Speech Emotion Recognition Using CapsNet

Sukanya K S, Leya Elizabeth Sunny

Abstract*: Extraction of emotion features is the key to emotion recognition from speech. Capsnet is an emerging neural network technology which gives better performance over convolution neural networks in feature extraction. This is the system which implement a speech emotion recognition system using Capsnet. With the gradual development of the new generation of man-machine interaction technology speech emotion recognition has attracted wide research attentions. In facing with the development trend of new technologies, speech interaction is go-ing to penetrate into thousands of households. Traditional machine learning method has achieved great progresses in speech emotion recognition. However, there are some problems: first, which features can reflect the differences between different emotions and the second, these artificially designed features rely highly on database and have low generalization ability. It takes long time to extract fea-tures from the speech. Deep learning can extract different layers of features from the original data through automatic learning. Capsule Network or Capsnet, is composed of a number of capsules in each layer as the name indicate. Each cap-sule is a group of neurons who work together to get a specific outcome for the capsule. Speech emotion recognition works based on the spectrogram constructed from the voice record. A spectrogram is the plot of the spectrum of frequencies of sound as they vary with time.

Index Terms: Capsnet, Neural networks, Speech emotion recognition..

I. INTRODUCTION

The fast and best regular way of communication between human is Speech. This reality inspire many investigators to reflect speech signal as a rapid and active process to interact among computer and human. It means that the computer should have sufficient information to recognize human voice and speech. The recognition of emotional speech is done by finding the condition of emotion of a person by providing their voice directly to the system. Speech emotion recognition is typically favorable for applications, which need human-computer communication. It is a big challenge to find the emotional condition from the speech signal. The first reason is that all methods are taking the finest features, which is sufficient to differentiate among dissimilar emotions. The existence of accent, sentences, various language, speaking style also makes additional trouble since there is some straight change in the extracted features. Moreover, it is probable to have more than one emotion at the same time in a single speech. Every emotion associate with a distinct fragment of speech. Hence, defines the boundaries among fragments of emotion is very puzzling task. An Artificial Neural Network is a group of associated units named artificial neurons, which model the neurons in an original brain. Each association of neurons can convey a signal from one neuron to another one. A neuron who gets the message or signal can process it and then indicate nearby neurons that are linked to it. In ANN implementations, normally the signal at the link between artificial neurons is a real number, and the output of each one is calculated by specific non-linear function of the sum of its inputs. The links among the neurons are termed as 'edges'. Artificial neurons and edges naturally have a weight which is adjusted as the learning process grows. The weight adjusts the strength of the signal at an edge. Artificial neurons will be having a threshold so that the signal is sent only if the cumulative signal crosses it. Usually, these neurons are combined into layers. Distinct layers perform distinct types of alterations on the inputs. Signals moves from the input layer, which is the first one to the last, output layer probably going through the other layers many times. The aim of introducing ANN was to imitate the human brain in solving the problems. But, as the time files it was clear that an artificial brain cannot be efficient as the real one. Then the researchers begun to use ANN for specific tasks which was quiet successive. The proposed method considers a Capsule network [1] or Capsnet for the speech emotion recognition. A Capsule Neural Network is a machine learning system which comes under ANN that are capable of building the model for categorized relations. The approach was to imitate the real neural network of brain. In capsnet, the idea is to introduce one more level of abstraction called capsules to the convolutional neural network (CNN), and to use output from these capsules in one layer to produce a constant input form for capsules in higher layers. The outcome is a vector containing the probability and a pose of an observation.

II. LITERATURE SURVEY

In traditional methods, speech emotion recognition rely on extraction of speech emotional features and mainly focus on extracting spectral features [2], prosodic features [3], and other features [4-5]. Recently Li Zheng and his colleagues proposed a new network model (CNN-RF) [6] based on convolution neural network combined with random forest. Initially, the convolution neural network is used as the feature extractor to extract the speech emotion feature from the normalized spectrogram, used random forest classification algorithm to classify the speech emotion features. As the result Nao robot can "try to figure out" a human's psychology through speech emotion recognition and also know about people's anger, happiness, sadness and joy, achieving a more intelligent human-computer interaction. The spectrogram was calculated from emotion speech samples through framing, windowing, short-time Fourier transform (STFT) and power spectral density (PSD) , and Speech emotion features were extracted by CNN and the output of CNN Flatten layer was input into RF[7]

Revised Manuscript Received on December 22, 2018.

