INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Improving Emotion Identification using Phone Posteriors in Raw Speech Waveform based DNN

Mousmita Sarma[1], *Pegah Ghahremani*[2], *Daniel Povey*[2 3], *Nagendra Kumar Goel*[4], *Kandarpa Kumar Sarma*[1], *Najim Dehak*[2]

[1]Department of Electronics and Communication Engineering, Gauhati University, India
[2]Center for Language and Speech Processing, Johns Hopkins University, USA
[3]Human Language Technology Center of Excellence, Johns Hopkins University, USA
[4]Go-Vivace Inc., USA

(mousmita.s, kandarpaks)@gauhati.ac.in, (pghahre1, ndehak3)@jhu.edu, dpovey@gmail.com, nagendra.goel@govivace.com

Abstract

We propose to exploit phone posteriors as an additional feature in Deep Neural Network (DNN) to recognize emotions from raw speech waveform. The proposed DNN setup uses a time domain approach of learning filters within the network. The frame-level phone posteriors are combined with the learned feature representation through the network. Appended learned time domain features and phone posteriors are used as an input to the temporal context modeling layers which interleaves TDNN-LSTM with time-restricted self-attention. We achieve 16.48% relative error rate improvement in IEMOCAP categorical problem (with a final weighted accuracy of 75.03%) using phone posteriors compared to DNN setup which uses only learned time domain features for temporal context modeling. Further, we study the effect of learning emotion categories leveraging dimensional primitives in multi-task learning DNN model.

Index Terms: Emotion recognition, phone posteriors, raw speech waveform, DNN

1. Introduction

Most of the speech based human to human or human to machine communication and cognition embeds the emotional state of the speaker. Speech based emotion recognition has particular importance in communication which happens over telephone lines and perhaps speech is the only attribute from which emotion can be detected in such environment.

Emotional descriptors are not only influenced by speech based variables, but also depends on information obtained from vision, context and prior knowledge about the speaker. Therefore, extensive research has been conducted in the past to derive acoustic and statistical feature set which is most robust to model emotional descriptors from speech. With the development of Deep Neural Networks (DNN) architecture in the recent time, DNN based approaches have been used for learning emotional content either from raw spectrogram or directly from raw waveform [1, 2, 3, 4, 5, 6]. Such an approach effectively removes the necessity of hard coded feature extraction outside the DNN. However, raw speech signal contains both linguistic and paralinguistic information. Recent psychological research like [7] suggest some degree of dependency of emotional content on linguistic cues. Therefore, DNN models which is designed to learn emotion specific feature from raw waveform may benefit with a guidance on the influence of language and message specific content. In this work, we propose to use phone posteriors extracted from a separately trained acoustic model as additional input to the DNN system. Phones are the smallest distinguishable linguistic elements of a speech signal and represents the basic sound units of the language. Hence, we postulate that frame level phone posterior vectors of speech signals should provide additional information to help the DNN to isolate emotion specific cues and provide better convergence of the training objective. The proposed DNN set up combines phone posteriors with learned feature through the convolutional layer just before the temporal context modeling layers.

In our previous work [6], we used time domain raw waveform front end layers to learn filters and experimented five different configurations for long temporal context modeling using temporal convolution in the form of time-delay neural network (TDNN) layers, unidirectional recurrent projected Long Short Term Memory (LSTM) [8] layers with time restricted attention mechanisms [9] as well as using TDNN with statistic extraction layers [10, 11, 12]. In the present work, we carry out the investigation further using additional phone posteriors information as input to the temporal modeling part and experimentally prove the effectiveness of the proposed method. In our best DNN set up, we obtain 16.48% relative reduction of error rate compared to non phone posterior based set up. In this work, we also extend our DNN set up to multi-task learning configuration, in order to study the effect of providing emotion dimensional information in emotion category recognition.

Rest of the paper is organized as follows. We briefly mention a few previous works that use DNN based feature learning for emotion recognition and the importance of phone posteriors information for emotion recognition in Section 1.1. Proposed DNN set up and related results are described in Section 2. The multi-task learning extension of the DNN and related results are described in Section 3. Section 4 provides summary of findings and subsequently concludes the description in Section 5.

