SMM19, Workshop on Speech, Music and Mind 2019
14 September 2019, Vienna, Austria

# Towards Feature-space Emotional Speech Adaptation for TDNN based Telugu ASR systems

*Vishnu Vidyadhara Raju V, Krishna Gurugubelli, Mirishkar Sai Ganesh and Anil Kumar Vuppala*

Speech Processing Lab, KCIS, IIIT Hyderabad

{vishnu.raju, krishna.gurugubelli, mirishkar.ganesh}@research.iiit.ac.in,
anil.vuppala@iiit.ac.in

## Abstract

The unavailability of speech corpora is one of the critical barriers for building a large vocabulary naturalistic Telugu automatic speech recognition (ASR) system. Hence, an effort is put towards the collection of both neutral and emotional speech samples created as Telugu naturalistic emotional speech corpus(IIIT-H TNESC). In this work, we investigate the feature-space adaptation approach to compensate the acoustic mismatch between neutral and emotional speech by using auxiliary features. The features derived from the maximum likelihood linear regression (fMLLR) of GMM models are used to perform the feature-space adaptation. The effectiveness of this adaptation is studied on deep neural network (DNN), time-delay neural network (TDNN) and combined TDNN with Long short-term memory (TDNN-LSTM) based acoustic models. Experimental results show that the feature-space adaptation approach has improved the performance of baseline by an average word error rate of 15.8%

**Index Terms**: ASR, auxiliary features, fMLLR,TDNN

## 1. Introduction

The presence of emotional content in the conversational speech is often a challenge to the neutral speech trained ASR systems. The performance of the ASR systems for such emotional speech degrades significantly. This degradation is due to the differences in the acoustic parameters under the influence of various emotions [1, 2]. Adaptation is an efficient approach for reducing these acoustic differences in the training and test conditions. Various acoustic model adaptation techniques were developed to overcome the problem of mismatch between the trained acoustic models and the data coming from a particular new speaker or channel. However, majority of the published works concentrated on various speaker adaptation techniques where there has been less focus towards these emotional speech adaptation methods [3, 4, 5]. Throughout this work, we, therefore, deal with the approach for adapting emotional speech towards robust emotional ASR.

The existing Gaussian mixture model based hidden Markov model based (GMM-HMM) acoustic model adaptation approaches are divided into two categories, namely model-space adaptation and feature-space adaptation. In the model-space adaptation, model parameters are transformed to fit the given input feature vectors better, for example, to maximize the scores of likelihood or posterior probabilities. Maximum a-posteriori (MAP) [6] and maximum likelihood linear regression (MLLR) [7] are the popular approches for model-space adaptation. Where as vocal tract length normalization (VTLN) [8] and fMLLR [7] are the popular feature-space adaptation approaches used. These approaches transform the acoustic feature vectors to fit better the acoustic model, which makes them suitable for real-time online ASR systems.

Most of the studies on emotion have mainly focused on the recognition of emotion from the speech [9, 10, 11, 12, 13, 14]. Existing emotional speech databases such as IEMOCAP [15], RECOLA [16], SEMAINE [17], FAU-AIBO [18], VAM [19], Berlin database EMO-DB [20] used for the task of emotion recognition and they lack the naturality in speech recordings as they are simulated corpora. There have been only a few attempts towards the collection of naturalistic emotional speech databases for building large vocabulary ASR systems [21], where the recording was done from podcast recordings. Similarly, most of the languages in the Indian scenario lack the amount of linguistic, text and speech resources such as transcriptions required to build deep learning based models for low resource languages. As a step towards building Standard ASR systems for low resource Indian languages, Microsoft has released speech corpus for three Indian languages of Telugu, Gujarathi and Tamil named as Microsoft speech corpus for Indian languages(MSCIL) [22]. The audio recordings provided for these three languages comprise of neutral speech. Also, there has been minimal efforts towards the collection of naturalistic emotional speech despite the availability of audio recordings in different shared websites. The existing telugu emotional speech corpora such as IITKGP-SESC [23] and IIIT-H Telugu [24] are simulated, and semi-natural emotion databases which confine to emotion recognition task.

In this paper, an effort is made towards the collection of naturalistic emotional speech from the various resources of youtube and facebook. Apart from data collection, an adaptation approach is proposed to improve the performance of the ASR system in emotional conditions. The evaluation of the collected emotional speech corpus is done on the neutral speech trained large vocabulary Telugu language ASR system of MSCIL corpus. This approach uses auxiliary feature based emotion adaptation on TDNN models and is based on the recently introduced GMM-derived features [25]. Here GMM log-likelihoods scores are used as features for training the ASR systems, with GMM-HMM based adaptation techniques.

