SIGTYP 2021 Shared Task: Robust Spoken Language Identification

Elizabeth Salesky[◇*] **Badr M. Abdullah**[‡*] **Sabrina J. Mielke**[◇*]
Elena Klyachko[◁,▷] **Oleg Serikov**[◁] **Edoardo Ponti**[□]
Ritesh Kumar[•] **Ryan Cotterell**[⊙] **Ekaterina Vylomova**[♯]
[◇]Johns Hopkins University [‡]Saarland University [◁]Higher School of Economics
[▷]The Institute of Linguistics RAS [□]Mila/McGill University Montreal
[•]Bhim Rao Ambedkar University [⊙]ETH Zürich [♯]University of Melbourne

Abstract

While language identification is a fundamental speech and language processing task, for many languages and language families it remains a challenging task. For many low-resource and endangered languages this is in part due to resource availability: where larger datasets exist, they may be single-speaker or have different domains than desired application scenarios, demanding a need for domain and speaker-invariant language identification systems. This year's shared task on robust spoken language identification sought to investigate just this scenario: systems were to be trained on largely single-speaker speech from one domain, but evaluated on data in other domains recorded from speakers under different recording circumstances, mimicking realistic low-resource scenarios. We see that domain and speaker mismatch proves very challenging for current methods which can perform above 95% accuracy in-domain, which domain adaptation can address to some degree, but that these conditions merit further investigation to make spoken language identification accessible in many scenarios.

1 Introduction

Depending on how we count, there are roughly 7000 languages spoken around the world today. The field of linguistic typology is concerned with the study and categorization of the world's languages based on their linguistic structural properties (Comrie, 1988; Croft, 2002). While two languages may share structural properties across some typological dimensions, they may vary across others. For example, two languages could have identical speech sounds in their phonetic inventory, yet be perceived as dissimilar because each has its own unique set of phonological rules governing possible sound combinations. This leads to tremendous variation and diversity in speech patterns across the world languages (Tucker and Wright, 2020), the effects of which are understudied across many downstream applications due in part to lack of available resources. Building robust speech technologies which are applicable to any language is crucial to equal access as well as the preservation, documentation, and categorization of the world's languages, especially for endangered languages with a declining speaker community.

Unfortunately, robust (spoken) language technologies are only available for a small number of languages, mainly for speaker communities with strong economic power. The main hurdle for the development of speech technologies for under-represented languages is the lack of high-quality transcribed speech resources (see Joshi et al. (2020) for a detailed discussion on linguistic diversity in language technology research). The largest multilingual speech resource in terms of language coverage is the CMU Wilderness dataset (Black, 2019), which consists of read speech segments from the Bible in ∼700 languages. Although this wide-coverage resource provides an opportunity to study many endangered and under-represented languages, it has a narrow domain and lacks speaker diversity as the vast majority of segments are recorded by low-pitch male speakers. It remains unclear whether such resources can be exploited to build generalizable speech technologies for under-resourced languages.

Spoken language identification (SLID) is an enabling technology for multilingual speech communication with a wide range of applications. Earlier SLID systems addressed the problem using the phonotactic approach whereby generative models are trained on sequences of phones transduced from the speech signal using an acoustic model (Lamel and Gauvain, 1994; Li and Ma, 2005). Most current state-of-the-art SLID systems are based on deep neural networks which are trained end-to-end from a spectral representation of the acoustic sig-

[*]Equal contribution

Proceedings of the Third Workshop on Computational Typology and Multilingual NLP, pages 122–129

nal (e.g., MFCC feature vectors) without any intermediate symbolic representations (Lopez-Moreno et al., 2014; Gonzalez-Dominguez et al., 2014). In addition to their ability to effectively learn to discriminate between closely related language varieties (Gelly et al., 2016; Shon et al., 2018), it has been shown that neural networks can capture the degree of relatedness and similarity between languages in their emergent representations (Abdullah et al., 2020).

