INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Self-supervised Phonotactic Representations for Language Identification

G. Ramesh, C. Shiva Kumar, K. Sri Rama Murty

Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
gundlururamesh720@gmail.com, ee19mtech01001@iith.ac.in, ksrm@ee.iith.ac.in

Abstract

Phonotactic constraints characterize the sequence of permissible phoneme structures in a language and hence form an important cue for language identification (LID) task. As phonotactic constraints span across multiple phonemes, the short-term spectral analysis (20-30 ms) alone is not sufficient to capture them. The speech signal has to be analyzed over longer contexts (100s of milliseconds) in order to extract features representing the phonotactic constraints. The supervised senone classifiers, aimed at modeling triphone context, have been used for extracting language-specific features for the LID task. However, it is difficult to get large amounts of manually labeled data to train the supervised models. In this work, we explore a self-supervised approach to extract long-term contextual features for the LID task. We have used wav2vec architecture to extract contextualized representations from multiple frames of the speech signal. The contextualized representations extracted from the pre-trained wav2vec model are used for the LID task. The performance of the proposed features is evaluated on a dataset containing 7 Indian languages. The proposed self-supervised embeddings achieved 23% absolute improvement over the acoustic features and 3% absolute improvement over their supervised counterparts.

Index Terms: Language identification, Bottleneck features, Deep neural networks, Self-supervised representations.

1. Introduction

Multilingual interoperability is an important issue for the deployment of speech systems in practice. The language of communication needs to be identified for loading the corresponding speech recognizer/synthesizer modules in the voice assistant and speech translation systems. Automatic identification of the language of a spoken utterance by a machine is referred to as a language identification (LID) task. The development of a robust LID system requires the extraction of language-specific features from the spoken utterance. Language-specific cues such as phonology, morphology, syntax and prosody play a vital role in discriminating between the languages [1, 2]. Depending on the choice of features, the LID approaches are categorized into lexical and prelexical approaches [3]. While the lexical approaches rely on higher-level features like word formation and syntax, the prelexical approaches use acoustic-phonetic, phonotactic and prosodic cues for building the LID system. In this paper, we explore a self-supervised approach to extract language-specific acoustic-phonetic and phonotactic features for the LID task.

The language-specific information can be extracted from the speech signal at different temporal resolutions. At the lowest level, analysis of 20-30 ms frames of speech signal captures the phonetic characteristics of the language. Short-time spectral features like mel-frequency cepstral coefficients (MFCC), linear prediction cepstral coefficients (LPCC) [4] and instantaneous frequency cosine coefficients (IFCC) [5] have been used for language recognition. At the next level, the phonotactic constraints of the language can be captured by analyzing speech signal over hundreds of milliseconds spanning a sequence of phonemes. Since speech is a nonstationary signal, spectral analysis techniques cannot be extended directly to the longer segments. Instead, temporal dynamics of the short-time spectral features are modeled over multiple frames to capture the phoneme-transitions. Bottleneck features extracted from a DNN phoneme classifier, trained over longer contexts, have been shown to offer significant improvements in LID performance [6, 7, 8, 9]. However, such an approach requires a large amount of manually transcribed speech data for training the phoneme classifier. The lexical approaches require large vocabulary continuous speech recognizers (LVCSR) of all languages to perform the LID task [10].

In this paper, we explore the efficiency of self-supervised approaches in extracting phonotactic features for the LID task. Representation learning with contrastive predictive coding has been shown to capture the long-term contextual information by using autoregressive models in the latent space to predict multiple time-steps in future [11] As speech signal is an outcome of a random process, the contrastive loss between the distribution of positive and negative examples was used to update the parameters of the network. Schneider *et al.* proposed a fully convolutional architecture wav2vec for contrastive predictive coding of speech signals, which can be easily parallelized over time [12]. The self-supervised representations extracted from these approaches were shown to improve the performance of speech recognition and speaker tasks with a limited amount of manually labeled data [12, 13].

