Interspeech 2018
2-6 September 2018, Hyderabad

Emotion Recognition from Human Speech Using Temporal Information and Deep Learning

John W. Kim[1], *Rif A. Saurous*[2]

[1]Menlo School, Atherton, CA, USA
[2]Google Inc. Mountain View, CA, USA
john.kim@menloschool.org, rif@google.com

Abstract

Emotion recognition by machine is a challenging task, but it has great potential to make empathic human-machine communications possible. In conventional approaches that consist of feature extraction and classifier stages, extensive studies have devoted their effort to developing good feature representations, but relatively little effort was made to make proper use of the important temporal information in these features. In this paper, we propose a model combining features known to be useful for emotion recognition and deep neural networks to exploit temporal information when recognizing emotion status. A benchmark evaluation on EMO-DB demonstrates that the proposed model achieves a state-of-the-art performance of 88.9% recognition rate.

Index Terms: emotion recognition, temporal information, deep learning, CNN, LSTM

1. Introduction

Human-machine speech communication is spreading into our daily lives, thanks to recent advances in accurate speech recognition and accompanying wide availability of speech recognition devices. However, the capabilities of such devices are limited to understanding only the word-level content of human speech; enabling machines to perceive our emotions would help more natural and empathic dialogue in human-machine interactions.

Typical speech emotion recognition systems consist of a feature extraction stage followed by classification stages. Extensive studies have investigated and extracted key features relevant to emotion status carried in speech waveforms. In [1], a large set of 6373 features, mainly derived from short-time waveform segments with sliding windows, is defined. Recently, Eyben et al. [2] proposed a minimalistic set of features called the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) consisting of 62 features, and extended GeMAPS consisting of 88 features as the baseline for evaluation of future research. These features, combined with static pattern classifiers such as Support Vector Machine (SVM), demonstrate reasonably good performance.

One of the major problems with the conventional approach is that much less effort has been made to make use of how the extracted features change over time. Numerous studies in psychology support the importance of temporal information of the features, where it is demonstrated that the pattern of stress and intonation is highly associated with emotion status [3]. A common method to employ temporal information in emotion recognition systems is to use the standard deviation of the sequence of the speech feature vectors and the mean vector to produce an input vector to a static classifier. However, this method can lead to the loss of key temporal information for emotion recognition. For example, a time-reversed version of feature vectors results in the same standard deviation as the original feature vectors, but they wouldn't necessarily correspond to the same emotion. To overcome this limitation, some systems append explicit temporal features to the input vector, such as mean and standard deviation of voiced and unvoiced speech duration, pseudo-syllable rate, etc. [2]. Yet, these features fail to convey all the key temporal information, such as multiple patterns of temporal changes of individual features spread in different times.

On the other hand, recent advances in deep neural networks (DNN) have demonstrated great success. Convolutional Neural Networks (CNN) for extracting higher-level representations effectively in speech recognition [4, 5] and Long Short-Term Memory (LSTM) models for sequence classification [6] are the most common examples. Recently, an extreme end-to-end approach using DNNs demonstrated good performance, where raw speech waveforms were used directly as inputs to build a model that learns feature extraction and classification together [7]. However, this end-to-end approach fails to take advantage of useful features established by experts in the past decades in the field of emotion recognition. Moreover, an end-to-end approach also requires a large network size because the model needs to jointly learn feature extraction and classification, and therefore needs an immense amount of data for good performance.

