Exploiting Foundation Models for Spoken Language Identification

Benedikt Augenstein[1], Darjan Salaj[2]

[1]Hochschule Aalen
[2]inovex GmbH

Abstract

Spoken Language Identification (SLID) is the task of identifying the language from speech audio recordings, which poses challenges due to the variability of speech recordings and the diverse properties of languages. Traditional SLID methods rely on labor-intensive feature engineering and classical machine learning algorithms. The emergence of deep learning has allowed for more efficient and automated SLID, but come at a much higher compute cost and data volume requirements. Despite the advancements, automated SLID remains limited in many applications, such as voice assistants, dictation software, and customer-facing services, especially for the underrepresented languages. To address the issue of improving SLID for the underrepresented languages with limited data availability we propose a fine-tuning ensemble approach that achieves higher SLID performance then the individually fine-tuned models. We further identify the core issue in training SLID models, and show through meta-analysis the critical flaw in evaluations and datasets of previous works. This insight suggests possible improvements to the quality and availability of datasets and benchmarks.

Keywords

Spoken Language Identification, Foundation Models, Whisper, Audio Analysis

1. Introduction

The task of identifying the language from speech audio recordings is known as Spoken Language Identification (SLID [12, 26], also abbreviated S-LID [2], LID [11, 27] or LI [1] which can be confused with the textual language identification in NLP). SLID has remained a challenge in speech signal processing research due to the difficulty of discerning relevant from irrelevant features in extremely varied speech audio recordings. The extreme variability of speech recordings is a consequence of variation in speakers' anatomy, age, sex, mood, and dialect, spoken contents and acoustic conditions [11, 12, 5]. In addition, the distinguishing properties of languages are highly varied. For example, some languages, despite using common phonetic sounds, can be classified as distinct due to a unique set of phonological rules. Earlier versions of SLID relied heavily on hand-crafted and domain knowledge informed feature extraction and selection. These include noise suppression, tuning and selection of acoustic, prosodic and phonotactic

Business Intelligence & Analytics (WSBIA 2023), October 09–11, 2022, Marburg, HE
benedikt.augenstein@ibm.com (B. Augenstein); darjan.salaj@inovex.de (D. Salaj)

features [2, 8]. The end classification is then realized with classical machine learning algorithms, such as Gaussian Mixed Model, Support Vector Machine, Hidden Markov Model, Vector Quantization, and Total Variation Warping [11]. To circumvent this labor-intensive feature engineering, novel SLID methods started to rely on deep learning, which subsumed all the feature engineering work [1, 3, 4, 6, 9, 12]. Recently, the advent of large foundation models has brought even more advantages and innovations to deep learning models in the language domain [10]. Such deep learning based solutions are part of the ever-growing number of use cases.

While the past decade has seen a growing number and growing quality of voice assistants and dictation software, the language setting has remained mostly manual. The simultaneous use of several languages is still only possible to a very limited extent. For example, a second language can be added manually to the Google Assistant[1], but the addition of more than three languages is not possible.[2] Amazon Alexa and Apple Siri can also process multiple languages to a very limited extent [4].[3] In the case of language assistants for private use, this automatic language identification might not be necessary, since the languages spoken usually do not change. However, for applications that are shared by many users (e.g. deployed in public spaces), reliable language identification would be essential. An example of this would be voice assistants at train stations or airports with the purpose of answering questions from a wide variety of people in different languages. SLID systems also save resources and improve quality in call centers when used to match and route international calls with the language compatible operators [7]. Extending this to emergency call systems, such as those in aviation, crisis management and law enforcement, could potentially have a direct impact on human lives [11].

Furthermore, a reliable implementation of an SLID system could be used for large video or streaming platforms, enabling the creation of automatically generated subtitles without a need for a prior determination of the spoken language by a user or support staff. Therefore, this software could help to decrease costs by automating this task as well as increase customer satisfaction by making subtitles possible, regardless of whether the content creator set the language for the video. It is thus evident that a solution for automatic language identification could be highly beneficial for improving user experience for many platforms or services that collect, store and process large amounts of unstructured, human-generated audio data.