In this work, we have used self-supervised embeddings extracted from wav2vec architecture for the LID task. Since the wav2vec architecture is trained to model longer context windows spanning multiple phonemes, it captures the phonotactic constraints required for LID. The performance of the proposed method is evaluated on 7 Indian languages, and compared with the supervised embeddings extracted from senone classifier [14]. It is observed that the proposed self-supervised embeddings performed better than or similar to the supervised embeddings. The rest of the paper is organized as follows. Section 2 describes the supervised and unsupervised approaches for extracting phonotactic features using deep neural networks (DNNs). In Section 3, we present the DNN-based statistical pooling techniques to obtain the utterance level representation for LID. Section 4 presents an experimental evaluation of the proposed method on 7 Indian languages. Section 5 summarizes the important contributions of this work and highlights possible future directions.

2. Extraction of Phonotactic Features

A good understanding of the acoustic-phonetic characteristics of basic sound units across the languages is useful in building a robust LID system [15]. Even though there may be a significant

http://dx.doi.org/10.21437/Interspeech.2021-1310

overlap between the set of phonemes across the languages, their allophonic variations and pronunciation may differ. Moreover, languages are subjected to phonotactic constraints, i.e., restrictions on the plausible sequence of phonemes in a language [16]. As a consequence of the language-specific phonotactic structure, the relative frequency of phoneme sequences differs significantly across the languages [17]. Among the language-specific features, the phonotactic constraints were shown to be the more powerful for LID task [18]. The phonotactic features have to be extracted from longer contexts covering a sequence of multiple phonemes. The phoneme transitions are manifested in the spectral domain as formant trajectories over 100s of milliseconds. In this work, we have used a self-supervised approach to extract long-term contextual features for the LID task.

2.1. Self-supervised approach: wav2vec

In the recent past, self-supervised learning has witnessed significant progress in extracting contextualized representations from multiple-time steps of the speech signal. Motivated by the success of wav2vec architecture in speech and speaker recognition tasks, we have used it for extracting contextual features for the LID task.

The wav2vec network consists of two components, viz., encoder network and aggregator network. The encoder network, $e : X \rightarrow Z$, maps raw audio X to a latent representation Z. Each latent vector $\mathbf{z}_i \in Z$ covers 30ms of raw speech samples with a stride of $10ms$. The aggregator network, $a : Z \rightarrow C$ maps a sequence of latent vectors to a contextualized representation $\mathbf{c}_i = a(\mathbf{z}_i, \mathbf{z}_{i-1} \ldots \mathbf{z}_{i-v})$. The encoder network consists of 7 convolution layers with strides of {5,4,2,2,2,1,1} and kernel sizes of {10,8,4,4,4,1,1}. The encoder network, with a receptive field of 30 ms, captures the short-time spectral characteristics of the speech signal. On the other hand, the aggregator networks consist of 12 convolutional layers. The aggregator network, with a receptive field of 810 ms, captures long-term contextual features representing the phonotactic constraints. Both the encoder and aggregator networks comprise of causal convolutional layers with 512 feature maps.

The wav2vec network is trained to distinguish the future latent representation in k steps ahead from the negative latent representations by minimizing Noise Contrastive Estimation (NCE) loss [19].

$$L_k = -\sum_{i=1}^{T-k} \Big(\log \sigma(z_{i+k}^T h_k(c_i)) + \lambda \underset{\hat{z} \sim p_n}{E} [\log \sigma(-\hat{z}^T h_k(c_i))] \Big) \quad (1)$$

Here $\sigma(.)$ is the sigmoid function, $h_k(c_i)$ is affine transformation and λ represents the number of negatives. $\sigma(z_{i+k}^T h_k(c_i))$ is the probability of z_{i+k} being true sample and $\sigma(-\hat{z}^T h_k(c_i))$ is the probability of $\hat{z}$ being negative sample. Here k represents the future steps from the context c_i. Total loss $L = \sum_{k=1}^{K} L_k$, obtained by summing over each step upto K=12, is optimized. The wav2vec network is implemented using the fairseq toolkit [20]. After training, the contextualized representations c_i are used as features for the LID task rather than the traditional MFCCs. Figure 1(c) shows 512-dimensional contextualized representations extracted from the speech waveform a Telugu utterance in Figure 1(a). Phoneme boundaries are marked with red lines. The neuronal activations of the encoder representation are clearly distinct across the phonemes and are sparser compared to the spectrogram representation in Figure 1(b). The performance of the self-supervised representations is compared with the supervised representations obtained from the senone classifier.