The rest of the paper is organized as follows. The details of the database collection and baseline ASR system are provided in Section 2. Overview of the TDNN models and the aspects of emotional speech adaptation is explained in Section 3 and Section 4 respectively. The experimental results are discussed in Section 5 and the conclusion of the paper is given in Section 6.

## 2. Experimental Speech Database and Baseline Telugu ASR System

This Section consists of two sub sections. Section 2.1 provides the details of the emotional speech corpus collected for



10.21437/SMM.2019-4

the study. The baseline ASR system performance in emotional conditions is shown in Section 2.2 .

### 2.1. Emotional Speech Database

This naturalistic emotional data in Telugu (Indian) language is collected from 362 speakers of (217 male and 145 female) in four basic (neutral, sad, angry and happy) emotions.This database is named as IIIT-H TNESC. These audio recordings are collected from the freely available multimedia content on the Internet such as YouTube, Facebook, etc. The audio segments with the emotionally balanced content covering these basic emotions were selected and downloaded. These selected recordings contain natural conversations from different topics such as political debates, movie reviews, election campaigns, parliament discussions, etc. Recordings which have more restrictions and fall under copyright licenses are not downloaded. Recordings having the creative common licenses are only considered, so that this database can be shared across broader communities. Also acted recordings are avoided inorder to maintain the naturalness in the corpus. All the selected recordings are converted with software sound exchange (SOX) to mono channel mode, maintaining a sampling rate of 16 kHz with 16 bit PCM. This database consists of 15506, 3320, 3526, 3950 utterances of neutral, happy, anger and sad emotions respectively. The average duration of these utterances is around 4 seconds. For this data, the utterances in neutral and emotional state are manually segregated. The total size of the corpus is 30 hours which comprises of 18 hours of neutral speech and 12 hours of emotional speech.

Questionnaire to annotate the emotional content of the corpus was given to 5 native language listeners. Scaling from 1 to 5 was given inorder of lowest to highest similarity to the target emotion. The subjective evaluation scores which are having a score of equal or greater than 4 are only considered and the corresponding emotional labels are given accordingly. Uniformity to the collected and annotated emotional speech is maintained by following the naming convention as (spkname1_spkname2_emotion_xxx.wav). The collected emotional speech contains the speech spoken from only one speaker spk1, the label 'spk1_spk1_emotion_001.wav' is followed, by either 'n,h,a,s' for the corresponding emotions of neutral, happy, anger and sadness, followed by wav file numbering. For the case of speech considered from dyadic interaction between two speakers speaker1 and speaker2 then naming is given 'spk1_spk2_emotion_002.wav'. Initially the collected audio emotional speech samples are transcribed using Google ASR engine. Further errors in the text transcriptions are corrected manually. Audio samples and description of the collected database are available at `https://researchweb.iiit.ac.in/~vishnu.raju/`

### 2.2. Baseline ASR frame work

#### *2.2.1. Training and testing corpora*

In this work, the training phase is accomplished by combining the train and test part of Telugu language from the MSCIL corpus. The total duration of the combined part is 68 hours where all the audio recordings are emotionally neutral.

The test phase involves the evaluation of emotional speech samples on the built, neutral speech trained ASR system. For that purpose phonetically-rich emotionally spoken sentences are considered from the IIIT-H TNESC corpus. One-hour data from each of four basic emotions of anger, happiness, sad and neutral speech is used for the evaluation. There is no overlap between the training and test data during the experimentation through out the study.

