Speech Analysis of Language Varieties in Italy

Moreno La Quatra*†, Alkis Koudounas*‡, Elena Baralis‡, Sabato Marco Siniscalchi††
†Kore University of Enna, ‡Politecnico di Torino, ††Università degli Studi di Palermo
†moreno.laquatra@unikore.it, ‡{name.surname}@polito.it, ††sabatomarco.siniscalchi@unipa.it

Abstract

Italy exhibits rich linguistic diversity across its territory due to the distinct regional languages spoken in different areas. Recent advances in self-supervised learning provide new opportunities to analyze Italy's linguistic varieties using speech data alone. This includes the potential to leverage representations learned from large amounts of data to better examine nuances between closely related linguistic varieties. In this study, we focus on automatically identifying the geographic region of origin of speech samples drawn from Italy's diverse language varieties. We leverage self-supervised learning models to tackle this task and analyze differences and similarities between Italy's regional languages. In doing so, we also seek to uncover new insights into the relationships among these diverse yet closely related varieties, which may help linguists understand their interconnected evolution and regional development over time and space. To improve the discriminative ability of learned representations, we evaluate several supervised contrastive learning objectives, both as pre-training steps and additional fine-tuning objectives. Experimental evidence shows that pre-trained self-supervised models can effectively identify regions from speech recording. Additionally, incorporating contrastive objectives during fine-tuning improves classification accuracy and yields embeddings that distinctly separate regional varieties, demonstrating the value of combining self-supervised pre-training and contrastive learning for this task.

Keywords: Italian Language Varieties, Spoken Language Identification, Linguistic Analysis

1. Introduction

The study of linguistic variation in the form of varieties, dialects, and regional languages is becoming an increasingly important area of research within the field of natural language processing (Zampieri et al., 2020). Analysis of the differences between related language forms at various levels, from lexical features to grammatical structures to phonological patterns, provides unique opportunities to advance NLP techniques. Examining the nuances between closely related varieties can help improve systems' abilities to handle diverse linguistic inputs and gain a more comprehensive understanding of a language landscape.

Italy represents a fascinating case study for investigating linguistic variety within a single nation (Maiden and Parry, 2006). This country exhibits a rich diversity of local languages concentrated within its geographic borders (Ramponi, 2022). This extensive linguistic heterogeneity arises from Italy's unique historical and cultural influences over time. Furthermore, the integration and use of regional languages alongside Standard Italian have contributed additional layers of complexity to the nation's linguistic landscape. Analyzing Italy's diverse yet interrelated regional languages allows for exploring the cultural and social factors shaping the development of linguistic variations across communities over time.

This paper explores linguistic variation within Italy using data-driven acoustic analysis of speech signals without resorting to intermediate textual transcriptions. We analyze the feasibility of automatically determining the geographic origin of speech samples based solely on their acoustic properties.

In this work, we refer to this task as region or language variety identification rather than dialect classification. This is because the regional languages, such as those spoken across different parts of Italy, are not necessarily distinct dialects of Standard Italian (Avolio, 2009; Ramponi, 2022, *inter alia*). Instead, they represent linguistic variations developed locally within different geographical areas and speech communities in Italy. The term *"language variety"* more accurately describes these regional forms of the language. By using this terminology, we aim to avoid potential limitations or ambiguities of the term *"dialect"*, while still focusing on the core task. The regional varieties are not dialects of Italian *per se* (Berruto et al., 2005), but represent locally developed forms of language usage.

To address the region identification task, we leverage VIVALDI (Tosques and Castellarin, 2013), an extensive collection of speech recordings from cities throughout the country. Uniquely, this dataset contains samples of local language varieties spoken in their native form across Italian regions.

From a methodological perspective, we investigate the use of contrastive learning objectives to both enhance the model's ability to accurately identify the geographic region of speech samples, as well as improve the quality of the learned acous-

[*] Both authors contributed equally to this work.

LREC-COLING 2024, pages 15147–15159
20-25 May, 2024.

tic representations. While contrastive learning is commonly used in a self-supervised setting (Chen et al., 2020; Al-Tahan and Mohsenzadeh, 2021; Giorgi et al., 2021), recent works have shown it can also be effective when applied in a supervised manner (Khosla et al., 2020a; Fu et al., 2022). Specifically, we apply contrastive losses: (i) as an additional pre-training step, (ii) as an auxiliary loss during fine-tuning, and (iii) by combining the two approaches. Experimental results aim to assess how supervised contrastive training can aid the model's ability to classify accurately while inducing meaningful separations between linguistic groups in the embedding space. Our experiments show that utilizing the contrastive objective as an auxiliary loss during fine-tuning leads to the largest improvements compared to other settings.

The main contributions of this work are threefold. First, it is the first effort at classifying Italian language varieties solely from speech data. Second, it explores the use of contrastive learning techniques to improve model accuracy in region identification, while also inducing clearer separations between linguistic groups in the embedded feature space. Third, it provides an in-depth analysis of how well models trained with different contrastive objectives capture relationships between data points from the same versus different Italian regions. The goal is to shed light on the strengths and limitations of the data and the task itself in differentiating subtler nuances between certain challenging regions.

Advancing the ability to accurately differentiate language varieties may enhance language understanding tasks by exploiting the knowledge of local language varieties. It also has promising educational and cultural applications. For example, games or language learning tools could teach aspects of different languages through speech-based interaction and feedback. Automatic speech-based region recognition could also help document and analyze lesser-known regional varieties facing endangerment. Finally, location-targeted personalized advertising could be used to better engage the audience using familiar local words and idioms to build trust and understanding.

2. Related Work

Research on fine-grained classification of language varieties has advanced in recent years. Notable developments include improved data collection and modeling approaches within the speech and natural language processing domains.

NLP approaches to model language varieties. Automatic region classification aims explicitly at predicting the region of origin for linguistic samples based on their textual or linguistic content (Gaman et al., 2020; Dunn and Wong, 2022, *inter alia*). This task differs from geolocation (Rahimi et al., 2017), which seeks to directly pinpoint the exact coordinates where a text sample was recorded rather than classifying the associated region based on linguistic attributes. It also differs from geo-characterization (Adams and McKenzie, 2018), which focuses on predicting descriptive attributes about a location rather than classifying the region associated with a speaker's linguistic content.