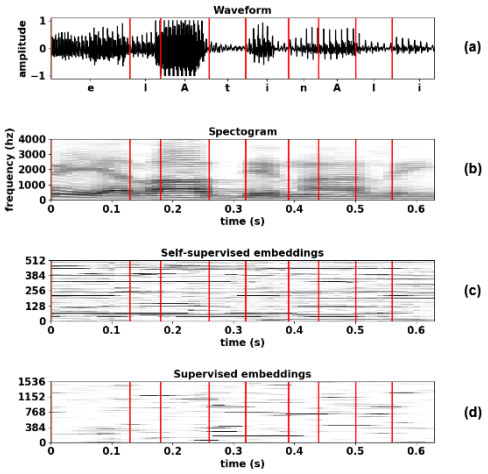


Figure 1: *Visualization of Telugu utterance waveform with phonemes and their boundaries, spectogram, self-supervised embeddings and supervised embeddings*

2.2. Supervised approach: Senone classifier

In supervised approaches, the spectral dynamics can be captured by modeling context-dependent phones like triphones and pentaphones. In the state-of-the-art speech recognition systems, the context-dependent phones are formed by using a sequence of senones [21], which are distinct short sound detectors. The major advantage of senones is their ability to predict and model even the unseen triphones. The bottleneck features extracted from the senone classifier were shown to offer significant performance improvement in the LID task [6, 7, 8, 9].

As it is difficult to get manually transcribed data for all the target languages in the LID task, the senone classifier is typically trained on a neutral resource-rich language. The bottleneck features from the neutral resource-rich senone classifier are used for the LID task.

In this work, the phoneme discriminator is trained on 960 hours of manually transcribed Librispeech data, and the 1536 dimensional bottleneck features extracted from this model are used to perform the LID task on 7 Indian languages. A time-delay neural network architecture with an effective context of 810 ms (40 frames on either side) is used to develop the senone classifier [14]. The inputs for the phoneme discriminator 40-dimensional high-resolution MFCCs concatenated with 100-dimensional i-vectors for speaker adaption.A pre-trained acoustic model from the Kaldi recipe for Librispeech (960 hours) is considered for extracting supervised phonotactic features [22].

3. Utterance-level Pooling of Features

The extracted input features contain the language-specific information at different levels. The decision cannot be taken directly from these features. We need to pool the input features to get utterance level representation. Initially, generative models have been used for pooling the raw-level features to obtain the compact utterance level representation in the LID task [23]. These

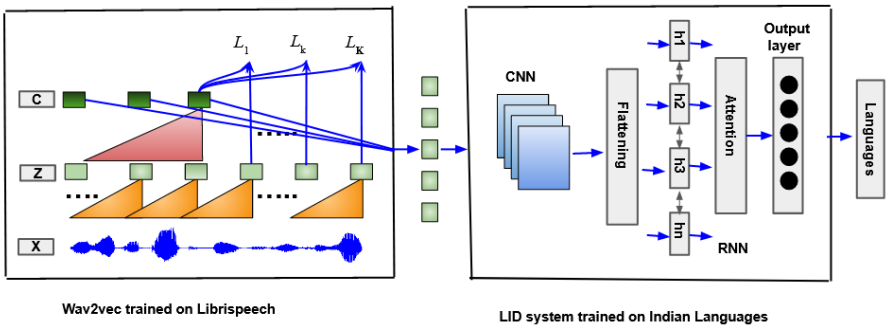


Figure 2: *Proposed architecture for language identification*

generative models learn the underlying data distribution but not the discriminative language patterns across the languages. The discriminative models, in contrast, achieve superior performance in the classification tasks by learning discriminative patterns. DNNs can learn nonlinear discriminative patterns and yield an utterance-level representation from the input features [24, 25, 26]. Based on these models, we have considered three different neural networks to obtain high-level frame representations explained in the following subsections.

3.1. Time-distributed feed-forward neural networks

This network uses a time-distributed layer to project the input features into high-level frame representations. We considered a single-layer, fully connected feed-forward neural network with 128 nodes and ReLu activation to reduce the network complexity. The weights are shared across time steps.

3.2. Sequential networks

The Recurrent neural networks (RNNs) are proposed to capture the temporal contextual information. They outperform the standard i-vector based approach and feed-forward neural networks [27, 28]. When the networks are deeper, we may encounter vanishing gradients. To address this, we added residual connections to bidirectional long short term memory (BLSTM) layers [29, 30]. The BLSTM network contains three BLSTM layers, and each layer is configured with 128 nodes.

