What LID Systems Say About Dialectal Variation. The Case Of Yiddish, Quechua and Mande

Johanna Cordova, Eric Jordan, Valentina Fedchenko
ERTIM - Inalco
2 rue de Lille, 75007 Paris, France
surname.name@inalco.fr

Abstract

This study investigates the ability of speech-based language identification (LID) systems to handle dialectal variation in low-resource settings and explores whether classification outcomes correspond to phonetic proximity, as well as an exploratory tool for dataset quality. We collected corpora for three macrolanguages Mande, Quechuan, and Yiddish, each presenting distinct internal variation, and evaluated three types of models: GMM, Whisper, and wav2vec2-based architectures. Models were tested both within language families and across the entire multilingual dataset to assess generalization. Layer-wise classifiers built on wav2vec2-XLSR embeddings were used to identify the layers most sensitive to phonetic or phonological features. Results show that simple GMM models can generalize well in small, highly similar datasets, while Whisper-based classifiers tend to overfit, particularly on closely related dialects. Wav2vec2-XLSR (layer 12 + MLP) captures better fine phonetic and prosodic distinctions, suggesting that embeddings encode nuanced pronunciation cues. For datasets with more diverse sources like Quechua, Whisper demonstrates better generalization. Overall, LID classifiers can both reveal linguistic patterns and highlight dataset quality issues, with model architecture and layer-specific representations shaping performance.

Keywords: Spoken Language Identification (SLID), dialectal variation, language classification

1. Introduction

While automatic speech processing achieves impressive results, including for an increasing number of low-resource languages, the more complex issue of handling dialectal variation remains largely underexplored and typically yields unsatisfactory results in large multilingual models (Joshi et al., 2024). The work presented in this article seeks to address two questions: is it possible to quantify the level of granularity in variation that speech processing systems based on embeddings are able to handle in a low-resource context? If so, can the classifications produced by these models be used as a new metric for measuring linguistic proximity between variants? To test these hypotheses, we begin by collecting corpora of variants for three highly distinct macrolanguages, each presenting different issues in terms of internal classification: the Mande languages, the Quechuan languages, and Yiddish dialects. Based on these corpora, we train or fine-tune three types of spoken language identification (SLID) models (GMM, wav2vec2, Whisper) in order to evaluate the classification and get an initial idea of the difficulty of the task with limited data. At the same time, we train a series of classifiers using embeddings extracted from each layer of an XLS-R model to determine whether the layers known to mainly encode phonological features (Pasad et al., 2021) yield higher classification accuracy. This will give us an idea of the extent to which the classification is based on phonetic/phonological criteria. The primary challenge of this task lies in the fact that numerous factors influence the results: the quality, type, quantity, and diversity of available data, as well as the impact of pretraining in the multilingual models used. Initially, our work will focus on identifying how these biases manifest in the languages being studied, thereby guiding more in-depth future research in the resulting directions. This work is therefore intended to be exploratory and preliminary, while providing the necessary data for further studies.

1.1. Quechuan languages

The Quechuan languages, spoken mainly in Peru, Bolivia, Ecuador and Colombia form a linguistic family with a complex internal structure. In terms of phylogenetic classification, the family is divided into two main branches: QI, grouping the central Peruvian languages, and QII (all other languages in the family) (Torero, 1970 [2002]; Cerrón-Palomino, 1987 [2003]; Blum et al., 2023). The number of distinct languages within the family remains unresolved. In Camacho Rios et al. (2024), it is proposed that between 12 and 17 languages can be distinguished, based on the criteria of mutual intelligibility, phonological and morphosyntactic distance, lexical distance, and sociolinguistic perception. The number of variants within these languages is especially difficult to estimate as the variations form a continuum (Mannheim, 2018). The work of SIL (Summer Institute of Linguistics) led to the creation of 44 ISO codes for modern Quechua languages, 43 of which are still spoken today. The criteria used to assign these ISO codes are unclear and many of them refer to linguistic variants

Proceedings of SPEAKABLE @ LREC 2026, pages 98–108

rather than distinct languages. Therefore, the current classification of Quechua languages leaves room for improvement. Any improvement on this classification should build on the existing work, ideally enhancing it through empirical foundations. We believe that the analysis of speech signal can be of particular interest to address this issue.

1.2. Yiddish dialects

Yiddish is a language of the Indo-European family, belonging to the Germanic branch, which arose through a historical language shift from Middle High German. Yiddish is considered by Ethnologue[1] as a "macrolanguage", divided into nearly extinct Western Yiddish (yih) and endangered Eastern Yiddish (ydd). Yiddish dialects are not assigned distinct ISO language codes. The present study focuses exclusively on Eastern Yiddish. Dialological research (Weinreich et al., 2008; Jacobs, 2005; Beider, 2015) has shown that the most salient isoglosses separating three major Eastern Yiddish dialects (Nothern-Eastern – NEY, Central-Eastern – CEY and Southern-Eastern – SEY) are primarily phonetic, particularly in the vowel system. Differences at the morphosyntactic and lexical levels are also attested, but they tend to be less systematic, less frequent, and less discriminative for dialect classification. Hasidic Yiddish is a somewhat special case, constituting a distinct sociolinguistic phenomenon: due to intense and ongoing contact with American English and Modern Hebrew, it exhibits a range of contact-induced features, including significant phonetic innovations (Nove, 2021).