Several SLID evaluation campaigns have been organized in the past, including the NIST Language Recognition Evaluation (Lee et al., 2016; Sadjadi et al., 2018), focusing on different aspects of this task including closely related languages, and typically used conversational telephone speech. However, the languages were not sampled according to typologically-aware criteria but rather were geographic or resource-driven choices. Therefore, while the NIST task languages may represent a diverse subset of the world's languages, there are many languages and language families which have not been observed in past tasks. In this shared task, we aim to address this limitation by broadening the language coverage to a set of typologically diverse languages across seven languages families. We also aim to assess the degree to which single-speaker speech resources from a narrow domain can be utilized to build robust speech language technologies.

2 Task Description

While language identification is a fundamental speech and language processing task, it remains a challenging task, especially when going beyond the small set of languages past evaluation has focused on. Further, for many low-resource and endangered languages, only single-speaker recordings may be available, demanding a need for domain and speaker-invariant language identification systems.

We selected 16 typologically diverse languages, some of which share phonological features, and others where these have been lost or gained due to language contact, to perform what we call robust language identification: systems were to be trained on largely single-speaker speech from one domain, but evaluated on data in other domains recorded from speakers under different recording circumstances, mimicking more realistic low-resource scenarios.

2.1 Provided Data

To train models, we provided participants with speech data from the CMU Wilderness dataset (Black, 2019), which contains utterance-aligned read speech from the Bible in 699 languages,[1] but predominantly recorded from a single speaker per language, typically male. Evaluation was conducted on data from other sources—in particular, multi-speaker datasets recorded in a variety of conditions, testing systems' capacity to generalize to new domains, new speakers, and new recording settings. Languages were chosen from the CMU Wilderness dataset given availability of additional data in a different setting, and include several language families as well as more closely-related challenge pairs such as Javanese and Sundanese. These included data from the Common Voice project (CV; Ardila et al., 2020) which is read speech typically recorded using built-in laptop microphones; radio news data (SLR24; Juan et al., 2014, 2015); crowd-sourced recordings using portable electronics (SLR35, SLR36; Kjartansson et al., 2018); cleanly recorded microphone data (SLR64, SLR65, SLR66, SLR79; He et al., 2020); and a collection of recordings from varied sources (SS; Shukla, 2020). Table 1 shows the task languages and their data sources for evaluation splits for the robust language identification task.

We strove to provide balanced data to ensure signal comes from salient information about the language rather than spurious correlations about e.g. utterance length. We selected and/or trimmed utterances from the CMU Wilderness dataset to between 3 to 7 seconds in length. Training data for all languages comprised 4,000 samples each. We selected evaluation sources for validation and blind test sets to ensure no possible overlap with CMU Wilderness speakers. We held out speakers between validation and test splits, and balanced speaker gender within splits to the degree possible where annotations were available. We note that the Marathi dataset is female-only. Validation and blind test sets each comprised 500 samples per language. We release the data as derivative MFCC features.

3 Evaluation

The robust language identification shared task allowed two kinds of submissions: first, *constrained* submissions, for which only the provided training

[1]Data source: bible.is

ISO	Wilderness ID	Language name	Family	Genus	Macroarea	Train	Eval
kab	KABCEB	Kabyle	Afro-Asiatic	Berber	Africa	Wilderness	CV
iba	IBATIV	Iban	Austronesian	Malayo-Sumbawan	Papunesia	Wilderness	SLR24
ind	INZTSI	Indonesian	Austronesian	Malayo-Sumbawan	Papunesia	Wilderness	CV
sun	SUNIBS	Sundanese	Austronesian	Malayo-Sumbawan	Papunesia	Wilderness	SLR36
jav	JAVNRF	Javanese	Austronesian	Javanese	Papunesia	Wilderness	SLR35
eus	EUSEAB	Euskara	Basque	Basque	Eurasia	Wilderness	CV
tam	TCVWTC	Tamil	Dravidian	Southern Dravidian	Eurasia	Wilderness	SLR65
kan	ERVWTC	Kannada	Dravidian	Southern Dravidian	Eurasia	Wilderness	SLR79
tel	TCWWTC	Telugu	Dravidian	South-Central Dravidian	Eurasia	Wilderness	SLR66
hin	HNDSKV	Hindi	Indo-European	Indic	Eurasia	Wilderness	SS
por	PORARA	Portuguese	Indo-European	Romance	Eurasia	Wilderness	CV
rus	RUSS76	Russian	Indo-European	Slavic	Eurasia	Wilderness	CV
eng	EN1NIV	English	Indo-European	Germanic	Eurasia	Wilderness	CV
mar	MARWTC	Marathi	Indo-European	Indic	Eurasia	Wilderness	SLR64
cnh	CNHBSM	Chin, Hakha	Niger-Congo	Gur	Africa	Wilderness	CV
tha	THATSV	Thai	Tai-Kadai	Kam-Tai	Eurasia	Wilderness	CV