Deep learning based SLID approaches, although offering better accuracy and lower engineering cost, need high volumes of data and associated compute cost for training [11]. More importantly, the data volumes required for high performance are often not available for languages spoken by smaller communities [32]. Also, researchers belonging to those less represented communities often lack the compute resources required to train large models, even if data would be available. We propose an ensemble approach that generalizes to previously unseen languages, circumvents the need for large data volumes

[1]Tested on Android version 13 Kernel version `5.10.149-android13-4-00002`

[2]see Google Assistant Help [35] and Google Nest Help [34]

[3]Amazon Alexa supports two languages [36], while Apple Siri supports only one language. Due to the extensive use of English words in Indian languages, the English words are also recognized when an Indian language is configured in Apple Siri [37].

of underrepresented languages, and is developed under constrained compute resources.

We hypothesize the core issue in training SLID models and perform a meta-analysis of previous SLID works. We show that many of the datasets suffer from a critical flaw making them uninformative about the generalization on the unseen data.

2. Related work

Traditional algorithms and feature engineering. Early SLID works were based on the handcrafted features combined with classical machine learning algorithms for classification. Examples of such methods are: detection of events of stability and rapid change in audio spectrum [16], Markov modelling over formant and prosodic features [17], polynomial classifier over Linear Predictive Coding (LPC) features [19], Vector Quantization over LPC features [21], hidden Markov models over Mel-scale cepstrum vectors [20]. In a more recent feature engineering attempt, authors developed the MFCC-2 [23] features to better support the multilingual speech processing. Another work in this category developed a novel feature selection method based on Harmony Search and Naked Mole-Rat algorithm [2].

Deep learning methods. Moving beyond the handcrafted feature engineering and classical machine learning algorithms, novel approaches make used of deep learning methods and general feature preprocessing. Convolutional Neural Networks (CNN) [3], Residual Networks (ResNet) [9] and Long Short-Term Memory (LSTM) networks [25] were used to classify languages from Mel spectrograms. The combination of previous methods dubbed Convolutional Recurrent Neural Network (CRNN) were also applied to SLID [4]. Conditional Generative Neural Networks (cGAN) were used to improve regularization and accuracy of jointly optimized classifiers [6]. Later the attention mechanism was combined with both CNNs and RNNs [1]. Capsule Networks (CapsNet) [27] were also applied to SLID and consistently outperformed other architectures with CNNs, RNNs, and attention. Under the low-resource setting, the 1D time-channel separable CNNs with Self-Attentive Pooling [33] achieved the state-of-the-art SLID results.

The original Time-Delay Neural Networks (TDNN) [30] approach was improved upon and applied to speaker recognition task [29]. X-vectors [28], a TDNN based approach to map variable-length audio to fixed-dimensional embeddings, made further advances in data augmentation. Authors of [26] further extend the TDNN approach to SLID task with unsupervised domain adaptation using optimal transport.

Ensemble methods. An ensemble method named FuzzyGCP [12] achieved a high SLID accuracy by combining Deep Dumb Multi Layer Perceptron (DDMLP), Deep Convolutional Neural Network (DCNN) and Semi-supervised Generative Adversarial Network (SSGAN).

Large foundation models. The latest trend of scaling up the models and training them on large datasets eventually arrived to the speech processing domain and resulted in

foundation models like Wav2Vec 2.0 [24], XLSR [13] and Whisper [14]. Wav2Vec 2.0 [24] introduced a novel way of applying self-supervised learning of representations to raw audio data. It also beat all previous methods in speech recognition accuracy without using a language models while being more data efficient. XLSR [13] is the extension of Wav2Vec 2.0 and it set the new state-of-the-art in speech recognition on 53 languages. Whisper [14] is a transformer based model with which the authors scaled up the weakly supervised-pretraining on multilingual multi-task data and set the new state-of-the-art in speech recognition.

3. Ensemble of foundation models for SLID

In this paper we develop a model that achieves a high accuracy across many languages, most importantly those that are underrepresented in datasets, while working under a minimal compute budget (a single GPU node). To avoid the need for high data volumes and compute costs associated with training the large deep learning models [11], we propose exploiting the foundation models already pretrained on available large datasets. This way we are able to make use of the high quality of multilingual embeddings provided by the foundation models. Further, we propose combining the embeddings (latent representation in penultimate layers) of multiple foundation models with the goal of having an even richer representation of the input from different perspectives. Finally, we propose training a readout model that classifies the combined embeddings to the target languages. Here we implicitly rely on multilingual foundation models to be able to give an informative embedding even when applied to previously unseen languages.