Sukanya K.S. , Computer Science, Mar Athanasius College of Engineering, Kothamangalam, India.

Leya Elizabeth Sunny, Computer Science, Mar Athanasius College of Engineering, Kothamangalam, India.

classifier as eigenvectors of speech emotion samples. Considering the present growth in classifiers for assessments in numerous modalities we can say that nowadays deep neural networks (e.g. [8], [9]) has got great attention. Additionally, these networks are also used in the emotion recognition from speech. Particularly for the explanation of pictorial affects a deep neural network forms shows incredible outcomes, specifically the Convolutional Neural Network (CNN). The method of speech emotion recognition by Spectral Estimates [10] studied CNNs for the recognition of emotions from the voice signals. Working on spectrograms is a best idea, but on different architectures. Especially, in CNNs a spectrogram denotes the visual form corresponding to the acoustic samples protecting the emotion content in the audio fragment. For this, CNNs are the best to be used on these samples.

Most of the works on speech emotion recognition using CNNs is done on EmoDB[11] . In that, the work is drawn-out to another benchmarks corpora[12] , namely eNTERFACE[13] and SUSAS[14]. In CNNs an internal illustration of the learned informations can be visualized, which will show how each network understands the input. This method by Google is called deep dreaming and is particularly used to highpoint portions of input images for which the CNN is sensitive. Human vision avoids unwanted information to guarantee that only a small portion of the optic array is always handled at the peak resolution. Each layer in the network will is fragmented to number of neurons groups called “capsules”. The work of the neurons in an active capsule denotes the several characteristics of a specific entity that existing in the input. These characteristics will be different types of instantiation parameter such as pose (orientation, size, and position), velocity, deformation, etc. A way to denote presence of an entity is by using a discrete logistic unit whose output is the probability that the entity exists. Here the total length of the vector of instantiation parameters is used to denote the presence of the entity. It should be ensured that the length of the output from a capsule shouldn’t exceed 1. It is accomplished by using a non-linearity function which makes the vector orientation unaffected but decreases its value. The dynamic routing by agreement is working smoothly since the output of the capsule is a vector with the probability of having a particular entity in the input. Primarily, the output vector is directed to all probable parents then it is scaled down by coupling coefficients which sums to 1. For all parent, the capsule calculates a “prediction vector” by multiplying its output with a weight matrix. If this prediction vector has a big scalar product with the output of a possible parent, there is top-down feedback which increases the coupling coefficient for that parent and decreasing it for other parents. This will increase the contribution that the capsule makes to that parent thus further increasing the scalar product of the capsule’s prediction with the parent’s output. This is called “routing-by-agreement” which is extra efficient when compared to routing implemented by max-pooling. The max-pooling routs all information to all neurons which is a wastage of energy.

III. PROBLEM DEFINITION

Speech emotion recognition is recognizing the patterns. So the stages in the pattern recognizing systems will also be present in this system in one another form. Recognizing the emotional condition from a speakers voice fragment is a most challenging task due to the given reasons: first, it is not clear that, to differentiate among the emotion, which feature should be taken. Another is the same speech may process dissimilar emotions. Every emotion may match to the distinct parts of the spoken word. So it is very difficult to distinguish these parts of sound. Another problem is the way of expressing the emotion condition is depending on the speaker and their culture and environment. As these two factors are changed, the speaking style will also be changed. There may be two or more types of emotions, long term emotion and transient one, so it is not clear which type of emotion the recognizer will detect. Capsnet will be a solution to all these problems. Routing by agreement allows the capsules to focus on particular emotions. It will make better accuracy on speech emotion recognition while comparing with the former neural networks.

