Study on the Effect of Emotional Speech on Language Identification

by

Priyam Jain, Krishna Gurugubelli, Anil Kumar Vuppala

in

NCC

Report No: IIIT/TR/2020/-1

Centre for Language Technologies Research Centre
International Institute of Information Technology
Hyderabad - 500 032, INDIA
February 2020

Study on the Effect of Emotional Speech on Language Identification

Priyam Jain
Speech Processing Laboratory
LTRC, KCIS, IIIT-H
Hyderabad, India
priyam.jain@research.iiit.ac.in

Krishna Gurugubelli
Speech Processing Laboratory
LTRC, KCIS, IIIT-H
Hyderabad, India
krishna.gurugubelli@research.iiit.ac.in

Anil Kumar Vuppala
Speech Processing Laboratory
LTRC, KCIS, IIIT-H
Hyderabad, India
anil.vuppala@iiit.ac.in

***Abstract*—Identifying language information from speech utterance is referred to as spoken language identification. Language Identification (LID) is essential in multilingual speech systems. The performance of LID systems have been studied for various adverse conditions such as background noise, telephonic channel, short utterances, so on. In contrast to these studies, for the first time in the literature, the present work investigated the impact of emotional speech on language identification. In this work, different emotional speech databases have been pooled to create the experimental setup. Additionally, state-of-art i-vectors, time-delay neural networks, long short term memory, and deep neural network x-vector systems have been considered to build the LID systems. Performance of the LID system has been evaluated for speech utterances of different emotions in terms of equal error rate and C_{avg}. The results of the study indicate that the speech utterances of anger and happy emotions degrades performance of LID systems more compared to the neutral and sad emotions.**

***Index Terms*—Language Identification, i-vector, TDNN, LSTM, DNN x-vector**

I. Introduction

Spoken language identification is the process of determining in which language the speech is pronounced. There are thousands of languages and hundreds of language families in the world. This leads to diversity in the human population. Due to globalisation and interaction of human population from different linguistic backgrounds, multilingual systems become a necessity [1]. Hence, Language Identification (LID) plays a key role in many applications, including multilingual speech recognition [2], multilingual communication systems [3], spoken document retrieval [4], and spoken language translation [5]. This has led to an increased amount of interest in the area of LID in recent years. We have seen challenges such as Language Recognition Evaluation (LRE) [6] and Oriental Language Recognition (OLR) [7] getting good attention and hence providing improved solutions to the society. Moreover, the advancement in computation power and machine learning, coupled with the signal processing knowledge has enabled rapid improvement in LID [8].

From literature, it is understood that i-vectors have demonstrated good performance for both LID and speaker recognition [9], [10]. Due to the ability to model sequential information from input, Long Short Term Memory (LSTM) networks have shown considerable performance gains over deep neural networks (DNN) [11]. In recent years, bidirectional LSTM (b-LSTM) networks have outperformed traditional LSTM, and when combined with the attention mechanism, result in powerful temporal modelling systems [12]. There has also been work on fusion of i-vectors with the attention based b-LSTM for language identification [13]. Time Delay Neural Network (TDNN) is a variation of DNN, which imposes a different temporal resolution at each layer and it increases from the first layer as we go deeper [14]. It has been used as an acoustic modelling technique in automatic speech recognition [15] as well as for LID [16]. DNN x-vector has been introduced recently for speaker and language recognition tasks [17], [18] and have given promising results.

From the studies in [19]–[21], state-of-art systems for LID have shown degradation in performance due to the variations such as short duration speech and multiple dialects of a language family among many more. There have been some attempts to see the effectiveness of LID systems for noise and telephone quality speech [22]–[25]. From the studies in [21], [26], [27], state-of-art LID systems have been shown to provide better performance on speech data collected in a neutral emotion state. However, there are not many attempts which study the performance of LID systems in the context of different emotional states. Recent studies on the impact of speakers' emotional state on speaker verification [28], speaker identification [29], and automatic speech recognition [30], [31] indicate that the variation in acoustic features due to different emotional states cause the degradation in the performance of these systems. All these together motivated us to investigate the impact of the speakers' emotional state on the performance of state-of-the-art LID systems.