The general idea behind this approach is that the different, pre-trained models most likely extract the relevant information from the audio files in different ways as they use different architectures and have been trained on different datasets. Using this approach, the differing ways of extracting information from audio of the two models can be reused, resulting in mixed embeddings containing more relevant information for the final classification than the embeddings of the individual models. This increases the amount of information that can be used to improve classification performance.

In the experiments section, we combine the Whisper [14] and the TDNN [30] model by taking the output vectors of the penultimate layers of the models, concatenating them, and using these combined vectors from each audio file as input vectors for the new classifier model. The classifier model is then trained on a smaller dataset containing languages previously unseen by the Whisper and the TDNN models. Instead of the output layers of each original foundation model, the newly trained model acts as an output layer which can classify combined embeddings from both foundation models.

The model contains 102 output neurons, matching the number of languages in the target FLEURS dataset [15]. As the extraction of relevant information already takes place in the layers of the foundation models, we choose a simple architecture for the classifier model. The architecture consists of only two dense layers using a rectified linear unit activation function and two dropout layers (preceding the dense layers) with a dropout rate of 40%. The first dense layers has 1000 neurons, while the second output layer has

102 neurons matching the number of languages, totally $615k$ trainable parameters. The model is trained using the Adam optimizer on the categorical cross entropy loss function, a batch size of 4096, and early stopping with up to 50 epochs.

4. FLEURS Dataset

In this paper we use the FLEURS (Few-shot Learning Evaluation of Universal Representations of Speech) dataset. It is a benchmark designed to enable development and evalution of speech processing methods in low-resource settings. FLEURS is derived from the FLoRes dev and devtest sets, containing 2009 n-way parallel sentences across 102 languages. The speakers in the training and evaluation sets are different. The dataset is grouped into seven geographical areas, allowing for analysis and comparison of results. With approximately 12 hours of speech supervision per language, FLEURS can be used for various speech tasks such as Automatic Speech Recognition, SLID, Translation, and Retrieval. The 102 languages are grouped into seven geographical areas:

- Western Europe: Asturian, Bosnian, Catalan, Croatian, Danish, Dutch, English, Finnish, French, Galician, German, Greek, Hungarian, Icelandic, Irish, Italian, Kabuverdianu, Luxembourgish, Maltese, Norwegian, Occitan, Portuguese, Spanish, Swedish, Welsh
- Eastern Europe: Armenian, Belarusian, Bulgarian, Czech, Estonian, Georgian, Latvian, Lithuanian, Macedonian, Polish, Romanian, Russian, Serbian, Slovak, Slovenian, Ukrainian
- Central-Asia/Middle-East/North-Africa: Arabic, Azerbaijani, Hebrew, Kazakh, Kyrgyz, Mongolian, Pashto, Persian, Sorani-Kurdish, Tajik, Turkish, Uzbek
- Sub-Saharan Africa: Afrikaans, Amharic, Fula, Ganda, Hausa, Igbo, Kamba, Lingala, Luo, Northern-Sotho, Nyanja, Oromo, Shona, Somali, Swahili, Umbundu, Wolof, Xhosa, Yoruba, Zulu
- South-Asia: Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Nepali, Oriya, Punjabi, Sindhi, Tamil, Telugu, Urdu
- South-East Asia: Burmese, Cebuano, Filipino, Indonesian, Javanese, Khmer, Lao, Malay, Maori, Thai, Vietnamese
- CJK languages: Cantonese and Mandarin Chinese, Japanese, Korean

5. Results

Unsuitability of datasets for evaluating generalization. The greatest challenge in training new models for SLID is effectively preventing models from learning the wrong objectives. More precisely, the model tends to learn the voices or specific audio characteristics of the audio files rather than the difference between the languages themselves. The gravity of this problem increases with a smaller dataset as a model can easily learn to differentiate between a small number of voices or other audio features like the characteristics of the specific microphones used.

Table 1
Test accuracy of past approaches from 2020-2023, with last column indicating if the training and test sets had disjoint speakers. Details about the datasets used in each of the approaches is listed in table 4.

