Emotion Recognition using Deep Neural Network

Devdarshan K R* , Mounica K V**,R Kumaraswamy* and Suryakanth V Gangashetty**
*Siddaganga Institute of Technology, Tumkur, India
Devdarshan06@gmail.com, hyrkswamy@sit.ac.in
**International Institute of Technology, Hyderabad, India
Mounica.kv@research.iiit.ac.in, svg@iiit.ac.in

Abstract: In this paper, we propose deep neural network (DNN) architecture with attention mechanism (DNN-WA) for emotion recognition (ER). DNN-WA is an utterance level classification mechanism which accounts for the long-term dependencies within an utterance unlike the conventional frame level classification. Mel-frequency cepstral coefficients (MFCC) are used to represent the emotion information within the spoken utterance. To incorporate additional temporal information at feature level, shifted delta cepstra (SDC) operation is performed on frame based MFCC features. The studies on ER are carried out using Berlin database. The results of our studies show that, the DNN outperforms the baseline GMM system indicating a better representation capability. Further, DNN-WA outperforms the DNN based system. From this, it is evident that DNN-WA indeed captures the contextual information which is essential for ER.

Keywords: Emotion recognition, deep neural networks, attention mechanism.

Introduction

Emotion recognition (ER) refers to the task of extracting the emotional states of a speaker using their speech segment [1], [2]. An ER system is necessary for any natural man-machine interactions. With the advent of internet and availability of technology, there are a plethora of applications that are coming up with human-machine interactions. An ER system could be used to keep a driver alert during the journey based on his mental state [3]. Computer tutorial applications can get more real and natural with an ER system. Speech emotion recognition can also be used in call centre applications and mobile applications. To effectively recognize the emotions from speech, features that represent emotional information within the utterance have to be extracted and modelled to be classified. Mel-frequency cepstral coefficients (MFCC) [4], linear prediction cepstral coefficients (LPCC) [5], real cepstral coefficients (RCC), pitch, formants, energy of a segment are some of the features that have been used for ER so far. MFCC, LPCC are the most popular spectral features while pitch, formants are the temporal features that play a major role in emotion recognition. Modulation spectral features that capture both spectral and temporal information have been explored in [6] for ER. Gaussian mixture models (GMM) have been used for classifying the speech samples into emotions [7]. In [8] and [9] hidden Markov models (HMM) were used as a classifier. In [10] support vector machines (SVM) were used for speech emotion recognition. MFCC and LPCC provide static information they do not provide any information about the trajectory of the signal. Speech tasks like speaker recognition (SR) and language identification (LID) have been using the trajectory information also called as the dynamic information as features. In [11] and [12], shifted delta cepstra (SDC) has been reported to capture the context over multiple frames. In [13], relation between SDC and prosodic features has been examined. The prosodic features considered in [13] are pitch and energy. Their results show correlation between SDC and prosodic features. In [13], prosodic features were proved suitable for ER, hence we explore SDC to check the validity of SDC in recognizing emotions from speech at feature level. In [14], state-of-the-art techniques applied for SR and LID has been explored for emotion recognition. Motivated by the success of deep neural networks (DNN) for speech recognition [15] and other speech related tasks, in this work, we propose to use a DNN architecture equipped with attention mechanism (DNN-WA) for emotion recognition. DNN-WA has been reported to outperform the stat-of-the-art iVector model for LID. In [16], DNN-WA has been used for LID task. The context vector built using DNN-WA is a vector of context summarization [17]. Similar in spirit with the idea of context summarization, in this work we explore DNN-WA for emotion recognition. Rest of the paper is organized as follows: In Section 2, we describe the database used for our studies. Section 3 gives a detailed description of the proposed method. Followed by experimental results in section 4. Finally, conclusions and scope for future work are presented in sections 5 and 6 respectively.

Description of database

Berlin Dataset was used for our experiments. The details of train and test examples are given in Table 1. There are four categories of emotions such as anger, fear, neutral and sadness.

Table 1: Description of the speech corpus used

Emotions	Train	Test
Anger	112	15
Fear	56	15
Neutral	64	15
Sad	47	15

Deep Neural Network with Attention

In this section, we describe the DNN architecture equipped with attention mechanism (see Figure 1). Attention mechanism was initially used in the fields of Neuroscience and Computational Neuroscience for neural processes involving attention [18], [19]. Later, attention in neural networks gained a wide popularity particularly in image recognition [20]. But only recently have they made their way towards Natural Language Processing (NLP) [20]. Recently, in [16], DNN-WA was used for a language identification task. Similar idea of focusing on specific parts of input has been applied in speech recognition, reasoning and visual identification of objects. Inspired from the success of DNN-WA in multiple domains, in this work, we propose to explore DNN-WA for emotion recognition.