The rest of this paper is organised as follows : Section II outlines and gives the detailed description of the modelling techniques that have given good results in recent times and are used in this work, Section III outlines the database and the experimental setup, Section IV reports the results obtained and Section V presents the summary and conclusion.

II. Baseline systems

A. UBM/GMM i-vector with LDA/PLDA reduction

I-vectors are used to learn fixed dimensional embeddings from spoken data [9]. They are low-rank vectors, representing the characteristics of the speaker and linguistic information in an utterance. The Gaussian Mixture Model (GMM) and the Universal Background Model (UBM) are modelled using Mel-frequency cepstral coefficients (MFCC), which are extracted from speech utterances. All GMM and UBM mean vectors are stacked together to obtain super vectors. The GMM supervector is then modelled as:

$$M = m + Tw + \epsilon. \tag{1}$$

Here M and m are GMM and UBM supervectors respectively, T is the total variability matrix [32], w is the i-vector having normal distribution and ϵ is the residual noise.

Furthermore, dimensionality reduction techniques such as Linear Discriminant Analysis (LDA) and Probabilistic Linear Discriminant Analysis (PLDA) are used to maximise the intra-class variance while minimising inter-class variance between features vectors [33], [34]. These techniques are applied to the i-vectors before scoring.

B. Time Delay Neural Network

Unlike traditional DNNs, TDNN learns long-term dependencies as the network becomes deep. Layers in the DNN are activated from the full temporal context of the previous layer, whereas in TDNN, each neuron is activated by a narrow context, coming from the previous layer, and gets accumulated as the network goes deep. Furthermore, the initial layers in a TDNN also learn from the whole temporal context due to back-propagation of the gradients. Thus lower layer activation become translation-invariant [14].

C. Long Short Term Memory

LSTM network is a special class of recurrent neural networks, used for sequential modelling tasks. The activation at each time-step is a mapping learnt from the input as well as previous time-steps. Each LSTM cell has three gates associated with it: input gate (for activation of input), output gate (for activating the output) and forget gate (for manipulating the cell state). In practice, several layers with each containing multiple LSTM cells are stacked together to form a LSTM network. Generally, the activation at the last time-step of the last state is considered for classification tasks, as it has been mapped from the complete temporal context [11].

D. DNN x-vector

Similar to i-vectors, x-vectors learnt using a DNN architecture are fixed dimension embeddings for variable length spoken utterances. Several hidden layers are stacked one after the other as in a traditional DNN, but the layers are not fully connected rather are more similar to that of a TDNN. Hence each layer is activated with a narrow temporal context of the previous layer along with its own temporal context, so it keeps getting accumulated, hence the deeper layers get a wide temporal context.

After the dense layers, a stat pooling layer is used to calculate mean and standard deviation of the previous layer output across all time-steps. Hence the stat pooling layer makes sure of the fixed dimensional embedding for variable duration utterances. Generally, we use the concatenated mean and standard deviation vectors from the stat pooling layer as input to subsequent dense layers. Embeddings can be extracted as the output of these subsequent dense layers.

III. Experimental Setup

In this work, we have used the kaldi toolkit [35] end to end, from feature extraction to classification models.

A. Database

In this work, the experiments have been performed by aggregating various datasets for the following 7 languages : Basque, English, German, Hindi, Serbian, Spanish and Telugu. For train data of English, Hindi and Telugu IIIT-ILSC dataset [36] has been used. German train data has been taken from [37], while Spanish and Basque train data has been taken from Mozilla Common Voice Project[1]. Emotional datasets for Basque, English, German, Hindi, and Telugu are taken from [38]–[42]. The Spanish[2], and Serbian[3] are collected from European language resources association. In this work, four basic emotions anger, happy, sad and neutral are considered.

In total 6 sets are prepared: training, test, test_angry, test_happy, test_neutral and test_sad with each containing utterances from the 7 considered languages. A part of neutral data from these emotion labelled datasets is also used in training to minimize the variance due to channel and recording environments. The dataset for train and test are split in such a way that no speaker or linguistic information is shared. Table I describes the database in terms of number of utterances and duration.