1.3. Mande languages

The Mande languages constitute a genealogically coherent family traditionally classified within the Niger–Congo phylum, although their precise position within Niger–Congo has been subject to debate due to their typological distinctiveness. Mande languages are spoken across a wide area of West Africa, extending from Senegal and Mali to Côte d'Ivoire, Guinea, and Burkina Faso. The family is internally divided into several branches, usually grouped into Western, Eastern, Southern, and Southwestern Mande (Kastenholz, 1996; Vydrine, 2004). Bambara (bam), Wan (wan), and Ngen (gnj), the languages considered in the present study, belong to different subgroups within Western and Southern Mande and exhibit substantial phonological and morphosyntactic diversity (Vydrine, 2004; Nikitina and Treis, 2020; Korol, 2022). Unlike Quechua and Yiddish, each Mande language is assigned a distinct ISO 639-3 code, reflecting substantial mutual unintelligibility and perception of clear linguistic boundaries between varieties. Nevertheless, dialect continua and contact-induced variation are also attested, particularly in regions of intense multilingualism.

Typologically, Mande languages are predominantly isolating to slightly agglutinative, display strict SOV word order, and are characterized by tonal systems that play a central role in lexical and grammatical distinctions. Tone inventories and tonal processes vary considerably across the family, contributing to significant phonetic and phonological distance between languages (Vydrine, 2004).

1.4. Linguistic diversity and ISO code attribution in our dataset

The presented dataset of chosen language groups is treated as a structured ensemble of linguistic varieties whose distances are approximated through their ISO language codes. ISO codes represent a formalized attempt to capture perceived distances between linguistic variants, yet the criteria underlying these assignments vary considerably. Linguistic distance is understood here as a complex qualitative notion, grounded in accumulated linguistic evidence (phonological, morphosyntactic, and lexical), mutual interpretability, and the social judgment of speaker communities and linguistic institutions.

The dataset includes three distinct configurations of ISO coding practices. First, in the case of the Mande languages, ISO codes are predominantly assigned to what are widely regarded as separate languages, corresponding to relatively large phonological and grammatical distances. Second, Quechuan languages presents a scenario, where ISO codes have been assigned inconsistently and often correspond to dialectal or sub-language-level variation within a dense dialect continuum. Third, Eastern Yiddish dialects constitute a case in which no separate ISO codes are assigned, despite well-established dialectological, mostly phonetic, distinctions.

From a typological perspective, the dataset brings together highly diverse phonetic and morphosyntactic systems. It includes tonal Mande languages, which rely heavily on pitch contrasts; agglutinative Quechua languages, characterized by complex suffixal morphology; and inflectional Yiddish dialects, whose phonetic profiles reflect Middle and Eastern European areal features. These typological differences have direct consequences for how the languages are represented in acoustic language models.

[1] https://www.ethnologue.com/language/yid/

2. Datasets

To ensure comparability, all datasets created contain 2.5 hours of audio data per ISO code or variant. The files have all been converted to 16kHz mono WAV format and are between 5 and 30 seconds long. All datasets have a train, validation and test split, each containing utterances produced by different speakers. We have taken particular care in preparing the data in order to reduce speaker- and microphone-recognition biases. The overall quality of the corpora was measured using the Signal-To-Noise Ratio; a visualisation of the results is provided in Appendix 7, Figure 9. The overall synthesis of the corpus, including the train-dev-test split distribution and the number of speakers per split, is presented in Figure 11.

2.1. Quechuan languages

Two main sources with ISO code identification are available for Quechuan languages:

- Mozilla Data Collective[2] (MDC). This successor to Common Voice brings together speech datasets developed by speaker communities. For Quechua languages, 17 datasets are available, covering 16 ISO codes. All but one of these datasets consist of scripted speech: speakers read very short sentences (3.5 words in average) with simple syntax.

- Recordings of the Bible and New Testament. Many variants have recordings available online, some of which are prepared for NLP purposes[3]. This is the main data source for Quechuan languages in the large multilingual models.

Outside of these corpora, the volume of available audio content in Quechua varies significantly depending on the language. Southern Quechua (QIIC linguistic group) stands out as the majority language in terms of both speakers number and media visibility (social networks, radio, television), which facilitates data collection. For the LID experiments, we will focus on this language, comparing 4 closely related variants, for which we have collected between 15 and 30 minutes of additional data from the media or from recordings in our possession. These variants are divided into two sub-groups: Chanka (`quy`) and Collao (`quz`, `qxp`, `quh`). At the phonological level, they differ in that the Collao variants feature ejective and aspirated consonants. This distinction allows for rapid identification, even with short utterances. Within the Collao group, the differences are more subtle, as the variants share the same phonological system; our experiments will seek to determine whether the models are able to achieve this level of granularity in distinguishing between the variants.

[2] https://datacollective.mozillafoundation.org/datasets?q=quechua

[3] https://huggingface.co/datasets/Flux9665/BibleMMS

2.2. Yiddish dialects

Two Yiddish datasets were constructed for the experiments: one comprising the full training pipeline (train, development, and test splits), and a separate dataset for out-of-domain evaluation. Both datasets consist of three Eastern Yiddish varieties: Northern Eastern Yiddish (yid_ney), Central Eastern Yiddish (yid_cey), and Southern Eastern Yiddish (yid_sey). Data were drawn from the Corpus of Spoken Yiddish in Europe (CSYE, Bleaman and Nove (2025)). As the corpus consists of post-Holocaust interviews with native Yiddish speakers living in the United States, a preprocessing pipeline was applied to filter out interviewer speech produced by non-native speakers, as well as segments containing Yiddish–English code-switching.