1.1. Related works and importance of phone posteriors

Raw speech waveforms were used by [1] in a deep CNN framework for speech emotion dimensional rating and reported results of valence and arousal estimation in Recola database. [2] used 257-dimensional magnitude Fourier log energy vectors

http://dx.doi.org/10.21437/Interspeech.2019-2093

in bidirectional LSTM and attention based DNN classifier for emotion recognition. [3] reported results using multi-task learning deep belief network (DBN), where either valence or activation dimensional information is used as auxiliary task to optimize categorical objective. A valence classifier was reported by [4] using spectrogram in deep convolutional generative adversarial network in a multi-task learning framework, where activation dimension is used as auxiliary task. [5] proposed attention pooling based representation learning method using speech spectrogram in CNN for improvised speech of IEMOCAP. Thus, many recent works have shown that feature learning through DNN is more effective than hand crafted feature set for emotion identification. However, it can be observed that influence of linguistic content on task like emotion identification has not been very widely studied so far. In the past [13] and [14] independently studied usefulness of acoustic features and emotional keywords fusion for various emotion classifiers designed using linear discriminant classifiers, k-NN classifiers, GMM/SVM based classifiers etc. and observed improvement. In this work, we propose an approach to provide linguistic knowledge as an additional information to DNN to optimize emotion identification objective. The linguistic information from speech can be extracted as a sequence of phone units. Therefore, phone posteriors obtained from a pre-trained acoustic model have been used to represent linguistic information and combined with learned feature representation prior to temporal context modeling layers. Such an approach should help the network to learn the emotion specific cues better.

2. Phone posterior based DNN for emotion recognition

The DNN configuration designed in this work can be partitioned into four parts. The feature learning layers, phone posteriors layer, temporal context modeling layers and output layer. The layout of the DNN is shown in Figure 1.

Time domain raw waveform front end uses convolution layer based feature learning, that attempts to learn filters within the DNN. In our previous work [6], we reported that time domain approach learns emotion specific cues better than MFCC features and frequency domain filter learning set up described in [17]. This raw waveform front end has a $1-d$ time convolution layer consisting of 100 FIR filters each of 31.25 ms dimension. The filter operates on 25 ms raw signal with step size 1.25 ms. Five consecutive raw speech frames are concatenated and passed to the input layer of DNN. Absolute logarithm of filter outputs are computed using a logarithmic layer and next filter outputs are aggregated using two trainable Network-in-Network (NIN) non linearity layers introduced in [10].

The phone posterior layers consist of two affine layers connected back to back. A pre-trained acoustic model is used to compute phone posteriors for the speech utterances. We train an English DNN acoustic model using multiple databases (Fisher, Switchboard, WSJ, HUB4 English Broadcast News, TED-LIUM and Librispeech) following the Kaldi $multi\_en$ recipe [1]. The posteriors of tied triphone states (a 3683 dimensional vector) obtained from the output layer of DNN acoustic model are used in the emotion identification set up. The frame level phone posterior vectors are used to compute the principal components and then used to initialize an affine transformation layer within the DNN. Raw phone posteriors of dimension 3683 are passed to the DNN as an additional input along with raw

[1] https://github.com/kaldi-asr/kaldi/tree/master/egs/multi_en/s5

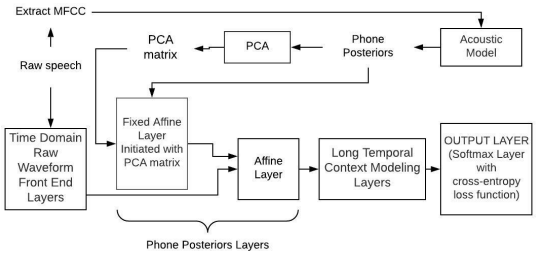

Figure 1: *Lay out of the proposed phone posterior based DNN system*

waveform. Principal component analysis (PCA) transformation of phone posteriors are obtained within the network using the affine layer initiated with PCA matrix. Next, we use another affine layer which combines the phone information with the learned feature output of time domain raw waveform front end layers. The time domain feature combined with phone posterior vector is next used as input to the temporal context modeling layers as described below.

