries where those languages are 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 focused primarily 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**.

4 Experiments

4.1 Geolocation

4.1.1 Pretrained Models

We first ran experiments to determine the sensitivity of the geolocation model to the underlying pretrained model. We train models 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. We also compare to the 1B parameter MMS model.

For these experiments we trained on 4 A100 GPUs using a batch size of 400s of audio for the 300M parameter models, and 200s of audio for the 1B parameter models. All radio segments longer than 10s were cut into 10s windows, and any chunks shorter than 2s were discarded. We used the **Radio Valid** set to determine when the model had converged. We trained using a learning rate of 3×10^{-6}. We updated all parameters except for the convolutional layers of the pretrained base, as in Baevski et al. (2020).

We froze the entire pretrained model for the first 1000 steps, training only the attention pooling module and classifiers using a fixed learning rate of

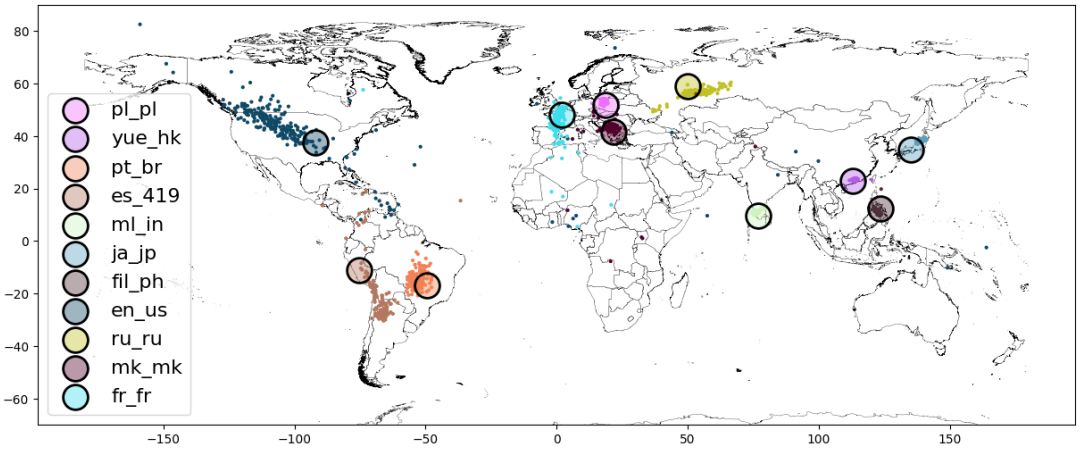


Figure 2: The geolocation model predictions (MMS-1b in Table 1) on FLEURS 11 Dev. Large circles are the ground-truth locations for each language, each point represents an utterance. The color shows the language.

1×10^{-5}. Then we use the OneCycle (Smith and Topin, 2019) learning-rate schedule, warming up the learning rate for the first 8% of steps. Models were trained for up to 800000 steps, but in practice they converged around 136000 steps, which is the checkpoint used to report results. Subsequently, we trained the 1-billion parameter model using the best configuration and fewer total steps (140000).

Pretrained Model	Radio Valid	FLEURS 11 Dev	FLEURS 11 Test
Random Sphere[‡]	10019 km	10019 km	10019 km
Random Train[§]	8624 km	8624 km	8624 km
(1) XLSR-53	3248 km	2345 km	2253 km
(2) XLS-R	2893 km	2576 km	2509 km
(3) MMS	2614 km	**774 km**	**741 km**
(4) MMS avg	2895 km	927 km	853 km
(5) MMS post avg	**2555 km**	987 km	909 km
(6) MMS-1B	**2355 km**	**684 km**	**627 km**

Table 1: Average prediction error (km) from ground-truth locations of geolocation models built starting from different pretrained models.

Table 1 shows the effect of the pretrained model on the geolocation performance. Bolded values are the best among the 300M parameter models, while bolded and underlined values indicate the best overall scores. To contextualize the performance we compare to a theoretical and a simulated random baseline, Random Sphere, and Random Train. All trained models significantly outperform these random baselines.

Among trained models, 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. Since the XLS-R model is trained primarily on European Parliamentary speech from the VoxPopuli corpus (Wang et al., 2021), it may be better suited for European languages, or 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 performance. The attention pooling mechanism seems to help performance slightly compared to average-pooling of embeddings (avg), or pooling of the output predictions (post avg). The 1-billion parameter model outperforms the smaller, 300-million parameter model. A further advantage of the pooling mechanism is that we can use it verify the model is focusing on regions of speech in the audio signal (See Figure 6a of Appendix A for an example).

