

A Transfer Learning Method for Speech Emotion Recognition from Automatic Speech Recognition

Recognition Technologies, Inc.

Technical Report: RTI-20200330-01

Sitong Zhou¹, Homayoon Beigi²

¹Columbia University

²Recognition Technologies, Inc. and Columbia University

¹sz2629@columbia.edu, ²beigi@recotechnologies.com

Abstract

This paper presents a transfer learning method in speech emotion recognition based on a Time-Delay Neural Network (TDNN) architecture. A major challenge in the current speech-based emotion detection research is data scarcity. The proposed method resolves this problem by applying transfer learning techniques in order to leverage data from the automatic speech recognition (ASR) task for which ample data is available. Our experiments also show the advantage of speaker-class adaptation modeling techniques by adopting identity-vector (i-vector) based features in addition to standard Mel-Frequency Cepstral Coefficient (MFCC) features.[1] We show the transfer learning models significantly outperform the other methods without pretraining on ASR. The experiments performed on the publicly available IEMOCAP dataset which provides 12 hours of emotional speech data. The transfer learning was initialized by using the Ted-Lium v.2 speech dataset providing 207 hours of audio with the corresponding transcripts. We achieve the highest significantly higher accuracy when compared to state-of-the-art, using five-fold cross validation. Using only speech, we obtain an accuracy 71.7% for anger, excitement, sadness, and neutrality emotion content.

Index Terms: transfer learning, emotion recognition, IEMOCAP, time-delay neural network

1. Introduction

Detecting emotions from speech has attracted attention for its usage in enhancing natural human-computer interaction. The ability of understanding human emotion status is helpful for machines to bring empathy in various applications.

Speech emotion recognition is suffering from insufficiency of labeled data. Though several emotion datasets have been released [2][3][4], the size of emotion datasets are relatively small due to the expensive collection costs compared with plentiful data for tasks like Automatic Speech Recognition(ASR)[5] and speaker recognition[6]. Most speech emotion detection models are trained from scratch within a single dataset[7][8][9][10], therefore cannot successfully adapt to novel scenarios which has not been encountered during training. One possible solution is to leverage knowledge acquired from large-scale datasets of relevant speech tasks to emotion recognition domain. Although some efforts were spent on transfer learning method for categorical emotion detection from other paralinguistic tasks, such as speaker, and gender recognition and effective emotional attributes prediction[11][12], however they don't choose

source domain with large-scale datasets, not shown significant improvement over non-transfer learning methods.

Previous research has shown that speech emotion detection can be improved after combined with textual data [13]. Multi-modal methods can significantly improve the emotion detection performance by incorporating lexical features from given transcripts with the acoustic features from audios [14][15][16]. However, in real application scenario, transcripts are often absent. Although ASR can provide transcripts in real time to emotional speech data[16], it requires large language models loaded, and is computational costly when decoding sequence. Therefore, transfer learning using ASR as the source domain might be an efficient solution in emotion detection to incorporate textual features through high level features extracted in ASR models.

Another challenge for emotion recognition is that speakers express emotions in different ways, in addition, environments can affect acoustic features. Speaker adaptation is useful to capsule speaker and environment specific information into acoustic features. iVector[17] based adaptation has been shown fast and efficient in speech recognition[18]. We employ i-vector based speaker adaptation in emotion detection.

This paper proposes a transfer learning method to adapt ASR models in emotion recognition domain. The model is pre-trained on Tedlium2 dataset[5], with over 207 hours data, and fine tuned on 12 hours of emotional speech. The model architecture is TDNN-based[19][1][18] with input as speaker adapted MFCC[1] features. Our experiments show the improvements in emotion detection using transfer learning from ASR to speech emotion recognition combined with speaker adaptation. The performance is evaluated on the benchmark dataset Interactive Emotional Dyadic Motion Capture (IEMOCAP)[2], with 71.7% unweighted accuracy among "angry", "happy", "sad" and "neutral" under the 5-fold cross validation strategy. Our method significantly outperforms the state-of-art strategy[9].