3.3. Convolutional recurrent neural networks

The performance of RNNs gets degraded when the input length is long. With the success of connecting the convolutional neural networks (CNNs) to RNNs in language identification [26, 31], we use convolutional recurrent neural networks (CRNNs) to extract the local patterns in the input sequence. The input features are projected into 64-dimensional space before fed to CNNs. Two convolutional layers of 32 filters each, with the kernel sizes of 10×10, 3×3 at each layer with a stride of 1×1 followed by batch normalization and max-pooling of size 2×2 are used. The CNN features are flattened and projected to 64 dimensions. These features are further fed to the BLSTM layer with 128 nodes.

All the obtained high-level frame representations from front-ends are aggregated by using attention mechanism [32]. The series of high-level frame representations are denoted by $H = \{h_1, h_2, .., h_t, .., h_T\}$. Where T is the number of frames and h_t is the representation of frame at t^{th} time instant. The attention layer takes all frame-level representations H of size $d_h \times T$, generates attention weights denoted by A, and then a weighted average is performed. The equations are given by

$$A = softmax(g(H^T W_1)W_2) \quad (2)$$

$$E = HA \quad (3)$$

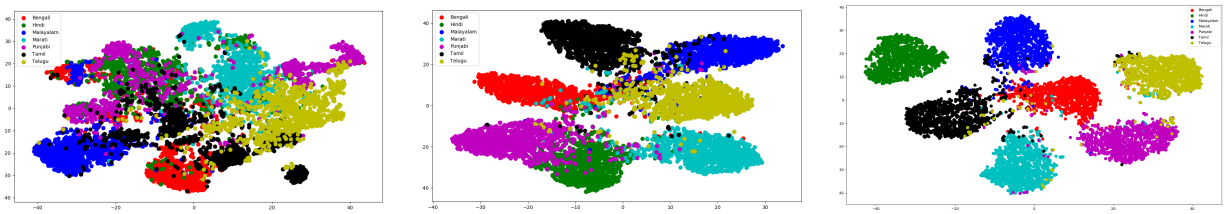
where W_1 with the shape $(d_h \times d_a)$ and W_2 with the shape $(d_a \times d_r)$ are learnable weight matrices, d_h, d_a and d_r represent the frame-level representation dimension, hidden node dimension and number of attention heads respectively. Here $g(\cdot)$ is ReLu activation. E in the equation 3 denotes the final utterance level representations for each head. We averaged across the heads to get the final utterance level representation and projected to the number of target languages. Figure 2 shows the proposed architecture for LID.

4. Experimental Evaluation

The language-discriminative capability of self-supervised representations from the wav2vec network is evaluated on an Indian language dataset containing speech data from 7 languages, viz., Hindi, Bengali, Punjabi, Malayalam, Marathi, Tamil, and Telugu. The dataset has been collected from a YouTube channel named Josh talks [1]. This channel contains motivational, business and real-life story talks from several speakers. The extracted audio tracks are converted resampled to 16kHz. The segments containing music, audience applause, and claps were manually removed from the utterances. From each speaker, 10 minutes of clean data is extracted. The dataset is divided into train and test sets. The train set for each language consists of approximately 9 hours of speech data from around 80 speakers, while the test set for each language consists of 1.5 hours of speech data from 12 speakers. It is ensured that there is no overlap of speakers between the train and test sets. During training, all the inputs are cropped to 3 seconds and fed to the network.

In order to ensure that the results from supervised and self-supervised embeddings are comparable, we have used 960 hours of Librispeech data to pre-train the senone classifier and wav2vec networks. However, it should be noted that the manual transcriptions of the Librispeech data are used only while training the senone classifier. The wav2vec approach, being unsupervised, does not require manual transcriptions. The pre-trained senone classifier is used to extract 1536-dimensional supervised embeddings as described in Section 2.2. Similarly, the pre-trained wav2vec network is used to extract 512-dimensional self-supervised embeddings as described in Section 2.1. We have also considered the 7-dimensional MFCCs as the baseline acoustic features for comparison [33].