Our primary focus is on using deep learning techniques to detect language variety from speech audio data, however, insights from dialectometry, as discussed by Goebl (1989, 2006); Wieling et al. (2014), suggest the possibility of quantifying the similarity between language varieties. While progress has been made in NLP and speech technologies for major languages, work explicitly tailored for Italy's language varieties is still relatively limited. Recent advancements in Italian-specific NLP models, including both sequence-to-sequence models tailored for Italian (Sarti and Nissim, 2022; La Quatra and Cagliero, 2023) and adaptations of decoder-only language models (Basile et al., 2023; Santilli and Rodolà, 2023), have yielded promising results across various tasks. However, most NLP research still adopts a monolithic view of Italian as Standard Italian alone, without representation of local languages and regional varieties that characterize the full linguistic landscape (Ramponi, 2022).

Additional work tailored to local languages and sociolinguistic factors could advance the field toward solutions that better reflect the full spectrum of language varieties spoken in diverse communities across Italy. Pioneering efforts in this direction include the DiatopIt corpus (Ramponi and Casula, 2023a), which stands as the first work focusing on collecting data including diatopic variation beyond Standard Italian as well as the related GeoLingIT shared task (Ramponi and Casula, 2023b), which aimed at identifying the geographic origin of tweets based on their linguistic content. Building upon this foundation, recent studies by Gallipoli et al. (2023b) and Koudounas et al. (2023a) explored novel multi-task learning strategies for enhancing a textual model ability to discriminate between Italian language varieties and solve geolocation challenges.

Recent work has also aimed at developing speech understanding systems for Italian, such as EMOVO (Costantini et al., 2014) targeting emotional speech, or IDEA (Marini et al., 2021) modeling dysarthric speech using isolated words. However, these datasets offer limited domain coverage or lack information on speakers' regional origins. ITALIC (Koudounas et al., 2023b) is the largest Italian speech dataset for intent classification. However, while it collects information on speakers' origins, the recordings are in Standard Italian rather than regional languages.

Identifying language varieties from speech. The task of automatically identifying the language of a speech audio recording is commonly known as spoken language recognition. Large-scale evaluations like the Language Recognition Evaluation (LRE) campaigns have addressed spoken language identification at a broader scale, evaluating systems on a wide range of languages from around the world (Lee et al., 2023; Sadjadi et al., 2018). Recent research has shown that convolutional neural networks and transformer-based models can both achieve high accuracy for language recognition tasks. CNN architectures have been shown to reach state-of-the-art performance even with limited training data (Sarni et al., 2023). On the other hand, transformers leverage self-supervised pre-training to capture robust linguistic patterns and learn broadly generalizable representations (Alumäe et al., 2023). Models employing the Wav2Vec 2.0 (Baevski et al., 2020) architecture can inherently capture language-discriminative information in their lower layers (Bartley et al., 2023). These models can classify unseen languages and adapt to new conditions without further training (Liu et al., 2022), suggesting such capabilities may also enable the models to distinguish closely related languages when provided with specific supervision. The latent, discretized representations learned by these models are shared across languages (Abdullah et al., 2023), allowing for effective language identification. Several studies have already extended this research line to the finer identification of linguistic variations across geographical regions. Developing modeling approaches that can distinguish local varieties requires targeted corpora representing regional linguistic variation. Moisio et al. (2023) introduced the *Lahjoita puhetta* corpus containing diverse samples of colloquial Finnish speech reflecting different sociolinguistic factors and dialects. Similarly, Rotaru et al. (2023) contributed to the RoDia dataset containing Romanian dialectal speech from various regions. These resources open opportunities for designing specific models to analyze curated speech corpora. Dobbriner and Jokisch (2019) demonstrated the effectiveness of combining selected spectral features with Gaussian mixture models for dialect discrimination and classification tasks. Hämäläinen et al. (2021) assessed both textual and multi-modal classifiers for Finnish dialects, highlighting the importance of leveraging the audio modality to discriminate nuanced dialectal differences. Additionally, Kakouros and Hiovain-Asikainen (2023) showcased self-supervised speech models' ability to distinguish between four variants of North Sámi. Together, these findings indicate the potential of pre-trained speech models for fine-grained language and variety discrimination tasks.

3. Methodology

To address the proposed task, we leverage multilingual pre-trained models trained on large datasets to learn general representations. We also investigate the use of contrastive learning objectives to enhance the fine-tuning process and better separate the embeddings of different regions. This section provides an in-depth examination of these two key aspects: Section 3.1 describes the contrastive learning objectives explored, while Section 3.2 explains how fine-tuning is performed to adapt the pre-trained models to the region identification task.

3.1. Contrastive Learning Objectives

Contrastive learning is a representation learning method that focuses on acquiring knowledge by comparing and contrasting positive and negative examples. In a nutshell, the model learns class-discriminative representations by maximizing the similarity between representations of positive examples and minimizing similarity for negatives. We evaluate several supervised contrastive loss functions to improve the model's ability to learn discriminative representations. Specifically, we examine supervised contrastive loss (SC), triplet margin loss (TM), and multi-similarity loss (MS) as additional training objectives.

Supervised contrastive loss (SC). Defined as a direct optimization of embedding similarities between positive and negative pairs, the supervised contrastive loss (Khosla et al., 2020a) aims to maximize agreement between samples from the same region while minimizing agreement between samples from different regions. Let $f(\mathbf{x}_i)$ be an encoder network that generates an embedding $\mathbf{z}_i = f(\mathbf{x}_i)$ for each audio sample $\mathbf{x}_i \in \mathcal{X}$. The SC loss is then defined as:

$$\mathcal{L}_{SC} = -\sum_{i \in \mathcal{I}} \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{n \in \mathcal{N}_i} \exp(\mathbf{z}_i \cdot \mathbf{z}_n / \tau)} \tag{1}$$

Where $\mathcal{I}$ is the set of all samples in a given batch, $\mathcal{P}_i$ is the set of positive samples for sample i, $\mathcal{N}_i$ is the set of all negative samples for sample i, and τ is a tunable parameter. The positive samples are defined as all samples in the same training batch having the same label (i.e., the same region of origin in our case), while the negative samples are all samples in the same batch having a different label. The temperature τ is a scaling fudge factor we set to 0.1, following the default value used in the original paper (Khosla et al., 2020a).

