The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages
29-31 August 2018, Gurugram, India

# IIITH-ILSC Speech Database for Indian Language Identification

*Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, Anil Kumar Vuppala*

Speech Processing Laboratory, LTRC, KCIS
International Institute of Information Technology, Hyderabad, India.
{ravikumar.v, krishna.gurugubelli, priyam.jain, hari.vydana}@research.iiit.ac.in,
anil.vuppala@iiit.ac.in

## Abstract

This work focuses on the development of speech data comprising 23 Indian languages for developing language identification (LID) systems. Large data is a pre-requisite for developing state-of-the-art LID systems. With this motivation, the task of developing multilingual speech corpus for Indian languages has been initiated. This paper describes the composition of the data and the performances of various LID systems developed using this data. In this paper, Mel frequency cepstral feature representation is used for language identification. In this work, various state-of-the-art LID systems are developed using i-vectors, deep neural network (DNN) and deep neural network with attention (DNN-WA) models. The performance of the LID system is observed in terms of the equal error rate for i-vector, DNN and DNN-WA is 17.77%, 17.95%, and 15.18% respectively. Deep neural network with attention model shows a better performance over i-vector and DNN models.
**Index Terms**: Language identification system, Deep neural network, Deep neural network with attention, i-vectors.

## 1. Introduction

The objective of language identification (LID) systems is to recognize the language of the spoken utterance. In the recent past, there has been significant interest in developing spoken dialog systems. Due to similarities in the origin and overlapping phone-sets, developing LID system for Indian languages is a challenging task. Despite the similarities in phone-sets, Indian languages differ in their phonotactics, prosody, and intonations. Developing spoken dialog systems for Indian scenarios has instigated the interest in developing LID systems. Language identification systems are used in various speech technologies such as multilingual dialog systems and information-query systems. In multilingual automatic speech recognition (MASR) system, LID is used as a front-end to switch between multiple monolingual automatic speech recognition systems. When a LID system is used as a front-end switch of a dialog system, the phonotactic constraints of the language can aid the dialog systems to operate more robustly. Similarly, the voice-operated applications such as automatic call routing use LID system to recognize the language from the incoming call, and it will be routed to the desired application [1].

Language ID systems can be broadly classified into implicit LID systems and explicit LID systems. The implicit approaches directly use the acoustic level information to predict the language of the spoken utterance [2]. A detailed review of LID systems in the perspective of speech features and models is presented in [1]. Various implicit approaches for developing LID systems are described in [2]. The explicit LID systems initially transform the acoustic sequences to an intermediate representation such as phones, Senones or tokens and the temporal relations among them is used for developing LID systems. The LID systems using phonotactics, phone frequency, and syntax are described in [3, 4]. LID systems using parallel phone recognizers followed by language model (PPRLM) is developed by modeling the temporal relations among the phones decoded using a language-independent phone recognizer [1]. Language models like SRILM and RNNLM have been explored to model the temporal relations among the tokens for developing large-scale LID systems [4]. In this work, implicit techniques for language identification in Indian scenario are investigated.

Different Modeling techniques are explored in the context of LID systems. Gaussian mixture models (GMM) and Gaussian mixture models with universal background modeling (GMM-UBM) have been explored for developing LID systems [5, 6]. Deep neural network based LID systems manifested superior performance due to the availability of large-sized dataset [7, 8]. I-vectors convert the variable length sequences to fixed dimension sequence and capture the temporal context LID systems developed using i-vectors showed superior performance [9]. Performance of LID systems degrades for shorter utterances. In [7], i-vectors have been modified for developing LID systems to operate on shorter duration utterances. Multilingual bottleneck features have been used for developing LID systems [10]. In [11, 12], RNNs and CNNs which can process the whole utterance have been employed for building LID systems. Though sequential models such as LSTM have performed well, they are not parallelizable. Recently a feed-forward architecture has been proposed in [13, 14], where a self-attention the mechanism is used to convert the variable length sequence to a fixed dimension vector and the fixed dimension representation is used to discriminate the languages. The whole network can be trained as a single-framework through back-propagation. In this work, i-vector, DNN, and DNN-WA modeling techniques are explored in building LID systems for Indian languages.