Given an input sequence, $X = \{x1, x2, \cdots, xN\}$, a hidden layer representation $H = \{h1, h2, \cdots, hN\}$, is computed for all the N frames of an utterance by forward pass through regular DNN and attention is computed over hidden features.

$$H = [h1\ h2\ \cdots\ hN]$$
$$\gamma = tanh(WaH + ba)$$
$$\alpha = softmax(\gamma) \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The attention mechanism is computed using a single layer perceptron with the hidden layer representations H as input that learns the weight to be given for each representation. Figure 1 gives the description of the attention mechanism used. The attention model a (hn) thus computes an attention vector α that further helps in building a context vector c as given in Equations (1) and (2) where, Wa, ba are the weights and bias of the attention model optimized along with the DNN using backpropagation algorithm.

$$c = H\alpha \ldots\ldots (2)$$

Using Equation (3), the output of a DNN-WA is computed by transforming the context vector c using output layer weights and bias (U, bo) followed by a softmax operation is performed to normalize the values between zero and one.

$$y = softmax(Uc + bo) \quad \ldots\ldots (3)$$

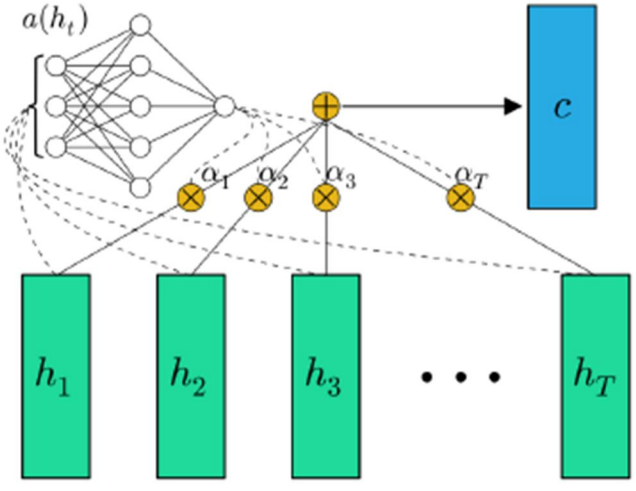

Figure 1:DNN-WA[20]

In this work, we used 1 and 3 hidden layers before the attention model. The final output y is a vector with probabilities of being each class. Thus we obtain a single decision per utterance unlike the conventional frame-based DNN where the average over all frame decision was considered as the final decision.

Experiment and Results

Extracting suitable features is one of the main aspects of the emotion recognition system. Mel-frequency cepstral coefficients (MFCC) are one of the most widely used features for both speech and emotional recognition. Dynamic information from MFCC could be obtained by including delta and delta delta features along with the static features obtained through short-term analysis of an utterance. An improved performance has been reported using SDC feature vectors created by stacking delta cepstra compounded across multiple speech frames in many speech tasks [11]. We validate the importance of SDC features for the ER task. The computation of SDC is illustrated in Figure 2. SDC features are specified by 4 parameters, N, d, P, k, where, N is the number of cepstral coefficients considered from each frame, d represents the time delay and time advance for delta computation. P represents the time shift between consecutive delta computed blocks and k represents the number of blocks, whose delta coefficients are concatenated together,

$$where\ i\ =\ 0,1,2..k-1.$$

$$\Delta c(t)\ =\ c(t\ +\ iP\ +\ d)\ -\ c(t\ +\ iP\ -\ d)\ \ldots\ldots\ldots(4)$$