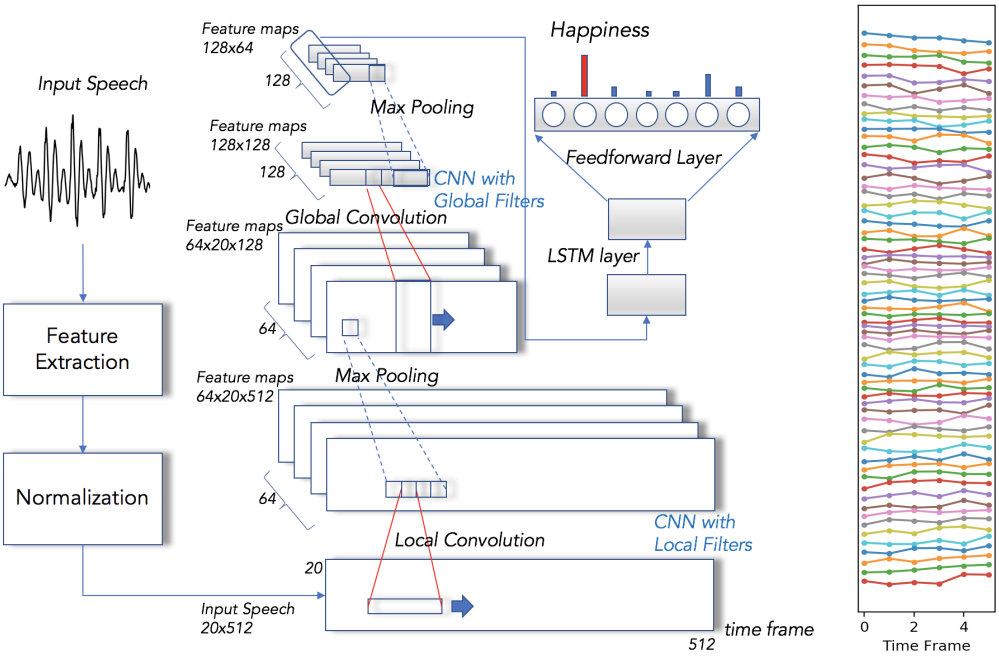
We propose EmNet, a model that combines 1) the use of common features useful for emotion recognition without eliminating their temporal information, and 2) DNN to extract higher level representations of temporal patterns of the features and relating them to the corresponding emotion status.

The rest of this paper is organized as follows: Section 2 presents the details of EmNet. Experimental validation on a common data corpus is presented in comparison with conventional approaches in Section 3, followed by conclusions in Section 4.

2. Proposed EmNet Model

Fig. 1 (a) shows the structure of EmNet for emotion recognition. The model consists of multiple stages of processing: feature extraction, feature normalization, CNN layer with local convolution, CNN layer with global convolution, and LSTM layers followed by feedforward layer.

10.21437/Interspeech.2018-1132



(a) (b)

Fig. 1: Proposed EmNet model – (a) model structure, and (b) trained 64 filter weights of the local convolution layer.

2.1. Feature extraction

The focus of this study is not on the extensive study of features themselves but on the usefulness of their temporal trajectory. Thus, we decided to use 20 features among the 88 features in the eGeMAPS [2] for simplicity. They are zero-crossing rate, log frame energy, frame energy entropy, spectral centroid, spectral spread, spectral entropy, spectral flux, spectral roll-off, Mel-Frequency Cepstral Coefficients (MFCC) C1 ~ C5, voicing probability, pitch, formant bandwidth, formant gain, and three harmonic energy ratios (ratio of log energy of the first harmonic to the log energy of the second through fourth harmonics).

These features are extracted from the input speech signal every 10 msec using a sliding 30-msec Hamming window, to form a time-sequence of 20-dimension feature vectors. To preserve temporal information, these feature vectors are used as the input to the network directly, instead of using their mean and standard deviation as an input.

2.2. Normalization

Following the typical feature normalization method used in the emotion recognition field [2, 8], the feature vectors are standardized (normalized) by the mean and standard deviation vectors obtained from the corresponding talker.

For easier use of keras deep learning package [13] for subsequent machine learning stages, the number of feature vectors along the time axis is either truncated or zero-padded to 512 frames, depending on the length of input speech utterance. The resulting feature vector is a 20x512 feature-time representation.