In Indian languages identification scenario, different approaches for developing implicit LID system have been studied in [15]. LID systems for Indian languages are developed using spectral and prosodic features [16, 17, 18, 19]. Magnitude and phase spectrum of speech are used for developing LID system for Indian languages [20]. LID systems developed using source and systems have been explored in [21]. LID for Indian languages for varying background and mobile environments are explored in [22, 23]. Existing speech corpus used for Indian LID is limited in terms of both the number of speakers and duration of speech [16]. The state-of-the-art models for LID are data-driven, hence require larger data. Recently, LID challenges like Oriental Language Recognition(OLR) [24] and LRE


10.21437/SLTU.2018-12

released a large speech corpus, but these are limited to world languages and includes only a few Indian languages. This work focuses on developing a large speech corpus for Indian language identification system named as, International Institute of Information Technology Hyderabad-Indian language speech corpus (IIITH-ILSC).

The paper is organized as follows: Databases collection process and demographic details of the database are described in section 2. Section 3 describes the language identification models used in this study. Analysis and results of various LID systems are discussed in Section 4. Conclusion and future scope are presented in section 5.

## 2. International Institute of Information Technology Hyderabad-Indian language speech corpus (IIITH-ILSC)

India is a multilingual society with various spoken languages, India has 22 official languages (according to 2011 census survey of India). Indian languages can be broadly classified into three categories, i.e., Indo-Aryan, Dravidian, and Sino-Tibetan languages which are spoken by 76.5%, 20.5% and 3% of the total population respectively.

Table 1: *Comparision of IITKGP-MLISC and IIITH-ILSC speech databases*

| | IITKGP-MLISC | IIITH-ILSC |
|---|---|---|
| No. of Languages | 27 | 23 |
| Avg. No. of speakers per Language | 10 | 50 |
| Mode of speech | Read Speech | Read and conversational speech |
| Quality of speech | Studio qualities | Studio qualities and realistic |
| Noise | No background noise | Both clean and moderately noisey |
| Channel variation | No channel noise | Channel noise is due various sources |
| No. of hours per Language | 1 hour | 4.5 hours |
| Total no. Of speakers | 300 | 1150 |
| Total no. Hrs of speech | 27 hours | 103.5 hours |

The IIITH-ILSC covers 23 Indian languages which include speech data from 22 official Indian Languages and English. The languages in the IIITH-ILSC are Assamese, Bodo, Dogri, Gujarati, Hindi, Kannada, Kashmiri, Konkani, Maithili, Malayalam, Manipuri, Marathi, Nepali, Odia, Punjabi, Sanskrit, Santali, Sindhi, Tamil, Telugu, Urdu and Indian English. Except for English and Urdu, all Indian languages share the same phonetic space and have an overlapping phone sets. The speech corpus comprises of 103.5 hours of data collected from 1150 speakers. The data are pooled from different sources such as archives of Prasar Bharati, All Indian Radio, TED-talks, conversational speech from broadcasts and speech recorded from students of IIITH, University of Hyderabad, Maulana Azad National Urdu University-Hyderabad, National Institute of Technology-Warangal, Goa University, Bodoland University and the University of Jammu. In IIITH-ILSC, data for each language is collected from 50 speakers (25 male and 25 female) with 5-10 min duration from each speaker. The data volume collected for 4.5 hours of data from each language, 3.5 hours is used for training, and 1 hour of data is used for testing. The data is comprised of both native and non-native speakers with a speech quality ranging from studio quality to mobile quality. The speech data collected has 60% of clean speech and 40% of speech with moderately low noise including channel variations and other environmental noises. The speech corpus is code mixed with English and Hindi languages. Other details about the speech corpus are described in Table 1 in comparison with IITKGP-MLISC speech corpus. After collecting the data each speech file is chopped into 5-10 sec utterances using Audacity software. While chopping the record, the utterance with background music, non-speech sounds are avoided in the preparation of database. Most of the data which is collected from the internet and broadcasts are verified by the native speaker of corresponding languages. The speech corpus is pooled from various sources with files of different formats and sampling rates. For the uniformity, all files are converted into .wav format with a sampling rate of 8000 samples/sec using Audacity software.