2. Related Work

Most efficient models in prior work are based on deep learning models, which can learn high-level features from low-level acoustic features. Lee's work[7] showed the importance of long-range context effect, and significantly improve the RNN results over DNN model. This work had been staying state-of-art for years with 63.89% unweighted accuracy(UA), until surpassed by a hybrid approach of convolution layers and LSTM convolutional LSTM[9] with 68.8% accuracy. Previous litera-

ture has seldomly discussed about TDNN architecture in emotion recognition. TDNN can efficiency capture temporal information as RNN and LSTM do, but is faster for its parallelization ability and lower computation costs during training[18], which is a desired property when training on large-scale ASR data.

Many approaches[7] [10]are speaker independent where features are normalized for individuals resulting in information loss, while our work is conducted in speaker dependent context using full MFCC raw features combined with iVectors containing speaker characteristics[17]. Peddinti[18] has proposed an efficient TDNN-based architecture for ASR with features as the combination of MFCC and iVectors, efficiently learned robust representations among various speakers and environments. Our study uses bottleneck layers of this TDNN architecture to use high-level feature representations that reflect insights from ASR tasks, and fine tunes on emotion datasets.

3. Method

3.1. Problem Definition

The emotion recognition problem is a classification problem when we represent emotion as categories rather than dimensional representations,

$$D = \{(X, z)\} \quad (1)$$

where where X are the acoustic features input and z is dimensional output corresponding to the emotion prediction. We want to find a function

$$D = \{f : X \rightarrow z\} \quad (2)$$

to map features to categories. This model is trained on frame-level labels, and predicts the utterance labels by aggregating frame-level predictions through max likelihood by summing up the results of frames.

3.2. Feature

Full MFCC features with all 40 coefficients are computed at each time index is used as input to neural network. Instead of mean normalization on MFCC, an 100-dimension iVector is appended to MFCC features at each frame to encode mean-offset information. The iVector extraction model is trained as described in [18].

3.3. TDNN for ASR

TDNN is designed for capturing long term temporal dependencies in lower computational costs compared to RNN. It operates similar to a feed forward DNN architecture where lower layers focus input content in narrow windows, and higher layers connects windows of selected previous layer nodes to process the information from a wider context. Therefore its deeper layers can learn effective long term temporal dependencies without recurrent connections which hurdle parallel computation.

The pretraining on ASR follows the Kaldi recipe for the TED-Lium tasks[20], where uses 13 TDNN layers and each layer consists of 1024 activation nodes. The time stride of each layer, which defines the window at which calculating over nodes at neighbor time steps in the past layer, is assigned as 0 for the 1st and 5th TDNN layer, as 1 from the 2nd and 4th layer, and as 3 for layers after since the 6th. A fully connected prefinal layer of 1024 dimension follows the 13th TDNN layer before decoding output sequences. The model is trained with a sequence-

level objective function named lattice-free version of the maximum mutual information (LF-MMI)[21], for maximising the log-likelihood of the correct sequences.

3.4. Training on Emotion Labels

Emotion labels are given for each utterance in the dataset. We label all the frames using the utterance label where the frames lie in. To train for emotion detection, the 12th and 13th TDNN layers as well as the ASR prefinal layer are selected to produce bottleneck embeddings as high level features learnt from ASR, a new fully connected dense layer is appended after the embedding layer for predicting the frame-level emotion labels, and a softmax layer with four dimension outputs is used to predict frame-level emotion. The model uses cross-entropy as the objective function for frame-level classification.

The output of this model is for frame unit rather than for utterance unit, we aggregate frame-level predictions, using maximum likelihood by adding the output vectors over frames, corresponding to the highest valued dimension of the sum of the softmax layer output over all frames within the utterance

4. Experiment

4.1. Dataset

This work uses IEMOCAP, which contains 12 hours of audio data with scripted and improvised speech, performed by ten actors, one male and one female as a pair in five sessions. In training and testing, four categories “angry”, “happy”, “sad” and “neutral” are selected out of ten categories, for a more balanced and efficient dataset, resulting in a final collection of 4936 utterances, each utterance has unique emotion label. This dataset consists of five sessions, and the category distribution is as in Table 1.