At the initial stage, our experiments included Hasidic Yiddish data extracted from the Mozilla Common Voice corpus (Ardila et al., 2020). However, the preliminary observations suggested a strong bias toward the Hasidic variety, which systematically attracted the majority of the model's predictions. We hypothesized that the model was relying on non-linguistic cues. Therefore, to reduce source-related effects, we restricted the dataset to European Yiddish dialects, which are more homogeneous both linguistically and with respect to the types of recordings available.

The second out-of-domain test dataset was drawn from a different source corpus—Reading Electronic Yiddish Documents (REYD, Webber et al. (2022))—for the `yid_ney` and `yid_cey` varieties. This dataset represents a distinct type of data, consisting of read speech from books, and involves different speakers and recording conditions. For `yid_sey`, we used a YouTube interview with a native speaker of the dialect, which was diarized in order to bring it into a format comparable to the other datasets. The composition of this out-of-domain evaluation dataset is presented in Table 1.

Dialect	No. of speakers	No. of segments
`yid_ney`	2	461
`yid_cey`	2	387
`yid_sey`	1	397

Table 1: Out-of-domain test corpus for Yiddish dialects

2.3. Mande languages

The dataset comprising the three Mande languages, Bambara, Wan, and Ngen, represents a subset of a continuous fieldwork effort spanning more than twenty years within Mande-speaking communities. The data were collected in naturalistic settings and cover a range of speech recordings, including spontaneous narratives, elicited vocabularies, and translations of isolated phrases. This diversity of data provides a broad sample of speech styles and communicative contexts, while also introducing variability in prosodic and segmental realization.

The recordings for Wan and Ngen originate from unpublished personal fieldwork corpora[4]. Prior to inclusion in the present study, these data underwent preliminary processing, including semi-automatic segmentation of the audio into phrase-level units. In addition, metadata were compiled on the basis of field notes provided by the original data collectors, including information on discourse type, topical domain, and the number of speakers involved in each recording. This metadata was used both to ensure internal consistency of the dataset and to support controlled experimental splits where possible.

The Bambara data are substantially more voluminous and heterogeneous. A portion of this dataset is publicly available through the CORPORAN platform (Corpus oraux annotés), hosted by TGIR Huma-Num[5] (Vydrin, 2013, 2014). These recordings include annotated oral corpora collected across multiple regions and communicative settings. For the purposes of the present experiments, a curated subset of the Bambara recordings was selected to ensure comparability with the Wan and Ngen datasets in terms of segment duration, recording conditions, and distribution of recording types.

ISO Code	Split	No. of speakers	No. of segments
quy	train	> 7	654
	dev	> 2	167
	test	> 11	100
quz	train	> 12	797
	dev	9	118
	test	12	100
qxp	train	9	840
	dev	1	202
	test	9	100
quh	train	> 8	711
	dev	> 2	128
	test	> 8	100
bam	train	9	594
	dev	3	148
	test	2	148
wan	train	2	509
	dev	1	113
	test	1	112
gnj	train	2	544
	dev	2	146
	test	1	136
yid_ney	train	25	977
	dev	9	209
	test	5	203
yid_cey	train	14	903
	dev	5	220
	test	5	216
yid_sey	train	26	1018
	dev	5	212
	test	5	229

Table 2: Numbers of speakers and segments across all datasets and splits

3. State of the art in SLID

The different languages included in this study do not receive the same amount of coverage in the training data of the state of the art acoustic large language models. For the Mande languages, the Whisper model (Radford et al., 2022) includes training data for Bambara but lacks coverage of Wan and Ngen. Omni-ASR (Omnilingual et al., 2025) does not include Wan or Ngen, although it does incorporate other Mande languages that are phonetically similar to Wan, such as Mwan (moa). Eastern Yiddish, treated as a single entity in most models, receives no dialect-specific acoustic modelling, despite the phonetic salience of dialectal distinctions.

Quechuan languages receive comparatively better coverage: for the LID task, the MMS model reports support for 34 ISO codes. To assess the actual performance of this model, we evaluate it using the test split of our dataset, supplemented with a few utterances from the train split to ensure 100 files per language. The overall accuracy is relatively low, with a micro-F1 score of 0.46. The confusion matrix in Figure 1 reveals that the language most distinct from the other three (quy) is the best predicted. The overprediction of the quh class, which absorbs similar variants, may be due to an imbalance in the training data. This potential imbalance must be accounted for when analysing the results produced by pretrained models.