## 3. Language identification models explored in this study

In this work, i-vector, DNN, and DNN-WA based LID system been developed, and these LID systems are briefly described in the following subsections. Spectral features are used for training an LID system. Performance of the LID system is presented in terms of equal error rate (EER).

### 3.1. i-vector

The State-of-the-art identity vector or i-vector based LID system was first introduced in [25] for speaker verification. With this inspiration, i-vector is applied for language identification [9]. A Universal Background Model (UBM) consisting of 2048 Gaussian components trained with 39-dimensional MFCC features. The Total Variability ($T$) matrix is derived from space of 400 dimensions. In this i-vector model, the $T$ is trained using a principal component analysis followed by 10 expectation maximization iterations. The i-vector can be represented in [9] as given in Equation. 1.

$$M_L = m_l + Tw_l \tag{1}$$

where $M_L$ is the super-vector created by stacking all the mean vectors from the GMM for a given utterance. The $m_l$ is the super-vector from the UBM model and $w_l$ refers to weight vector with a standard normal prior to the speech utterance.

### 3.2. Deep Neural Networks

Deep neural networks have been explored for developing LID systems [8]. In this work, DNNs are explored with different hidden layers each comprising of rectified linear units (ReLU). A softmax output layer with categorical-cross entropy loss function has been used. The network is optimized using ADAM and stochastic gradient descent(SGD) optimizer. A decrease in the validation accuracy in three successive epochs is considered as an early stopping criterion. Learning rate is halved whenever a minimum increase in validation accuracy is less than 0.5 between successive epochs.

### 3.3. Deep Neural Network with Attention

The decision on language is taken at every frame in the DNN based LID system, while language identification is on the ba-

sis of the whole utterance. Long-term temporal patterns of speech shown better performance in language identification. Whereas, DNNs failed in capturing the long-term temporal patterns. However, in this paper [26] the LID system can be modeled using long-short-term memory networks (LSTMs). But due to the sequential nature of these LSTM networks, they are slow in processing and are not parallelizable. The temporal context is captured by DNN-WA network operating in a feed-forward architecture.

Figure. 1 shows the DNN with attention architecture, which

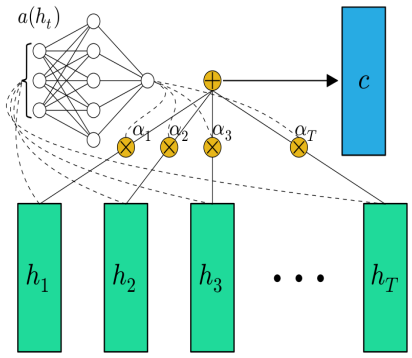


Figure 1: *Deep Neural Network with attention model [14]*

is a simple DNN implemented with attention mechanism. In DNN with attention mechanism, the adaptation was made such as the attention is computed just by adapting the input feature vectors as in contrast to adapting both input and output feature vectors [27]. The DNN-WA model described in [13] is implemented for Indian language identification. The network comprises of 4 hidden layers with each layer of ReLU activation functions. The attention layer comprises of a multilayer perceptron (MLP) with a single hidden layer. The entire network is trained end-to-end with categorical entropy function. The network is trained with stochastic gradient descent optimizer with different learning rates. The network is trained with variable batch size. The batch size is the same as the length of an acoustic sequence. The output function for the hidden layer is given by

$$H_T = f(x_t, h_t) \tag{2}$$

where $x_t$ is sequence of input feature vectors $\{x_1, x_2, ...., x_T\}$ and the sequence of hidden state vectors is $\{h_1, h_2, ....h_T\}$. The output of hidden layer $H_T$, is computed by forward pass through regular DNN and a self-attention is computed on these hidden features. Where $a(h_t)$ is computed using a single layer perceptron. The output layer, a softmax operation is performed to normalize the values between zero and one.

