FEATURE LEARNING BETWEEN TRADITIONAL SPEECH EMOTION RECOGNITION SYSTEMS AND RECURRENT NEURAL NETWORK MODEL

Prawesh Dahal

Department of Electrical Engineering, Columbia University, NY, USA

ABSTRACT

In recent years, human-machine interactions have become highly representative of a realistic interpersonal interaction. Speech analysis has played an integral role in reducing this gap between physical and digital world. Since a lot of information in human speech is conveyed through emotional cues, there has been a growing interest in the subfield of emotion recognition. However, automatic speech emotion recognition (SER) is a challenging task as it heavily depends on the effectiveness of the features used for classification. The primary objective of this project is to compare the performance of two classifiers: (i) conventional classifiers such as Support Vector Machine (SVM) that rely on the hand-picked features provided as input (ii) deep learning models such as a Recurrent Neural Network (RNN) that can automatically discover emotionally relevant features from speech. The proposed solution was evaluated on the RAVDESS corpus and the classification result verified that RNN model provides better accuracy than that achieved by using conventional machine learning classification methods.

Index Terms— speech, emotion, SVM, RNN, RAVDESS

1. INTRODUCTION

Speech not only serves as the most natural and effective way of communication but also carries of the most expressive modalities for human emotions. Emotions play an important role in the interactions between human beings as it influences most aspects of communication such as facial expressions, body gestures, voice and tonal properties and the linguistic content [1]. For an efficient interaction, we need to recognize and understand the correct emotion of the other person and be able to deliver an appropriate reaction. The role of emotional expressions is not limited only to the human world. Automatic speech emotion recognition has recently become a vast research field and has found applications in areas like psychology, psychiatry, behavioral science, artificial intelligence, computer vision and human-machine interactions [1-3].

However, one of the fundamental challenges in automatic speech emotion recognition has been the identification and extraction of appropriate features from speech. There have been many endeavors taken to discover speech features that can be indicative of different emotions [4-5]. Though both short-term frame-level as well as long-term utterance level features have been proposed, shown in Table I, there is still no definitive answer for which features are the appropriate descriptors for emotions.

TABLE I. COMMON LOW-LEVEL DESCRIPTORS (LLD) AND HIGH-STATISTICAL FUNCTIONS (HSF) FOR SER

LLDs	Pitch, voicing probability, frame energy, zero-crossing rate, MFCCs, formant locations, formant bandwidths, harmonics-to-noise ratio, spectral rolloff, spectral centroid, jitter, etc.
HSFs	mean, variance, min, max, median, quartiles, higher order moments (skewness, kurtosis), etc.

In this project, the performance of speech emotion recognition is compared between two methods. Conventional classifiers that uses machine learning algorithms has been used for decades in recognizing emotions from speech. However, in recent years, deep learning methods have taken the center stage and have gained popularity for their ability to perform well without any input hand-crafted features. Speech emotion sets obtained from RAVDESS corpus is classified using a conventionally used Support Vector Machine (SVM) and its performance is compared to that of a bidirectional long short-term memory (LSTM). The inspiration behind this project comes from the studies done by [6] which couples a recurrent neural network (RNN) with an attention mechanism to enable the model to focus on emotionally salient part of the sentence, significantly improving the accuracy to that of SVM.

2. RELEVANT WORK

Traditionally, the state-of-the-art approach used to handle this problem has been to extract large number of low-level descriptors (LLDs) from short time frames and aggregate this information over time using high-statistical functions (HSFs). To obtain a compact representation of the feature sets, dimension reduction techniques such as Principal Component Analysis (PCA) are also usually implemented. A standard machine learning algorithm is then applied onto the

final feature sets obtained to perform classification. Several studies such as in [5, 7-10] have employed this approach using methods such as Support Vector Machine (SVM), Gradient Boosting, K-Nearest Neighbor (KNN), Random Forest and Hidden Markov Models (HMM).