#### *2.2.2. Feature extraction and Acoustic model training*

39-dimensional mel-frequency cepstral coefficients(MFCCs) are used as baseline features for training the ASR system.The opensource Kaldi toolkit [26] is used for the experiments presented in this paper. Four different acoustic models of increasing complexity are considered. The procedure adopted for training such models is the same as that used for the original Kaldi voxforge s5 recipe. The first baseline mono is characterized by 56 context-independent phones, each modeled by a three state left-to-right HMM (overall using 1000 gaussians). The speaker adaptive training (SAT) is based on a context-dependent phone modeling for the Overall 2.5k tied states with 15k gaussians. DNNs are initialized with Boltzmann machines using layer by layer generative pre-training. Cross-entropy objective function is used for the DNNs with an initial learning rate of 0.015 to final learning rate of 0.002. The softmax activation function is used at the output layer with a dimension 0f 300. The total number of epochs are set to 20 with a minimum batch size of 128. The same trigram language model is used throughout the experimental results reported in Table 1. Performance of ASR systems in terms of word error rate (WER) for given neutral, anger, happy and sadness emotions is shown in Table 1. WER for the corresponding monophone, triphone (combination of linear discriminant analysis (LDA), maximum log-likelihood transform (MLLT) and speaker adaptive training (SAT)) and DNN acoustic models is reported in Columns 2,3,4,5 and 6 respectively. The best performance is observed for the case of neutral speech and the highest degradation is observed for the case of anger emotion for DNN acoustic models.

Table 1: *Neutral speech trained ASR system performance in presence of different emotions (neutral, anger, happy and sad)*

| Emotion | WER(%) | | | | |
|---|---|---|---|---|---|
| | Mono | LDA | LDA +MLLT | LDA +MLLT +SAT | DNN |
| Neutral | 43.5 | 29.9 | 28.0 | 29.1 | 27.2 |
| Anger | 78.1 | 61.7 | 57.2 | 59.4 | 45.3 |
| Happy | 77.3 | 52.2 | 51.3 | 52.8 | 43.2 |
| Sad | 65.0 | 41.7 | 38.7 | 40.5 | 30.5 |

## 3. Interleaved TDNN-LSTM Models

One of the initial major steps in the usage of temporal convolution for modeling the future temporal context was proposed in [27] where the convolutional layers (TDNNs) are combined with recurrent layers (LSTM) for better acoustic modeling. Lattice-free maximum mutual information (MMI) objective function [28] is used for the phone-level sequence training without involving any frame level pretraining. In these neural networks the combination of the structured temporal convolution is done with LSTMs with a number of inter-leaving. The computation of TDNN-LSTM network is shown in Figure 1. The experimental set up for the TDNN-LSTM architecture is explained in detail in Section 5.1.

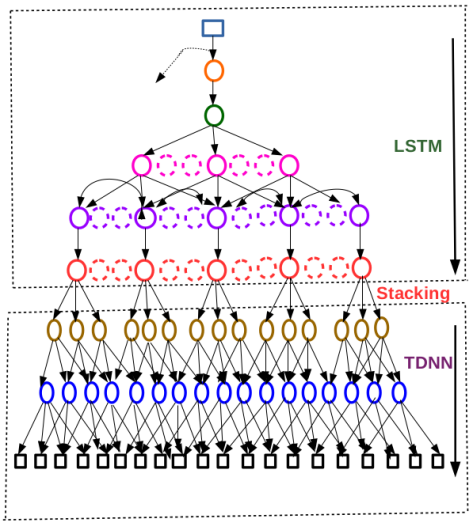


Figure 1: *Stacked TDNN-LSTM network having the dependencies among the activations with interleaved temporal convolutions.*

## 4. Auxiliary Feature-based Emotion Adaptation

The main focus of this paper is to perform the feature space emotional speech adaptation for the TDNN-LSTM acoustic models. fMLLR adaptation for the recently proposed GMM-derived features [25] was explored in this study for TDNN-LSTM models.

### 4.1. GMM-derived features for TDNN-LSTM models

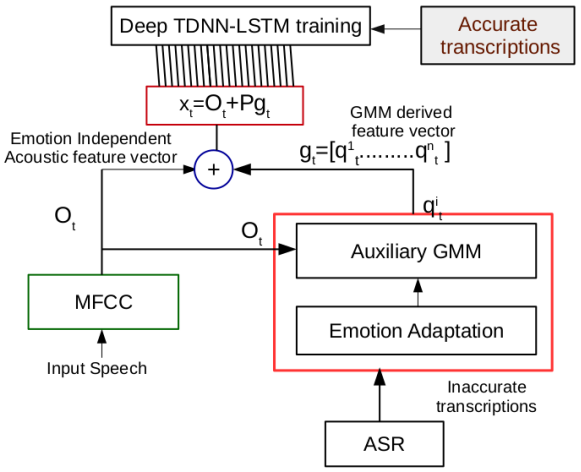


Figure 2: *Emotion adapted training for TDNN-LSTM acoustic model using fMLLR adapted GMM derived features*