$$\beta_t = tanh\left(W_{wa} h_t + b_{wa}\right) \tag{3}$$

$$\alpha_t = softmax\left(\beta_t\right) \tag{4}$$

In the equation 4, $\alpha_t$ is referred to as attention vector, with hidden weights $W_{wa}, b_{wa}$. The entire network is optimized along with other parameters of using the backpropagation algorithm. The attention-based model computes a "context vector " $C_T$

$$C_T = \sum_{t=1}^{T} \alpha_t H_T \tag{5}$$

where $C_T$ is context weighted mean, and T is a total number of time steps in the input sequence. The output of DNN-WA is computed by transforming the context vector $C_T$ using output layer weight $V_l$ followed by softmax operation

$$y_o = softmax\left(V_l C_T + b_o\right) \tag{6}$$

where $b_o$is the output layer bias. Note that for the entire input utterance $x_t$, only a single decision vector $y_o$ is predicted.

## 4. Experiments and Results

In the development of baseline language identification systems for IIITH-ILSC, various state-of-the-art modeling techniques such as i-vector, DNNs, and DNN-WA are explored. As a part of this, 39-dimension MFCC feature ($13static + \nabla + \nabla\nabla$ ) is used. In extracting the MFCC features, voice activity detector is used to remove the silence in speech. Then MFCC features are extracted from speech with a 20 msec frame size and a 10 msec overlap. To minimize the channel variances cepstral mean-variance normalization is incorporated after extracting the MFCCs. The architecture details of DNN and DNN-WA: the input layer fed with 39-dimensions MFCC coefficients and the output layer is soft-max with 23 (Number of languages) units are connected. Each hidden layers uses rectified linear units (ReLU/R) as an activation function. To minimize the loss on training set cross-entropy cost function is used. The experiments are done using different hyperparameters and the results are demonstrated in Table. 2, Table. 3 and Table. 4.

The performance of DNN based LID systems using different

Table 2: *Performance of LID systems developed using different optimizers such as ADAM and SGD for DNN*

| Deep Neural Network | | | | |
|---|---|---|---|---|
| | ADAM | | SGD | |
| Learning rates($\eta$) | 0.001 | 0.0001 | 0.001 | 0.01 |
| 2H | 22.45 | 27.45 | 19.18 | 21.54 |
| 4H | 20.16 | 25.15 | **17.95** | 18.96 |
| 6H | 21.14 | 25.95 | 20.01 | 20.77 |

hyperparameters has been presented in Table. 2. In this study, DNN trained with 2, 4, 6 hidden layers with 1024 rectified linear units (ReLU) in each hidden layer. The different learning rates with ADAM and stochastic gradient descent(SGD) network optimizers are used. It has been observed that the performance of LID systems developed using DNNs is optimal using a depth of 4 hidden layers. Further, in the study DNNs of 4 hidden layers have been used for developing LID systems.

The performance of DNN-WA based LID systems using differ-

Table 3: *Performance of LID systems developed using different optimizers such as ADAM and SGD for DNN-WA*

| Deep neural network with attention | | | | |
|---|---|---|---|---|
| | ADAM | | SGD | |
| Learning rates($\eta$) | 0.001 | 0.0001 | 0.0001 | 0.001 |
| 700R500R200R100R | 17.82 | 15.92 | 18.86 | 19.71 |
| 100R200R500R700R | 16.21 | **15.18** | 19.23 | 17.50 |

ent hyperparameters has been presented in Table. 3. DNN-WA

trained using a mini-batch stochastic gradient descent (SGD) with classical momentum and ADAM optimizer. The hyper-parameters are adjusted using a validation dataset. The utterances are randomized in the DNN-WA, but not the frames within the utterance. The mini-batch size is equal to the length of the utterance given as input to the network. The DNN-WA has been studied with different models such as architecture with an equal number of ReLU units (1024R1024R1024R1024R), architecture with increased ReLU units (700R500R200R100R), and architecture with decreased ReLU units in each hidden layer (100R200R500R700R) respectively. It has been observed that the performance of LID systems developed using DNN-WA is optimal using 15.15% of EER.