Despite these existing conventional techniques, the challenge of selecting good features still holds and the optimization can be a complicated process, often being a time-consuming effort in research, development and validation. However, deep neural architectures can bypass this cumbersome process of using hand-crafted features in traditional approaches. Neural network models can share low-level representations and naturally progress to high-level structures [11]. Therefore, such deep learning techniques can learn relevant features by stacking network layers and have taken a center role in speech processing and analysis. In particular, research conducted by [12] proposes a convolutional neural network (CNN) based SER system that learns salient emotion features using semi-CNNs. The authors in [13] used a Deep Neural Network (DNN) along with traditional frame-level statistical features to improve classification accuracy. Other works like [6] that uses RNN account for the long-term contextual effect in emotional speech and deal with the uncertainty of emotion labels.

3. MATERIALS AND METHODS

3.1. RAVDESS Database

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset was used for this project. It consists of audio, facial and multimodal data. Only audio speech files were used in this project. The project dataset consists of 1440 audio files from 24 subjects (12 male and 12 female actors) which encompasses eight emotion states, namely neutral, calm, happiness, sad, anger, fear, disgust and surprise. The speech files have annotated emotion and intensity labels, are around 4 seconds long, and consist of a single sentence expressing a single emotional state, read in North American English. All emotion states have 192 files, except for neutral which only has 96. Hence, this class imbalance was taken into account both during the implementation of SVM and RNN.

3.2 Pre-processing

All the audio recordings were zero-padded to make them all the same length. Sample raw traces and log-Mel spectrogram for the speech recording with statement – "Kids are talking by the door", uttered with four different emotions are presented in Fig. 1. For RNN, the spectrogram was computed with a window frame of 50 ms and a hop length of 10 ms which provided a spectrogram of size 128 frequency × 388 time frames. For SVM, a smaller window frame of 25 ms was used.

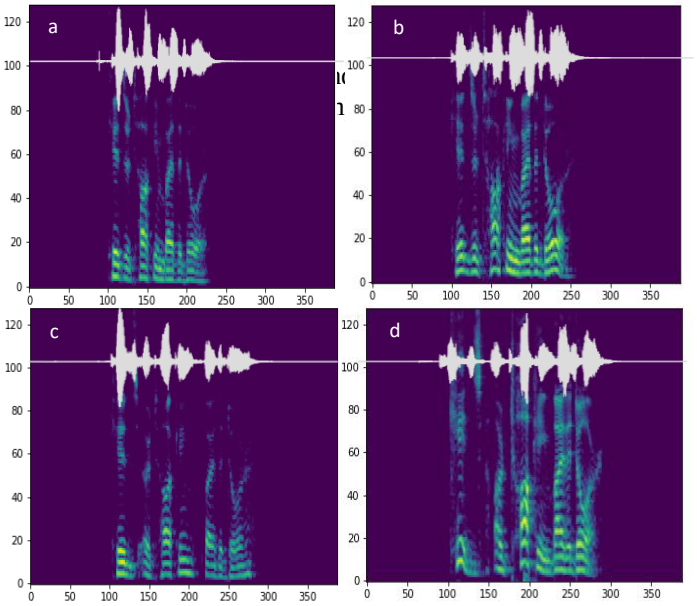

Fig. 1. Sample raw traces for a single speech statement – *kids are talking by the door* - uttered with four different emotion states, (a) neutral, (b) happy, (c) sad and (d) angry.

3.3. Classifier I: Support Vector Machine (SVM)

As a baseline SER system, a conventional SVM classifier was implemented on the emotion speech. SVM is a supervised machine learning algorithm that is used for both regression and classification. The key concept of this algorithm is the construction of a separating hyperplane in the n-dimensional input space that maximizes the between class margin to better separate the classes. For data like ours that is not linearly separable, the feature space is mapped into a high dimensional space using kernel functions like polynomial and radial basis function (RBF). A RBF kernel was used in the SVM multiclass classifier for this project. The framework of SVM Classifier used in this project is presented in Fig. 2.