B. Feature Extraction

This work uses two most common and well known features as input to the classifiers: 40 dimensional (40D) filter bank (fbank) and 39 dimensional (39D) Mel Frequency Cepstral Coefficients (MFCC). The cepstral mean variance normalisation and energy based Voice Activity Detection (VAD) have been applied to MFCC features to normalize and remove the features corresponding to non speech frames. For x-vector extraction, the VAD output of MFCC is adapted to remove non speech frames from the fbanks as well.

C. I-vector with LDA/PLDA

For i-vector extraction, this work use 39D MFCC features to train the UBM/GMM. 1800 utterances are selected from the train data to train the UBM, and then 3600 utterances are selected to train the GMM using the Expectation Maximization technique. 400 dimensional i-vectors are extracted using 2048 Gaussian mixtures. LDA and PLDA transforms are applied, where LDA reduces the dimensions of the i-vectors to 150.

As noted in [43], cosine scoring provides similar performance as SVM for i-vectors. Hence we use cosine scoring

[1] http://voice.mozilla.org/

[2] http://www.islrn.org/resources/477-238-467-792-9/

[3] http://www.islrn.org/resources/462-780-920-598-3/

TABLE I
DETAILS OF THE DATABASE USED FOR THE STUDY

	Basque	English	Hindi	German	Serbian	Spanish	Telugu	Total Duration	Total Utterances
Train	1014	960	2380	1208	292	780	1520	12.9	8154
Test	309	540	650	588	36	285	645	5.26	3053
Test Neutral	140	50	150	90	36	85	150	0.71	701
Test Angry	140	50	150	110	24	76	150	0.66	700
Test Happy	140	50	150	70	24	84	150	0.7	668
Test Sad	140	50	150	70	24	79	150	0.78	663

to evaluate the performance of the i-vectors. Cosine scoring is done by taking the cosine distance between the test vector and the mean vector of target language from the train set.

D. TDNN and LSTM

Input to both TDNN and LSTM are 40D fbanks, described in Section III-B. For TDNN, we are using 6 hidden layer architecture similar to [7]. The temporal context of the layers are [-2,2], [-1,1], [-1,1] and {-6,-3,0}. No splicing is used for the second and last hidden layers, which has been decided empirically. The first layer learns an affine transformation. Each layers contains 650 units with Relu activation.

In this work LSTM network uses the first layer as the affine transformation layer with similar temporal context as the first layer of TDNN architecture. Then a stacked LSTM layer with 512 cells is used. After the sixth layer of TDNN and the LSTM layer in the LSTM network, an output layer with size equal to the number of classes is used in this work.

E. DNN x-vector

The architecture for extracting x-vector embeddings is similar to [18], which is demonstrated in Fig. 1. The DNN x-vector extraction uses 40D fbank features, as described in Section III-B. In Fig. 1, F (40) is the input feature dimension, T (depends on the utterance length) is number of frames and C (7) is the number of classes. The x-vector embeddings are taken as the output of segment6, which implies that we are considering 512 dimensional embeddings. These embeddings can be scored in the similar way as i-vectors, but in this work, we are scoring them directly from the softmax layer.

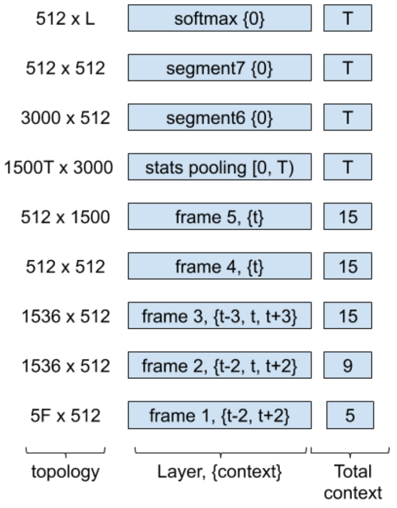

Fig. 1. DNN x-vector architecture used for extracting embeddings

F. Evaluation Metrics

In this work, Equal Error Rate (EER) and C_{avg} are used as metrics for evaluating the performance of LID systems. As these have been the primary metrics for evaluating LID systems in recent OLR and LRE challenges [6], [7]. The calculation of C_{avg} uses a pairwise loss between target and non-target classes. The pairwise loss is obtained as:

$$C(L_t, L_n) = P_{Target} P_{Miss}(L_t) + (1 - P_{Target}) P_{FA}(L_t, L_n) \tag{2}$$

where L_t and L_n are target and non-target languages, respectively and P_{Target} is the prior probability of the target language. During evaluation, the value of P_{Target} is set to 0.5. Further, the C_{avg} is defined as average of this pairwise loss and is given by:

$$C_{avg} = \frac{1}{N} \sum_{L_t} \sum_{L_n} C(L_t, L_n) \tag{3}$$

where N is the number of languages.

IV. RESULTS AND DISCUSSION

This study aims to investigate the effect of emotional state present in the speech utterance, on the performance of a LID

TABLE II
PERFORMANCE (IN TERMS OF EER AND C_{avg}) COMPARISON BETWEEN THE STATE-OF-THE-ART SYSTEMS FOR LANGUAGE IDENTIFICATION FOR THE SPEECH UTTERANCE OF MATCHED AND MISMATCHED CONDITIONS DUE TO EMOTIONAL STATES.

	Matched		Mismatched	
	EER	C_{avg}	EER	C_{avg}
i-vector	15.29	0.15	23.54	0.25
i-vector + LDA	13.13	0.13	19.45	0.23
i-vector + PLDA	11.17	0.12	17.77	0.22
TDNN	17.38	0.18	16.69	0.2
LSTM	18.10	0.27	23.69	0.28
DNN x-vector	8.12	0.08	17.18	0.22

TABLE III
PERFORMANCE (IN TERMS OF EER AND C_{avg}) OF THE STATE-OF-THE-ART SYSTEMS FOR LANGUAGE IDENTIFICATION FOR THE SPEECH UTTERANCES OF DIFFERENT EMOTIONAL STATES.

	i-vector		i-vector + LDA		i-vector + PLDA		TDNN		LSTM		DNN x-vector	
	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}	EER	C_{avg}
Test Neutral	22.11	0.21	17.97	0.19	15.41	0.17	7.13	0.12	17.26	0.20	8.42	0.10
Test Anger	25.00	0.26	22.86	0.26	22.00	0.25	22.14	0.29	23.86	0.39	19.29	0.24
Test Happy	24.10	0.27	20.51	0.26	18.41	0.24	20.96	0.24	27.10	0.31	21.41	0.25
Test Sad	21.42	0.21	16.89	0.18	14.03	0.17	5.9	0.22	19.91	0.22	12.22	0.15

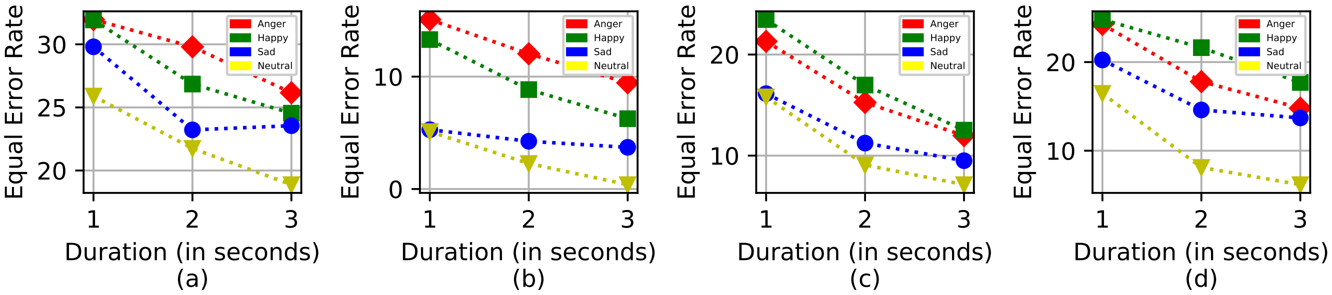

Fig. 2. Effect of speech utterance duration of different emotions, on the performance of state-of-the-art language identification systems. (a) i-vector + PLDA, (b) TDNN, (c) LSTM, and (d) DNN x-vector.