Triplet margin loss (TM). Similar to supervised contrastive loss, the triplet margin loss (Balntas et al., 2016) aims to maximize similarities between embeddings of positive samples and minimize similarities between negatives. However, it operates using triplets of samples during training rather than pairs. Given a triplet consisting of an anchor sample $\mathbf{x}_a$, a positive sample $\mathbf{x}_p$ and a negative sample $\mathbf{x}_n$, the TM loss is defined as:

$$\mathcal{L}_{TM} = \max(0, d(\mathbf{z}_a, \mathbf{z}_p) - d(\mathbf{z}_a, \mathbf{z}_n) + \mu) \quad (2)$$

Where $d(\cdot, \cdot)$ is a distance function measuring the distance between embeddings; we use the L2 distance here. The margin parameter μ regulates the desired separation between positive and negative samples, with a higher value resulting in greater separation. In our experiments, we set $\mu = 0.05$. Triplets are generated using all combinations of samples within each batch, where the anchor and its positives belong to the same class, and the negatives belong to a different class. Intuitively, this loss function aims to minimize distances between positive embeddings while simultaneously enforcing a minimum margin between negative embeddings relative to the anchor.

Multi-similarity loss (MS). The multi-similarity loss (Wang et al., 2019) is a pair-based contrastive loss that specifically models pair selection and weighting to pick the most informative pairs for training. It optimizes pair selection and weighting by considering the similarity of a given sample to: (i) itself, (ii) other samples from the same class, and (iii) samples from different classes. Once pairs are selected, the loss is computed as the sum of positive and negative terms:

$$\begin{aligned}\mathcal{L}_{MS} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\alpha}\log[1 + \sum_{p\in\mathcal{P}_i} e^{-\alpha(S_{ip}-\lambda)}] \\ + \frac{1}{\beta}\log[1 + \sum_{n\in\mathcal{N}_i} e^{\beta(S_{in}-\lambda)}]\end{aligned} \quad (3)$$

Where m is the batch size, $\mathcal{P}_i$ and $\mathcal{N}_i$ denote the sets of positive and negative samples for the anchor i, S_{ip} and S_{in} are the similarities between i and its positive/negative pairs, and α, β, λ control pair weighting. The multi-similarity loss aims to select the most informative sample pairs during training by adaptively weighting pairs based on their similarities, prioritizing the most discriminative contrasts.

We investigate the use of each contrastive objective in three different settings: (i) the contrastive loss is used as the sole objective function to optimize the model parameters during an additional pre-training phase, (ii) the contrastive loss is added as an extra term to the overall loss function during the fine-tuning stage, and (iii) a combination where the model first undergoes pre-training where contrastive loss is minimized, then during the fine-tuning stage both contrastive and task-specific classification loss terms are jointly optimized.

Split	# Samples	# Minutes	# Cities	Avg # per city
Train	81279	2253.88	237	344.40
Val	8242	228.98	24	343.42
Test	8241	227.52	24	343.38

Table 1: VIVALDI dataset statistics.

3.2. Model fine-tuning

To adapt the pre-trained models for the region identification task, we fine-tune them on VIVALDI dataset, which is better described in Section 4.1. Most of the models tested in this analysis generate a high-level vector representation for each audio frame (i.e., typically 20 ms) after applying a contextualized encoder to raw audio. We perform average pooling over all frame embeddings to obtain a single embedding representing the full audio recording. Specifically, for a given recording composed of T audio frames with corresponding embeddings $e_1, e_2, ..., e_T$, we obtain: $e = \frac{1}{T}\sum_t e_t$, where e is the overall recording embedding. Taking the average in this manner aggregates the linguistic information captured across all frames into a single fixed-dimensional vector for the recording. Other methods of obtaining a recording-level representation from frame embeddings (e.g., attention pooling) could also be explored. Investigating alternative pooling strategies is outside the scope of the current work but may be considered in future explorations seeking improved performance.

Each model is trained end-to-end on the labeled audio collection by minimizing the cross-entropy loss between the model's region predictions and reference labels, directly optimizing its weights for the region identification task. Optionally, a contrastive learning objective (see Section 3.1) may also be included to simultaneously address the classification task and enrich the quality of data representation.

4. Experiments

To evaluate the proposed approach, we conduct a series of experiments to separately analyze the performance of different pre-trained models and the impact of the contrastive learning objectives.

4.1. VIVALDI data collection

The VIVALDI dataset (Tosques and Castellarin, 2013), short for "Vivaio Acustico delle Lingue e dei Dialetti d'Italia", is a comprehensive collection of spoken utterances from nearly all regions of Italy (Batter, 1995; Kattenbusch and Köhler, 2004; Kattenbusch et al., 2011). Italy comprises 20 administrative regions, and VIVALDI covers recordings from 19 of them, including cities from all regions except *Marche*. While most regions contain samples from 3 or more cities, two regions, Lazio and Campania, are each represented by data from a single city only. Despite variations in city-level coverage between regions, to the best of our knowledge, this is the only data collection including extensive speech recordings in local languages from a wide range of locations across the country, allowing investigation at an unprecedented scale.

The VIVALDI collection comprises recordings of a specific set of sentences, averaging approximately 343 repetitions per city. Each sample included in the collection is geolocated, providing information about the specific city where it was recorded. In most cases, each city is represented by recordings from a single speaker, with the sole exception being *Aidone* in *Sicily* region where two different speakers were sampled. Unfortunately, due to the unavailability of speaker-related information, conducting specific analyses or differentiating results based on demographic factors is currently unfeasible. Nonetheless, the dataset's geographical context remains pivotal for our analysis of regional linguistic variations.