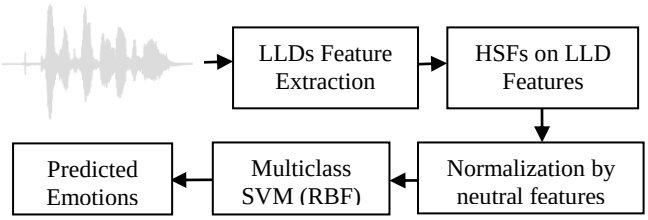

Fig. 2. SVM Classifier Framework

SVM requires hand-crafted features from the speech as input to the classifier. Feature extraction was performed in two stages. First, acoustic LLD features that are believed to be influenced by emotions are extracted from window

frames of 25 ms with a overlap of 10 ms. 13 Mel-frequency Cepstral Coefficients (MFCC), together with spectral rolloff, spectral centroid, zero-crossing rate and frame energy makes a 17-dimensional LLDs for each time frame. Next, statistical aggregation functions (HSFs), particularly mean, similar to [6], was applied to each of the LLDs over the duration of the utterance to roughly describe the temporal contours and variations of the different features during the speech. These final features were normalized by the global mean and standard deviations of neutral speech features in the training set. Since the training data is imbalanced with respect to the number of emotional classes, a cost-sensitive training strategy was applied where the cost of each emotion sample is scaled in accordance to the number of samples in that category.

3.4. Classifier II: Bi-directional LSTM

As a second classifier, a bi-directional LSTM recurrent layers with 128 memory cells was implemented where the memory cells are useful for learning the temporal aggregation. This project investigates on both fully connected BLSTM as well as attention-based BLSTM. A 30% dropout was implemented on all layers during training to avoid over-fitting. Each architecture concludes with a SoftMax layer for classification to one of the eight labels for emotion recognition

The training, validation and test sets were randomly sampled in the ratio 7:2:1 respectively from *each emotion set* such that the final sets each have same ratio of files of the 8 emotion classes. The input to the BLSTM was the log-Mel spectrogram of size 128 × 388, where all spectrograms were normalized by the mean and variance of the training set.

To calculate the loss, the Focal Loss function from [14] was adopted.

$$\mathrm{FL}(p_t) = -(1 - p_t)^{\gamma} \log(p_t)$$

This loss adds a factor of $(1 - p_t)^{\gamma}$ to the standard cross entropy criterion where setting $\gamma > 0$ reduces the relative loss for well-classified examples and applies more focus on hard, misclassified examples [14].

In the first RNN architecture, a final-frame (many-to-one) training was implemented. Instead of training the RNN frame-wise by assigning an overall emotion to each and every frame within the utterance, only the final RNN hidden representation at the last frame was picked and passed through the softmax layer. The errors calculated were then back-propagated to the beginning of the utterance. To implement attention-based RNN, the procedure in [6] was partially adopted. The product of attention parameter and the final output of RNN was computed as a score of the contribution of frames. The final attention weights were used with the final outputs to get utterance-level representation.

4. RESULTS

4.1. Classifier I: Support Vector Machine (SVM)

The confusion matrices obtained from SVM classification is shown below. Fig. 3 presents the confusion matrix for the classification of 4 emotions (neutral, happy, sad and angry). The reasoning behind performing analysis on these particular emotions was to compare the findings of the project to that in the reference paper [6]. Fig. 4 presents the classification performed on the entire eight states of emotion.