Model	Languages	Acc	Disjoint speaker set
FRESH + ANN [41]	6	99.94%	No
PLDA logistic regression [31]	3	96%	No
Self-attentive pooling decoder [33]	6	92.5%	No
CNN [39]	11	97.35%	No
CapsNet [27]	11	98.2%	No
CRNN ResNet50 DenseNet121 [38]	6	89%	No
CRNN ResNet50 DenseNet121 [38]	6	45%	Yes
CNN ResNet50 RNN [32]	16	53%	Yes
Whisper [14]	102 (99)	64.5%	Yes
mSLAM [42]	102	77.7%	Yes

This effect is evident when analyzing the performance of models in related works, shown in table 1. The table shows previous approaches to the SLID task, including the used model architecture, the number of the analyzed languages, and the accuracy on the test set. Most importantly, the table shows if there is a speaker overlap between the training and test sets of the given datasets, i.e. if the training and test speaker sets are disjoint. It can be seen that, whenever this has been the case, the evaluation accuracy has been significantly lower. Our hypothesis is that the models, in the cases where the training and test speaker sets are not disjoint, learn to recognize the voices and map them to the languages, instead of learning to recognize the languages themselves. This explains the misleading high test accuracy of such setups, in contrast to the approaches where they did have a disjoint speaker set between the training and test sets.

We have also observed this effect during training and testing of our models. When using a part of a training dataset as a validation dataset, the training and validation accuracy were constantly very high. However, when testing the model on a different dataset which did not include the same voices but the same languages, the accuracy regularly dropped to an accuracy only marginally higher than an accuracy which would be expected for a random classifier. This indicates that the models are overfitting on the training data and that the language-specific features have not or only to a very limited extent been learned.

To facilitate a truthful evaluation, we recommend using the FLEURS [15] or similar datasets which include strictly different speakers in training, development and test sets. Additionally, the FLEURS dataset consists of 102 languages across different language families, allowing for a more comprehensive evaluation and development of more general SLID models.[4]

Evaluation on FLEURS dataset. We trained the model described in section 3 on the FLEURS dataset and present the results in table 2 and 3. First we evaluated the two models individually and combined on six European (German, English, Spanish, French,

[4]Available at Hugging Face https://huggingface.co/datasets/google/fleurs

Russian and Italian) and six South Asian languages (Hindi, Urdu, Marathi, Malayalam, Bengali and Gujarati) to compare the performance within different language groups, and to test if the combination of the models is beneficial even on smaller and more constrained datasets, see table 2.

The results of this approach show that by combining two models and training a new output layer, a higher accuracy than the ones of the individual models can be achieved for certain language groups. This can, for example, be seen by looking at the results for the South Asian languages. The combination of the existing models with the newly trained readout significantly outperforms the classification accuracy of the models when they are being used individually for the task of language identification.

Furthermore, the results show that the accuracy does not decrease if the individual models already achieve a very high accuracy. For the European languages, the existing models achieved an accuracy of 100%. As the method of combining the existing models with a new model did not deteriorate those results but rather kept the accuracy at the same level, the method can be considered to be stable across a wide variety of languages. Therefore, we suggest that this method can be potentially generalized to the identification of all languages, as it improves the results regardless of the performance already achieved by existing models.

These results can be confirmed by the evaluation of this method on all of the 102 languages of the FLEURS dataset. Table 3 shows that our ensemble approach beats the reference models, including the two models serving as a basis for the ensemble, by a significant margin.

Another advantage of this approach is that, as the models have previously been trained on larger datasets, the final output layer can be trained on a relatively small amount of data and still achieve better results. This is due to the fact that the used foundation models have already achieved the ability to extract the relevant information from audio into useful embeddings for speech processing. This is especially useful in situations where available data or hardware resources are limited.

However, it must be noted that when using the approach in production, the time for each classification increases due to combined inference time of the foundation and classifier models. This is because for each input audio file, the forward propagation for each of the two models must be executed in order to retrieve both output vectors of the penultimate layers, which can then be used as an input for the new model.

Ablation study. To investigate if the accuracy gains are a result of the combination of the foundation models or if they solely stem from the training of the classifier model, we performed an ablation study. Instead of the combined foundation models, we trained the classifier model separately on top of each of the foundation models, see rows with "Retrained readout" in table 3. The results on the Whisper model indicate that the largest increase in the classification accuracy comes from the training of the readout model on the new languages, as is expected. Nonetheless, the factor of combining the embedding from multiple foundation models made another significant increase in the SLID accuracy.