Table 1: *Emotion category distribution in IEMOCAP*

| session | ang | exc | neu | sad | total |
|---------|------|------|------|------|-------|
| ses1 | 229 | 143 | 384 | 194 | 950 |
| ses2 | 137 | 210 | 362 | 197 | 906 |
| ses3 | 240 | 151 | 320 | 305 | 1016 |
| ses4 | 327 | 238 | 258 | 143 | 966 |
| ses5 | 170 | 299 | 384 | 245 | 1098 |
| total | 1103 | 1041 | 1708 | 1084 | 4936 |

4.2. Pretraining on ASR

For pretraining on ASR, we use the feed-forward TDNN to capture long term temporal dependencies from short term feature representations. Hidden activations are sub-sampled in order to speed up the training[18]. The model is pre-trained on ASR data, with 13 TDNN layers and output layers for decoding sequence based on acoustic models. As neighboring activations shares largely overlapped input contexts, sub-sampling on activations can reduce computational costs without sacrificing the coverage range over input frames. The hyper parameters for model architecture and training are chosen according to Kaldi Tedlium2 TDNN recipe[20], which has been tuned properly on Tedlium2 dataset, achieving 7.6 word error rate(WER) on test dataset after six epochs training. The parameters are optimized through preconditioned stochastic gradient descent (SGD) updates, following the training recipe detailed in [22].

4.3. Training for Emotion Classification

After obtaining the pretraining model, we use 12th, 13th and the prefinal layer as the bottleneck features for the appended fully connected layer and the softmax layer for predicting frame emotions. We test on session 5 after training on session 1-4, and find the 12th TDNN layer the best performance, therefore we use the 12th layer output as the bottleneck embeddings for later experiments.

4.4. Evaluation Method

For parallel comparison with other methods[7][9][10], we train under a 5-fold cross validation strategy where each time we choose a session from IEMOCAP for testing, and the other four for training. The results are evaluated by the average of unweighted accuracy over five cross validation experiment runs.

5. Results

5.1. Bottleneck Layer Selection

Compare the prefinal, 12th and 13th layer as in Table 2, we found that the 12th layer has the best performance and prefinal have the worst. We hypothesis that is because in ASR, prefinal and 13th layer are more specialized in speech recognition as they are closer to the final output layer, while the 12th learns the general high-level acoustic features that helps emotion recognition.

Table 2: Test Accuracy on Session 5 with different bottleneck layers

| Bottleneck Layer | Test Accuracy on Ses 5 (%) |
|-----------------------|----------------------------|
| Prefinal | 63.4 |
| 13 th TDNN | 69.3 |
| 12 th TDNN | 66.7 |

5.2. Model performance

Our model using the 12th TDNN layer outperforms other current state-of-art methods. Table 3 to our best knowledge. The 5-fold cross validation unweighted accuracy is improved from 68.8%[9] to 71.7%.

Table 3: Model Comparison on Unweighted Accuracy(UA) in % from Five-fold Cross Validation

| Model | Single Session UA | 5-fold CV UA |
|-----------------------|-------------------|--------------|
| ASR Transfer Learning | ses1 | 65.3 |
| | ses2 | 78.9 |
| | ses3 | 71.1 |
| | ses4 | 73.9 |
| | ses5 | 69.3 |
| | | 71.7 |
| RNN-ELM[7] | | 63.9 |
| Conv-LSTM[9] | | 68.8 |
| CTC-BLSTM[10] | | 54 |

5.3. Error analysis

To study the model performance within each emotion category, we present a confusion matrix, shown in Figure 1, by calculat-

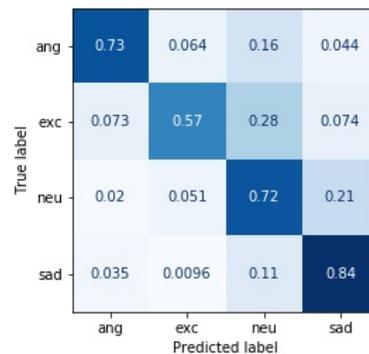


Figure 1: Confusion Matrix for Transfer Learning Method Predictions

ing the average confusion matrix of five cross validation experiments for our best model architecture, using TDNN 12th layer as the bottleneck layer. The excitement category has a lower accuracy compared to other categories, in which 28% samples are confused with neutral utterances. We observed that the neutral emotion is the most likely wrong prediction from all non-neutral categories. It might due to the fact that non-neutral utterances usually consist of a large proportion of frames carrying no emotional content. We also found the model confuses neutrality as sadness, that may because many neutral utterances are in low tone, as in sadness.