In this paper we attempt to incorporate these emotionally adapted GMM-derived features for training TDNN-LSTM acoustic models.The process of performing the emotional speech adaptation using auxiliary GMM features is shown in Figure 2.We also investigate the conventional fMLLR based adaptation algorithm in this approach. Cepstral mean variance normalization (cmvn) is performed over the input 39-dimensional MFCCs features for each emotion utterance. Log-likelihood scores from the GMM model are used for training. An auxiliary triphone GMM-HMM model is used for generating the transformed log-likelihood vectors from the input acoustic feature vectors. During the training step, the emotion adaptation of the auxiliary speaker independent GMM-HMM triphone model is performed on per-emotion basis in the training corpus to create the emotion adapted GMM-HMM model.

For the input acoustic feature vector $o_t$ at the given time instant $t$, the emotion adapted GMM-derived feature vector $g_t$ is calculated as:

$$g_t = [q_t^1, ...., q_t^n], \tag{1}$$

where $n$ represents the number of states in the auxiliary triphone GMM-HMM model. For each state probability, the log-likelihood scores estimated from the GMM-HMM model is given below:.

$$q_t^i = log(Q(o_t|s_t = i)) \tag{2}$$

where $s_t$ represents the state index at time $t$. Equation 2 denotes the log-likelihood estimated using GMM-HMM. The adapted GMM-derived feature vector $g_t$ is added with the original input vector $o_t$ to obtain the resultant vector $x_t$ which are used for training the TDNN-LSTM acoustic models. $x_t$ is given by:

$$x_t = o_t + Pg_t \tag{3}$$

$P$ represents the respective projection matrix which is used in mapping the auxiliary features $g_t$ and the given input feature $o_t$.

### 4.2. fMLLR based adaptation

In this work, we use fMLLR based adaptation algorithm [7] to adapt the speaker independent GMM-HMM model. Emotion adaptation of hybrid DNN and TDNN models which are built on GMM-derived features are performed through fMLLR adaptation of the auxiliary GMM-HMM model. This is further used for the calculation of GMM-derived features. fMLLR requires only the estimation of a single transform matrix and bias vector, where the implementation of it is done through a linear feature space transformation.

$$\hat{O}_t = AO_t + b = W\xi_t \tag{4}$$

where $O_t$ is a input feature vector with $N$-dimension at time $t$ in the original feature space and $\hat{O}_t$ is the transformed feature. $W = [b\ A]$ is a matrix with dimension $N*(N+1)$ which is used to maximize the likelihood of the available adaptation data. $A$ is the $N*N$ transformation matrix and $b$ is the $N*1$ bias term. The extended observation vector is given as $\xi_t = [1 \quad O_t^T]^T$.

## 5. Experiments and Discussion

Prior to emotion adaptation, we reported the baseline performance of acoustic models trained on neutral speech in Section 2.2. The kaldi voxforge recipe (except for GMM derived and adaptation) is used and the study is extended to DNN, Subspace gaussian mixture modeling (SGMM), TDNN and interleaved TDNN-LSTM acoustic models.

### 5.1. Adaptation for DNN, TDNN and TDNN-LSTM acoustic models:

As a first step towards an improvement in the performance over the existing models, emotion adaptation is done with the available adaptation data for the given input 39-dimensional MFCC features and the results are reported in Table 2. From the IIIT-H TNESC corpus, 9 hours of the emotional data is used for the purpose of adaptation. The adaptation experiments were conducted on the data spoken from 200 speakers of the four considered emotions of anger, happy, neutral and sadness. Sequence-discriminating training is done on per-emotion basis by optimizing the state minimum Bayes risk (sMBR). The emotional data from IIIT-H TNESC corpus is used to perform sMBR training. For each emotion, we have approximately 3 hours of training data, corresponding to 3 to 4 minutes of speech data per speaker. Experiments are ensured that there is no overlap of the adaptation data with the test data.

Five different acoustic models of increasing complexity from SAT, SGMM, DNN, TDNN and TDNN-LSTM are considered and shown in Table 2. The number of leaves and gaussians used in SGMM models are 1800 and 9000 respectively. The number of sub-states used in SGMM is 6000 and dimension of the diagonal UBM (universal back ground model) is 100. i-vectors of 100-dimensional size were extracted over the input MFCC features with out any splicing of frames. We performed phone-level sequence training for the DNN and TDNN networks without the initial frame level pretraining, using the lattice-free MMI objective function instead of cross entropy.