Table 1: Provided data with language family and macroarea information. **ISO** shows ISO 639-3 codes. Training data (**Train**) for all languages is taken from CMU Wilderness dataset; validation and evaluation data (**Eval**) is derived from multiple data sources.

data was used; and second, *unconstrained* submissions, in which the training data may be extended with any external source of information (e.g. pretrained models, additional data, etc.).

3.1 Evaluation Metrics

We evaluate task performance using precision, recall, and F_1. For each metric we report both micro-averages, meaning that the metric average is computed equally-weighted across all samples for all languages, and macro-averages, meaning that we first computed the metric for each language and then averaged these aggregates to see whether submissions behave differently on different languages. Participant submissions were ranked according to macro-averaged F_1.

3.2 Baseline

For our baseline SLID system, we use a deep convolutional neural network (CNN) as sequence classification model. The model can be viewed as two components trained end-to-end: a segment-level feature extractor (f) and a language classifier (g). Given as input a speech segment parametrized as sequence of MFCC frames $\mathbf{x}_{1:T} = (\mathbf{x}_1, \ldots, \mathbf{x}_T) \in \mathbb{R}^{k \times T}$, where T is the number of frames and k is the number of the spectral coefficients, the segment-level feature extractor first transforms $\mathbf{x}_{1:T}$ into a segment-level representation as $\mathbf{u} = f(\mathbf{x}_{1:T}; \boldsymbol{\theta}_f) \in \mathbb{R}^d$. Then, the language classifier transforms $\mathbf{u}$ into a logit vector $\hat{\mathbf{y}} \in \mathbb{R}^{|\mathcal{Y}|}$, where $\mathcal{Y}$ is the set of languages, through a series of non-linear transformations as $\hat{\mathbf{y}} = g(\mathbf{u}; \boldsymbol{\theta}_g)$. The logit vector $\hat{\mathbf{y}}$ is then fed to a softmax function to get a probability distribution over the languages.

The segment-level feature extractor consists of three 1-dimensional, temporal convolution layers with 64, 128, 256 filters of widths 16, 32, 48 for each layer and a fixed stride of 1 step. Following each convolutional operation, we apply batch normalization, ReLU non-linearity, and unit dropout with probability which was tuned over $\{0.0, 0.4, 0.6\}$. We apply average pooling to downsample the representation only at the end of the convolution block, which yields a segment representation $\mathbf{u} \in \mathbb{R}^{256}$. The language classifier consists of 3 fully-connected layers ($256 \rightarrow 256 \rightarrow 256 \rightarrow 16$), with a unit dropout with probability 0.4 between the layers, before the softmax layer. The model is trained with the ADAM optimizer with a batch size of 256 for 50 epochs. We report the results of the best epoch on the validation set as our baseline results.

3.3 Submissions

We received three constrained submissions from three teams, as described below.

Anlirika (Shcherbakov et al., 2021, composite) The submitted system (constrained) consists of several recurrent, convolutional, and dense layers. The neural architecture starts with a dense layer that is designed to remove sound harmonics from a raw spectral pattern. This is followed by a 1D convolutional layer that extracts audio frequency patterns

(features). Then the features are fed into a stack of LSTMs that focuses on 'local' temporal constructs. The output of the stack of LSTMs is then additionally concatenated with the CNN features and is fed into one more LSTM module. Using the resulting representation, the final (dense) layer evaluates a categorical loss across 16 classes. The network was trained with Adam optimizer, the learning rate was set to be 5×10^{-4}. In addition, similar to Lipsia, the team implemented a data augmentation strategy: samples from validation set have been added to the training data.