6. Conclusion

Our study shows transfer learning from ASR is a good strategy for emotion classification, and indicates potential feature overlap between speech-to-text and emotion recognition. Our method is limited in frame-level prediction, where frames are predicted first then aggregated into utterance level labels. The frame-level structure results in the ignorance of sequential information in emotion labels decoding. In future, we expect sequence models can predict at utterance-level and bring further performance improvements by considering sequential information for sequence decoding.

7. References

- [1] H. Beigi, *Fundamentals of Speaker Recognition*. New York: Springer, 2011, ISBN: 978-0-387-77591-3, <http://www.fundamentalsofspeakerrecognition.org>.
- [2] 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, pp. 335–359, 2008, springerlink Online: DOI 10.1007/s10579-008-9076.
- [3] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendmeier, and B. Weiss, "A database of german emotional speech," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, vol. 5, Lisbon, Portugal, 2006.
- [4] P. Jackson and S. ul Haq, "Surrey audio-visual expressed emotion (SAVEE) database," 04 2011.
- [5] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM corpus with selected data for language modeling and more TED talks," 05 2014.
- [6] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and

- hearing faces: Cross-modal biometric matching,” Jun. 18-22 2018.
- [7] J. Lee and I. Tashev, “High-level feature representation using recurrent neural network for speech emotion recognition,” Sep.
 - [8] Y. Kim, H. Lee, and E. M. Provost, “Deep learning for robust feature generation in audiovisual emotion recognition,” pp. 3687–3691, May 26-31 2013.
 - [9] A. Satt, S. Rozenberg, and R. Hoory, “Efficient emotion recognition from speech using deep learning on spectrograms,” Aug. 20-24.
 - [10] V. Chernykh, G. Sterling, and P. Prihodko, “Emotion recognition from speech with recurrent neural networks,” *arXiv preprint arXiv:1701.08071*, Jul. 2018.
 - [11] S. Latif, R. Rana, S. Younis, J. Qadir, and J. Epps, “Transfer learning for improving speech emotion classification accuracy,” Sep. 2-6.
 - [12] S. Ghosh, E. Laksana, L.-P. Morency, and S. Scherer, “Representation learning for speech emotion recognition,” in *Interspeech*, Sep. 2016.
 - [13] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, “A review of affective computing: From unimodal analysis to multimodal fusion,” *Information Fusion*, vol. 37, pp. 98–126, Sep. 2017.
 - [14] S. Tripathi and H. Beigi, “Multi-modal emotion recognition on iemocap dataset using deep learning,” *arXiv:1804.05788v1 [cs.AI]*, Apr 2018.
 - [15] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller, and S. Zafeiriou, “End-to-End multimodal emotion recognition using deep neural networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, pp. 1301–1309, Sep. 2017.
 - [16] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, “Deep neural networks for emotion recognition combining audio and transcripts,” Sep. 2-6.
 - [17] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” vol. 19, no. 4, pp. 788–798, May 2011.
 - [18] V. Peddinti, D. Povey, and S. Khudanpur, “A time delay neural network architecture for efficient modeling of long temporal contexts,” in *Interspeech*, Sep. 2015.
 - [19] M. Sugiyama, H. Sawai, and A. Waibel, “Review of tdnn (time delay neural network) architectures for speech recognition,” vol. 1, Jun 1991, pp. 582–585.
 - [20] I. GitHub, “A TDNN recipe for automatic speech recognition training on ted-lium,” Website. [Online]. Available: https://github.com/kaldi-asr/kaldi/blob/master/egs/tedlium/s5_r2/local/chain/tuning/run.tdnn.lg.sh
 - [21] D. Povey, V. Peddinti, D. Galvez, P. Ghahmani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, “Purely sequence-trained neural networks for asr based on lattice-free mmi,” in *Interspeech*, Sep. 2016.
 - [22] X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, “Improving deep neural network acoustic models using generalized maxout networks,” May 4–9.