Finally, while quantitative analysis of the geolocation task is difficult, it is very amenable to qualitative analysis. Figure 2 shows the geolocation predictions for speech from the FLEURS 11 dev set. These unseen utterances are well-localized, which indicates that these languages were present in the radio training data and corroborates that the geolocation predictions are based on language or accent and *not* content or channel artifacts.

[‡]Theoretical result: The average distance between two randomly sampled points on a sphere is $\rho\frac{\pi}{2}$, where ρ is the radius of the sphere.

[§]Simulated result: The average distance between and two randomly selected points from the training data.

4.1.2 Objective Functions

We then explored training using different objective functions. We used the best 300M parameter training configuration from the pretrained model experiments and swapped out objective functions. We trained using cross-entropy (CE) and a single ground-truth location, and using binary cross-entropy (BCE) where either the 1, 3, or 10 closest neighbors were considered to be the ground truth locations. The classifier produced 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 4 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.

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

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.

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 (BCE) using a single ground-truth location were the worst performing models. We noticed that model training was somewhat unstable, and we hypothesized that this may be due to the difficulty of distinguishing between nearby locations with nearly identical linguistic profiles. 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 the single mostly likely location, but does not outperform the best regression-based approach (bottom row). An example visualization of the multi-label predictions can be seen in 6b (see Appendix B).

	$\mathbf{y}^*$			μ_{geo}			$\mu_{\bar{\mathbf{h}}}$		
Lang	P	R	F	P	R	F	P	R	F
en_us	93	57.7	71.2	93.7	58.7	**72.2**	96.0	48.3	64.3
es_419	99.2	95	97.1	99.1	97.0	**98.0**	92.7	98.5	95.5
pt_br	94.8	84	89.1	96.3	86.7	**91.2**	95.4	88.4	91.8
fr_fr	68	75.6	**71.6**	52.2	91.0	66.3	51.9	90.7	66.0
pl_pl	56.2	31.7	40.5	45.0	44.9	**44.9**	37.9	38.3	38.1
mk_mk	44.6	45	**44.8**	58.3	30.3	39.9	58.6	20.3	30.2
ru_ru	42	93.9	58.0	74.1	77.3	**75.7**	62.3	77.8	69.2
ml_in	77.7	98.7	87.0	82.9	98.4	**90.0**	82.2	98.5	89.6
yue_hk	78.7	98	87.3	83.2	99.2	90.5	84.9	99.2	**91.5**
fil_ph	98.5	80.3	88.5	97.9	87.3	92.3	96.1	90.9	**93.4**
ja_jp	52.5	3.2	6.0	23.2	24.9	24.0	27.6	27.1	**27.3**
avg	73.2	69.4	71.2	72.3	72.3	**72.3**	71.4	70.7	71.0

Table 3: Precision (P), recall (R), and F-score (F), of Language ID on the subset of the FLEURS languages when using the geographical means (Geo-mean) and Calibrating Embeding. The best F-scores for each language are in **bold**.

4.2 Zero-shot Language Identification

A potential use for geolocation models is as a strong initialization for LID models and we ran several experiments exploring this use-case. 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 had previously trained on the first data collection ($\sim$ 400 hr see Section 3).

Our first approach was to use the point locations for each language as fixed parameters in a nearest neighbor classifier. We refer to this approach as $\mathbf{y}^*$. We can improve slightly on this approach by calibrating our point locations on some small amount of data, in this case the FLEURS 11 dev set, and updating point estimates according to our model's predictions on the entire development set. This enables us to correct 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 estimate language-specific mean embeddings. We use $\mu_{\bar{\mathbf{h}}}$ to denote this approach since in this case we reëstimate *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 compute the precision, P, recall, R, and F-score, F, for each language using language-specific one-versus-all binary classifiers.

	FL 11 Dev	FL 11 Test	FL 102 Dev	FL 102 Test
(1) MMS 11	89.4	89.6	-	-
(2) Geo 11	**99.1**	**99.4**	-	-
(3) FS MMS 11	21.7	13.5	-	-
(4) FS Geo 11	**86.4**	**86.1**	-	-
(5) MMS	93.3	93.4	84.3	84.7
(6) Geo	**98.1**	**98.3**	**89.4**	**89.8**
(7) MMS 1B	**99.7***	**99.8***	95.9	96.1
(8) Geo 1B	99.4	99.6	**96.4**	**96.7**
(Pratap et al., 2023)	-	-	-	96.2

Table 4: LID accuracy using geolocation based pretrained models. MMS is the MMS-300M model. FL stands for FLEURS. Geo is our best performing geolocation model (either 300M or 1B parameter). FS, stands for few-shot. MMS/Geo 11 indicates that the model is trained only on the 11 languages in the FLEURS 11 Dev / Test sets. * denotes results that were not statistically significant (Mcnemar's test $p > 0.05$).