2.3. Local convolution layer

In image processing, the input of CNN is organized as a two-dimensional receptive field to capture the patterns along horizontal and vertical coordinates. However, this geometric locality is not applicable in EmNet structure. In the 20x512 input speech representation (Fig. 1 (a)), different features arranged along the vertical coordinate are different quantities and thus are not directly related to each other. For this reason, we build a local convolution layer with 1x6 filters, by which convolution operation is performed along the time axis only for each feature component. Then, a ReLU activation function is applied to the output of each filter to produce 64 feature maps (we found empirically that 64 filters produce good performance). Finally, max pooling with a pool size of 4 is performed for each feature map on the output of local convolutional layer.

Once trained properly, we expect that individual filters are tuned to detect important temporal patterns leading to the direction to improve emotion recognition accuracy. As an example, 64 filter shapes, trained on the database described in Sec. 3.1, are shown in Fig. 1 (b).

2.4. Global convolution layer

From the feature maps created by the local convolution layer, a global convolution layer is designed to extract higher-level information using a receptive field that spans across the 20-dimension features for two time frames (corresponding to 80 msec). We found that 128 is a good number for the number of filters, and ReLU activation function is applied to the output of filter. Max pooling is then applied on each feature map with a pool size of 2 to produce a temporal sequence of 128 features, as shown in Fig. 1 (a). Here, the time interval between consecutive features becomes 160 msec.

2.5. LSTM and feedforward layer

The output of the global convolution layer is processed by a 2-layer LSTM network with 48 cells each, where we used a 0.25 dropout rate. The output of the LSTM layers is fed to a dense/feedforward layer with softmax activation units to classify the input onto one of the 7 emotion categories.

3. Experiments

3.1. Database

The performance of EmNet is investigated on the Berlin Emotion Speech Database (EMO-DB) [8], which is one of the most widely used databases for emotion recognition. It contains 535 speech wave files and consists of 10 short sentences spoken by 5 female and 5 male talkers with acted emotion. Each file is labeled with one of the seven emotions: anger, happiness, sadness, neutral, boredom, disgust, and fear.

3.2. Model training and validation

Experiments are performed using Leave-One-Speaker-Out (LOSO) cross-validation, where the model is trained with the data of nine speakers and evaluated with the samples of the remaining one speaker. By changing the speaker, we can obtain 10 recognition rates, and the final performance metric is the average of these values [2].

We built a baseline static recognition system based on the same feature set used in EmNet to investigate the effectiveness of EmNet in incorporating temporal information. For each speech file, a 40-dimensional vector, obtained from the mean and standard deviation of the temporal sequence of the 20-dimensional feature vectors (described in Sec. 2.1), is used as an input vector to a SVM classifier.

The training of EmNet is performed based on the ADAM optimizer [10] using keras/theano, and the batch size is set to 64. It is impractical to run an exhaustive search for the optimal selection of parameters consisting of the combination of filter size, number of filters and pooling size of the local and global convolution layers and the number of LSTM cells. We thus explored the parameters one by one in a reasonable search range derived from domain-specific intuition, and used the best or top 2 parameter values in the search process of the next parameter. In total, 98 different network configurations were evaluated to find the best performing network configuration.

3.3. Results and discussion

Fig. 2 shows the recognition rate as a function of the number of free parameters of EmNet across the 98 trials. Eleven surpassed the state-of-the-art performance of 86% from the INTERSPEECH 2013 Computational Paralinguistics Challenge (ComParE) and a SVM classifier using 6373 features [1, 2]. The highest performance is marked as X.

Table 1 summarizes the reported performance of different models on EMO-DB. The baseline SVM system we built based on only 20 features (Sec. 3.2) achieves a 77.3% recognition rate, which is much worse than the 86.0% in [2], most likely because of the much smaller feature sets used (20 vs. 6373). The proposed EmNet model achieves as high as 88.9% recognition rate with the same 20 features. This is a 51.1% error reduction rate compared to baseline, demonstrating that the temporal information is utilized and outperforming the previous state-of-the-art.