Table 2
Comparison of the language identification accuracy of the different SLID methods for different language groups consisting of 6 languages. European languages subset consists of German, English, Spanish, French, Russian and Italian. South Asian languages subset consists of Hindi, Urdu, Marathi, Malayalam, Bengali and Gujarati.

Model	Retrained readout	European languages acc.	South Asian languages acc.
Whisper		100.00%	76.67%
TDNN		100.0%	85.33%
Whisper + TDNN	✓	100.0%	88.00%

Table 3
Comparison of the "Combination with training"-method with existing models, evaluated on the complete FLEURS-test-dataset, partially derived from A. Radford et al. (2022, p. 8).

Model	Retrained readout	Accuracy
w2v-bert-51		71.4%
mSLAM-CTC		77.7%
Zero-shot TDNN		77.62%
TDNN	✓	76.6%
Zero-shot Whisper		64.5%
Whisper	✓	87.53%
Whisper + TDNN	✓	90.9%

6. Discussion

In this paper, we tackled the issue of training SLID models on many languages under the constrained data and compute resources. This scenario is relevant for the underrepresented languages for which the large data volumes are not available and for the researchers belonging to less represented communities, which often lack the compute resources for training large models from scratch. With our experimental results, we made two main contributions.

First, in the analysis of previous works, we show that many of the datasets used in SLID suffer from a critical flaw of having the same speakers included in both the training and test sets. Table 1 illustrates this issue and shows a large discrepancy between approaches that are evaluated on those datasets which suffer, and those that do not. This insight is critical for designing new datasets and benchmarks for SLID, and suggests which of the datasets gives test accuracies that are meaningful and representative of the generalization on the unseen data.

Second, we show that a simple ensemble method of combining penultimate layers of large pre-trained models and training the readout can lead to better accuracy on new languages with minimal compute costs. This relies on two assumptions: the large pre-trained models are producing somewhat language-independent embeddings of input audio, different pre-trained models have a different and complementary perspective (embeddings) of the input audio signals.

We hypothesize that this ensemble method of combining penultimate layers can be

scaled up with more pre-trained models to achieve a higher accuracy. Also, we believe that it can be extended to other domains, such as vision or graphs, where large pre-trained embedding generators could be easily reused for more constrained tasks.

Limitations. Due to the reuse of the large pre-trained models, the ensemble method requires the execution of the forward pass of all pre-trained models for every audio input. This limits the deployment of the ensemble model to cloud platforms of relatively powerful single nodes. Ideally, a high-precision universal SLID model should be deployable to embedded Edge AI devices due to the privacy concerns. Unfortunately, this is still not possible because the number of supported languages in SLID model directly correlates with the model size and consequently the memory and compute requirements.

Is is worth mentioning that if data and compute resources are not a limiting factor, the best accuracy can be achieved by training a network from scratch. For example, a recently published model called Massively Multilingual Speech (MMS) by Meta can perform SLID on 4017 languages [40]. We recommend applying knowledge distillation or other model compression techniques on the MMS model, for fine tuning on the underrepresented languages, while reducing the model size and enabling deployment on the edge.

References

[1] S. Shukla, G. Mittal, Spoken language identification using convnets, in: Ambient Intelligence: 15th European Conference, AmI 2019, Rome, Italy, November 13–15, 2019, Proceedings 15, Springer, 2019, pp. 252–265.

[2] S. Guha, A. Das, P. K. Singh, A. Ahmadian, N. Senu, R. Sarkar, Hybrid feature selection method based on harmony search and naked mole-rat algorithms for spoken language identification from audio signals, IEEE Access 8 (2020) 182868–182887.

[3] G. Singh, S. Sharma, V. Kumar, M. Kaur, M. Baz, M. Masud, Spoken language identification using deep learning, Computational Intelligence and Neuroscience 2021 (2021).

[4] C. Bartz, T. Herold, H. Yang, C. Meinel, Language identification using deep convolutional recurrent neural networks, in: Neural Information Processing: 24th International Conference, ICONIP 2017, Guangzhou, China, November 14–18, 2017, Proceedings, Part VI 24, Springer, 2017, pp. 880–889.

[5] R. P. Hafen, M. J. Henry, Speech information retrieval: a review, Multimedia systems 18 (2012) 499–518.

[6] P. Shen, X. Lu, S. Li, H. Kawai, Conditional generative adversarial nets classifier for spoken language identification., in: Interspeech, 2017, pp. 2814–2818.