State-of-the-art dialect classification systems are

[4]A short fragment of Ngen corpus was published in Pangloss Collection of oral data: https://pangloss.cnrs.fr/corpus/Ngen. In written form, part of the Wan corpus has been incorporated into the Speech Reporting Corpus (Nikitina, 2023), a curated collection of narrative texts focusing on discourse reporting strategies in storytelling

[5]https://corporan.huma-num.fr/

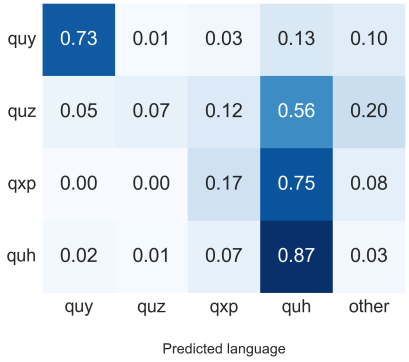


Figure 1: MMS LID classification result

often built on unsupervised acoustic representations such as traditional spectral features (Mel-Frequency Cepstral Coefficients), temporal derivatives (shifted-delta cepstra), or high-resolution signal processing outputs such as Single Frequency Filtering (SFF) and Zero-Time Windowing (ZTW), to capture phonetic and prosodic variability. When paired with robust classifiers, either classical statistical models (GMM, SVM) or neural architectures (Temporal Convolutional Networks, Time Delay Neural Networks, and variants like ECAPA-TDNN), the resulting systems achieve satisfactory performance on fine-grained dialect tasks: such combinations typically produce accuracies in the 80 % range according to multiple benchmarks in the literature (Agrawal et al., 2016; Chittaragi et al., 2018; Tibi and Messaoud, 2025; Kethireddy et al., 2022). Some research has already been conducted to understand which features these systems primarily rely on for classification decisions. For instance, (Bafna and Wiesner, 2025)'s work indicates that neural systems such as ECAPA-TDNN particularly encode accent-related information.

Conversely, handling dialectal variation is poorly suited to the architectural design of large language models: firstly, the lack of available corpora for non-standard varieties deprives tools of training data. Furthermore, the architectures of widely used models are designed to smooth out variations to ensure robustness against noise, often leading to outputs that are homogenised toward the closest majority variant. Consequently, attempts to identify dialects using models that encode the signal through embeddings yield mixed results, with a rapid decline in accuracy as dialectal similarity and the number of classes to discriminate increase (Joshi et al., 2024; Lonergan et al., 2023). The work of Van Dinh et al. (2024) on Vietnamese dialects serves as a particularly compelling example: the authors compiled a corpus of 102 hours of audio curated from publicly available sources for 63 regional variants. From this, they fine-tuned several flavours of pre-trained models (wav2vec2 and Whisper) for dialect identification, producing classification models with varying granularity. Each series of models is trained to predict 1) the region (3-class); 2) the dialect within each region (19 to 25-class); 3) the dialect at the global level (63-class). The results show high accuracy for the first task, but performance drops dramatically as the classification grows more complex. Significant F1-score variations in the performance of the 19-class models across regions suggest that there is a threshold of linguistic similarity beyond which dialectal variations are no longer recognised.

Understanding how embedding systems encode acoustic signals is a particularly important challenge, and several studies have addressed this question (Pasad et al., 2021; English et al., 2024), showing that:

- Early layers focus more on acoustic features (low-level signal properties),
- Middle layers begin to encode phonetic structure as well as tone and stress (de la Fuente and Jurafsky, 2024),
- Higher layers reflect more abstract linguistic units.

4. Experiment and results

Several speech-based language identification models were evaluated in this study in order to compare architectures with different inductive biases and to reveal possible correlations with linguistic distances and the system of ISO codes. The tested models include: (i) a Gaussian Mixture Model (GMM) baseline; (ii) *Whisper*, a transformer-based encoder–decoder model originally designed for automatic speech recognition (ASR) (Radford et al., 2022); and (iii) self-supervised speech representations based on `wav2vec2-base` (Baevski et al., 2020).

The models were evaluated under two experimental configurations: (a) separately within each language or dialect group, and (b) jointly on the entire multilingual dataset. This design allows us to assess model generalization under varying degrees of linguistic similarity and data scarcity. Specifically, we examine performance on: (i) highly similar and largely mutually intelligible dialects (Eastern Yiddish and Quechua) (ii) non-interintelligible languages within a genetically homogeneous family (Mande languages); and (iii) macro-groups of genetically and typologically diverse languages (Mande vs. Quechua vs. Yiddish).

4.1. GMM-based language identification

To establish a baseline free from biases induced by model pretraining, we first train a Gaussian

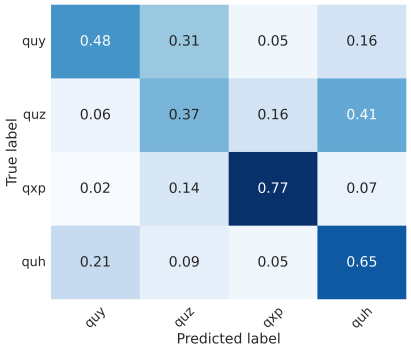
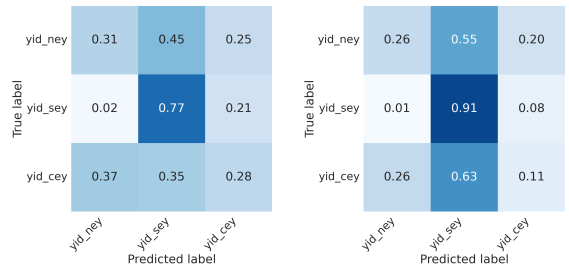


Figure 2: GMM predictions for Quechua variants

Mixture Model (GMM)-based classification system (Torres-Carrasquillo et al., 2004), using MFCC feature vectors extracted at the frame level from each speech utterance. Features are standardised using a scaler. The normalized MFCC vectors are then used to train class-specific GMMs via maximum likelihood estimation, whilst model hyperparameters (number of mixture components, covariance structure and covariance regularisation) are optimized through grid search.