IV. PROPOSED SYSTEM

The emotion recognition form speech is very important in man-machine interactions and it is an interesting area of development. The key to speech emotion recognition is extraction of speech emotion features. In capsule networks it replaces the scalar-output feature detectors of CNNs with vector-output capsules and max-pooling with routing-by agreement. To achieve this the last layer of capsules is made to be convolutional. The proposed method of speech emotion recognition by using capsnet will provide better performance according to convolutional neural network, since the max pooling concept is replaced. Max pooling have the disadvantage that a feature is shared with all the neurons present in the layer. The capsule network is based on a concept that, an information should be forwarded to only the relevant person. Capsnet is composed of number of capsules, where each capsule is a group of neuron. Here the max pooling is replaced by routing-by-agreement.

The vector output capsules helps to implement the routing by agreement which ensures that the output of a capsule will be routed to only the correct parent in the higher layer. Initially, the output is directed to all probable parents but is scaled down by coupling coefficient that sum to 1. For each probable parents, the capsule calculates a “prediction vector” by multiplying its own output by a weight matrix. If this prediction has a high scalar product with the output of a possible parent, there is top-down feedback which increases the coupling coefficient for that parent and it will be decreased for others in the same layer. This will increment the contribution to that parent by the capsule from the layer below and also further increases the scalar product of the capsule’s prediction with the parent’s output. This type of “routing-by-agreement” is more operational than the primitive form of routing implemented by max-pooling. A spectrogram is a visual representation of the spectrum of frequencies of sound as they vary with time. This spectrograms of each voice clips will be fed into the capsule network. The capsnet is designed to have

three layers. First the Convolution layer which extracts features from the spectrogram. Second Primary capslayer, which is a group of capsules. Each capsule finds a matching with the extracted features. Third Emotcapslayer, it have 6 capsules since I have considered 6 emotions angry, neutral, surprise, happy, fear and sad.

The length of the vector output of the capsule represents the presence or possibility of existence of an entity or a feature in the input fed to that capsule. Thus it use a non-linear "squashing" function to confirm that small vectors get shrunk to almost zero and long vectors get shrunk to a length slightly below 1. Then permits it to discriminative learning to create good use of it.

$$O_j = \frac{||T_j||^2}{1+||T_j||^2} \frac{T_j}{||T_j||^2} \quad (1)$$

Where O_j is the vector output of capsule j and T_j is the total input. Except the first layers of capsule, the whole input to a capsule T_j is the weighted sum of all "prediction vectors" $\hat{u}_{ij}$ from the capsules in the layer below and is produced by multiplying each output u_i of a capsule in the layer below by a weight matrix

$$T_j = \sum_i c_{ij} \hat{u}_{ji} \quad , \quad \hat{u}_{j/i} = W_{ij} u_i \quad (2)$$

Where the c_{ij} is coupling coefficients that are founded by the routing by agreement process. The coupling coefficients between capsule i and all the capsules in the layer above sums to 1 and are found by a "routing softmax " whose initial values b_{ij} are the log prior probabilities that capsule i should be coupled to capsule j.

$$c_{ij} = \frac{\exp(b_{ij})}{\sum_k \exp(b_{ik})} \quad (3)$$

The log priors can be learned at the same time as all the other weights are learned. They depends on the location and type of the two capsules but not on the current input. The initial coupling coefficients are then iteratively changed by measuring the agreement between the current output O_j of each capsule j, in the layer above and the prediction $\hat{u}_{j/i}$ made by capsule i.

$$a_{ij} = O_j . \hat{u}_{ij} \quad (4)$$

This agreement will be added to b_{ij} before computing the new values for all the coupling coefficients linking capsule i to higher level capsules.

The routing between each layer works by Routing by agreement between capsules.

Routing algorithm.