[7] P. Kumar, A. Biswas, A. N. Mishra, M. Chandra, Spoken language identification using hybrid feature extraction methods, arXiv preprint arXiv:1003.5623 (2010).

[8] Y. Obuchi, N. Sato, Language identification using phonetic and prosodic hmms with feature normalization, in: Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005., volume 1, IEEE, 2005, pp. I–569.

[9] S. Revay, M. Teschke, Multiclass language identification using deep learning on spectral images of audio signals, arXiv preprint arXiv:1905.04348 (2019).
[10] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models, arXiv preprint arXiv:2108.07258 (2021).
[11] I. A. Thukroo, R. Bashir, K. J. Giri, A review into deep learning techniques for spoken language identification, Multimedia Tools and Applications 81 (2022) 32593–32624.
[12] A. Garain, P. K. Singh, R. Sarkar, Fuzzygcp: A deep learning architecture for automatic spoken language identification from speech signals, Expert Systems with Applications 168 (2021) 114416.
[13] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, M. Auli, Unsupervised cross-lingual representation learning for speech recognition, arXiv preprint arXiv:2006.13979 (2020).
[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, arXiv preprint arXiv:2212.04356 (2022).
[15] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, A. Bapna, Fleurs: Few-shot learning evaluation of universal representations of speech, in: 2022 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2023, pp. 798–805.
[16] G. Leonard, Language Recognition Test and Evaluation., Technical Report, TEXAS INSTRUMENTS INC DALLAS CENTRAL RESEARCH LABS, 1980.
[17] J. Foil, Language identification using noisy speech, in: ICASSP'86. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 11, IEEE, 1986, pp. 861–864.
[18] F. J. Goodman, A. F. Martin, R. E. Wohlford, Improved automatic language identification in noisy speech, in: International Conference on Acoustics, Speech, and Signal Processing,, IEEE, 1989, pp. 528–531.
[19] D. Cimarusti, R. Ives, Development of an automatic identification system of spoken languages: Phase i, in: ICASSP'82. IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 7, IEEE, 1982, pp. 1661–1663.
[20] M. A. Zissman, Automatic language identification using gaussian mixture and hidden markov models, in: 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 2, IEEE, 1993, pp. 399–402.
[21] M. Sugiyama, Automatic language recognition using acoustic features, in: [Proceedings] ICASSP 91: 1991 International Conference on Acoustics, Speech, and Signal Processing, IEEE, 1991, pp. 813–816.
[22] V. Gazeau, C. Varol, Automatic spoken language recognition with neural networks, Int. J. Inf. Technol. Comput. Sci.(IJITCS) 10 (2018) 11–17.
[23] H. Mukherjee, S. M. Obaidullah, K. Santosh, S. Phadikar, K. Roy, A lazy learning-based language identification from speech using mfcc-2 features, International Journal of Machine Learning and Cybernetics 11 (2020) 1–14.
[24] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-

supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.
[25] R.-H. A. Lee, J.-S. R. Jang, A syllable structure approach to spoken language recognition, in: Statistical Language and Speech Processing: 6th International Conference, SLSP 2018, Mons, Belgium, October 15–16, 2018, Proceedings 6, Springer, 2018, pp. 56–66.
[26] X. Lu, P. Shen, Y. Tsao, H. Kawai, Unsupervised neural adaptation model based on optimal transport for spoken language identification, in: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2021, pp. 7213–7217.
[27] M. Verma, A. B. Buduru, Fine-grained language identification with multilingual capsnet model, in: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), IEEE, 2020, pp. 94–102.
[28] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, S. Khudanpur, X-vectors: Robust dnn embeddings for speaker recognition, in: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2018, pp. 5329–5333.
[29] D. Snyder, D. Garcia-Romero, D. Povey, Time delay deep neural network-based universal background models for speaker recognition, in: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), IEEE, 2015, pp. 92–97.
[30] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. J. Lang, Phoneme recognition using time-delay neural networks, IEEE transactions on acoustics, speech, and signal processing 37 (1989) 328–339.
[31] A. I. Abdurrahman, A. Zahra, Spoken language identification using i-vectors, x-vectors, plda and logistic regression, Bulletin of Electrical Engineering and Informatics 10 (2021) 2237–2244.
[32] E. Salesky, B. M. Abdullah, S. J. Mielke, E. Klyachko, O. Serikov, E. Ponti, R. Kumar, R. Cotterell, E. Vylomova, Sigtyp 2021 shared task: robust spoken language identification, arXiv preprint arXiv:2106.03895 (2021).
[33] R. Bedyakin, N. Mikhaylovskiy, Low-resource spoken language identification using self-attentive pooling and deep 1d time-channel separable convolutions, arXiv preprint arXiv:2106.00052 (2021).
[34] G. N. Help, Change the language of google assistant, 2023. URL: https://support.google.com/googlenest/answer/7550584?hl=en&co=GENIE.Platform%3DAndroid.
[35] G. A. Help, Change your language or use multiple languages, 2023. URL: https://support.google.com/assistant/answer/7394513?hl=en&co=GENIE.Platform%3DAndroid#zippy=%5C%2Cphone-or-tablet, last accessed 10 September 2023.
[36] A. Help, Ask alexa to speak in multiple languages, 2023. URL: https://www.amazon.com/gp/help/customer/display.html?nodeId=G9JFV7VRANZDKKWG, last accessed 10 September 2023.
[37] A. Support, Use multiple languages to speak to siri in india, 2023. URL: https://support.apple.com/en-in/HT212537, last accessed 10 September 2023.
[38] R. van der Merwe, Triplet entropy loss: improving the generalisation of short speech language identification systems, arXiv preprint arXiv:2012.03775 (2020).
[39] B. M. Abdullah, J. Kudera, T. Avgustinova, B. Möbius, D. Klakow, Rediscovering

the slavic continuum in representations emerging from neural models of spoken language identification, arXiv preprint arXiv:2010.11973 (2020).

[40] V. Pratap, A. Tjandra, B. Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, A. Baevski, Y. Adi, X. Zhang, W.-N. Hsu, A. Conneau, M. Auli, Scaling speech technology to 1,000+ languages, arXiv (2023).

[41] M. Biswas, S. Rahaman, A. Ahmadian, K. Subari, P. K. Singh, Automatic spoken language identification using mfcc based time series features, Multimedia Tools and Applications 82 (2023) 9565–9595.

[42] A. Bapna, C. Cherry, Y. Zhang, Y. Jia, M. Johnson, Y. Cheng, S. Khanuja, J. Riesa, A. Conneau, mslam: Massively multilingual joint pre-training for speech and text, arXiv preprint arXiv:2202.01374 (2022).

[43] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, S. Watanabe, Improving massively multilingual asr with auxiliary ctc objectives, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.

[44] F. He, S. H. C. Chu, O. Kjartansson, C. E. Rivera, A. Katanova, A. Gutkin, I. Demirsahin, C. C. Johny, M. Jansche, S. Sarin, K. Pipatsrisawat, Open-source multi-speaker speech corpora for building gujarati, kannada, malayalam, marathi, tamil and telugu speech synthesis systems, in: Proc. 12th Language Resources and Evaluation Conference (LREC 2020), 11–16 May, Marseille, France, 2020, pp. 6494–6503. URL: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.800.pdf.

Table 4: Test accuracy of past approaches from 2020-2023, with last column indicating if the training and test sets had disjoint speakers.

Model	Languages	Acc	Dataset	Disjoint speaker set
FRESH + ANN [41]	6	99.94%	IIIT-H Indic Speech Database	No
PLDA logistic regression [31]	3	96%	Combined OpenSLR, YouTube	No
Self-attentive pooling decoder [33]	6	92.5%	VoxForge	No
CNN [39]	11	97.35%	Combined Radio Broadcast Speech and GlobalPhone Read Speech	No
CapsNet [27]	11	98.2%	YouTube	No
CRNN ResNet50 DenseNet121 [38]	6	89%	NCHLT	No
CRNN ResNet50 DenseNet121 [38]	6	45%	Trained on NCHLT, tested on Lwazi	Yes
CNN ResNet50 RNN [32]	16	53%	combined CMU Wilderness dataset, Common Voice, radio news, crowd-sourced recordings as well as other microphone data [44] and "collections of recordings from varied sources"[32] (github link).	Yes
Whisper [14]	102 (99)	64.5%	Own & FLEURS	Yes
mSLAM [42]	102	77.7%	FLEURS has been used for the evaluation of the language identification task. For the pre-training, however, multiple other datasets have been used.	Yes