For some languages (Spanish, Portuguese, Filipino, Malayalam, and Hong Kong Yue) these approaches worked well. Calibrating the mean geographic location improved recall, and F-score in most cases. These methods gave an accuracy of $\sim$ 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 scarcity of Japanese radio in the training data. The low precision and recall for many of the European languages is likely due to the close proximity of these test languages to each other which can cause false positive and negatives. Because English is a global language, English appears in most locations in the training data, which could explain the higher variance in model predictions. This in turn, may explain the low recall in English.

4.2.1 Geolocation as LID Pretraining

Finally, we explore using our best geolocation models as an initialization for LID systems. Table 4 shows the results of these experiments on the FLEURS languages. We compare pairs of LID models trained on the FLEURS training data. The first model in each pair (MMS), initializes its encoder with an MMS encoder. The second model in each pair (Geo), initializes the encoder instead with a geolocation model.

We first trained 300-M parameter models (rows 1-6) using the same architecture as used for the geolocation models. In rows 1 and 2, we trained only on the FLEURS 11 training set. In all cases we use a max learning rate of 1×10^{-5}, freezing the pretrained model for the first 1000 steps during which we use a fixed learning rate of 1×10^{-5}. We also initialized $\mathbf{e}_{\text{loc}}$ with the value from the geolocation model where applicable. All segments 20s or longer were discarded from training.

We then trained the same models in a few-shot scenario (rows 3, 4) on 10 minutes of randomly selected training data per language. Rows 5, and 6 show the results of models trained on 102 FLEURS languages. Rows 7 and 8 of Table 4 show the performance of a 1B parameter model trained on all 102 FLEURS languages. For the 1B parameter model, we roughly reïmplemented the FLEURS-only baseline from Pratap et al. (2023), using average pooling of predictions rather than attention pooling. We used a maximum learning rate of 5e-06, and trained for 20k updates with 200s of speech per mini-batch.

Convergence was slower when fine-tuning the MMS model (row 1), and resulted in $\sim$90% accuracy on the FLEURS 11 evaluation sets. However, when fine-tuning the geolocation model (row 2), the model converged quickly to near 100% accuracy. Figure 5 demonstrates this behavior. In the few show scenario, (rows 3, 4), the quality of the MMS fine-tuned model deteriorated significantly. However, fine-tuning the geolocation model performed comparably to the fine-tuned MMS model trained on *all* the data. Speech geolocation pretraining **significantly reduced the need for labeled data**. In rows 5 and 6, we see that scaling to all 102 languages minimally degraded performance on the 11-language subset, and in fact improved the MMS model performance. Using the 1B parameter model (rows 7, 8) improved performance. In *almost all* cases initializing LID models with geolocation models helped. The fine-tuned 1B parameter geolocation is state-of-the-art.

5 Conclusions

We have demonstrated[¶] that radio stations with geolocation can be harvested and used to geolocate speaker traits such as language or dialect at a global scale. Furthermore, because geolocation and language are so correlated, training models using geolocations can be used to initialize language ID models, and work especially well in few-shot settings. Future work should examine their application to accent recognition, and integration with multilingual ASR systems.

[¶] https://geolocation-from-speech-demo.github.io

6 Ethical Considerations and Limitations

The primary limitation of our work is the availability of publicly accessible 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. We can, however, upon request, make available the scripts needed to collect data in the same manner, and potentially the radio data used for this work if it is clearly for academic purposes, but even broad release of tools designed to record radio at scale could be construed as violating copyright law.

As speech technologies have improved, people and governments have recognized that recordings of a person's voice constitute personal data, and the storing or release of such data in anything other than purely academic contexts could be a violation of various data privacy laws. It is specifically for this reason that we hesitate to release the data even though we believe it would be incredibly useful for the academic community. We believe, unfortuantely, that the best compromise is to describe our data collection method in detail (as was done in Section 3) so that others may recreate this setup. Our goal was to work with web-scale data, but for such data, especially audio data, it is impossible to ascertain the copyright of every recording. We suspect that industry could benefit from approaches detailed in this work, as they *do* have access to large amounts of audio labeled with geolocation. We have at least described a mechanism by which academic institutions can work on industry-relevant problems. Being able to train LID, or even accent or dialect recognition systems with significantly smaller amounts of labeled data would be incredibly useful for people in regions of the world specifically addressed in this paper, including sub-Saharan Africa, south-east Asia, and Latin America.