The results of this classification are consistent across the tested languages: for languages with marked differences, such as Mande languages, the classification has a 100% accuracy. For Quechua variants, the accuracy drops to 0.57, with disparities among classes (see Figure 2): the `quz` variant is frequently confused with `quh`, with which it shares the same phonological system. The strong performance for `qxp` is likely due to the higher quality of its data and its lower internal variability. Accuracy is even lower for Yiddish dialects (see Figure 3) : 0.46 with in-domain test set and 0.43 when adding out-of-domain samples.



(a) In-domain test (b) Out-of-domain test

Figure 3: GMM predictions for Yiddish dialects

4.2. Whisper-based language identification

Whisper is primarily designed as an ASR model and not as a dedicated LID system. In its original architecture, language identification is performed implicitly via the generation of a single language token at the beginning of decoding. As noted in prior work, this design introduces structural limitations for standalone LID tasks: because the language label corresponds to only the first generated token, the model relies predominantly on acoustic cues and makes only limited use of deeper linguistic context. As a consequence, Whisper's LID accuracy has been reported to reach approximately 80.3% in some evaluations, despite its strong ASR performance (Radford et al., 2022; Shen et al., 2023).

We tested Whisper in two configurations: (1) native language identification using the built-in LID classifier, and (2) LID based on acoustic embeddings extracted from intermediate layers of Whisper-small (layer 8) and Whisper-large (layer 20) (Yue et al., 2026), followed by an MLP classifier.

For the first configuration (1), we started by investigating Whisper-based LID on the Yiddish dataset, which represents a particularly challenging scenario due to the high phonetic similarity between dialects and the limited number of classes (three). An initial training run yielded an extremely high F1-score after the very first epoch (0.83), clearly indicating severe overfitting. To encourage the model to generalize over linguistic properties, we tested several increasingly constrained fine-tuning configurations on Whisper-small, always starting from pretrained weights. Our strategies varied in the degree of parameter freezing, from partial encoder fine-tuning to adaptation of the language identification (LID) head only. These progressively restrictive configurations effectively reduced the F1-score observed after the first training epoch to 0.45, indicating a decrease in overfitting and a reduced reliance on speaker- or corpus-specific cues.

Model training was conducted using a learning rate of 1×10^{-3}, with early stopping applied based on development set performance and a patience value of 3 epochs. To reduce overfitting and improve optimization stability, weight decay was set to 0.01 in the optimizer. Given the low-resource nature of the datasets and the resulting imbalance between language and dialect classes, a weighted cross-entropy loss function was employed.

The second configuration, particularly the MLP classifier built on representations from layer 20 of Whisper-large, shows reduced overfitting, it effectively separates more distantly related Mande languages, but still struggles to distinguish linguistically very similar varieties, such as Yiddish dialects. It nevertheless yields insightful results on the Quechua dataset, as illustrated in Figure 4.

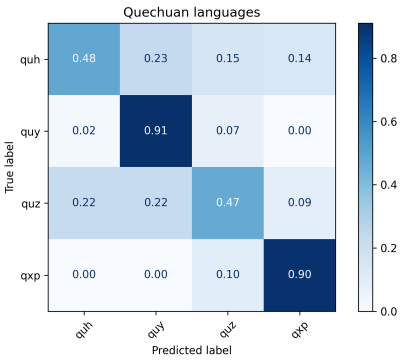


Figure 4: Confusion Matrix for Whisper-based LID (layer 20 + MLP) for Quechuan variants

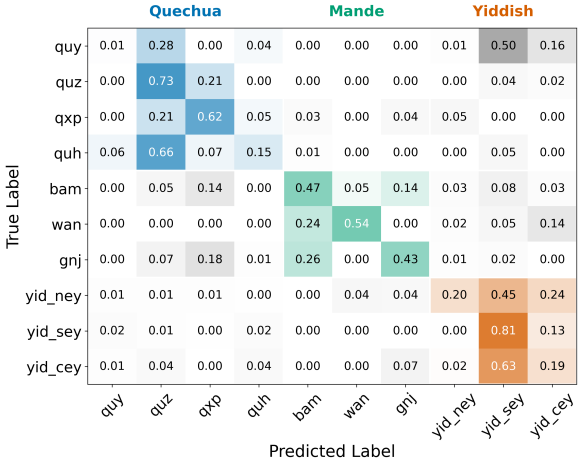


Figure 5: Confusion Matrix for Classifications on Macro test

4.3. Wav2Vec-based language identification

We used the widely adopted wav2vec2 architecture in 2 configurations. Firstly, the `wav2vec2-base` adapted for our task by fine-tuning the model alongside a classification head on our data (using the HuggingFace Wav2VecForSequenceClassification [6] config) and secondly, using the xlsr model extracting different layers of the model and training a classification head on the extracted embeddings.