For para-linguistic information extraction task like emotion recognition, it is important to preserve long temporal context since such information perhaps mostly resides over a longer span of time. In this work, we use temporal convolution in the form of TDNN layers along with LSTM layers for long temporal context modeling. Interleaving of temporal convolution with unidirectional LSTM is reported to outperform LSTM-attention and TDNN-Statistics pooling set up for emotion identification [6]. We further investigate LSTM-attention and TDNN-LSTM-attention layers for the proposed DNN. In LSTM-attention set up we found 128 cell dimension optimal for the current emotion identification task with the recurrence dimension of size 32 and the LSTM output of dimension 64. The LSTMs operates with a recurrence that spans 3 time steps. The layer wise context of temporal modeling block for TDNN-LSTM-attention layer is shown in Table 1. The dimensions of projection and the recurrence in LSTM layers are one quarter of the cell dimension. Time-restricted attention layer [9] is used as the last layer in both the set up along with TDNN and unidirectional LSTM layer. In such time-restricted self-attention mechanism the input and output sequence lengths are same and at particular frame it attends to input from a limited number of frames to the left and right. Use of an attention layer also helps the network to put less weight in non-speech or non relevant regions of the signal. Removing non speech frames using fixed speech activity detection (SAD) module beforehand may create a discontinuity in the temporal content. However, DNN setups with an attention layer become more attentive towards the salient regions of speech ignoring non speech regions. The attention layer used here has 12 heads, a context of $[-5, 2]$, a key-dimension of 40 and value dimension of 60.

We use softmax layer with cross-entropy loss as the output layer of DNN, modeling each emotion state as a separate output class. In our previous work [6] we experimented two different ways of aggregating time information over the chunk of speech frames: approaches with a single label per chunk with time aggregation inside the network using a statistics pooling layer and approaches where the label is repeated for each frame of the chunk. It has been observed that approaches where the label is repeated for each frame gives better performance. Therefore, in this work we use label per frame approach of training. To

Table 1: *Layer wise context of temporal modeling block (LSTM: delay time=-3)*

Layer	1	2	3	4	5	6	7	8	9
Layer-type	TDNN	LSTM	TDNN	TDNN	LSTM	TDNN	TDNN	LSTM	Attention
Context	[-1, 0, 1]	[0]	[-3, 0, 3]	[-3, 0, 3]	[0]	[-3, 0, 3]	[-3, 0, 3]	[0]	[-5,2]

obtain time aggregation over the frames of a chunk, we average frame posteriors outside the network and extract segment level aggregate.

Table 2: *Results of emotion category recognition using single task learning model*

Phone Posteriors	Temporal modeling	WA	UA
no	LSTM-attention	66.96	57.80
	TDNN-LSTM-attention	70.1	60.7
yes	LSTM-attention	68.09	64.93
	TDNN-LSTM-attention	**75.03**	**65.12**

2.1. Results

All our experiments are done using Kaldi toolkit [18]. We have used the most popular four class (neutral, angry, happy and sad) emotion category problem of Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [15] for this study. The database consists of about 12 hours of audiovisual data from five mixed gender pairs of male and female actors, at two recording scenarios- scripted and improvised speech. We use both scripted and improvised data from two male and female speakers from session four as test data and rest of the speakers are used for trainings and annotated segment boundaries are used to create utterance. To increase the amount of data in the training set we perform data augmentation by means of amplitude and speed perturbation. For each speech signal 10 different amplitude modulated versions are created initially. Speed perturbation [16] is performed on the amplitude modulated signals with speed factors 0.9, 1.0 and 1.1. Performance is reported using two parameters, weighted accuracy (WA) which is the overall classification accuracy and unweighted accuracy (UA) which is the average recall over the emotion categories.