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 spoken by men. The segmentation system used, which also predicts gender, estimates that between 65-75% of recordings are of male speech. These biases in data may effect model predictions and is something we have not studied in this paper, but that readers should be aware of. However, it is a problem also faced by similar efforts (Pratap et al., 2023) in massively multilingual speech processing, where a similar bias was present, and was found not to result in gender-biased ASR performance.

While identifying an individual's origins from speech is an interesting linguistic question, it could cause issues related to data privacy. Bad actors could be tempted to use such information to infer additional personally identifiable information about an individual. It is therefore important to consider and highlight this issue as it raises questions about the security of, for instance, password recovery questions, such as, In what city were you born?

The speech signal is deeply personal and geographic origin, accent, dialect, and language are not only correlated with each other, as shown in this paper, but likely also with other personal information such as political and religious beliefs, race, and educational background. Extreme care should be taken when deploying these models to not reinforce existing societal biases. These models could potentially be used to circumvent anti-discrimination practices. We note, however, that the above problems are not unique to geolocation of speech, but also language/dialect ID, given that spoken language is clearly correlated with a speaker's geographic origin.

This work could also have broad applications for social good in forensic analysis of speech in legal settings, biometric based security, and most importantly enabling speech technologies for speakers of under-resourced languages, unwritten languages, non-standard dialects, and historically marginalized communities.

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*.

Peter Bell, Catherine Lai, Clare Llewellyn, Alexandra Birch, and Mark Sinclair. 2015. A system for automatic broadcast news summarisation, geolocation and translation. In *Sixteenth Annual Conference of the International Speech Communication Association*.

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.

Oldrich Plchot, Valiantsina Hubeika, Lukáš Burget, Petr Schwarz, Pavel Matejka, and J Cernocky. 2009. "acquisition of telephone data from radio broadcasts with applications to language recognition: Technical report. *URL: http://www. nist. gov/speech/tests/lre/2009/radio broadcasts. pdf*.

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*.

Claytone Sikasote, Kalinda Siaminwe, Stanly Mwape, Bangiwe Zulu, Mofya Phiri, Martin Phiri, David Zulu, Mayumbo Nyirenda, and Antonios Anastasopoulos. 2023. Zambezi voice: A multilingual speech corpus for zambian languages. *arXiv preprint arXiv:2306.04428*.

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.

A Additional plots

Figure 3 shows the distribution of the 4,000 hr radio data used in most of the experiments in this paper. Figure 4 shows the results of the hyper-parameter sweep over the number of nearest adjacent locations to predict for models trained with cross-entropy objectives.

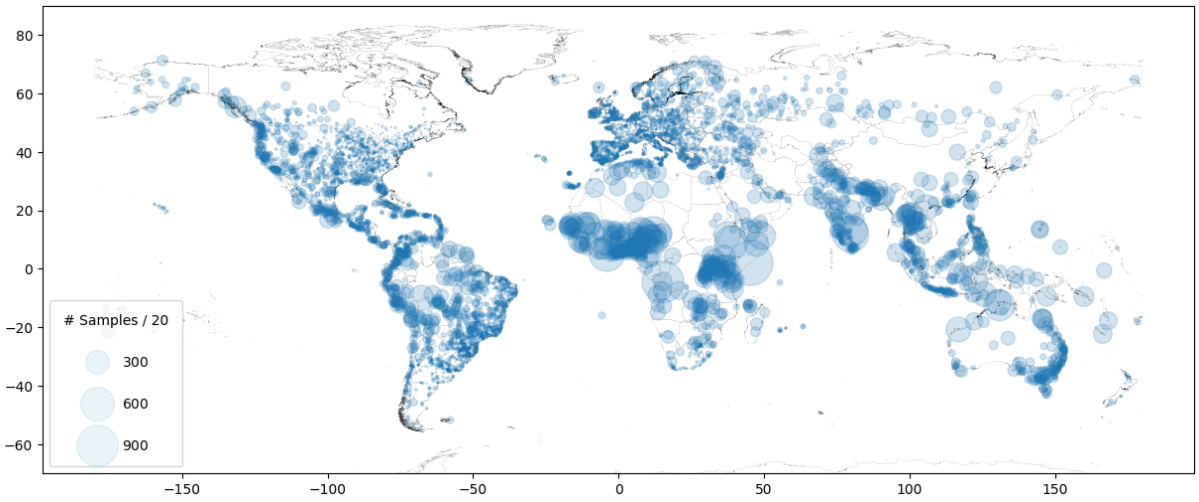


Figure 3: The distribution of the radio training data. Each circle represents a location with at least one training sample. The size of the circle is proportional to the number of utterances from a particular location.