Several configurations were tested for `wav2vec2-base`, however given the limited amount of data available from our datasets configuring the training process to allow the models to learn without overfitting the training data was of the utmost importance. With this in mind several configurations of hyperparameters and freezing of layers were tested. The most effective training configuration was obtained by freezing the CNN feature extractor and all but the final (12th) layer of the model.

The best balance between learning from the data while avoiding overfitting was found using a learning rate of 3×10^{-6} while performing a warmup on the first 6% of steps and using a weight decay of 0.1. The training was run for a total of 20 epochs with an early stopping patience of 5 epochs.

The most interesting result obtained for this model was on the macro test with all classes, where the model achieved an overall accuracy of 0.42 across the ten classes. However, perhaps more interesting than the pure performance of the model is how errors in classification tended to stay within language families as shown in Figure 5. This suggests that, in most cases, some common features within each of the language families is being learnt by the model. While the intra-family classification described above and this macro-performance indicate some linguistic criteria are playing a role in the classification performance, other factors of convergence cannot be entirely ruled out. Despite efforts to homogenize the data across language families, it must be noted that greater similarity in recording types (i.e. fieldwork data for Mande languages or Bible readings for Quechua) are also potential explanations for this intra-family classification performance.

4.4. Classifying from layer-extracted embeddings

We extracted hidden representations from each of the 24 transformer layers of `wav2vec2-xls-r-300m` and performed mean pooling over the temporal dimension to obtain a one-dimension output. These layer-wise embeddings were then used as input to a simple multilayer perceptron classifier (PyTorch SimpleClassifier) with a dropout rate of 0.3. We trained a separate classifier for each layer's embeddings and evaluated their performance on the test split, in order to investigate whether the layers that encode the richest phonological information are also those that yield the highest dialect classification accuracy.

The results for Yiddish, shown in Figure 6 support this theory: a peak in performance is indeed observed for all variants at the twelfth layer of the model. This trend, however, is not observed for Quechua (Figure 7), where accuracy remains stable (and high from the earliest layers), possibly due to the greater presence of this language in the model's pretraining data.

5. Concluding discussion

These results suggest that, in a low-resource setting involving a small number of highly similar linguistic varieties, such as the three Eastern Yiddish

[6] https://huggingface.co/docs/transformers/en/model_doc/wav2vec2

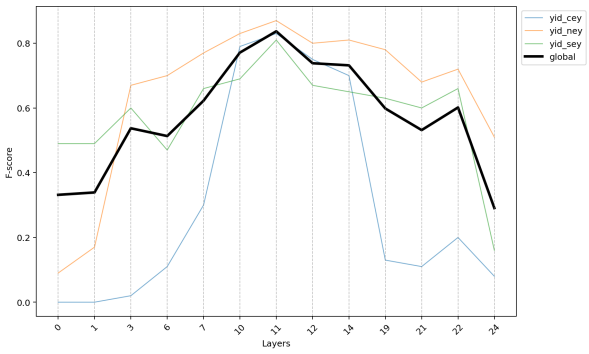


Figure 6: Fscore evolution through layers for Yiddish

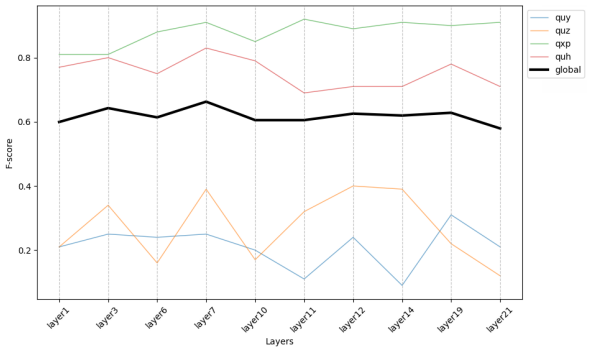
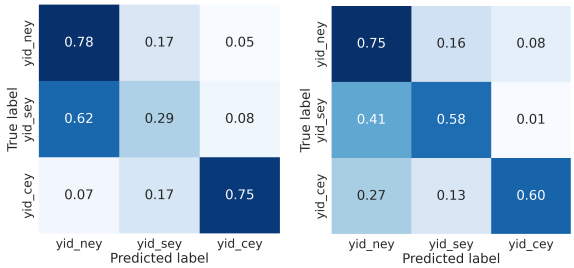


Figure 7: Fscore evolution through layers for Quechua



(a) In-domain test (b) Out-of-domain test

Figure 8: MLP predictions for Yiddish by classifier trained with embeddings from the 12th layer

dialects, the relatively simple GMM architecture exhibits a strong capacity for generalization. In contrast, Whisper-based classifiers appear to rely on a wider range of cues that may include extra-linguistic factors, limiting their ability to capture fine-grained phonetic distinctions between closely related varieties. The most granular analysis is provided by the wav2vec2-xlsr model (layer 12 + MLP). The model appears to rely more on finer phonetic and prosodic features than on coarse phonological cues: for instance, `yid_sey` speakers in our test dataset often exhibit what could be described as a "Russian accent" and are frequently bilingual in Russian, whereas `yid_cey` speakers are typically bilingual in Polish, a pattern not shared with `yid_ney` speakers. The wav2vec2-based classifier may capture these subtleties in pronunciation.