Dataset statistics. Table 1 summarizes the key dataset statistics, including the number of samples, total minutes of audio, cities covered, and the average number of samples per city for the three splits: training, validation, and test. The dataset isn't provided as a single unified collection on the official website, and to the best of our knowledge, no predefined splits are available. Consequently, we design these splits so that each city, and each speaker accordingly, is allocated to only one of the splits (train, validation, or test), to ensure that models can not rely on identifying individuals to determine a sample's region. Instead, the task is centered on classifying each recording based on its linguistic features. While an 80/10/10 train/validation/test split was initially intended, we applied some constraints to enforce that each split contains samples from at least one city of each region. Applying this constraint, regions with data from fewer than three cities were excluded from the dataset, resulting in a final dataset representing 17 out of the 20 Italian regions. This split makes the classification task very challenging as models must generalize to unseen cities within a region that may also exhibit local linguistic variations.

4.2. Experimental settings

As evaluation metrics, we consider accuracy and macro F1 score. Accuracy is defined as the percentage of samples correctly classified, while macro F1 score represents the unweighted average of the F1 scores for each class. Given the imbalanced nature of the dataset, the macro F1 score provides a more reliable metric for evaluating the model's performance. We report the mean and standard deviation of results across three independent runs for each experiment.

4.3. Model selection

We explore the following speech models to assess their performance on the VIVALDI dataset.

`WavLM` (Sanyuan and et al., 2022): a model trained on 84,000 hours of audio from Libri-Light (Kahn et al., 2020), GigaSpeech (Chen et al., 2021), and VoxPopuli (Wang et al., 2021) corpora. As the model was trained primarily on English data, it includes only limited multilingual capabilities.

`XLSR-53` (Conneau et al., 2021): a multilingual model trained on 53 languages from CommonVoice (Ardila et al., 2020), Multilingual LibriSpeech (Pratap et al., 2020), and BABEL (Gales et al., 2014) datasets.

`XLSR-128` (Babu and et al., 2022): a multilingual model trained on 128 languages including the same datasets as `XLSR-53` in addition to VoxPopuli (Wang et al., 2021) and VoxLingua107 (Valk and Alumäe, 2021).

`ECAPA` (Desplanques et al., 2020): a CNN-based spoken language identification model pretrained on 107 languages from VoxLingua107 (Valk and Alumäe, 2021) dataset using SpeechBrain toolkit (Ravanelli et al., 2021).

For the `XLSR-53` and `XLSR-128` models, we also evaluate versions that have been additionally fine-tuned for Italian ASR tasks on the CommonVoice (Ardila et al., 2020) dataset (i.e., `XLSR-53-ITA` and `XLSR-128-ITA`, respectively). Those models may better capture the linguistic variation across Italy's regional forms.

All models are fine-tuned using the same training and validation splits described in Section 4.1 and following the same training procedure. To this end, we use the AdamW optimizer with 10% of warm-up steps and a linear learning rate decay schedule. The maximum learning rate is set to 10^{-4}, and the weight decay is set to 10^{-2}. We train each model with a batch size of 32 for a maximum of 10 epochs, selecting the best model based on the validation loss. Specific details about the models used in this analysis, together with the corresponding hyperparameters and the code to reproduce the experiments, are available on the project's repository[1].

[1] https://github.com/MorenoLaQuatra/SALVI

Model	ITA-FT	Accuracy	F1 Macro
`ECAPA`	✗	23.53±0.39	18.42±0.18
`WavLM-L`	✗	53.35±1.62	43.76±1.14
`XLSR-53`	✗	56.99±0.61	48.02±1.13
`XLSR-128`	✗	52.85±2.03	44.95±2.28
`XLSR-53-ITA`	✓	**60.18±0.55**	**49.84±0.57**
`XLSR-128-ITA`	✓	55.62±2.24	47.83±2.33

Table 2: Model selection. Mean and standard deviation results across three independent runs considering different models with a classic fine-tuning approach. The ITA-FT column indicates whether models were fine-tuned for Italian ASR (✓) or not (✗). Best results are highlighted in bold.

Ctr-PT	Ctr-FT	Clf-FT	Accuracy	F1 Macro
✗	✗	✓	60.18±0.55	49.84±0.57
Supervised Contrastive Loss				
✗	✓	✓	59.02±1.26	49.31±1.33
✓	✗	✓	58.98±0.67	48.82±0.44
✓	✓	✓	57.46±2.10	47.53±1.44
Triplet Margin Loss				
✗	✓	✓	60.23±1.53	50.68±1.71
✓	✗	✓	59.87±0.58	50.56±0.50
✓	✓	✓	58.19±0.64	49.92±1.14
Multi-Similarity Loss				
✗	✓	✓	**60.49±0.88**	**51.29±1.36**
✓	✗	✓	58.92±1.35	51.07±0.61
✓	✓	✓	59.86±0.83	50.98±0.35

Table 3: Mean and standard deviation results across three runs considering pre-training (**Ctr-PT**) and fine-tuning approaches (**Ctr-FT** and **Clf-FT**) with different contrastive losses over the best-performing model. The best results are in bold.