1: procedure Routing ($\hat{u}_{ij}$, r, l)
2: for all capsule i in layer l and capsule j in layer (l + 1) : b_{ij} ← 0.
3: for r iterations do
4: for all capsule i in layer l: c_i ← softmax(b_i) .
5: for all capsule j in layer (l + 1): $T_j \leftarrow p_i \, c_{ij} \, \hat{u}_{ji}$
6: for all capsule j in layer (l + 1): v_j ← squash(T_j) .
7: for all capsule i in layer l and capsule j in layer (l + 1): b_{ij} ←b_{ij} + O_j.
return O_j

The proposed Capsnet architecture can be visualized as

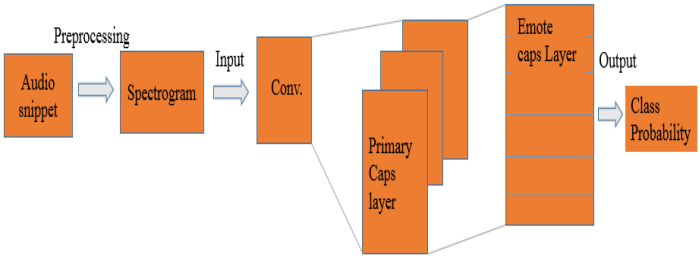

Fig 4.1: Architecture of the proposed system

The proposed system of speech emotion recognition works based on the routing algorithm which is considered for routing by agreement. The system contains three layers
1. Input layer: Which is a convolution layer which extracts the features from the given spectrograms.
2. Primary capslayer: Which is a group of capsules that routs based on routing by agreement. The spectrograms of specific emotion are routed to specific capsules.
3. Emotcapslayer: Which have 6 capsules representing the six emotions considered here ats as output layer, where the output is the prediction that the spectrogram or audio snippet comes under which emotion.

The input is preprocessed for converting the audio snippet into the visualized format i.e. the spectrogram. A spectrogram is a plot of frequency to the time of an audio snippet. Spectrogram generation can be done with Fourier transform or by using band pass filters. Then this is given as the input to the capsule Network. The capsule network work based on the three layers as described above. The system will accurately predict the emotion of an individual based on the audio snippet. The speech emotion recognition is based on the spectrogram generated from the audio snippet so the system can accurately predict the emotion. This will be grate goal in man machine interaction as we are moving to a world where machines think as man.

V. RESULT

The speech emotion recognition system is implemented to find the emotion of the speaker from the audio. The system succeeded to find emotion from the audio. The training and testing of the Capsnet is done using the datasets EmoDB and Ravdess. On testing the system correctly predicted the emotion of the speaker. Some of the screenshots are :