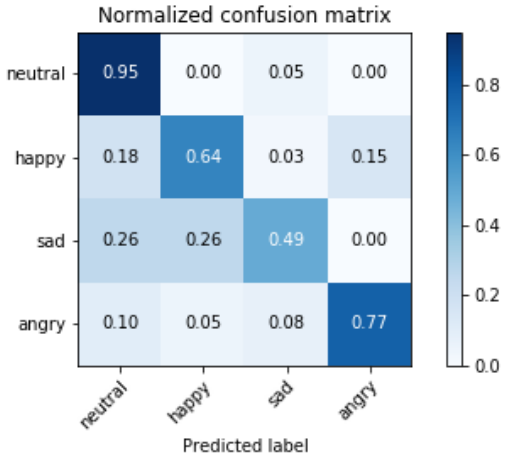

Fig. 3. Confusion Matrix of RAVDESS 4 emotions states

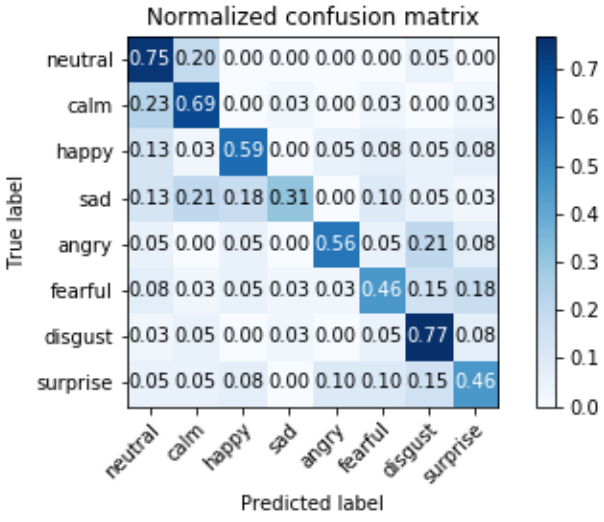

Fig. 4. Confusion matrix of RAVDESS all emotions states

TABLE II. COMPARISON OF CLASSIFIER RESULTS WITH PAPER [6]

Results	Features	Classifier	HSF	Accuracy
Mirsamadi *et al.* (2017) (4 emotions)	32 LLDs	SVM	Mean	53.3%
Project (4 emotions)	17 LLDs	SVM	Mean	**61.2%**
Project (8 emotions)	17 LLDs	SVM	Mean	**56.3%**

The SVM classifier provides an accuracy of 61.2% on four emotions, achieving a +7.9% absolute improvement compared to the reference literature for this project. When classifying eight emotions, the model gives an accuracy of 56.3%. One of the most important lesson learned from SVM classification is the relevance of features. In the initial phase of this experiment, the same features used by [6] were used to perform classification on the RAVDESS dataset. However, the SVM gave an accuracy of chance and was not being trained properly. It was then realized that depending on the database, features relevant to IEMOCAP emotion dataset used by [6] may not be suitable to the emotion dataset obtained for this project. Only by the elimination of some features, mainly the single and double-derivatives of MFCC, which had proved successful for [6], the accuracy of the SVM was improved.

4.1. Classifier II: Bi-directional LSTM

The training and validation loss and accuracy of the final-frame (many-to-one) frame work RNN training is shown below. The best model obtained from the validation set before overfitting was used to evaluate the test dataset.

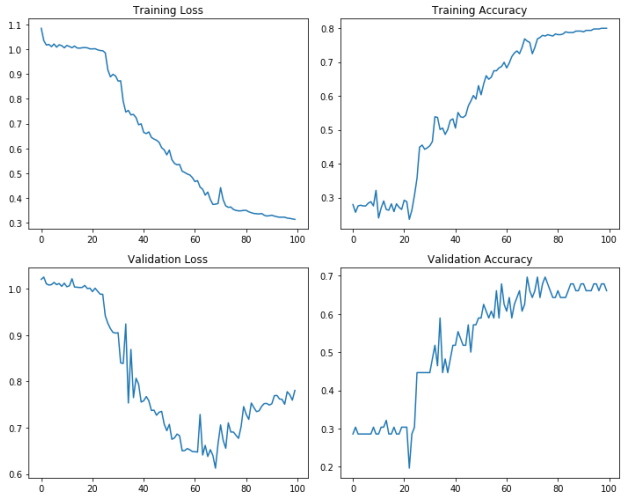

Fig. 5. Performance of BLSTM, with 30% dropout

TABLE III. COMPARISON OF LSTM CLASSIFIER RESULTS WITH PAPER [6]

Results in:	Features	Temporal aggregation	Test Accuracy
Mirsamadi *et al.* (2017) (4 emotions)	Raw spectral	RNN-final frame RNN-attention	54.4% 61.8%
Project (4 emotions)	Mel spectral	RNN-final frame RNN-attention	**66.07%** **67.86%**
Project (8 emotions)	Mel spectral	RNN-final frame RNN-attention	**52.50%** **54.10%**