4.4. Quantitative Results

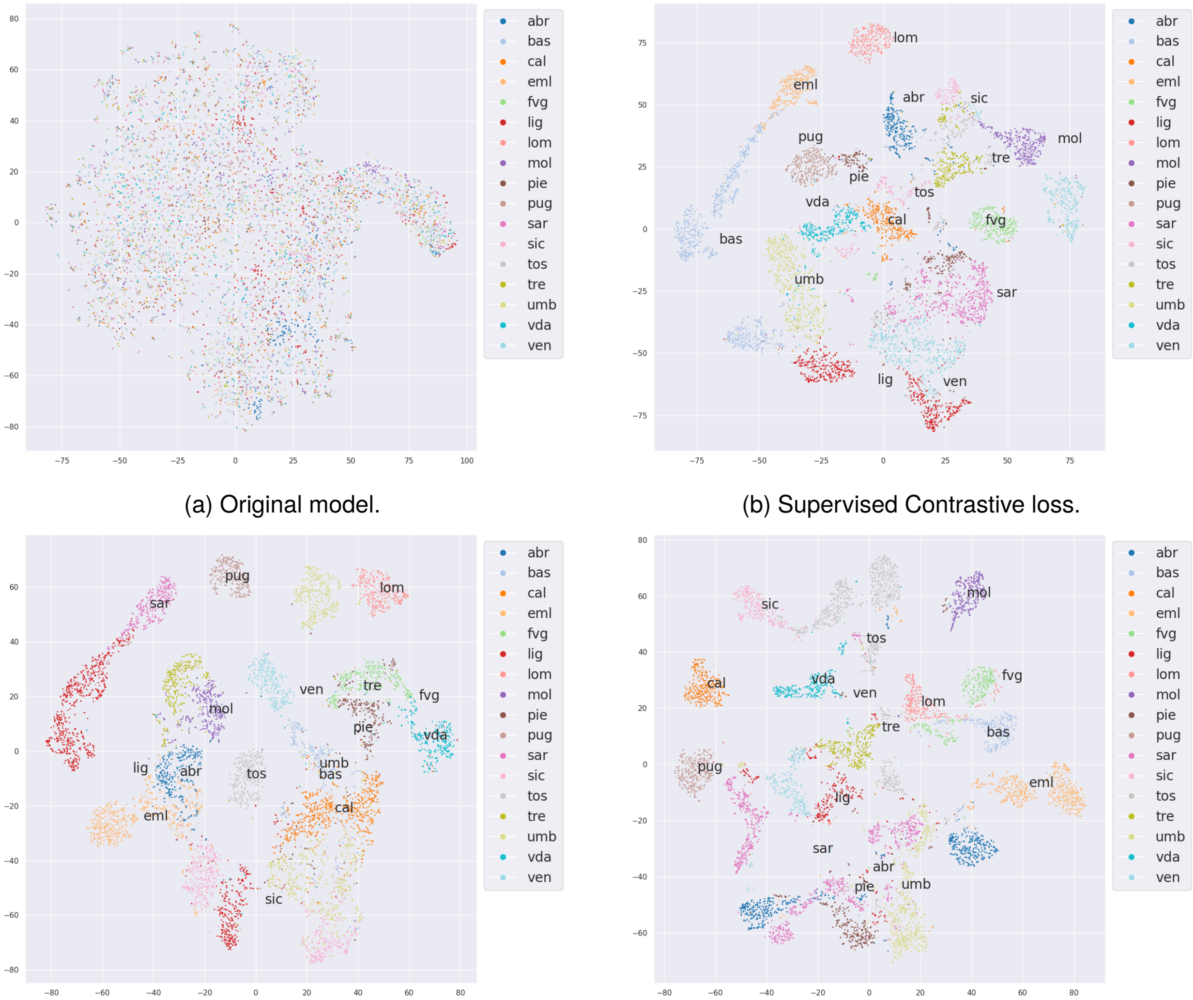
Standard fine-tuning. Table 2 reports the mean and standard deviation of accuracy and macro F1 scores for models trained using standard fine-tuning for the downstream task of region classification. Overall, the best-performing model is `XLSR-53-ITA`, achieving a macro F1 score close to 50%. Interestingly, the raw version of `XLSR-53` without Italian fine-tuning outperforms even the `XLSR-128-ITA` model and its pre-trained version without additional fine-tuning. This is likely because while the model architectures are identical, `XLSR-128` covers a much higher number of languages in its pre-training and thus may lack the capacity to capture fine-grained linguistic variations. `WavLM` does not achieve good results, because it was only partially pre-trained on multilingual data, which is insufficient for capturing the specific nuances information needed to solve the non-trivial region classification problem on the VIVALDI data collection. `ECAPA` is the only model leveraging a CNN-based architecture and is designed for spoken language identification and trained on 107 languages (including Italian). It performs worst among the tested models with a macro F1 score below 20%.

Contrastive learning. We experiment with three different supervised contrastive losses used in three different settings as explained in Section 3. Table 3 shows the results of our approach using the previously best-performing model, `XLSR-53-ITA`. We separately indicate the use of standard classification objective during fine-tuning (**Clf-FT**), contrastive objective as additional fine-tuning task (**Ctr-FT**), and contrastive objective as additional pre-training step (**Ctr-PT**). Applying multi-similarity and triplet margin losses as additional contrastive objectives helps improve performance on the downstream task in all tested configurations. Models trained using multi-similarity objective achieve the best overall results in all three configurations, with the highest macro F1 score of 51.29% in the multi-task fine-tuning scenario. Conversely, applying the supervised contrastive loss consistently results in decreased performance, especially as a pre-training step. The multi-similarity loss prioritizes the most informative contrasts during training by adaptively weighting sample pairs based on their similarities. This fine-grained approach allows the model to better learn representations that can capture subtle linguistic variations between regions, compared to the standard supervised contrastive loss, which treats all negative pairs equally.

Regardless of the specific objective, the best results are achieved using the additional contrastive loss for multi-task fine-tuning only. Combining objectives uniquely during fine-tuning leverages their complementary strengths by directly shaping the representations for the downstream task. Earlier-stage usage lacks this targeted optimization and results to slightly worsened final performance on the identification task.

4.5. Region Correlation Analysis

Pre-training objectives. We analyze the structure of models' high-dimensional feature spaces to gain additional insights into how effectively the pre-training objectives capture relationships between data points. Specifically, we apply t-Distributed Stochastic Neighbor Embedding (t-SNE) (Van der Maaten and Hinton, 2008), a widely used nonlinear dimensionality reduction technique, to project the embeddings into a two-dimensional space for visualization while attempting to preserve the local geometric structure of the original data. The resulting projection examines how each model distributes data points relative to their ground-truth



(a) Original model.

(b) Supervised Contrastive loss.

(c) Triplet loss.

(d) Multi-similarity loss.

