Available online at http://scik.org

J. Math. Comput. Sci. 11 (2021), No. 4, 4535-4551

https://doi.org/10.28919/jmcs/5760

ISSN: 1927-5307

OPTIMIZED RECURRENT NEURAL NETWORK BASED EMOTION RECOGNITION USING SPEECH FOR ASSAMESE LANGUAGE

NUPUR CHOUDHURY[1,*], UZZAL SHARMA[2]

[1]Department of Computer Science & Engineering, Assam Don Bosco University, Guwahati 781017, India

[2]Department of Computer Applications, Assam Don Bosco University, Guwahati 781017, India

Abstract: This paper deals with the design and development of a dataset for emotion recognition in Assamese language and its analysis using Machine Learning. This dataset comprises of spoken sentences in Assamese language which relates 7 different emotions of the speakers. It also focusses on the analysis of the dataset with Deep Learning architectures and classification algorithm like Recurrent Neural Network (RNN). In total these 7 emotions included calm/neutral, anger, sad, fear, disgust, happy and surprise. The performance of RNN is improved using Representation Learning where training of the model is done using the glottal flow signal to investigate the effect of speaker and phonetic invariant features on classification performance. The speech samples were experimented with different combinations of features and a variety of results were obtained for each of them. In this paper 2 experiments are performed to improve the performance of RNN training. The first one is the representation learning where the training is modelled on the glottal flow signal which is used to investigate the effect of the speaker and phonetic distinct features on the performance of the classification. The second is the transfer learning based RNN training which is done on the valence and activation that is adapted to a 7-category emotion classification. On the Assamese dataset the

*Corresponding author

E-mail address: nupur.choudhury@dbuniversity.ac.in

Received March 23, 2021

experimented approach results in a performance which is comparable and similar to the existing state of art systems for emotion recognition using speech.

Keywords: recurrent neural network; representation Learning; transfer learning; deep neural network; autoencoders etc.

2010 AMS Subject Classification: 62M45.

1. INTRODUCTION

All references in the list must be cite in the Text (by number(s) in square brackets), e.g., Zhao [1], Pecaric, Perusic and Vukelic [2]. As deep neural networks have evolved in past decade, significant progress is witnessed in the domain of pattern recognition as well as various problem-solving approaches specially in the field of paralinguistics. In order to facilitate further research a range of neural network architectures like CNN, autoencoder networks, Long Short-Term Memory (LSTM) models [1]. Variety of research works have demonstrated selected properties of these multiple networks for speeches [2] with minimum human interaction [3]. However, most of these models utilize generic features as input features like the spectral features, frequency, pitch, formants and features related to energy and then apply classification in order to predict various emotions which have displayed robustness for a wide variety of speech domains [4][5]. Some of the works also mention about conversion of the raw and unprocessed data directly into input signal which leads to improved performance. Existing approaches primarily depend upon utilisation of acoustic features which are taken as standards like pitch, MFCC (Mel-Frequency cepstral coefficients), Energy etc. in addition to that the temporal features of the samples are extracted using statistical functions that would be used as segment descriptors for speech segments or detection of utterance of from emotional speech [6]. This stage is followed by classification of the samples with various classifiers like Support Vector Machines (SVM), Hidden Markov Models (HMM), Artificial Neural Networks, Deep Neural Networks etc. In this work limited efforts are being given in achieving an effective representation learning for emotion recognition using speech where training a neural network requires a temporal waveform or a spectrogram. The features that would represent

the signal are extracted by sampling the waveform or from the spectral representation [7]. The unseen data is well generalised without additional requirement for feature extraction. Moreover, the entire process of data collection and annotations is expensive which has led the researchers to experiment with limited labelled data and a few numbers of volunteers for the same. Various training paradigms like transfer learning [8] and semi supervised learning [9] has been experimented with to improve the accuracy of classification, but even then, extensive research is needed as expression of emotion through a language form would also require the language of expression and not all classifier work in the same manner for different types of dataset that is used even if the methodology remains same. In this paper, a novel dataset is used which is based on the Assamese Language. Assamese is also known as Asamiya is the official language of northeastern region of Assam, India and is derived from Indo-Aryan language. Speech Spectrogram features as well as spectrogram related to glottal volume velocity is also experimented with. It includes if the performance of classification can be improved by removing unnecessary variation factors like identity of the speaker and phonemes from the speech signals. For this a model is derived known as Bidirectional Long-Short Term Memory (BSLTM)- RNN that deals with emotions at the utterance level and makes use of features which are extracted by the training of stacked denoising autoencoders from the spectrograms.

A study of the scenario based on transfer learning is also done where the additional utterances are leveraged that are not labelled using the emotions that are being experimented on by training up an RNN based on the activation labels and valence that would consist of all the utterances which are present in the dataset. Finally, it is followed by adaptation of a network that is trained to the classification of the mentioned emotions. There is also an attempt of learning the affect of salient features for Speech Emotion Recognition (SER) by making use of Convolutional Neural Networks (CNN). This entire process involves 2 processes. The first step is to identify the local invariant features by using the samples that are not labelled along with a type of sparse autoencoder which is based on penalization at reconstruction. The next step involves using the local invariant features for feature extraction which would be used to learn noticeable effect, distinct features which makes

use of an objective function that emphasizes on noticeable features, discrimination and orthogonality. The experimental results on the new dataset shows that the approach results in well-built and robust performance in terms of recognition even in complicated scenarios as well as proves to be better than several other approaches from this domain.