The effective learning rate of TDNNs are set at 0.0005 with splice-indexes of {-1,0,1}, {-2,1} and {-4,2,0} for the dimension of 2000 input and 250 outputs. The initial and final learning rates of the LSTM network are 0.0003 and 0.00003 respectively with a batch size of 128. A recurrent scaling factor of 0.85 is prefered in the LSTM layers for generalization. The number of epochs is four and the number of training samples considered per each iteration are 20,000 for the LSTMs.

Table 2: *Evaluation of ASR system after performing the Emotion adaptation for the given input 39-dimensional MFCC features*

| Emotion | WER (%) Emotion Adaptation for input 39-D MFCC | | | | |
|---|---|---|---|---|---|
| | LDA+ MLLT+ SAT | SGMM | DNN | TDNN | TDNN -LSTM |
| Neutral | 25.2 | 22.6 | 20.8 | 17.1 | **16.2** |
| Anger | 40.9 | 41.0 | 35.6 | 28.1 | **24.7** |
| Happy | 40.7 | 37.2 | 38.5 | 32.1 | **26.9** |
| Sad | 27.2 | 25.0 | 24.4 | 18.0 | **15.1** |

Improvement in the performance of ASR is observed for all emotions across all the acoustic models using lattice-free MMI objective function. The main advantage of MMI objective function over the cross-entropy function used in baseline system of Table 1, it corresponds to a sequence discriminative training which takes the utterance as a whole into account instead of a frame-level used in cross-entropy. The best performance with the adaptation among these emotions is reported for TDNN-LSTM based acoustic models in Column 6 of Table 2.

### 5.2. Feature combination

The maximum improvement in the performance with the available emotion data is shown in Table 2. In this experiment, an attempt was made to improve the performance of ASR system by combining the input MFCC features with fMLLR based emotion adapted features. For this purpose 13-dimensional MFCC features are concatenated with 40-dimensional fMLLR features to form a 53-dimensional feature vector. Also considering the delta, delta-delta coefficients of the MFCC features resulted in a 159-dimensional feature vector. The performance of the ASR system for this input combination is reported in Table 3. for SGMM, DNN, TDNN and TDNN-LSTM acoustic models. The best performance is observed for the TDNN-LSTM acoustic models reported in Column 5 of Table 3. There has been a absolute improvement in the performance for anger and happy emotions by 4.6% and 5.6% respectively.

Table 3: *Evaluation of ASR system after performing the emotion adaptation for the given combination of MFCC features fMLLR adapted GMM-derived features.*

| Emotion | WER (%) for MFCC+ fMLLR based GMM-derived features(159-D) | | | |
|---|---|---|---|---|
| | SGMM | DNN | TDNN | TDNN-LSTM |
| Neutral | 21.2 | 18.7 | 15.4 | **14.5** |
| Anger | 39.6 | 26.2 | 24.8 | **20.1** |
| Happy | 35.8 | 26.8 | 27.2 | **21.3** |
| Sad | 24.5 | 16.6 | 17.1 | **14.9** |

Hence, from the experimental results it is observed that the auxiliary fMLLR based GMM-derived features are capable of handling the emotion specific information. It is also noticed that feature-space adaptation performed over the combination of MFCC with GMM-derived features proves effective when there is a limited amount of data for adaptation.

## 6. Conclusion

This paper has explored the need for having a naturalistic emotional speech corpus to build robust emotional ASR technologies for low-resource languages like Telugu and how DNN and TDNN based acoustic models benefit from the emotion adaptation methods. This paper also focused at the adaptation of neutrally trained acoustic models for emotional speech where the auxiliary feature based adaptation is investigated in detail. fMLLR based auxiliary GMM-derived features were effective in handling the emotion specific information for building the ASR systems. Combination of MFCC features with emotion adapted auxiliary GMM-derived features has yielded a better performance. In future work, we aim to extend this study towards end-to-end deep learning architectures such as recurrent neural networks with connectionist temporal classification (RNN-CTC), encoder-decoder models etc. Further, investigation of the effect of emotional speech on the language model seems to be an interesting topic to be explored.

## 7. Acknowledgements

The first author would like to thank Tata Consultancy Services (TCS), India for supporting the PhD program.

## 8. References

[1] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Does affect affect automatic recognition of children's speech?" in *Proc. 1st*

*Workshop on Child, Computer and Interaction, Chania, Crete, Greece*, 2008.