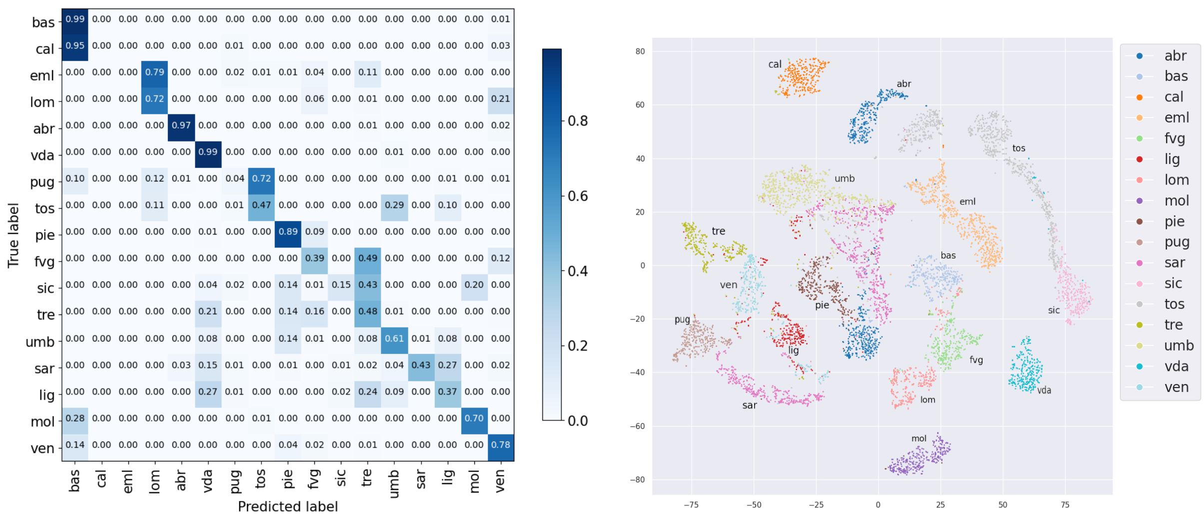
Figure 1: t-SNE visualization of the original `XLSR-53-ITA` model (a) and the corresponding pre-trained versions with the three different contrastive learning objectives: supervised contrastive loss (b), triplet-margin loss (c), and multi-similarity loss (d).

region labels in the lower-dimensional space. An ideal clustering would show clear separations between different regions, with points from the same region closely located together. This analysis is performed using only test set data and aims to characterize each method's capacity to learn discriminative representations that properly distinguish between regions.

Figure 1 shows t-SNE projections of the embeddings from the original `XLSR-53-ITA` model (Fig. 1a) and the same model pre-trained with different contrastive objectives: supervised contrastive loss (Fig. 1b), triplet margin loss (Fig. 1c), and multi-similarity loss (Fig. 1d). The data points are color-coded by their ground-truth region labels.

In Figure 1a, the representations show no clear clustering, as expected, since the model was not trained for this specific task and does not encode *a priori* the differences between language varieties. On the other hand, contrastive pre-training facilitates the formation of more distinct clusters across regions. Figure 1b with SC loss exhibits the most overlapping clusters. Figures 1c and 1d with TM and MS losses exhibit more precise separation of clusters corresponding to different regions, with multi-similarity (Fig. 1d) displaying slightly more well-defined clusters. This corroborates the quantitative results shown in Table 3, where multi-similarity training led to the best performance gains. The projections indicate that contrastive pre-training helps to learn embeddings that more distinctly encode relationships between data points, especially with margin-based losses.

Classification task. Figure 2 finally displays the confusion matrix (Fig. 2a) and t-SNE projection (Fig. 2b) of the best performing model, `XLSR-53-ITA` that underwent multi-task fine-tuning using the multi-similarity contrastive objective (i.e., second-



(a) Confusion matrix.

(b) t-SNE visualization.

Figure 2: Confusion Matrix (a) and t-SNE (b) of the `XLSR-53-ITA` model w/ multi-task fine-tuning using the multi-similarity contrastive objective.

last row of Table 3).

The confusion matrix (Figure 2a) indicates that the model achieves a relatively good correlation, on average, between the true and predicted region labels. However, there is significant confusion between specific, geographically close, region pairs, such as Basilicata (`bas`) and Calabria (`cal`), and Emilia-Romagna (`eml`) and Lombardia (`lom`)[2]. This suggests proximity likely leads to more linguistic similarity that the model struggles to distinguish, especially considering there may also be linguistic variations within regions. There is also some more unexpected confusion between regions without clear geographic explanation, like Sicilia (`sic`) with Trentino-Alto Adige (`tre`), and Liguria (`lig`) with both Trentino-Alto Adige (`tre`) and Valle d'Aosta (`vda`). An initial qualitative analysis of select recordings revealed that some contained a stronger presence of Standard Italian with only slight cues of the local variety. This suggested that some samples showed more language mixing than expected, which could make it harder to clearly distinguish between varieties.

Overall, while the model achieves good average performance, the confusion matrix highlights where it still struggles to reliably differentiate nuances between certain challenging regions' pairs.

The t-SNE visualization (Fig. 2b) further demonstrates the model's ability to discriminate between samples from different Italian regions. Distinct clusters are clearly formed corresponding to each region's data points. In contrast to the projections in Figures 1a (the *raw* `XLSR-53-ITA` model) and 1d (the version pre-trained with the MS loss), this fine-tuned model is highly effective at grouping points from the same region while separating those from other regions. Incorporating multi-task fine-tuning with a multi-similarity contrastive objective enables this model to far surpass the original `XLSR-53-ITA` version in learning representations that properly encode relationships between samples according to their Italian regional provenance. This explains its superior quantitative performance on the region classification task.

Interestingly, while most clusters are tightly formed, a few regions display some overlap or dispersion of points. This suggests the model has more difficulty fully disentangling subtler differences in certain language varieties but still captures significant inter-regional variances overall. This aligns with previous findings in the NLP domain (Ramponi and Casula, 2023a), which also observed similar phenomena, with specific regional varieties exhibiting more confounded representations than others, when analyzing textual region identification tasks.

5. Conclusions and Future Directions

This work presented an analysis of the linguistic variation across Italy's many regional language varieties directly from speech data. By leveraging the VIVALDI dataset, we assessed the performance of pre-trained speech models on automatic region identification. Specifically, we examined how effectively different models could distinguish the regional origins of speech samples based solely on their intrinsic features. The experimental results demonstrate that modern pre-trained models, particularly those fine-tuned on Italian ASR tasks, can capture meaningful differences between Italy's diverse yet closely related regional languages. Addition-

[2]Region acronyms are provided in our repository[1].

ally, we showed that contrastive learning objectives can enhance the discriminative ability of learned representations when applied as auxiliary training criteria during fine-tuning. However, even the top-performing model were confused when dealing with specific challenging region pairs. This indicates the intrinsic difficulty of the task. Further modeling improvements are still needed to fully disentangle the subtle linguistic variations that differentiate regional varieties. The analysis of the confusion matrix and t-SNE projections revealed that, while many regions were distinctly clustered, some overlap remained.