[1] https://www.youtube.com/c/JoshTalksLive/channels



(a) Acoustic features (b) Supervised phonotactic embeddings (c) Self-supervised embeddings

Figure 3: *t-SNE plots for the embeddings of the CRNN system with various features.*

4.1. Results and discussion

The frame-level features are used in conjunction with the stat-pooling networks discussed in Section 3 to identify the language of an input utterance. The performance interms of micro-average accuracy of different feature and classifier combinations is given in Table 1. As the supervised and self-supervised embeddings capture the long-term contextual information, their performance is significantly better than the acoustic features. The performance of the acoustic features improved with the increased classifier complexity, i.e., BLSTM and CRNN networks performed better than the time-distributed network. On the other hand, supervised and self-supervised embeddings achieved near 90% accuracy even with a simple one-layer time-distributed network. From this, we can conclude that the long-term contextual representations are good at capturing language-specific information. It is also observed that the self-supervised embeddings consistently performed better than their supervised counterparts. The CRNN classifier trained on self-supervised embeddings outperformed all the other systems with an accuracy of 93.89%. This observation is also reflected in the t-SNE visualizations of the utterance level embeddings extracted from the CRNN network shown in Figure 3. The self-supervised embeddings are better separable compared to the acoustic and supervised counterparts.

The effect of test utterance duration on the three features with the CRNN classifier is shown in Figure 4. The performance of the acoustic features declined for longer duration utterances. As acoustic features capture only short-term spectral characteristics, this decline may be attributed to longer pauses or influences in the utterances. However, this hypothesis requires a deeper analysis in the future. The performance of the supervised and self-supervised embeddings improved for longer duration utterances. On the shorter test utterances (2-3 s), the self-supervised embeddings achieved a 6% absolute improvement in the performance compared to the supervised embeddings. The consistent performance of self-supervised embeddings can be attributed to their ability in phonotactic constraints even from shorter duration utterance. Finally, the confusion matrix of the self-supervised embeddings is shown in Figure 5. The higher confusion between Hindi and Punjabi can be attributed to the close similarity between the two languages.

Table 1: *LID accuracy of different model and feature pairs*

Feature	Acoustic	Supervised	Self-supervised
Time-distributed	42.15	89.90	**91.67**
BLSTM	67.09	90.1	**93.66**
CRNN	70.69	90.73	**93.89**

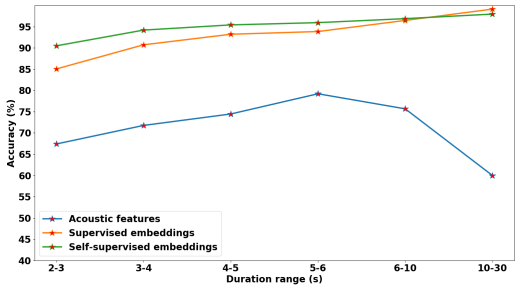


Figure 4: *Effect of duration on performance of different features*

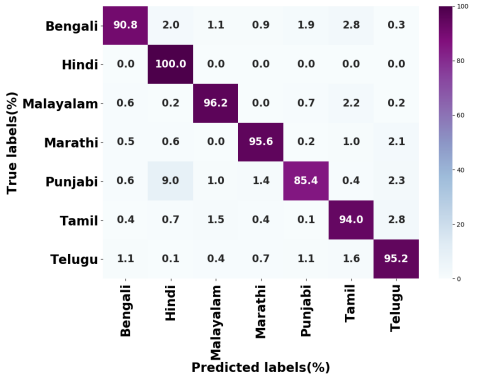


Figure 5: *Confusion matrix for self-supervised embeddings*

5. Summary & Conclusions

In this study, we established the language-specific nature of contextualized representations extracted from the wav2vec network. The self-supervised embeddings extracted from the wav2vec network performed better than their supervised counterparts. It is also observed the wav2vec network trained on a neutral language (American English) is good enough to extract language-specific phonotactic constraints for the target languages, i.e., the 7 Indian languages. However, in the future, it is interesting to study the effect of the training data used for building the wav2vec network, i.e., neutral language vs a combination of target languages. In the future, we will also conduct studies on quantifying the capacity of wav2vec embeddings, i.e., the number of target languages they can distinguish when trained with a neutral language.

6. References

[1] L. Mary and B. Yegnanarayana, "Autoassociative neural network models for language identification," in *International Conference on Intelligent Sensing and Information Processing, 2004. Proceedings of.* IEEE, 2004, pp. 317–320.