Lipsia (Celano, 2021, Universität Leipzig) submitted a constrained system based on the ResNet-50 (He et al., 2016), a deep (50 layers) CNN-based neural architecture. The choice of the model is supported by a comparative analysis with more shallow architectures such as ResNet-34 and a 3-layer CNNs that all were shown to overfit to the training data. In addition, the authors proposed transforming MFCC features into corresponding 640x480 spectrograms since this data format is more suitable for CNNs. The output layer of the network is dense and evaluates the probabilities of 16 language classes.[2] Finally, the authors augmented the training data with 60% of the samples from the validation set because the training set did not present enough variety in terms of domains and speakers while the validation data included significantly more. Use of the validation data in this way seems to have greatly improved generalization ability of the model.

The model performed relatively well with no fine-tuning or transfer-learning applied after augmentation.[3]

NTR (Bedyakin and Mikhaylovskiy, 2021, NTR Labs composite), submitted an essentially constrained[4] system which uses a CNN with a self-attentive pooling layer. The architecture of the network was QuartzNet ASR following Kriman et al. (2020), with the decoder mechanism replaced with a linear classification mechanism. The authors also used a similar approach in another challenge on low-resource ASR, Dialog-2021 ASR[5]. They applied several augmentation techniques, namely shifting samples in range (-5ms; +5ms), MFCC perturbations (SpecAugment; Park et al., 2019), and adding background noise.

4 Results and Analysis

The main results in Table 2 show all systems greatly varying in performance, with the Lipsia system clearly coming out on top, boasting best accuracy and average F_1 score, and reaching the best F_1 score for nearly each language individually.[6]

All four systems' performance varies greatly on average, but nevertheless some interesting overall trends emerge. Figure 1 shows that while the Anlirika and Lipsia systems' performance on the different languages do not correlate with the baseline system (linear fit with Pearson's $R^2 = 0.00$ and $p > 0.8$ and $R^2 = 0.02$ and $p > 0.5$, respectively), the NTR system's struggle correlates at least somewhat with the same languages that the baseline system struggles with: a linear fit has $R^2 = 0.15$ with $p > 0.1$. More interestingly, in correlations amongst themselves, the Anlirika and Lipsia systems do clearly correlate ($R^2 = 0.57$ and $p < 0.001$), and the NTR system correlates again at least somewhat with the Anlirika system ($R^2 = 0.11$ and $p > 0.2$) and the Lipsia system ($R^2 = 0.19$ and $p > 0.05$).

Note that most systems submitted are powerful enough to fit the training data: our baseline achieves a macro-averaged F_1 score of .98 ($\pm$.01) on the training data, the Lipsia system similarly achieves .97 ($\pm$.03), the NTR system reaches a score of .99 ($\pm$.02). An outlier, the Anlirika system reaches only .75 ($\pm$.09). On held-out data from CMU Wilderness which matches the training data domain, the baseline achieves .96 F1. This suggests an inability to generalize across domains and/or speakers without additional data for adaptation.

Diving deeper into performance on different languages and families, Figure 2 shows confusion matrices for precision and recall, grouped by language family. We can see the superiority of the Lipsia

[2]The submitted system actually predicts one out of 18 classes as two other languages that weren't part of the eventual test set were included. The system predicted these two languages for 27 of 8000 test examples, i.e., $\approx$ 0.34%.

[3]The authors trained ResNet-50 from scratch.

[4]Although technically external noise data was used when augmenting the dataset, no language-specific resources were.

[5]http://www.dialog-21.ru/en/evaluation/

[6]Each of the "wins" indicated by boldface in Table 2 is statistically significant under a paired-permutation significance test (note that as we are not in a multiple-hypothesis testing setting, we do not apply Bonferroni or similar corrections). There are no significant differences between the baseline and the Anlirika system for kab, ind, por, rus, and eng; between the baseline and the Lipsia system for sun; between the baseline and the NTR system for ind, iba, and cnh; between Anlirika and Lipsia on rus; between Lipsia and NTR on rus; between Anlirika and NTR on ind and rus.