Future work will address two primary directions. First, additional data collection efforts are needed to better cover underrepresented regions and capture intra-regional diversity through multiple speakers per location. Second, methodological advances like tailored contrastive objectives and new pooling strategies may help extract maximally informative representations from pre-trained speech models. Addressing dataset imbalances and pushing modeling capabilities could advance the state-of-the-art on this linguistically fascinating and technologically important task.

6. Ethical statement

This research aims to advance the understanding of linguistic diversity and promote the preservation of understudied language varieties through technology. However, developing speech-based models also raises ethical concerns that may deserve careful consideration.

Language data reflects social and cultural norms. Automatic classification models could unintentionally encode harmful stereotypes or biases if data is imbalanced or fails to capture intra-group diversity, for example by not adequately representing all cities and local subgroups (Koudounas et al., 2024). Expanding data collection through local collaboration helps provide more representation of regional communities. Also, automated speech analysis is far from perfect and may pose risks of misclassification that could impact individuals if not interpreted carefully. Human oversight is crucial for any research involving human data or interactions. We intend neither to offend any subgroups nor make claims pertaining to cultural or personal identities with this technical work.

If conducted carefully and with input from language communities, this research has the potential to aid in documenting regional varieties and raising awareness of the richness of Italy's diverse linguistic landscape. The well-being of language communities should always be the top priority in efforts focused on language preservation.

7. Limitations

A key limitation of this work stems from variations in coverage across regions within the VIVALDI dataset. Some regions have relatively few or even no speech recordings included. As a result, not all areas of Italy have equal representation in the models' training and evaluation. Models may struggle more with fine-grained distinctions in undersampled regions compared to others. While the dataset remains extremely valuable for its scope, efforts to expand data collection from underrepresented regions or social groups would help address these limitations.
Additionally, similarly to prior NLP research (Ramponi and Casula, 2023b), this work treats regional languages as unified varieties without modeling the fine-grained linguistic variation that exists between local forms even within the same administrative region (Pellegrini, 1977; Andreose and Renzi, 2013). In some cases, linguistic variation occurs at a hyper-local level, which is not captured by coarse regional classifications. While useful as a first approximation, modeling linguistic variation at a finer granularity could provide greater insight by linking each city to its administrative region and geographical areas of shared linguistic roots.

8. Bibliographical References

Badr M. Abdullah, Mohammed Maqsood Shaik, and Dietrich Klakow. 2023. On the nature of discrete speech representations in multilingual self-supervised models. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 159–161, Dubrovnik, Croatia. Association for Computational Linguistics.

Benjamin Adams and Grant McKenzie. 2018. Crowdsourcing the character of a place: Character-level convolutional networks for multilingual geographic text classification. *Transactions in GIS*, 22(2):394–408.

Haider Al-Tahan and Yalda Mohsenzadeh. 2021. Clar: Contrastive learning of auditory representations. In *International Conference on Artificial Intelligence and Statistics*, pages 2530–2538. PMLR.

Tanel Alumäe, Kunnar Kukk, Viet-Bac Le, Claude Barras, Abdel Messaoudi, and Waad Ben Kheder. 2023. Exploring the Impact of Pretrained Models and Web-Scraped Data for the 2022 NIST Language Recognition Evaluation. In *Proc. INTERSPEECH 2023*, pages 516–520.

Alvise Andreose and Lorenzo Renzi. 2013. *Geography and distribution of the Romance languages in Europe*, volume 2, page 283–334. Cambridge University Press.

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.

Francesco Avolio. 2009. Lingue e dialetti d'italia.

Arun Babu and et al. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*.

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.

Vassileios Balntas, Edgar Riba, Daniel Ponsa, and Krystian Mikolajczyk. 2016. Learning local feature descriptors with triplets and shallow convolutional neural networks. In *Bmvc*, volume 1, page 3.

Travis M. Bartley, Fei Jia, Krishna C. Puvvada, Samuel Kriman, and Boris Ginsburg. 2023. Accidental learners: Spoken language identification in multilingual self-supervised models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Matteo Bartoli, Ugo Pellis, and Lorenzo Massobrio. 1995. *Atlante Linguistico Italiano*. Istituto Poligrafico e Zecca dello Stato.

Pierpaolo Basile, Elio Musacchio, Marco Polignano, Lucia Siciliani, Giuseppe Fiameni, and Giovanni Semeraro. 2023. Llamantino: Llama 2 models for effective text generation in italian language. *arXiv preprint arXiv:2312.09993*.

Roland Batter. 1995. Vivaldi-sicilia. documentazione sonora dei dialetti siciliani.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-slu: A new dataset for SLU in italian. *CoRR*, abs/1907.07526.

Gaetano Berruto et al. 2005. Dialect/standard convergence, mixing, and models of language contact: the case of italy. *Dialect change. Convergence and divergence in European languages*, pages 81–97.

Massimo Cerruti, Riccardo Regis, et al. 2005. Code switching'e teoria linguistica: la situazione italoromanza. *Italian Journal of Linguistics*, 17(1):179.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Giovanni Costantini, Iacopo Iaderola, Andrea Paoloni, Massimiliano Todisco, et al. 2014. Emovo corpus: an italian emotional speech database. In *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, pages 3501–3504. European Language Resources Association (ELRA).

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834.

Johanna Dobbriner and Oliver Jokisch. 2019. Towards a dialect classification in german speech samples. In *Speech and Computer: 21st International Conference, SPECOM 2019, Istanbul, Turkey, August 20–25, 2019, Proceedings 21*, pages 64–74. Springer.

Jonathan Dunn and Sidney Wong. 2022. Stability of syntactic dialect classification over space and time. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 26–36, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Li Fu, Xiaoxiao Li, Runyu Wang, Lu Fan, Zhengchen Zhang, Meng Chen, Youzheng Wu, and Xiaodong He. 2022. SCaLa: Supervised Contrastive Learning for End-to-End Speech Recognition. In *Proc. Interspeech 2022*, pages 1006–1010.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Giuseppe Gallipoli, Moreno La Quatra, Daniele Rege Cambrin, Salvatore Greco, and Luca Cagliero. 2023a. Dante at geolingit: Dialect-aware multi-granularity pre-training for locating tweets within italy. In *Proceedings of the Eighth Evaluation Campaign of Natural LanguageProcessing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473. CEUR-WS. org.