Table 4: *Performance of LID systems developed using different models i-vector, DNN and DNN-WA*

| Language | i-vector | DNN | DNN-WA |
|---|---|---|---|
| Assamese | 11.50 | 12.65 | 8.23 |
| Bengali | 25.54 | 22.27 | 14.89 |
| Bodo | 4.14 | 2.05 | 6.80 |
| Dogri | 15.24 | 21.98 | 11.35 |
| Gujarati | 6.34 | 14.08 | 20.97 |
| Hindi | 19.22 | 21.19 | 14.20 |
| Indian English | 21.52 | 29.72 | 18.60 |
| Kannada | 15.73 | 15.31 | 16.33 |
| Kashmiri | 21.00 | 13.16 | 26.40 |
| Konkani | 14.52 | 20.87 | 10.11 |
| Maithili | 32.30 | 37.28 | 12.61 |
| Manipuri | 15.29 | 13.35 | 12.58 |
| Malyalam | 16.50 | 18.46 | 21.43 |
| Marathi | 12.01 | 18.65 | 6.63 |
| Nepali | 31.14 | 21.29 | 24.71 |
| Oriya | 6.79 | 4.27 | 5.94 |
| Punjabi | 24.00 | 27.55 | 20.58 |
| Sanskrit | 21.00 | 19.73 | 19.90 |
| Santali | 24.50 | 24.90 | 13.98 |
| Sindhi | 28.50 | 26.29 | 23.28 |
| Tamil | 15.00 | 18.67 | 12.54 |
| Telugu | 8.91 | 2.89 | 5.01 |
| Urdu | 8.00 | 6.15 | 12.18 |
| Average | **17.77** | **17.95** | **15.18** |

The baseline LID results of IIITH-ILSC are presented in Table. 4. The baseline results are compared over different modeling techniques such as i-vector, DNN, and DNN-WA. The performance of baseline LID systems is represented in terms of equal error rate for all individual languages an average for the same is computed. The lower the EER is, the better is the systems performance. From the average performances, it can be noticed that DNN-WA shows superior performance, 15.92% of EER over i-vector and DNN with 17.77% and 17.95% EERs respectively.

## 5. Summary and Conclusions

The similarity in origin and overlapping phone sets pose a significant challenge in developing LID system in the Indian scenario. This motivated us to develop LID systems for Indian languages. In this work, a speech corpus namely International Institute of Information Technology Hyderabad Indian language speech corpus (IIITH-ILSC) is developed. IIITH-ILSC covers 23 Indian languages speech data which includes 22 official Indian languages and English. At present, IIITH-ILSC is one of the largest multilingual speech corpora in terms of both the number of speakers and hours of speech. In developing LID system for Indian languages, i-vectors, DNN, and DNN-WA models are explored. A 39 dimension($13static + \nabla + \nabla\nabla$ ) Mel frequency cepstral features representation of speech is used. The performance of LID systems with different architectures are compared using a metric equal error rate. The LID system using DNN-WA architecture exhibited better performance than the LID systems developed using i-vector and DNN models. The LID system with DNN-WA for Indian languages shows an equal error rate of 15.18%.

The performance of the baseline system can be further improved by considering more sophisticated language specific features from the speech and modeling techniques. Hybrid models are a choice to capture better language-specific information for LID system. Robust approaches for LID systems to model both short and long-term temporal patterns can be explored. Performances of LID systems can be explored for noisy environments.

## 6. Acknowledgements

The authors would like to thank Science & Engineering Research Board (SERB) for funding Language Identification in Practical Environments (YSS/2014/000933) project.

## 7. References

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

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

[3] M. A. Zissman and E. Singer, "Automatic language identification of telephone speech messages using phoneme recognition and n-gram modeling," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1. IEEE, 1994, pp. I–305.