This interpretation is further supported by the layer-wise visualization of F1 scores (6), which shows that the F1 trajectory for `yid_cey` diverges from that of `yid_ney` and `yid_sey`, which are closer to each other (6). There is no reason to believe that the model was exposed disproportionately to `yid_ney` or `yid_sey` during training, since `yid_cey` is actually better represented in publicly available online Yiddish data than `yid_sey`. The layer-wise analysis of wav2vec2 F1 scores 6 indicates that performance is highest at layer 12, a layer often associated with phonetic sensitivity in probing studies (Pasad et al., 2021; San et al., 2024). While not conclusive, this pattern suggests that the wav2vec2-based classifier may rely primarily on phonetic representations when performing language identification generalization.

Conversely, for datasets with greater internal diversity of domain, such as the Quechuan languages, the Whisper-based model demonstrates superior generalization capabilities, likely benefiting from its broader representational capacity.

Finally, despite efforts to reduce bias introduced by the audio data available it cannot be excluded that the results from the macro-language classification may in fact be influenced by similarities between the files and their origins (fieldwork data, bible recordings etc.). Further work will aim to investigate the influence of these categories of data on classification performance. Nonetheless, this result indicates the promise of applying these models when processing even quite diverse corpora to determine possible similarities between audio files.

Future work will extend the present study to a broader range of varieties within the considered language groups, with particular emphasis on Quechuan and Mande languages. We plan to deepen the layerwise analysis of model predictions in order to better understand which linguistic properties are captured at different representational levels. In addition, we will systematically investigate the relationship between data quality and diversity and their impact on model performance. Finally, we aim to explore methods for extracting more fine-grained linguistic knowledge from speech models, with the goal of improving their usefulness for dialectological research.

6. Acknowledgments

We sincerely thank the linguists who generously shared their fieldwork datasets with our team for the purposes of this experiment: Tatiana Korol, Tatiana Nikitina and Valentin Vydrine. The work is supported by the French National Research Agency

and Ministry of Higher Education, Research and Innovation (MESR).

7. Bibliographical References

S. Agrawal, Aruna Jain, and S. Sinha. 2016. Analysis and modeling of acoustic information for automatic dialect classification. *International Journal of Speech Technology*, 19:593–609.

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

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Niyati Bafna and Matthew Wiesner. 2025. Lid models are actually accent classifiers: Implications and solutions for lid on accented speech. *arXiv preprint arXiv:2506.00628*.

A. Beider. 2015. *Origins of Yiddish Dialects*. Oxford Linguistics. Oxford University Press.

Isaac L Bleaman and Chaya R Nove. 2025. The corpus of spoken yiddish in europe: Goals, methods, and applications. *Language Documentation & Conservation*, 19.

Frederic Blum, Carlos Barrientos, Adriano Ingunza, and Zoe Poirier. 2023. A phylolinguistic classification of the quechua language family. *Indiana*, 40(1).

Gladys Camacho Rios, Simeon Floyd, and Félix Julca Guerrero. 2024. ¿cuántas lenguas quechuas hay? una estimación del número de lenguas quechuas. *Lexis*, 48(1):34–77.

Rodolfo Cerrón-Palomino. 1987 [2003]. *Lingüística quechua*. Centro de estudios rurales andinos" Bartolomé de Las Casas".

Nagaratna B Chittaragi, Ambareesh Prakash, and Shashidhar G Koolagudi. 2018. Dialect identification using spectral and prosodic features on single and ensemble classifiers. *Arabian Journal for Science and Engineering*, 43(8):4289–4302.

Antón de la Fuente and Dan Jurafsky. 2024. A layer-wise analysis of mandarin and english suprasegmentals in ssl speech models. In *Interspeech 2024*.

Patrick Cormac English, Erfan A. Shams, John D. Kelleher, and Julie Carson-Berndsen. 2024. Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6685–6689.

Neil G. Jacobs. 2005. *Yiddish: A linguistic introduction*. Cambridge University Press.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57:1 – 37.

Raimund Kastenholz. 1996. Sprachgeschichte im west-mande. köln: Rüdiger köppe verlag.

Rashmi Kethireddy, Sudarsana Reddy Kadiri, and S. Gangashetty. 2022. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations. *The Journal of the Acoustical Society of America*, 151 2:1077.

Tatiana Korol. 2022. Preliminary description of ngen pronominal elements. *Mandenkan*, 68:43–58.

Liam Lonergan, Mengjie Qian, Neasa Ní Chiaráin, Christer Gobl, and Ailbhe Ní Chasaide. 2023. Towards spoken dialect identification of irish. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages*.

Bruce Mannheim. 2018. Three axes of variability in quechua. *The Andean World*, pages 507–523.

Tatiana Nikitina. 2023. *A narrative corpus of Wan*. CNRS-LLACAN & LACITO.

Tatiana Nikitina and Yvonne Treis. 2020. The use of manner demonstratives in discourse: A contrastive study of Wan (Mande) and Kambaata (Cushitic). In Åshild Næss, Anna Margetts, and Yvonne Treis, editors, *Demonstratives in discourse*. Language Science Press.

Chaya R. Nove. 2021. *Outcomes of language contact in New York Hasidic Yiddish*, pages 43–71. Berlin: Language Science Press.

Team Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe

Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921. IEEE.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*.

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

Peng Shen, Xuguang Lu, and Hisashi Kawai. 2023. Generative linguistic representation for spoken language identification.