ISO	Anlirika		Baseline		Lipsia		NTR	
	Valid.	Test	Valid.	Test	Valid.	Test	Valid.	Test
Family: Afro-Asiatic	*.329*	.214	*.181*	.235	*.670*	**.453**	*.102*	.082
kab	*.329*	.214	*.181*	.235	*.670*	**.453**	*.102*	.082
Family: Austronesian	*.429*	.368	*.082*	.094	*.578*	**.498**	*.065*	.060
iba	*.692*	.696	*.029*	.018	*.980*	**.968**	*.020*	.031
ind	*.350*	.108	*.033*	.105	*.700*	**.338**	*.096*	.074
sun	*.406*	**.369**	*.160*	.149	*.090*	.140	*.086*	.082
jav	*.267*	.300	*.106*	.106	*.540*	**.547**	*.059*	.053
Family: Basque	*.565*	.405	*.100*	.090	*.850*	**.792**	*.077*	.016
eus	*.565*	.405	*.100*	.090	*.850*	**.792**	*.077*	.016
Family: Dravidian	*.351*	.246	*.202*	.138	*.807*	**.572**	*.074*	.053
tam	*.342*	.272	*.348*	.204	*.800*	**.609**	*.172*	.046
kan	*.188*	.168	*.000*	.042	*.820*	**.557**	*.004*	.015
tel	*.523*	.298	*.259*	.168	*.800*	**.550**	*.046*	.097
Family: Indo-European	*.439*	.225	*.130*	.144	*.722*	**.402**	*.114*	.047
hin	*.458*	.378	*.091*	.099	*.780*	**.635**	*.021*	.011
por	*.211*	.143	*.157*	.166	*.550*	**.358**	*.102*	.068
rus	*.630*	**.034**	*.014*	.014	*.900*	**.065**	*.050*	**.049**
eng	*.194*	.148	*.161*	.179	*.460*	**.414**	*.270*	.099
mar	*.701*	.423	*.229*	.263	*.920*	**.539**	*.126*	.010
Family: Niger-Congo	*.516*	.403	*.138*	.063	*.860*	**.763**	*.122*	.038
cnh	*.516*	.403	*.138*	.063	*.860*	**.763**	*.122*	.038
Family: Tai-Kadai	*.362*	.156	*.086*	.052	*.780*	**.401**	*.025*	.015
tha	*.362*	.156	*.086*	.052	*.780*	**.401**	*.025*	.015
F1, Macro Avg.	*.421*	.282	*.131*	.122	*.719*	**.508**	*.086*	.049
F1, Micro Avg.	*.436*	.298	*.145*	.137		**.532**		.063
Accuracy		29.9%		13.7%		**53.1%**		6.3%

Table 2: F_1 scores, their macro-averages per family, and overall accuracies of submitted predictions on validation and test data (validation results are self-reported by participants). The Lipsia system performed best across nearly all languages and consistently achieves the highest averages.

system and to a lesser degree the Anlirika system over the generally more noisy and unreliable baseline system and the NTR system which was likely overtrained: it classifies 23% of examples as tel, 20% as kab, and 16% as eng, with the remaining 41% spread across the remaining 13 languages (so $\approx$ 3.2% per language).

Interestingly, the other three systems all struggle to tell apart sun and jav, the Anlirika and baseline systems classifying both mostly as sun and the Lipsia system classifying both mostly as jav. Note that the baseline system tends to label many languages' examples as sun (most notably mar, the test data for which contains only female speakers), eus (most notably also rus), and eng (most notably also iba), despite balanced training data. In a similar pattern, the Anlirika predicts tam for many languages, in particular ind, the other two Dravidian languages kan and tel, por, rus, eng, cnh, and tha.

Looking more closely at the clearly best-performing system, the Lipsia system, and its performance and confusions, we furthermore find that the biggest divergence from the diagonal after the sun/jav confusion is a tendency to label rus as por, and the second biggest divergence is that mar examples are also sometimes labeled as kan and tel; while the first one is within the same family, in the second case, these are neighbouring languages in