system. The main obstruction to such studies is the lack of a standard database, in which case we have designed a database (discussed in Section III (A)) to study the performance of LID systems in the context of various emotional states such as neutral, anger, happiness, and sadness. In this regard, state-of-the-art systems such as i-vector, TDNN, LSTM and DNN x-vector architectures have been studied to determine how emotional states in speech affect the performance of LID systems. The LID systems have been trained using the speech utterances of neutral emotional state. In this work, evaluation of LID systems with speech utterances of a neutral emotional state is referred to as a matched test condition, and the evaluation of LID systems with speech utterances of angry, happy, and sad emotional state is referred to as mismatched test condition. The results have been reported in Table II, Table III and Figure 2.

Table II shows the performance comparison between state-of-art LID systems for matched and mismatched (which is an aggregated set of speech utterances of anger, happy and sad emotional states) testing conditions. For matched test data DNN x-vectors shows best performance and i-vector with PLDA system shows comparable results in terms of EER and C_{avg}. The LSTM based LID system shows poor performance in both matched and mismatched test conditions. For mismatched test data, TDNN shows best performance, and the DNN x-vector and i-vector PLDA systems show comparable results. However, the results suggest that there is a degradation in performance of all LID systems except TDNN for mismatch test data. Table III shows the performance of LID systems for speech utterances of different emotional states. The results show that the performance of LID systems for speech of sad emotional state is comparable to that of neutral emotional state. The effect of sadness on the performance of LID systems is less compared to anger and happiness. Though the DNN x-vector system showed the best performance under matched conditions, it shows degradation in performance under mismatched emotional conditions.

Figure 2 shows the performance under 4 emotions for different test utterance durations. Each utterance, from all 4 emotions, is split into 1 second, 2 second and 3 seconds utterances to study the effect of duration on the trained LID systems. For a typical LID system, the duration v/s EER would be a monotonically decreasing curve. Similar behavior can be observed, from Figure 2, for the mismatched conditions discussed in this paper.

V. SUMMARY AND CONCLUSION

This paper studied the effect of emotional speech on the performance of state-of-art systems for language identification (LID). To accomplish this, database was pooled from various sources. Additionally, this study used the state-of-art systems such as i-vector, TDNN, LSTM, and DNN x-vector architectures to built the LID systems. The results shows the similarity in performance for neutral-sad and happy-angry emotion pairs. Further, for high arousal speech (anger and happy) the LID systems shows poor performance. On the other hand, the LID systems showed better performance for low arousal speech (neutral and sad). Though the linguistic information in speech utterance is not varying with the speakers' emotional state, the state-of-art systems for LID shows poor performance for high arousal speech. In this case, normalization of acoustic features may improve the performance of LID systems for high arousal speech. Hence, in future works, we intend to investigate different adaptation methods to improve the performance of LID systems for the mismatch due to speakers' emotional state.

References

[1] Alex Waibel, Petra Geutner, L Mayfield Tomokiyo, Tanja Schultz, and Monika Woszczyna, "Multilinguality in speech and spoken language systems," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1297–1313, 2000.

[2] Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao, "Multilingual speech recognition with a single end-to-end model," in *Proc. ICASSP*. IEEE, 2018, pp. 4904–4908.

[3] Arbi Haza Nasution, Nesi Syafitri, Panji Rachmat Setiawan, and Des Suryani, "Pivot-based hybrid machine translation to support multilingual communication," in *Proc. International Conference on Culture and Computing*. IEEE, 2017, pp. 147–148.

[4] Zheng-Yu Wu, Li-Phen Yen, and Kuan-Yu Chen, "Generating pseudo-relevant representations for spoken document retrieval," in *Proc. ICASSP*. IEEE, 2019, pp. 7370–7374.