[2] M. A. Zissman and K. M. Berkling, "Automatic language identification," *speech communication*, vol. 35, no. 1-2, pp. 115–124, 2001.

[3] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[4] E. Wong and S. Sridharan, "Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing. ISIMP 2001 (IEEE Cat. No. 01EX489)*. IEEE, 2001, pp. 95–98.

[5] K. Vijayan, H. Li, H. Sun, and K. A. Lee, "On the importance of analytic phase of speech signals in spoken language recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5194–5198.

[6] F. Richardson, D. Reynolds, and N. Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE signal processing letters*, vol. 22, no. 10, pp. 1671–1675, 2015.

[7] S. Ganapathy, K. Han, S. Thomas, M. Omar, M. V. Segbroeck, and S. S. Narayanan, "Robust language identification using convolutional neural network features," in *Fifteenth annual conference of the international speech communication association*, 2014.

[8] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions." in *Odyssey*, 2014.

[9] Z. Tang, D. Wang, Y. Chen, L. Li, and A. Abel, "Phonetic temporal neural model for language identification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 1, pp. 134–144, 2017.

[10] J. L. Hieronymus and S. Kadambe, "Spoken language identification using large vocabulary speech recognition," in *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, vol. 3. IEEE, 1996, pp. 1780–1783.

[11] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[12] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[13] M. Ravanelli and Y. Bengio, "Learning speaker representations with mutual information," *arXiv preprint arXiv:1812.00271*, 2018.

[14] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[15] K. N. Stevens, *Acoustic phonetics*. MIT press, 2000, vol. 30.

[16] M. Goldrick, "Phonological features and phonotactic constraints in speech production," *Journal of Memory and Language*, vol. 51, no. 4, pp. 586–603, 2004.

[17] R. Futrell, A. Albright, P. Graff, and T. J. O'Donnell, "A generative model of phonotactics," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 73–86, 2017.

[18] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Transactions on speech and audio processing*, vol. 4, no. 1, p. 31, 1996.

[19] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 2010, pp. 297–304.

[20] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[21] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy modelling," in *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*, 1994.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[23] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Twelfth annual conference of the international speech communication association*, 2011.

[24] W. Geng, W. Wang, Y. Zhao, X. Cai, B. Xu, C. Xinyuan *et al.*, "End-to-end language identification using attention-based recurrent neural networks." in *Interspeech*, 2016, pp. 2944–2948.

[25] A. Lozano-Diez, R. Zazo-Candil, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "An end-to-end approach to language identification in short utterances using convolutional neural networks," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[26] W. Cai, D. Cai, S. Huang, and M. Li, "Utterance-level end-to-end language identification using attention-based cnn-blstm," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5991–5995.

[27] J. Gonzalez-Dominguez, I. Lopez-Moreno, H. Sak, J. Gonzalez-Rodriguez, and P. J. Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Fifteenth annual conference of the international speech communication association*, 2014.

[28] R. Zazo, A. Lozano-Diez, J. Gonzalez-Dominguez, D. T. Toledano, and J. Gonzalez-Rodriguez, "Language identification in short utterances using long short-term memory (lstm) recurrent neural networks," *PloS one*, vol. 11, no. 1, p. e0146917, 2016.

[29] C. C. Bhanja, D. Bisharad, and R. H. Laskar, "Deep residual networks for pre-classification based indian language identification," *Journal of Intelligent & Fuzzy Systems*, vol. 36, no. 3, pp. 2207–2218, 2019.

[30] R. K. Vuddagiri, H. K. Vydana, and A. K. Vuppala, "Improved language identification using stacked sdc features and residual neural network." in *SLTU*, 2018, pp. 210–214.

[31] C. Bartz, T. Herold, H. Yang, and C. Meinel, "Language identification using deep convolutional recurrent neural networks," in *International conference on neural information processing*. Springer, 2017, pp. 880–889.

[32] Y. Zhu, T. Ko, D. Snyder, B. Mak, and D. Povey, "Self-attentive speaker embeddings for text-independent speaker verification." in *Interspeech*, vol. 2018, 2018, pp. 3573–3577.

[33] B. Yin, "Language identification with language and feature dependency." Ph.D. dissertation, University of New South Wales, Sydney, Australia, 2009.