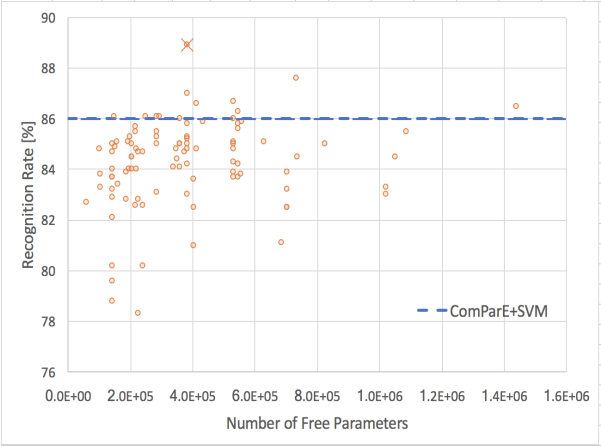


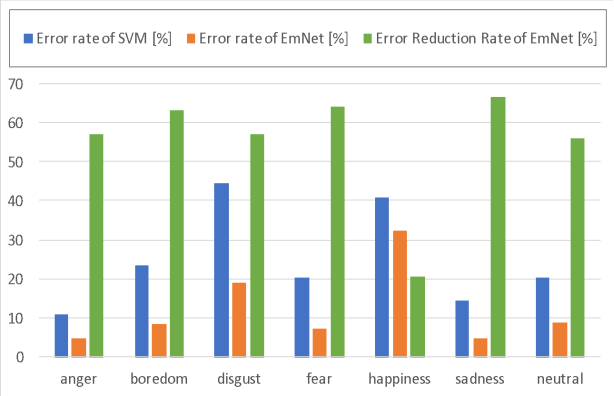
Fig. 2: Recognition rate versus the number of free parameters of EmNet.

Table 1: Emotion recognition results on EMO-DB

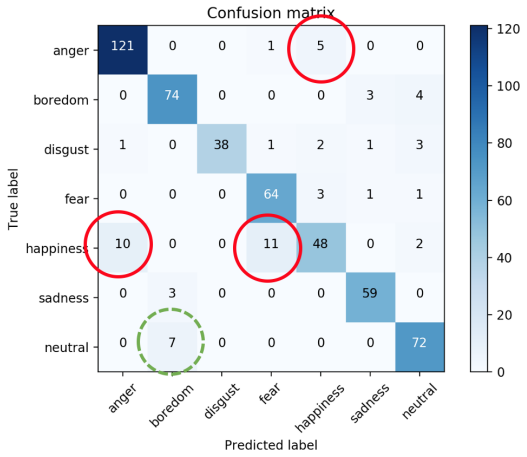
Model	Recognition Rate [%]
Chaspari, 2014 [7]	79.8
Kalinli, 2016 [10]	82.7
ComParE + SVM, 2016 [1, 2]	86.0
Lotfidereshgi, 2017 [11]	82.4
Baseline SVM (this paper)	77.3
Proposed model: EmNet	88.9

Fig. 3 (a) compares more in-depth results of EmNet, where we can see the effectiveness of the proposed model compared with the static baseline SVM for each emotional category. EmNet reduces recognition error rates significantly by more than 50-60% across all emotional categories except for happiness. The error reduction rate for happiness reaches only about 20%. The confusion matrix of recognition performance in Fig. 3 (b) provides a better insight into the misclassification of happiness. Almost half of the errors are due to the misclassification to anger and the other half to fear. Similarly, anger is sometimes confused with happiness. These errors, depicted as red circles, especially need to be reduced because they are more critical than the those between neutral and boredom (depicted as a green circle) in practical use.