[4] B. M. L. Srivastava, H. Vydana, A. K. Vuppala, and M. Shrivastava, "Significance of neural phonotactic models for large-scale spoken language identification," in *Proc. Int. Joint Conf. Neural Networks*. IEEE, 2017, pp. 2144–2151.

[5] P. A. Torres-Carrasquillo, D. A. Reynolds, and J. R. Deller, "Language identification using Gaussian mixture model tokenization," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*, vol. 1. IEEE, 2002, pp. I–757.

[6] E. Wong and S. Sridharan, "Methods to improve Gaussian mixture model based language identification system," in *Proc. Seventh International Conference on Spoken Language Processing*, 2002.

[7] F. Richardson, D. Reynolds, and N. Dehak, "A unified deep neural network for speaker and language recognition," *arXiv preprint arXiv:1504.00923*, 2015.

[8] I. Lopez-Moreno, J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez, and P. Moreno, "Automatic language identification using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing*. IEEE, 2014, pp. 5337–5341.

[9] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. INTERSPEECH*, 2011.

[10] R. Fér, P. Matějka, F. Grézl, O. Plchot, and J. Černockỳ, "Multilingual bottleneck features for language recognition," in *Proc. Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[11] Y. Lei, L. Ferrer, A. Lawson, M. McLaren, and N. Scheffer, "Application of convolutional neural networks to language identification in noisy conditions," *in Proc. Odyssey-14, Joensuu, Finland*, 2014.

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

[13] K. V. Mounika, S. Achanta, H. Lakshmi, S. V. Gangashetty, and A. K. Vuppala, "An investigation of deep neural network architectures for language recognition in Indian languages." in *Proc. INTERSPEECH*, 2016, pp. 2930–2933.

[14] C. Raffel and D. P. Ellis, "Feed-forward networks with attention can solve some long-term memory problems," *arXiv preprint arXiv:1512.08756*, 2015.

[15] T. Nagarajan, "Implicit systems for spoken language identification," Ph.D. dissertation, IIT, MADRAS, 2004.

[16] S. Maity, A. K. Vuppala, K. S. Rao, and D. Nandi, "IITKGP-MLILSC speech database for language identification," in *Communications (NCC), 2012 National Conference on*. IEEE, 2012, pp. 1–5.

[17] K. S. Rao, S. Maity, and V. R. Reddy, "Pitch synchronous and glottal closure based speech analysis for language recognition," *International Journal of Speech Technology*, vol. 16, no. 4, pp. 413–430, 2013.

[18] L. Mary and B. Yegnanarayana, "Extraction and representation of prosodic features for language and speaker recognition," *Speech communication*, vol. 50, no. 10, pp. 782–796, 2008.

[19] V. R. Reddy, S. Maity, and K. S. Rao, "Identification of Indian languages using multi-level spectral and prosodic features," *International Journal of Speech Technology*, vol. 16, no. 4, pp. 489–511, 2013.

[20] D. Nandi, D. Pati, and K. S. Rao, "Language identification using hilbert envelope and phase information of linear prediction residual," in *Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), 2013 International Conference*. IEEE, 2013, pp. 1–6.

[21] K. V. Mounika, V. Ravi Kumar, S. V. Gangashetty, and A. K. Vuppala, "Combining evidences from excitation source and vocal tract system features for Indian language identification using deep neural networks," *International Journal of Speech Technology*, pp. 1–8, 2017.

[22] V. Ravi Kumar, H. K. Vydana, J. V. Bhupathiraju, S. V. Gangashetty, and A. K. Vuppala, "Improved language identification in presence of speech coding," in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2015, pp. 312–322.

[23] V. Ravi Kumar, H. K. Vydana, and A. K. Vuppala, "Curriculum learning based approach for noise robust language identification using DNN with attention," *Expert Systems with Applications*, 2018. [Online]. Available: https://doi.org/10.1016/j.eswa.2018.06.004

[24] Z. Tang, D. Wang, Y. Chen, and Q. Chen, "AP17-OLR challenge: Data, plan, and baseline," *arXiv preprint arXiv:1706.09742*, 2017.

[25] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[26] 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 *Proc. Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[27] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.