Giuseppe Gallipoli, Moreno La Quatra, Daniele Rege Cambrin, Salvatore Greco, and Luca Cagliero. 2023b. DANTE at geolingit: Dialect-aware multi-granularity pre-training for locating tweets within italy. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473 of *CEUR Workshop Proceedings*.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895.

Hans Goebl. 1989. *Problèmes et méthodes de la dialectométrie*, pages 165–184. De Gruyter Mouton, Berlin, Boston.

Hans Goebl. 2006. Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing*, 21(4):411–435.

Mika Hämäläinen, Khalid Alnajjar, Niko Partanen, and Jack Rueter. 2021. Finnish dialect identification: The effect of audio and text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8777–8783, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Sofoklis Kakouros and Katri Hiovain-Asikainen. 2023. North Sámi Dialect Identification with Self-supervised Speech Models. In *Proc. INTERSPEECH 2023*, pages 5306–5310.

Dieter Kattenbusch and Carola Köhler. 2004. La sardegna nel progetto vivaldi. *GRIMALDI, LUCIA & MENSCHING, GUIDO (a cura di), Su sardu. Limba de Sardigna e limba de Europa. Cagliari: Cuec*, pages 193–203.

Dieter Kattenbusch, Fabio Tosques, and Andreas Rauher. 2011. „umbria dialettale “. *Claudia Schlaak, Lena Busse (Hg.), Sprachkontakte, Sprachvariation und Sprachwandel, Tübingen*, pages 443–460.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020a. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020b. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673.

Alkis Koudounas, Flavio Giobergia, Irene Benedetto, Simone Monaco, Luca Cagliero, Daniele Apiletti, and Elena Baralis. 2023a. baρtti at geolingit: Beyond boundaries, enhancing geolocation prediction and dialect classification on social media in italy. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, September 7th-8th, 2023*, volume 3473 of *CEUR Workshop Proceedings*.

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana

Pastor, Luca Cagliero, and Elena Baralis. 2023b. ITALIC: An Italian Intent Classification Dataset. In *Proc. INTERSPEECH 2023*, pages 2153–2157.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2024. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:1468–1480.

Moreno La Quatra and Luca Cagliero. 2023. Bart-it: An efficient sequence-to-sequence model for italian text summarization. *Future Internet*, 15(1).

Yooyoung Lee, Craig Greenberg, Eliot Godard, Asad A Butt, Elliot Singer, Trang Nguyen, Lisa Mason, and Douglas Reynolds. 2023. The 2022 nist language recognition evaluation. *arXiv preprint arXiv:2302.14624*.

Hexin Liu, Leibny Paola Garcia Perera, Andy W. H. Khong, Eng Siong Chng, Suzy J. Styles, and Sanjeev Khudanpur. 2022. Efficient self-supervised learning representations for spoken language identification. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1296–1307.

Martin Maiden and Mair Parry. 2006. *The dialects of Italy*. Routledge.

Marco Marini, Mauro Viganò, Massimo Corbo, Marina Zettin, Gloria Simoncini, Bruno Fattori, Clelia D'Anna, Massimiliano Donati, and Luca Fanucci. 2021. Idea: An italian dysarthric speech database. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 1086–1093.

Anssi Moisio, Dejan Porjazovski, Aku Rouhe, Yaroslav Getman, Anja Virkkunen, Ragheb Al-Ghezi, Mietta Lennes, Tamás Grósz, Krister Lindén, and Mikko Kurimo. 2023. Lahjoita puhetta: a large-scale corpus of spoken finnish with some benchmarks. *Language Resources and Evaluation*, 57(3):1295–1327.

Giovan Battista Pellegrini. 1977. *Carta dei dialetti d'Italia*. Pacini, Pisa. Illustrated, includes 1 folded geographic map (125x99 cm.).

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Afshin Rahimi, Timothy Baldwin, and Trevor Cohn. 2017. Continuous representation of location for geolocation and lexical dialectology using mixture density networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 167–176, Copenhagen, Denmark. Association for Computational Linguistics.

Alan Ramponi. 2022. Nlp for language varieties of italy: Challenges and the path forward. *arXiv preprint arXiv:2209.09757*.

Alan Ramponi and Camilla Casula. 2023a. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Alan Ramponi and Camilla Casula. 2023b. Geolingit at evalita 2023: Overview of the geolocation of linguistic variation in italy task. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR. org, Parma, Italy*.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Codrut Rotaru, Nicolae-Catalin Ristea, and Radu Tudor Ionescu. 2023. Rodia: A new dataset for romanian dialect identification from speech. *arXiv preprint arXiv:2309.03378*.

Seyed Omid Sadjadi, Timothee Kheyrkhah, Craig Greenberg, Elliot Singer, Douglas Reynolds, Lisa Mason, and Jaime Hernandez-Cordero. 2018. Performance Analysis of the 2017 NIST Language Recognition Evaluation. In *Proc. Interspeech 2018*, pages 1798–1802.

Andrea Santilli and Emanuele Rodolà. 2023. Camoscio: An italian instruction-tuned llama. *arXiv preprint arXiv:2307.16456*.

Chen Sanyuan and et al. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.*

Salvatore Sarni, Sandro Cumani, Sabato Marco Siniscalchi, and Andrea Bottino. 2023. Description and analysis of the KPT system for NIST Language Recognition Evaluation 2022. In *Proc. INTERSPEECH 2023*, pages 1933–1937.

Gabriele Sarti and Malvina Nissim. 2022. It5: Large-scale text-to-text pretraining for italian language understanding and generation. *arXiv preprint arXiv:2203.03759*.

Fabio Tosques and Michele Castellarin. 2013. Das vivaio acustico delle lingue e dei dialetti d'italia (vivaldi).

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.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5022–5030.

Martijn Wieling, John Nerbonne, Simonetta Montemagni, and R. Harald Baayen. 2014. Lexical differences between tuscan dialects and standard italian: Accounting for geographic and sociodemographic variation using generalized additive mixed modeling. *Language*, 90(3):669–692.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.