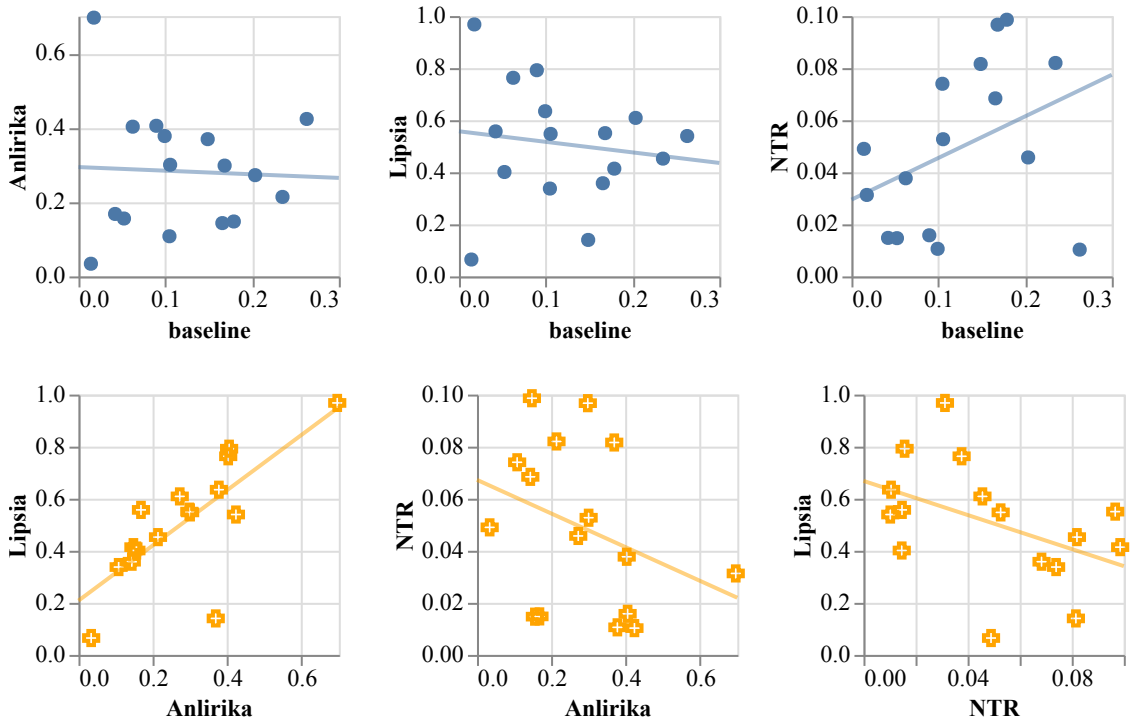

Figure 1: Correlating submitted systems' F_1 scores for our 16 languages on the test set. The lines are linear regressions as described in Section 4.

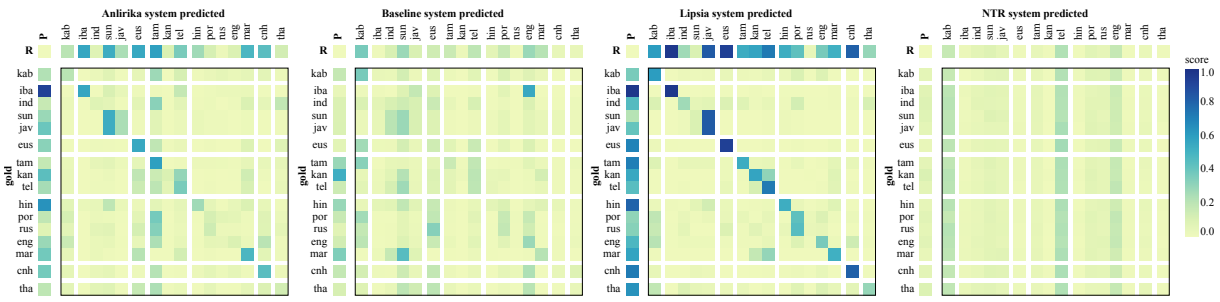

Figure 2: Visualization of Precision (P), Recall (R), and confusion matrices (scores are counts normalized by number of gold entries) for the Anlirika, baseline, Lipsia, and NTR system, grouped by language families.

contact and mar shares some typological properties with kan (and kan and tel belong to the same language family).

5 Conclusion

This paper describes the SIGTYP shared task on robust spoken language identification (SLID). This task investigated the ability of current SLID models to generalize across speakers and domains. The best system achieved a macro-averaged accuracy of 53% by training on validation data, indicating that even then the task is far from solved. Further exploration of few-shot domain and speaker adaptation is necessary for SLID systems to be applied outside typical well-matched data scenarios.