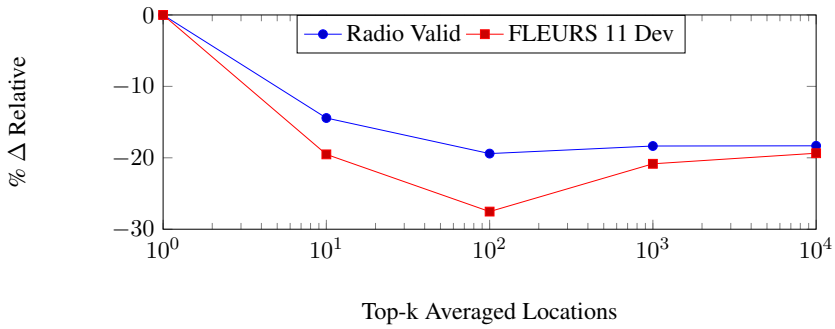


Figure 4: Averaging the Multi-label predictions

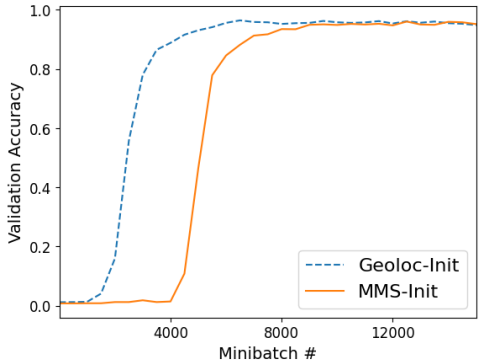
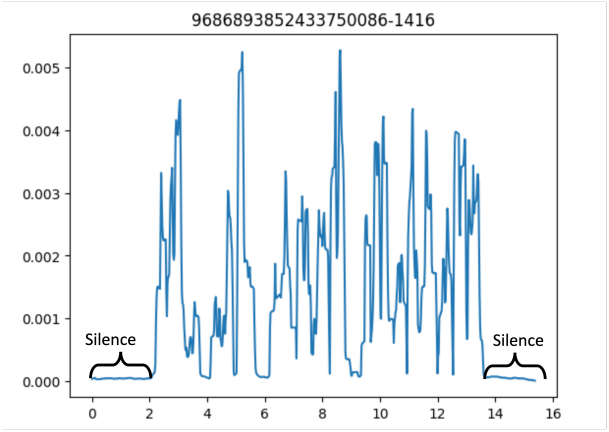
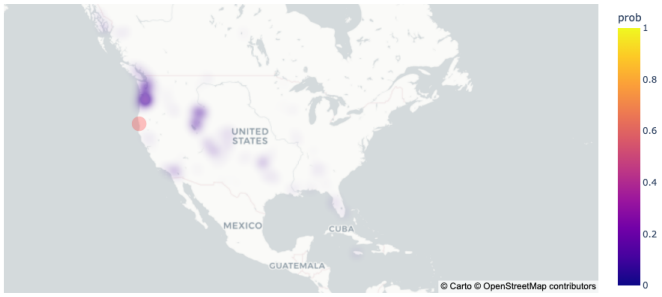


Figure 5: Convergence of LID models. Initializing from geolocation models (Geoloc-init), helps LID models to converge more quickly.

Figure 5 shows the language identification validation accuracy during training of models initialized from pretrained geolocation models and MMS models.



(a) Attention Weights



(b) Heatmap

Figure 6: (a) The attention weights for positions of an utterance from the FLEURS corpus. Note that regions of silence before and after the utterance, but also between words, are visible and can be seen where the attention weights are 0. (b) The binary-cross entropy geolocation model predictions on a radio station in the Radio Valid set. The red dot marks the broadcast location.

B Qualitative Evaluation of Geolocation

B.1 Attention Pooling

An advantage of the attention pooling is that we can ensure that the geolocation model is actually focusing on regions of speech to make predictions and not on channel artifacts, or background music. We see an example of this in Figure 6a.

B.2 Cross-Entropy Heatmaps

One advantage of the cross-entropy based models is that they can be used to create heatmaps, which attempt to answer the question, where are you from? This is a natural way to depict predictions for this task. An example is shown in Figure 6b. Areas outside of North American had no probability mass for this example and so we zoomed in on North America to make visualization easier.