Nejib Tibi and Mohamed Anouar Ben Messaoud. 2025. Arabic dialect classification using an adaptive deep learning model. *Bulletin of Electrical Engineering and Informatics*.

Alfredo Torero. 1970 [2002]. *Idiomas de los Andes. Lingüística e historia*. Editorial horizonte.

Pedro A Torres-Carrasquillo, Terry P Gleason, and Douglas A Reynolds. 2004. Dialect identification using gaussian mixture models. In *Odyssey*, pages 297–300.

Nguyen Van Dinh, Thanh Chi Dang, Luan Thanh Nguyen, and Kiet Van Nguyen. 2024. Multi-dialect vietnamese: Task, dataset, baseline models and challenges. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7498.

Valentin Vydrin. 2013. Bamana Reference Corpus (BRC). *Procedia - Social and Behavioral Sciences*, 95:75–80.

Valentin Vydrin. 2014. Bambara and maninka manding languages written corpora project ("projet des corpus écrits des langues manding : le bambara, le maninka") [in French]. In *TALN-RECITAL 2014 Workshop TALAf 2014 : Traitement Automatique des Langues Africaines (TALAf 2014: African Language Processing)*, pages 109–113, Marseille, France. Association pour le Traitement Automatique des Langues.

Valentin Vydrine. 2004. Areal and genetic features in west mande and south mande phonology: In what sense did mande languages evolve? *Journal of West African Languages*, XXX(2):113–125.

Jacob Webber, Samuel K. Lo, and Isaac L. Bleaman. 2022. Reyd – the first yiddish text-to-speech dataset and system. In *Interspeech 2022*, pages 2363–2367.

M. Weinreich, P. Glasser, P.E. Glasser, Y.I.J. Research, and S. Noble. 2008. *History of the Yiddish Language*. Number 1 in History of the Yiddish Language. Yale University Press.

Zhengjun Yue, Devendra Kayande, Zoran Cvetkovic, and Loweimi Erfan. 2026. Probing whisper for dysarthric speech in detection and assessment. In *2026 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2026*. IEEE.

Appendix

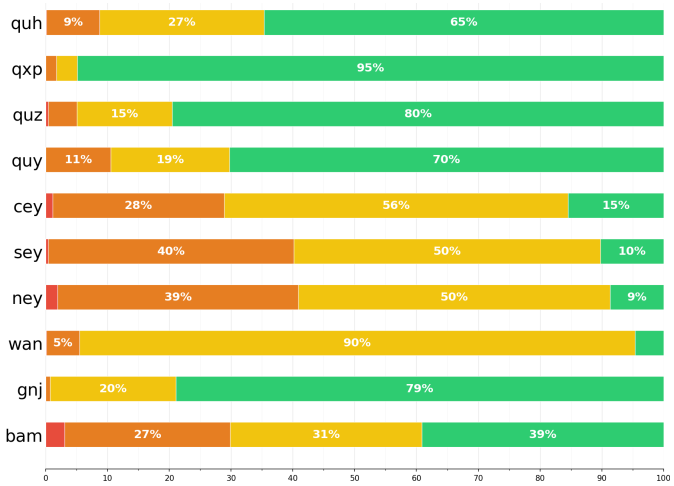


Figure 9: Estimation of corpus quality using the Signal-To-Noise Ratio (SNR).

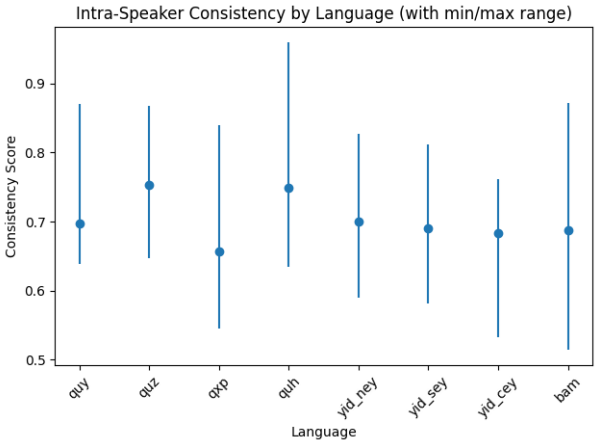


Figure 10: Intra-Speaker Temporal Consistency across the corpus. Data are not provided for Ngen and Wan because the speakers were not manually identified.

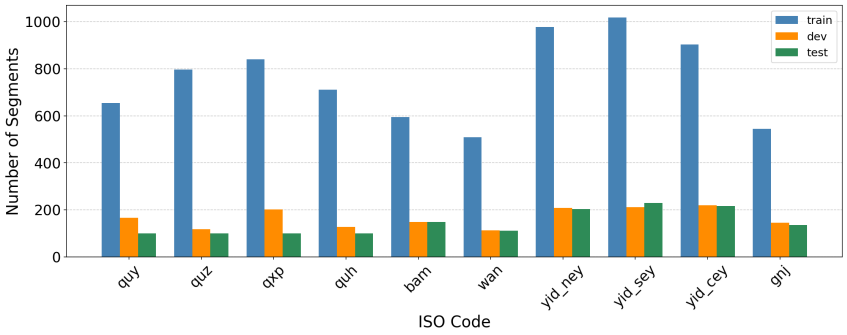


Figure 11: Segment distribution across all languages / dialects and splits