[5] Albert Haque, Michelle Guo, and Prateek Verma, "Conditional end-to-end audio transforms," *arXiv preprint arXiv:1804.00047*, 2018.

[6] Seyed Omid Sadjadi, Timothee Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, "The 2017 NIST language recognition evaluation," in *Proc. Odyssey*, 2018, pp. 82–89.

[7] Zhiyuan Tang, Dong Wang, and Qing Chen, "AP18-OLR challenge: Three tasks and their baselines," in *Proc. APSIPA ASC*. IEEE, 2018, pp. 596–600.

[8] Seyed Omid Sadjadi, Timothee Kheyrkhah, Craig S Greenberg, Elliot Singer, Douglas A Reynolds, Lisa P Mason, and Jaime Hernandez-Cordero, "Performance analysis of the 2017 NIST language recognition evaluation," in *Proc. INTERSPEECH*, 2018, pp. 1798–1802.

[9] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds, and Reda Dehak, "Language recognition via i-vectors and dimensionality reduction," in *Proc. INTERSPEECH*, 2011, pp. 857–860.

[10] Noor Salwani Ibrahim and Dzati Athiar Ramli, "I-vector extraction for speaker recognition based on dimensionality reduction," *Procedia Computer Science*, vol. 126, pp. 1534–1540, 2018.

[11] Javier Gonzalez-Dominguez, Ignacio Lopez-Moreno, Haşim Sak, Joaquin Gonzalez-Rodriguez, and Pedro J Moreno, "Automatic language identification using long short-term memory recurrent neural networks," in *Proc. INTERSPEECH*, 2014, pp. 2155–2159.

[12] Bharat Padi, Anand Mohan, and Sriram Ganapathy, "End-to-end language recognition using attention based hierarchical gated recurrent unit models," in *Proc. ICASSP*. IEEE, 2019, pp. 5966–5970.

[13] Bharat Padi, Anand Mohan, and Sriram Ganapathy, "Attention based hybrid i-vector blstm model for language recognition," *Proc. INTERSPEECH 2019*, pp. 1263–1267, 2019.

[14] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. INTERSPEECH*, 2015, pp. 3214–3218.

[15] Noor Fathima, Tanvina Patel, C Mahima, and Anuroop Iyengar, "TDNN-based multilingual speech recognition system for low resource Indian languages," in *Proc. INTERSPEECH*, 2018, pp. 3197–3201.

[16] Tirusha Mandava and Anil Kumar Vuppala, "Attention based residual-time delay neural network for Indian language identification," in *Proc. Twelfth International Conference on Contemporary Computing (IC3)*. IEEE, 2019, pp. 1–5.

[17] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5329–5333.

[18] David Snyder, Daniel Garcia-Romero, Alan McCree, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "Spoken language recognition using x-vectors," in *Proc. Odyssey*, 2018, pp. 105–111.

[19] Eliathamby Ambikairajah, Haizhou Li, Liang Wang, Bo Yin, and Vidhyasaharan Sethu, "Language identification: A tutorial," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 82–108, 2011.

[20] Meng-Ge Wang, Yan Song, Bing Jiang, Li-Rong Dai, and Ian McLoughlin, "Exemplar based language recognition method for short-duration speech segments," in *Proc. ICASSP*. IEEE, 2013, pp. 7354–7358.

[21] Ruchir Travadi, Maarten Van Segbroeck, and Shrikanth S Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *Proc. INTERSPEECH*, 2014, pp. 3037–3041.

[22] Seyed Omid Sadjadi and John HL Hansen, "Mean Hilbert envelope coefficients (mhec) for robust speaker and language identification," *Speech Communication*, vol. 72, pp. 138–148, 2015.

[23] Ravi Kumar Vuddagiri, Hari Krishna Vydana, and Anil Kumar Vuppala, "Curriculum learning based approach for noise robust language identification using dnn with attention," *Expert Systems with Applications*, vol. 110, pp. 290–297, 2018.

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

[25] Yeshwant M, Kay B, T Arai, Ronald Cole, and Etienne Barnard, "A comparison of approaches to automatic language identification using telephone speech," in *Proc. Third European Conference on Speech Communication and Technology*, 1993, pp. 1307–1310.