We compare LSTM-attention and TDNN-LSTM-attention setups, with and without using the phone posteriors as shown in Table 2. It can be observed that for both the setups phone posteriors significantly improves WA and UA. We train all the networks using fixed chunk length and with a fixed number of chunks per mini batch. The experiment in row 2 is trained with 100 frames per chunk and 64 chunks per mini batch and trained for 30 epochs. The experiment in row 4 is trained using fixed length chunk of 50 frames and 128 chunks per mini batch. The learning rate is set to be gradually decreased from 1×10^{-3} to 1×10^{-4} over the course of 8 epochs. It has been observed that with phone posteriors the network converges faster than without phone posteriors and also the final objective value is observed to be slightly better in the case of phone posteriors for both LSTM-attention and TDNN-LSTM-attention set up. The training log convergence curves for the TDNN-LSTM-attention models are shown in Figure 2. We trained both the models beyond 8 epoch and 30 epochs to obtain more stable curve, however we observe decrease in the accuracy beyond 8 epochs. Table 3 shows the results of these experiments for TDNN-LSTM-attention set up. When we use phone posteriors layers, the network is effectively

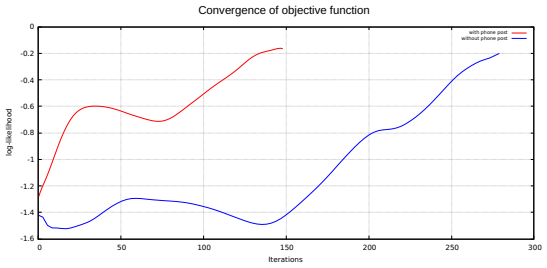

Figure 2: *Training objective convergence with and without phone posteriors*

Table 3: *Effect of chunk per minibatch and epochs in TDNN-LSTM-attention set up with phone posteriors (chunk length 50 frames)*

Number of chunks per minibatch	epcohs	WA	UA
128	30	74.02	65.03
128	10	73.01	61.66
128	8	**75.03**	**65.12**
64	30	74.02	65.73
64	10	72.25	63.29
64	8	71.1	61.52

much deeper and the stochastic gradient descent gets a better guidance to converge. Thus, we obtain a 3.42% relative reduction of error rate in LSTM-attention set up and 16.48% in case of TDNN-LSTM-attention set up.

For phone posterior based DNN we use PCA to reduce the output dimension of phone posterior layers. To select the dimension of PCA without losing relevant information, we trained the network using three different dimensions as shown in Table 4 and found the dimension of 300 as an optimal value for the present task. All results reported in Table 2, 3 and 5 use PCA dimension 300.

Table 4: *Effect of dimension of PCA in TDNN-LSTM-Attention set up*

Dimension of PCA	WA	UA
100	74.52	62.96
300	**75.03**	**65.12**
500	73.51	62.14

3. Multi-task learning DNN model using 3-dimensional information

Recent psychological research provides an alternative approach of describing emotions in continuous 3-D space called valence-activation-dominance (VAD) space. In the dimension of va-

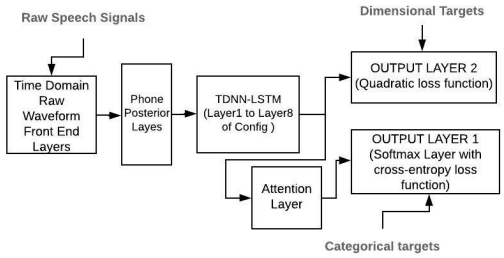

Figure 3: *Learning with dimensional information*

lence, emotional state can range from negative to positive; in the dimension of activation, emotion range from low to high and in the dimension of dominance, emotion can range from weak to strong [19, 20]. Emotion dimensional primitives are interrelated with categorical representation and knowledge of emotional description in continuous space may help machine learning system to set discrimination boundary of emotion categories more precisely. Work by Grimm et al, 2006 and Xia & Liu, 2017 previously suggested that using dimensional information in emotion category recognition helps to some extent [21, 3]. Therefore, we design multi-task learning (MTL) DNN to recognize emotion categories, where dimensional information is provided through an auxiliary output layer (Figure 3). Both the objectives are optimized through shared hidden representation. Sharing representations between related tasks is known to help the model generalize better. Since, both the tasks provides evidence for a relevant and irrelevant features, model gets a better guidance while learning. In the proposed MTL setup, feature learning and temporal context modeling layers are jointly trained and hidden activations are used to train two output layers. One output layer treats emotion as regression like problem, whereas the other output layer treats the emotion as classification problem. We jointly optimize the cross entropy loss objective of categorical emotion classification and the quadratic objective of emotion dimensional target.