Internal representations of EmNet lead to better visualizations of geometric relationships between different emotion states. Fig. 4 shows a 2-dimensional emotion space obtained by applying a t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to the LSTM output vector at the last time step of individual speech utterances. t-SNE is a useful data visualization tool that focuses on preserving the local distances of the original data and its low-dimensional projections [14]. It is interesting to see that high arousal emotions (happiness, fear and anger) are clustered together in close proximity, while valence states are difficult to separate—happiness (+) vs. fear and anger (-). Similar tendency was observed with different values of perplexity of t-SNE between 5 and 80, and the perplexity of 30 is used in Fig. 4.



(a)



(b)

Fig. 3: Detailed performance analysis of EmNet – (a) categorical performance comparison with the baseline SVM, and (b) confusion matrix.

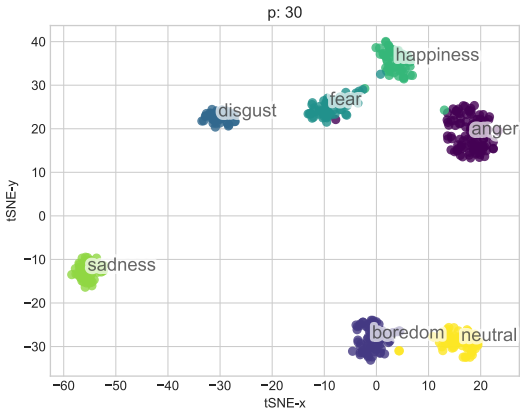


Fig. 4: t-SNE visualization of EmNet emotion space.

4. Conclusions

In this paper, we propose EmNet for emotion recognition of speech. The model combines 1) a feature extraction stage known to be useful for emotion recognition task and 2) a DNN to model the unknown mechanism in recognizing emotion status from the temporal sequence of feature vectors, where the DNN consists of two CNN layers for local and global convolution and LSTM layers. The proposed model, evaluated on the EMO-DB, demonstrates the state-of-the-art performance.

One important note is that although we followed the common practice of LOSO cross validation, it seems probable that we overfit the architecture because no separate development set was used. All other works might suffer from a similar problem where hyperparameters are effectively tuned on the test set. We suggest that collecting a new dataset large enough to support a separate development set would be a valuable contribution. Further remaining works include the use of extended feature set, e.g., 88 features in eGeMAPS [2] instead of the current 20 features, performance evaluation on different databases, and the performance evaluation of the end-to-end DNN approach [7] on EMO-DB.

5. References

[1] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani et al., "The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social Signals, Conflict, Emotion, Autism," in Proc. INTERSPEECH. Lyon, France, 2013.

[2] F. Eyben, K. Scherer, B. Schuller, J. Sundberg, E. Andre, C. Busso, L. Devillers, J. Epps, P. Laukka, S. Narayanan, and K. Truong, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," IEEE Transactions on Affective Computing, vol. 7, no. 2, 2016.

[3] I. Murray and J. Arnott, "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion," J. Acoust. Soc. Am., vol. 32, no. 2, pp. 1097-1108, 1993.

[4] G. Hinton, L. Deng, Y. Dong, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82– 97, November 2012.

[5] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in Proc. ICASSP, Brisbane, Australia, pp. 4580–4584, April 2015.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[7] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. ICASSP, pp. 5200-5204, 2016.

[8] T. Chaspari, D. Dimitriadis, and P. Maragos, "Emotion classification of speech using modulation features," in Proc. European Signal Processing Conference (EUSIPCO), 2014.

[9] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in Proc. INTERSPEECH, Lisbon, Portugal, pp. 1517–1520, 2005.

[10] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. ICLR, San Diego, USA, 2015.

[11] O. Kalinli, "Analysis of Multi-Lingual Emotion Recognition Using Auditory Attention Features," in Proc. INTERSPEECH, 2016.

[12] R. Lotfidereshgi and P. Gournay, "Biologically Inspired Speech Emotion Recognition," in Proc. ICASSP, 2016.

[13] F. Chollet, keras, in GitHub, GitHub repository, https://github.com/fchollet/keras, 2015.

[14] L. Van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," J. Machine Learning Research, vol. 9, pp. 2579-2605, 2008.