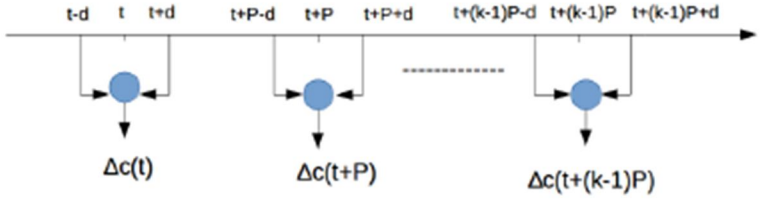

Figure 2: Computation of SDC feature vector at frame t for configuration N-d-P-k

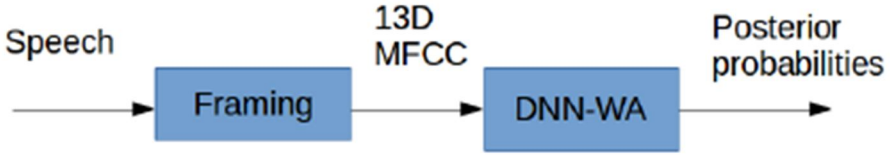

Figure 3: Block diagram of ER system using 13D MFCC features

Figure 4: Block diagram of ER system using MFCC dynamic features

Figure 5: Block diagram of ER system using shifted delta cepstral features

Given a spoken utterance, SDC feature vector is computed for every frame time t considering the context across P*k frames based on the configuration. For instance, a 7-1-3-7 configuration captures the context over 21 frames. We use an end-to-end DNN and a modified DNN with attention mechanism for the task of ER. Several experiments were conducted to select the best dimension and type of feature i.e static vs dynamic vs SDC. Block diagram of ER system based on static, dynamic and SDC feature extraction has been given in Figures 3, 4 and 5 respectively. From the observations, static MFCC features outperform the dynamic and SDC features for ER task as given in Table 2. Stacking multiple frames does not help in emotion recognition because emotion cannot sustain for a longer period. Using static MFCC features we compare our results as GMM vs DNN vs DNN-WA for ER task.

Table 2: Performance of Deep Neural Network with Attention mechanism (DNN-WA) Emotion Recognition system based on feature selection

Architecture	13D(in %)	39D(in %)	56D(in %)
700R500R200R100R	88.337	78.667	72
500R500R500R500R	90	81.667	79

GMM based Emotion Recognition

As a baseline system, we trained a GMM classifier using MFCCs as the acoustic features. GMM is a statistical way of modelling the feature vector in an unsupervised way for classification task. This model looks at the data as a linear combination of many Gaussian models. The GMM aims to model the data in terms of mean, variance and mixture coefficients. Expectation Maximization (EM) algorithm was used to train the GMMs. In this work we vary the mixtures up to 64 Gaussian components. Increasing the number of Gaussian components improved the performance by helping in better fitting the input data. The results are shown in Table 3.

Table 3: Performance of Gaussian Mixture Model (GMM) based Emotion Recognition

Number of Gaussian Components	Accuracy (in %)
2	70
4	70
8	81.667
16	85
32	86.667
64	86.667

DNN based emotion recognition

DNN

Initially, a series of experiments were performed to determine the depth and breadth of the DNN for ER. DNNs were initialized using Normalized initialization [21] and trained using the mini-batch stochastic gradient descent with classical momentum (SDG-CM). The input layer has based on the feature dimension, 13, 39 and 56 units for static, dynamic and SDC respectively. The output layer has four units each representing one of the four emotions i.e anger, happiness, sadness and neutral. The architectural choices and the resultant accuracy have been shown in Table 4.

Table 4: Performance of Deep Neural Network based Emotion Recognition

Architecture	Accuracy (in %)
DNN700R	71.667
DNN500R	75
DNN700R500R	75
DNN1500R500R	87

DNN-WA

From the experiments, as we observed that the dynamic features did not help over the static features, we further continued our studies using static MFCCs with DNN-WA. Input and output layer units dimension remains the same as used in DNN but with a single layer perceptron before the last hidden layer. This perceptron gets as input the hidden layer representations of an utterance and learns the weightage to be given to each frame's hidden representation. In this way a context vector for the entire utterance is built and hence the context summarization. This context vector is further used as input to the final hidden layer and then the output layer follows. The output is a single vector for an utterance that gives the posterior probabilities for each class. Table 5 shows the results obtained using DNN-WA for ER.

Although there is a little improvement in ER accuracy moving from GMM to DNN, a significant improvement is observed using DNN-WA. DNN-WA is capable of summarizing the entire utterance into a single context vector that further helps in giving an utterance level classification. Unlike other recognition tasks like speaker and language identification where the speaker and language information exists throughout the utterance, emotion does not sustain for a longer period. Thus, although stacking multiple frames did not help in identifying emotion, the context of an utterance over multiple frames helps

Table 5: Performance of Deep Neural Network with Attention mechanism (DNN-WA) based Emotion Recognition

Architecture	Accuracy (in %)
DNN700R	78.33
DNN500R	87
DNN700R500R	88.333
DNN1500R500R	90

in ER. In an ebullient situation, a person may not scream on top of his voice out of joy for long time, but the content of his conversation could reveal his state-of-mind. This kind of context summarization is achieved using DNN-WA for ER.