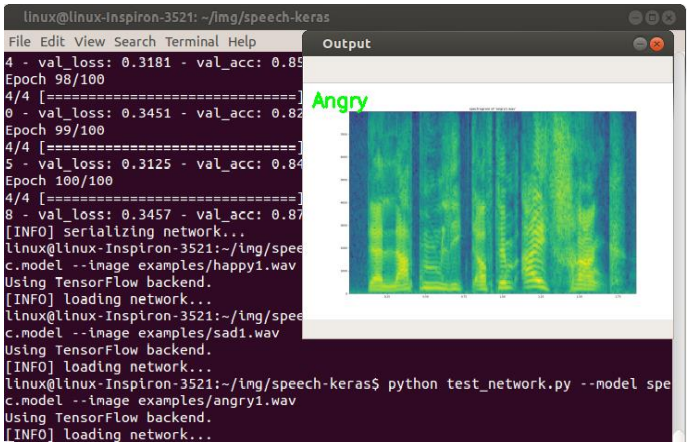

Fig 5.1: Detecting emotion Angry

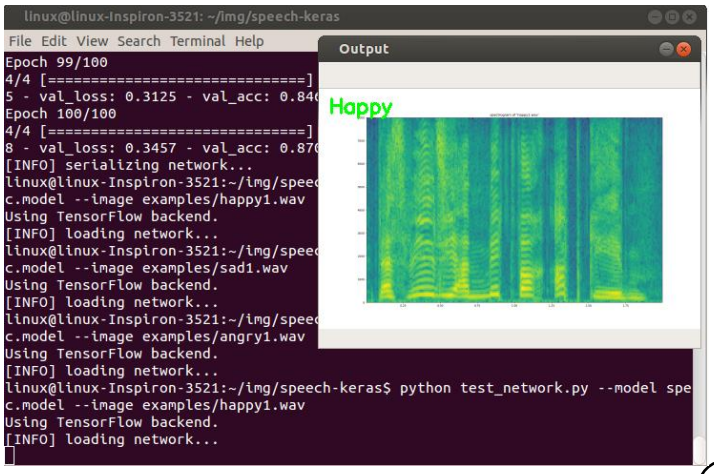

Fig 5.2: Detecting emotion Happy

VI. CONCLUSION

For the last thirty years, the speech recognition used hidden Markov models with Gaussian mixtures as output distributions. Then it started to use ANN which have the ability to learn and model non-linear and complex relationships. As they can generalize after learning from the initial inputs and their relationships, it can infer unseen relationships on unseen data as well, thus making the model generalize and predict on unseen data. So the trend stuck on CNNs. Now Capsules advantages the old methods by replacing the max-pooling with routing by agreement. The proposed system of speech emotion recognition rely on Capsule networks. Swapping the unwanted routing in max-pooling with Routing by agreement. Routing by agreement will route the information only to the needful capsule so that avoiding the spreading of unwanted information among all.

REFERENCES

1. Sara Sabour, Nicholas Frosst, Geoffrey E.Hinton “Dynamic Routing Between Capsules” (NIPS 2017).
2. Juan Pablo Arias, Carlos Busso, Nestor Becerra Yom. Shape-based modeling of the fundamental frequency contour for emotion detection in speech[J]. Computer Speech & Language, 2014, 28(1):278-294.
3. Yongming Huang, Guobao Zhang, Yue Li, et.al. Improved Emotion Recognition with Novel Task-Oriented Wavelet Packet Features [J]. Intelligent Computing Theory, 2014, 8588:706-714.
4. H.M.Teager Sc.D, S.M.Teager. Allen. Evidence for nonlinear production mechanisms in the vocal tract [J]. Speech Production and Speech Modelling, 1990, 55:241-261.
5. Chong Feng, Chunhui Zhao. Voice activity detection based on ensemble empirical mode decomposition and teager kurtosis [A]. International Conference on Signal Processing[C]. USA: IEEE, 2014:455-460.
6. Li Zheng, Qiao Li, Hua Ban , Shuhua Liu, “Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest” 978-1-5386-1243-9/18/$31.00 c 2018IEEE
7. Liaw A, Wiener M. “Classification and regression by random Forest “ news, 2002, 2(3): 18-22.
8. G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. 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, 2012.
9. L. Deng, G. Hinton, and B. Kingsbury, “New types of deep neural network learning for speech recognition and related applications: an overview,” in Proc. of 2013 Intern. Conf. on Acoustics, Speech & Signal Processing. Vancouver, Canada: IEEE, 2013, pp. 8599–8603.
10. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, “A database of german emotional speech,” in Proc. of INTERSPEECH-2005. Lisbon, Portugal: ISCA, 2005, pp. 1517–1520.
11. B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, “Acoustic emotion recognition: A benchmark comparison of performances,” in Proc. of the ASRU 2009. Merano, Italy: IEEE, 2009, pp. 552–557.
12. O. Martin, I. Kotsia, B. Macq, and I. Pitas, “The eNTERFACE’05 audio-visual emotion database,” in Proc. of the Workshop on Multimedia Database Management. Atlanta, USA: IEEE, 2006, s.p.
13. J. Hansen and S. Bou-Ghazale, “Getting started with SUSAS: A speech under simulated and actual stress database,” in Proc. of EUROSPEECH-1997

AUTHORS PROFILE

Sukanya K. S. is currently studying M-Tech degree with Computer Science specialization at Mar Athanasius college of Engineering under Kerala Technological University. The main area of research is Machine learning and Artificial Intellegance. She got her BTech in Computer Science and Engineering from Malabar college of Engineering and Technology under Calicut University.

Leya Elizabeth Sunny is currently working as Assistant Professor in the Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala. She was finished M-Tech degree in the field of Information Systems Security.