3.1. Results

For this experiment we consider TDNN-LSTM-attention set up which is found to be better in the single task learning experiment described in previous section. We independently study the effect of all three dimensions using any of valence, activation or dominance as auxiliary task. We also consider all three dimensions as a 3-D vector and use that as the auxiliary task. Results of all four experiments with and without using the phone posteriors are shown in Table 3. It can be observed that using phone posteriors improves all four set ups. Use of valence and activation information could not reach the best single-task learning accuracy obtained above. In terms of WA we don't get any advantage compared to single task learning set up in all four set ups. However, we can see the improvement of UA using dominance and 3-D vector in MTL set up (row 6 and row 7 of Table 5). It means use of dimensional information improves accuracy within a particular class compared to single task learning. The reason for this improvement is quite natural since the learning process gets more clues about discrimination of the four classes through dimensional knowledge. Further, it is very important to notice that compared to valence and activation, dominance dimension is helping more to improve UA. On the other experiment, using all three dimension as 3-D vector provides almost equal UA with the experiment where only dominance is used. Perhaps providing both valence and activation is creating more confusions because of which WA reduces, although we need to investigate this further.

Table 5: *Results of emotion category recognition using Multi-task learning (MTL) model (effect of learning with dimensional information)*

Auxiliary task	Phone Posteriors	WA	UA
Valence	no	59.39	55.99
	yes	74.14	64.70
Activation	no	57.12	52.08
	yes	73.13	64.33
Dominance	no	60.27	54.97
	yes	72.88	**67.40**
Emotion 3-D Vector	no	69.98	59.81
	yes	70.99	**67.19**

4. Summary of Findings

In this work, we propose to use phone posteriors as an additional input to temporal modeling layers in raw speech waveform based end to end DNN for categorical emotion identification. We describe experimental results in two parts: a single task learning DNN set up and a multi-task learning DNN set up that leverages dimensional information. In single-task learning DNN set up we achieve 3.42% and 16.48% relative reduction of error rate respectively, in LSTM-attention and TDNN-LSTM-attention setups using phone posteriors input. The second part of experimental works attempts to study the effect of using dimensional information in multi-task learning DNN to recognize emotion categories. Here, we experiment using valence, activation and dominance information independently in multi-task learning set up and we also experiment using a single 3-D emotion vector. Experimental results show that dimensional information balances the individual recognition rate of the categories and improves UA by nearly 6.7% compared to non phone posterior set up and 2.28% compared to the single task learning set up that uses phone posteriors. Further holding our initial observation, here also we observe that for all four experiments, phone posteriors helps to improve both UA and WA compared to non phone posterior counterpart. Our present results outperforms our previously reported results on IEMOCAP emotion category recognition problem. Further, we are not aware about any such previously reported works, where linguistic information is combined with learned time domain feature representation through DNN to optimize emotion recognition objectives.

5. Conclusions

Influence of linguistic content of speech signals in paralinguistic task like emotion identification has not been very well studied. We postulated that it is important to provide knowledge of linguistic content to the DNN which learns feature representation within the network from raw waveform. We propose to supply phone posteriors as additional input along with learned time domain feature to the temporal modeling layers of the DNN and observed significant improvement in accuracy. We also studied the effect of providing emotion dimensional information in multi-task learning DNN model.

6. References

[1] G. Trigeorgis , F. Ringeval , R. Brueckner , E. Marchi , M. A. Nicolaou , B. Schuller, S. Zafeiriou, "Adieu Features? End-to-End Speech Emotion Recognition using a Deep Convolutional Recurrent Network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2016, March 20-25, Shanghai, China, Proceedings, Proceedings*, 2016.

[2] S. Mirsamadi, E. Barsoum and C. Zhang, "Automatic Speech Emotion Recognition using Recurrent Neural Networks with Local Attention," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2017, March 5-9, New Orleans, LA, USA,Proceedings, Proceedings*, 2017

[3] R. Xia and Y. Liu, "A Multi-Task Learning Framework for Emotion Recognition Using 2D Continuous Space," in *IEEE Transactions on Affective Computing,* vol. 8, no. 1, Jan-March, 2017.