Summary and Conclusion

In this paper, we have explored the DNN-WA for performing the utterance level ER. Our results indicate that, the proposed architecture is well suited for the task of ER. Through our experiments, we conclude that, the SDC features or the dynamic features are not necessary for emotion recognition unlike other speech tasks like SR and LID. Thus, adding the context over multiple frames at feature level did not help in improving the performance of an ER system. Emotion cannot sustain for longer durations, and our results comparing the static and dynamic features prove the same. At classification level, using DNN-WA helps in context summarization utterance level. This is different context in a way that DNN here is trying to decide based on the content being spoken at different levels of abstraction rather than simply looking for the energy levels over few frames to identify the state of mind. We compare GMM, deep neural network and deep neural network with attention. DNN outperformed the GMM based ER system and DNN-WA in turn has better capability in capturing the context which further helps in emotion recognition.

Acknowledgment

The authors wish to thank Dr. R Kumaraswamy, Prof. Siddaganga Institute of Technology and Dr. Suryakanth V Gangashetty, Assistant Professor, IIIT Hyderabad.

References

[1] D. Ververidis and C. Kotropoulos, “Emotional speech recognition: Resources, features, and methods,” Speech Communication, vol. 48, no. 9, pp. 1162–1181, 2006.
[2] D. Ververidis, C. Kotropoulos, and I. Pitas, “Automatic emotional speech classification,” in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 2004, pp. I–593.
[3] 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 Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, vol. 1, 2004, pp. I–577.
[4] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357–366, 1980.
[5] B. S. Atal, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” the Journal of the Acoustical Society of America, vol. 55, no. 6, pp. 1304–1312, 1974.
[6] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech communication, vol. 53, no. 5, pp. 768–785, 2011.
[7] X. Cheng and Q. Duan, “Speech emotion recognition using Gaussian mixture model,” in Proc. The 2nd International Conference on Computer Application and System Modeling, 2012.
[8] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden markov models,” Speech Communication, vol. 41, no. 4, pp. 603–623, 2003.
[9] A. Nogueiras, A. Moreno, A. Bonafonte, and J. B. Mari˜no, “Speech emotion recognition using hidden markov models.” In Proc. INTERSPEECH, 2001, pp. 2679–2682.
[10] P. Shen, Z. Changjun, and X. Chen, “Automatic speech emotion recognition using support vector machine,” in Electronic and Mechanical Engineering and Information Technology (EMEIT), vol. 2. Proc. IEEE, 2011, pp. 621–625.
[11] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J. R. Deller Jr, “Approaches to language identification using Gaussian mixture models and shifted delta cepstral features,” in Interspeech, 2002.
[12] J. R. Calvo, R. Fern´andez, and G. Hern´andez, “Channel/handset mismatch evaluation in a biometric speaker verification using shifted delta cepstral features,” in Iberoamerican Congress on Pattern Recognition. Springer, 2007, pp. 96–105.
[13] J. R. Calvo, D. Ribas, R. Fern´andez, and G. Hern´andez, “Evaluation of lineal relation between shifted delta cepstral features and prosodic features in speaker verification,” in Proc. Iberoamerican Congress on Pattern Recognition. Springer, 2008, pp. 112–119.
[14] M. Kockmann, L. Burget et al., “Application of speaker-and language identification state-of-the-art techniques for emotion recognition,” Speech Communication, vol. 53, no. 9, pp. 1172–1185, 2011.

[15] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath et al., "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, 2012.
[16] Mounika.K.V, Sivanand Achanta, Lakshmi HR, Suryakanth V Gangashetty, and Anil Kumar Vuppala, "An investigation of deep neural network architectures for language recognition in Indian languages," in Proc. INTERSPEECH, 2016, pp. 2930–2933.
[17] K. Vesel`y, S. Watanabe, K. ˇ Zmol´ıkov´a, M. Karafi´at, L. Burget, and J. H. Cˇ ernocky`, "Sequence summarizing neural network for speaker adaptation," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 5315–5319.
[18] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.
[19] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," Annual Review of Neuroscience, vol. 18, no. 1, pp. 193–222, 1995.
[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in Proc. International Conference on Learning Representations(ICLR), 2014.
[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in Proc. International conference on artificial intelligence and statistics, 2010, pp. 249–256