[26] Bharat Padi, Anand Mohan, and Sriram Ganapathy, "Attention Based Hybrid i-Vector BLSTM Model for Language Recognition," in *Proc. INTERSPEECH*, 2019, pp. 1263–1267.

[27] Xiaoxiao Miao, Ian McLoughlin, and Yonghong Yan, "A New Time-Frequency Attention Mechanism for TDNN and CNN-LSTM-TDNN, with Application to Language Identification," in *Proc. INTERSPEECH*, 2019, pp. 4080–4084.

[28] Klaus Scherer, Tom Johnstone, and Tanja Bänziger, "Automatic verification of emotionally stressed speakers: The problem of individual differences," in *Proc. SPECOM*, 1998, pp. 233–238.

[29] Marius Vasile Ghiurcau, Corneliu Rusu, and Jaakko Astola, "A study of the effect of emotional state upon text-independent speaker identification," in *Proc. ICASSP*. IEEE, 2011, pp. 4944–4947.

[30] Bogdan Vlasenko, Dmytro Prylipko, and Andreas Wendemuth, "Towards robust spontaneous speech recognition with emotional speech adapted acoustic models," in *Proc. 35th German Conference on Artificial Intelligence, Saarbrücken, Germany*. Citeseer, 2012, pp. 103–107.

[31] Vishnu Vidyadhara Raju V, Krishna Gurugubelli, Mirishkar Sai Ganesh, and Anil Kumar Vuppala, "Towards Feature-space Emotional Speech Adaptation for TDNN based Telugu ASR systems," in *Proc. SMM19, Workshop on Speech, Music and Mind*, 2019, pp. 16–20.

[32] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[33] Suresh Balakrishnama and Aravind Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, vol. 18, pp. 1–8, 1998.

[34] Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[35] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *Proc. IEEE workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[36] Ravi Kumar Vuddagiri, Krishna Gurugubelli, Priyam Jain, Hari Krishna Vydana, and Anil Kumar Vuppala, "IIITH-ILSC speech database for Indain language identification," in *Proc. SLTU*, 2018, pp. 56–60.

[37] Stephan Radeck-Arneth, Benjamin Milde, Arvid Lange, Evandro Gouvêa, Stefan Radomski, Max Mühlhäuser, and Chris Biemann, "Open source German distant speech recognition: Corpus and acoustic model," in *Proc. International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 480–488.

[38] Ibon Saratxaga, Eva Navas, Inmaculada Hernáez, and Iker Luengo, "Designing and recording an emotional speech database for corpus based synthesis in Basque," in *Proc. LREC*, 2006, pp. 2126–2129.

[39] Sungbok Lee, Serdar Yildirim, Abe Kazemzadeh, and Shrikanth Narayanan, "An articulatory study of emotional speech production," in *Proc. Ninth European Conference on Speech Communication and Technology*, 2005.

[40] Felix Burkhardt, Astrid Paeschke, Miriam Rolfes, Walter F Sendlmeier, and Benjamin Weiss, "A database of German emotional speech," in *Proc.*

Ninth European Conference on Speech Communication and Technology, 2005.

[41] Shashidhar G Koolagudi, Ramu Reddy, Jainath Yadav, and K Sreenivasa Rao, “IITKGP-SEHSC: Hindi speech corpus for emotion analysis,” in *Proc. International conference on devices and communications*. IEEE, 2011, pp. 1–5.

[42] Shashidhar G Koolagudi, Sudhamay Maity, Vuppala Anil Kumar, Saswat Chakrabarti, and K Sreenivasa Rao, “IITKGP-SESC: speech database for emotion analysis,” in *Proc. International conference on contemporary computing*. Springer, 2009, pp. 485–492.

[43] Ahilan Kanagasundaram, Robbie Vogt, David B Dean, Sridha Sridharan, and Michael W Mason, “I-vector based speaker recognition on short utterances,” in *Proc. INTERSPEECH*, 2011, pp. 2341–2344.