[4] J. Chang and S. Scherer, "Learning Representation of Emotional Speech with Deep Convolutional Generative Adversarial Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, March 5-9 , New Orleans, LA, USA,Proceedings, Proceedings*, 2017

[5] P. Li, Y. Song, I. McLoughlin, W. Guo, L. Dai "An Attention Pooling based Representation Learning Method for Speech Emotion Recognition," in *INERSPEECH 2018, The 19th Annual Conference of the International Speech Communication Association, 2018, 2-6 September, Hyderabad, India, Proceedings, Proceedings*, 2018.

[6] M. Sarma, P. Ghahremani, D. Povey, N. K. Goel, K. K. Sarma and N. Dehak, "Emotion Identification from raw speech signals using DNNs," in *INERSPEECH 2018, The The 19th Annual Conference of the International Speech Communication Association, 2018, 2-6 September, Hyderabad, India, Proceedings, Proceedings*, 2018.

[7] K. A. Lindquist, J. K. MacCormack and H. Shablack, "The role of language in emotion: predictions from psychological constructionism," in *Frontiers in Psychology, Hypothesis and Theory,* vol. 6, April, 2015.

[8] H. Sak, A. Senior and F. Beaufays, "Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling," in *Interspeech 2014, The 15thAnnual Conference of the International Speech Communication Association, September 14-18, Singapore,Proceedings, Proceedings*, 2014.

[9] D. Povey, H. Hadian, P. Ghahremani, Ke. Li and S. Khudanpur, "A Time- Restricted Self-attention Layer for ASR," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, April 15-20, Calgary, Alberta, Canada, Proceedings, Proceedings*, 2018.

[10] P. Ghahremani, V. Manohar, D. Povey and S. Khudanpur, "Acoustic modelling from the signal domain using CNNs," in *INERSPEECH 2016, 17th Annual Conference of the International Speech Communication Association, September 8-12, San Francisco, CA, USA, Proceedings, Proceedings*,2016.

[11] D. Snyder, D. Garcia-Romero, D. Povey and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *INERSPEECH 2017, The 18th Annual Conference of the International Speech Communication Association, August 20-24, Stockholm, Sweden,Proceedings, Proceedings*, 2017.

[12] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-VECTORS: Robust DNN Embeddings for Speaker Recognition," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2018, April 15-20, Calgary, Alberta, Canada, Proceedings, Proceedings*,2018.

[13] C. M. Lee, S. S. Narayanan and R. Pieraccini, "Combining acoustic and language information for emotion recognition," in *Interspeech 2002, The 7th Annual Conference of the International Speech Communication Association, September 16-20, Denver, Colorado, USA, Proceedings, Proceedings*, 2002.

[14] B. Schuller, G. Rigoll and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2004, 17-21 May, Montreal, Que., Canada , Proceedings, Proceedings*, 2004.

[15] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," in *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[16] T. Ko, V. Peddinti, D. Povey and S. Khudanpur, "Audio Augmentation for Speech Recognition," in *INERSPEECH 2015, The 16th Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, Proceedings, Proceedings*,2015.

[17] P. Ghahremani, H. Hadian. L. Hang, D. Povey, S. Khudanpur "Acoustic Modeling from Frequency-Domain Representations of Speech," in *INERSPEECH 2018, The 19th Annual Conference of the International Speech Communication Association, September 2-8, Hyderabad, India, Proceedings, Proceedings*, 2018.

[18] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek , N. Goel , M. Hannemann, P. Motlcek , Y. Qian , P. Schwarz, J. Silovsky, G. Stemmer, K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Hilton Waikoloa Village, Big Island, Hawaii, USA, Proceedings, Proceedings*, 2011.

[19] J. A. Russell, "Core Affect and the Psychological Construction of Emotion," in *Psychological Review*, vol. 110, no. 1, pp. 145–172, 2003.

[20] R. Cowie and R. R. Cornelius "Describing the emotional states expressed in speech," in *Speech Communication*, vol. 40, issues 1-2, pp. 5–32, 2003.

[21] M. Grimm, E. Mower, K. Kroschel and S. S. Narayanan, "Combining categorical and primitives-based emotion recognition," in *The European Signal Processing Conference, Florence, Italy, Proceedings, Proceedings*, September 2006.