References

Badr M. Abdullah, Jacek Kudera, Tania Avgustinova, Bernd Möbius, and Dietrich Klakow. 2020. Rediscovering the Slavic continuum in representations emerging from neural models of spoken language identification. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 128–139, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

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 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Roman Bedyakin and Nikolay Mikhaylovskiy. 2021. Language ID Prediction from Speech Using Self-Attentive Pooling and 1D-Convolutions. In *Proceedings of the Third Workshop on Computational Research in Linguistic Typology*. North American Association for Computational Linguistics.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Giuseppe Celano. 2021. A ResNet-50-based Convolutional Neural Network Model for Language ID Identification from Speech Recordings. In *Proceedings of the Third Workshop on Computational Research in Linguistic Typology*. North American Association for Computational Linguistics.

Bernard Comrie. 1988. Linguistic typology. *Annual Review of Anthropology*, 17:145–159.

William Croft. 2002. *Typology and Universals*. Cambridge University Press.

Grégory Gelly, Jean-Luc Gauvain, Lori Lamel, Antoine Laurent, Viet Bac Le, and Abdel Messaoudi. 2016. Language recognition for dialects and closely related languages. In *Odyssey*, pages 124–131.

Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez, and Pedro J Moreno. 2014. Automatic language identification using long short-term memory recurrent neural networks. In *Fifteenth Annual Conference of the International Speech Communication Association*.

Fei He, Shan-Hui Cathy Chu, Oddur Kjartansson, Clara Rivera, Anna Katanova, Alexander Gutkin, Isin Demirsahin, Cibu Johny, Martin Jansche, Supheakmungkol Sarin, and Knot Pipatsrisawat. 2020. Open-source Multi-speaker Speech Corpora for Building Gujarati, Kannada, Malayalam, Marathi, Tamil and Telugu Speech Synthesis Systems. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6494–6503, Marseille, France. European Language Resources Association (ELRA).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Mohamed Dyab. 2015. Using resources from a closely-related language to develop asr for a very under-resourced language: A case study for iban. In *Proceedings of INTERSPEECH*, Dresden, Germany.

Sarah Samson Juan, Laurent Besacier, and Solange Rossato. 2014. Semi-supervised g2p bootstrapping and its application to asr for a very under-resourced language: Iban. In *Proceedings of Workshop for Spoken Language Technology for Under-resourced (SLTU)*.

Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. 2018. Crowd-Sourced Speech Corpora for Javanese, Sundanese, Sinhala, Nepali, and Bangladeshi Bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 52–55, Gurugram, India.

Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2020. Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6124–6128. IEEE.

Lori F Lamel and Jean-Luc Gauvain. 1994. Language identification using phone-based acoustic likelihoods. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I–293. IEEE.

Kong Aik Lee, Haizhou Li, Li Deng, Ville Hautamäki, Wei Rao, Xiong Xiao, Anthony Larcher, Hanwu Sun, Trung Nguyen, Guangsen Wang, et al. 2016. The 2015 nist language recognition evaluation: the shared view of i2r, fantastic4 and singams. In *Interspeech 2016*, volume 2016, pages 3211–3215.

Haizhou Li and Bin Ma. 2005. A phonotactic language model for spoken language identification. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 515–522.

Ignacio Lopez-Moreno, Javier Gonzalez-Dominguez, Oldrich Plchot, David Martinez, Joaquin Gonzalez-Rodriguez, and Pedro Moreno. 2014. Automatic language identification using deep neural networks. In *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5337–5341. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Seyed Omid Sadjadi, Timothee Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero. 2018. The 2017 nist language recognition evaluation. In *Odyssey*, pages 82–89.

Andrei Shcherbakov, Liam Whittle, Ritesh Kumar, Siddharth Singh, Matthew Coleman, and Ekaterina Vylomova. 2021. Anlirika: an LSTM–CNN Flow Twister for Language ID Prediction. In *Proceedings of the Third Workshop on Computational Research in Linguistic Typology*. North American Association for Computational Linguistics.

Suwon Shon, Ahmed Ali, and James Glass. 2018. Convolutional neural network and language embeddings for end-to-end dialect recognition. In *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 98–104.

Shivam Shukla. 2020. Speech dataset in hindi language.

Benjamin V Tucker and Richard Wright. 2020. Introduction to the special issue on the phonetics of underdocumented languages. *The Journal of the Acoustical Society of America*, 147(4):2741–2744.