[2] B. Schuller, J. Stadermann, and G. Rigoll, "Affect-robust speech recognition by dynamic emotional adaptation," in *Proc. Speech Prosody, Dresden*, 2006.

[3] B. Vlasenko, D. Prylipko, and A. Wendemuth, "Towards robust spontaneous speech recognition with emotional speech adapted acoustic models," in *Proc. Conference on Artificial Intelligence, Saarbrücken, Germany*. Citeseer, 2012, pp. 103–107.

[4] S. Steidl, A. Batliner, D. Seppi, and B. Schuller, "On the impact of children's emotional speech on acoustic and language models," *EURASIP Journal on Audio, Speech, and Music Processing*, p. 6, 2010.

[5] V. V. Raju, P. Gangamohan, S. V. Gangashetty, and A. Kumar Vuppala, "Application of prosody modification for speech recognition in different emotion conditions," in *Proc. TENCON*. IEEE, 2016, pp. 951–954.

[6] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *IEEE Transactions on Speech and Audio processing*, vol. 2, no. 2, pp. 291–298, 1994.

[7] M. J. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, 1998.

[8] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. International Conference on Acoustics, Speech, and Signal (ICASSP)*, vol. 1. IEEE, 1996, pp. 353–356.

[9] J. Kim and R. A. Saurous, "Emotion recognition from human speech using temporal information and deep learning," in *Proc. Interspeech*, 2018, pp. 937–940.

[10] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," in *Proc. Interspeech*, 2018, pp. 247–251.

[11] S. Sahu, R. Gupta, and C. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Proc. Interspeech*, 2018, pp. 3693–3697.

[12] P. Yenigalla, A. Kumar, S. Tripathi, C. Singh, S. Kar, and J. Vepa, "Speech emotion recognition using spectrogram & phoneme embedding," in *Proc. Interspeech*, 2018, pp. 3688–3692.

[13] X. Ma, Z. Wu, J. Jia, M. Xu, H. Meng, and L. Cai, "Emotion recognition from variable-length speech segments using deep learning on spectrograms," in *Proc. Interspeech*, 2018, pp. 3683–3687.

[14] S. Parthasarathy and C. Busso, "Ladder networks for emotion recognition: Using unsupervised auxiliary tasks to improve predictions of emotional attributes," in *Proc. Interspeech*, 2018, pp. 3698–3702.

[15] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, p. 335, 2008.

[16] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *proc. of International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 2013, pp. 1–8.

[17] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.

[18] S. Steidl, *Automatic classification of emotion related user states in spontaneous children's speech*. University of Erlangen-Nuremberg Erlangen, Germany, 2009.

[19] M. Grimm, K. Kroschel, and S. Narayanan, "The vera am mittag german audio-visual emotional speech database," in *IEEE International conference on Multimedia and Expo*. IEEE, 2008, pp. 865–868.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[21] R. Lotfian and C. Busso, "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings," *IEEE Transactions on Affective Computing*, 2017.

[22] B. M. L. Srivastava, S. Sitaram, R.-K. Mehta, K.-D. Mohan, P. Matani, S. Satpal, K. Bali, R. Srikanth, and N. Nayak, "Interspeech 2018 low resource automatic speech recognition challenge for indian languages," in *Proc. International Workshop on Spoken Language Technologies for Under-Resourced Languages*, 2018.

[23] S. G. Koolagudi, S. Maity, V. A. Kumar, S. Chakrabarti, and K. S. Rao, "Iitkgp-sesc: speech database for emotion analysis," in *International conference on Contemporary Computing*. Springer, 2009, pp. 485–492.

[24] P. Gangamohan, S. R. Kadiri, and B. Yegnanarayana, "Analysis of emotional speech at subsegmental level." in *Proc. Interspeech*, 2013, pp. 1916–1920.

[25] N. Tomashenko, Y. Khokhlov, and Y. Estve, "Speaker adaptive training and mixup regularization for neural network acoustic models in automatic speech recognition," in *Proc. Interspeech*, 2018, pp. 2414–2418.

[26] 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," IEEE Signal Processing Society, Tech. Rep., 2011.

[27] V. Peddinti, Y. Wang, D. Povey, and S. Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, vol. 25, no. 3, pp. 373–377, 2018.

[28] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free mmi." in *Interspeech*, 2016, pp. 2751–2755.