Compared with the reference paper [6], the BLSTM network created in this project achieved better classification test accuracy in both types of final-frame and attention-based RNN architecture. More importantly, compared with traditional SVM solution, the RNN algorithm achieved better results. In particular, in the classification of four emotions, RNN-final frame and RNN-attention based network achieved +4.87% and +6.66% absolute improvements in the weighted test accuracy. Within the two RNN architectures, it was observed that an attention-based network can provide improved results. In particular, in the classification of four and eight emotions, LSTM-attention performed +1.79% and +1.60% better.

4. CONCLUSION

This paper presented two different approaches for feature learning in speech emotion recognition. It was shown that using conventional approach such as SVM can be used to classify emotion states. However, this is a cumbersome process, especially due to the task of hand-picking suitable features from emotion. From the experiment in this project, it was identified that the types of LLDs used as inputs to the SVM can have a major effect on the classification accuracy. However, using deep RNN, it was shown that not only the task of identifying and extracting fixed designated features can be avoided, but these networks also perform better than conventional SVMs. Through RNN, we can learn frame-level characterization and aggregate the information temporally into longer time spans. In addition, the RNN attention-based strategy outperforms all other training methods in this project. Though final-frame outputs was used in one of the training methods, relying on the final frame of the sequence may not fully capture the intended emotion. One possible improvement to this technique can be to perform a mean pooling over time on the RNN outputs and pass the result to a softmax layer. With more training data, the parameters of short-term characterization, long-term temporal aggregation and the attention model and all be jointly optimized for maximum performance.

-

5. REFERENCES

[1] Y. Gao, B. Li, N. Wang, and T. Zhu, “Speech Emotion Recognition Using Local and Global Features,” in Brain Informatics, 2017, pp. 3–13.

[2] A. Iqbal and K. Barua, “A Real-time Emotion Recognition from Speech using Gradient Boosting,” in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019, pp. 1–5.

[3] W. Minker, J. Pittermann, A. Pittermann, P.-M. Strauß, and D. Bühler, “Challenges in speech-based human–computer interfaces,” Int J Speech Technol, vol. 10, no. 2, pp. 109–119, Sep. 2007.

[4] M. Tahon and L. Devillers, “Towards a Small Set of Robust Acoustic Features for Emotion Recognition: Challenges,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 16–28, Jan. 2016.

[5] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, “Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge,” Speech Communication, vol. 53, no. 9, pp. 1062–1087, Nov. 2011.

[6] S. Mirsamadi, E. Barsoum, and C. Zhang, “Automatic speech emotion recognition using recurrent neural networks with local attention,” in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017, pp. 2227–2231.

[7] M. Ghai, S. Lal, S. Duggal, and S. Manik, “Emotion recognition on speech signals using machine learning,” 2017 Int. Conf. Big Data Anal. Comput. Intell., pp. 34–39, 2017.

[8] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process. A Rev. J., vol. 22, no. 6, pp. 1154–1160, 2012.

[9] C. N. Anagnostopoulos and T. Iliou, “Towards emotion recognition from speech: Definition, problems and the materials of research,” Stud. Comput. Intell., vol. 279, pp. 127–143, 2010.

[10] O. Kwon, K. Chan, J. Hao, and T. Lee, “Emotion Recognition by Speech Signals,” in In Proceedings of International Conference EUROSPEECH, 2003, pp. 125–128.

[11] W. Lim, D. Jang, and T. Lee, “Speech emotion recognition using convolutional and Recurrent Neural Networks,” in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016, pp. 1–4.

[12] Huang, Zhengwei, et al. "Speech emotion recognition using CNN." Proceedings of the ACM International Conference on Multimedia. ACM, 2014.

[13] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, “Deep neural networks for acoustic emotion recognition: Raising the benchmarks,” 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal Loss for Dense Object Detection,” arXiv:1708.02002 [cs], Aug. 2017.