The experiments in this paper depend upon initial work in the domain of Deep Neural Network (DNN) and representation learning which typically deals with affect and emotion recognition. Jaitly et. Al [10] used datasets like Arctic and TIMIT and transformed autoencoders for learning acoustic event from them. Graves et. Al [11] used RNN in their experiments and received a test error of 17.7% using TIMIT. Deep Learning approaches have been one of the primary domains for research in this area. Kim et. al worked on IEMOCAP dataset using audio-visual emotions and multimodal deep learning methods. The same dataset is explored by Han et.al [12] by combining Extreme Learning Machine and Deep Learning where they achieved 20% relative accuracy as compared to similar approaches. Lee et. al [13] used RNN in IEMOCAP dataset and modelled each frame as a random variable sequence where the improvement of the weighted accuracy as compared to a DNN is around 12%. In this work investigation is being done to find out if the speaker and the invariant phonetic representations represents distinct emotion and if the transfer learning could improve the performance of the classification

2. MODEL

2.1. INITIAL TRAINING WITH DENOISING AUTOENCODERS

A neural network which is trained to Learn the distributed, lower dimensional representation of the input samples is known as autoencoder [16]. In this process a feed forward neural network is used which has 1 hidden layer having the activations

$$\{y_i\}_{i=1}^{i=N} \text{ where } y_i = \tanh(Wx_i+a) \qquad (1)$$

The input dataset comprises of N samples which is represented by $\{x_i\}_{i=1}^{i=N}$. Output given by the encoder is represented as

$$Z_i = \mathrm{W}^{`\prime} y_i + b' \quad (2)$$

which is generated from the hidden layer. The training of the encoder is done via backpropagation. The training is done using Sum of squared loss which is represented using

$$\mathrm{L} = \sum_{i=1}^{i=N} ||x_i - Z_i||^2 \quad (3)$$

Vincent et.al worked with denoising encoders where a fraction of the elements is set to 0 that leads to a corrupted data point x_i which produces input $\widetilde{x_i}$ and sets the original x_i to Z_i which is then reconstructed using the autoencoder. Here a distinct stacking approach is used for the autoencoders which is pyramidal in nature and the number of neurons is generally halved or are significantly lesser in number. The size of the output layer generated by the autoencoder remains same so that the feature sets follow a similar methodology.

2.2. BLSTM-RNN CLASSIFICATION

RNN are generally used for experimenting with temporal data and their correlations. However they suffer from a problem known as vanishing gradient problem which is directly proportional to the length of the training samples. In-order to address this problem Hochreiter et al [17] proposed Long Short Term Memory(LSTM-RNN) which models the dependencies of the long term time series data. In this paper BLSTM-RNN(Bidirectional Long Short Term Memory) Long Short Term Memory is used with replication scheme for the targets for classification. Here each time step is taken into consideration and is assigned to the category of emotion for every speech utterance. The category of predicted emotions for the input sequences is obtained through a voting scheme where the majority votes are considered for the predictions related to each time step in the utterance sequence related to the frames of context. This RNN consists of 2 layers where size of each cell is 40 and having a 4 D softmax layer on top of it for classification of emotions. This is followed by experiments on validation to identify the best possible model in which the hyper parameters are extracted by random sampling on the following segments:

- Batch size: 1786 utterances
- Learning Rate : [8e-6,6e-6,2e-5,1e-5,3e-5]

- Momentum : [0.6,0.5,0.6]
- Input noise variance : [0.0,0.1,0.2,0.3]
- Weight noise variance : [0.0,0.05,0.1,0.15,0.2]
- No. of Epochs : 200

In every epoch, random noise is added to the input sample sequences and model weights that can be controlled using distinct noise hyperparameters.

3. DATASET AND FEATURE EXTRACTION

For the experiments in this paper, a novel dataset created by Nupur et.al[18] that involves our prior work and is based on Assamese Language is used which has been validated and verified using standard measure and methodologies through certified professionals and speakers. The dataset is generated in a simulated and elicited manner where a team of speakers which belongs to Assamese origin contributed to the work. It consists of around 1750 audio samples each of around 3-6 seconds speech with short periods of silence before and after the utterance. It is collected from 25 human speakers (10 male and 15 female) and is labelled with 7 emotional annotators like Neutral/calm, Happy, Sad, Fear, Angry, Surprise and Disgust along with their respective dimensional labels like Activations and Valence. These standard categories are chosen as they are the most common forms of emotional categories. The valence, activation and dominance are basically annotated using a Likert Scale (1-5) where the various ratings of dimensions are averaged across the 3 different annotators. The performance of the experiment would be analysed in comparison to the improvisational utterances of the same dataset. These utterances are spontaneous and nearly equivalent to the natural speech and is generally does not have much influence from the predefined script. Some of the samples of the predefined script includes the following samples from table 1.

Emotion	Sentences	English Translation
Neutral/calm	মই ভালে আছো । তোমাৰ কি খবৰ । অলপ ফুৰিব যাব উলাইছো ।	I am good How are you? I am about to go somewhere
Happy	বাহ চাকৰী পাই গলো । তোমাৰ বিয়াৰ কথা শুনি খুব ভাল লাগিছে । এইটো বিৰাট ভাল খবৰ ।	Wow I got the Job I am happy to hear about your marriage This is great news
Sad	মোক যে বিৰাট দুখ দিলা । ই পৰীক্ষাত ফেইল কৰিলে । মনটো কিবা বেয়া লাগি আছে ।	You have hurt me tremendously He failed in his exams I am not feeling good today
Fear	ইমান ৰাতিকে কেনেকৈ যাও, ভয় লাগি আছে । আন্ধাৰ ৰুমটোৰ ভিতৰত কি কি আছে । মোৰ বেমাৰ হব যেন লাগি আছে ।	How do I go there late at night, I am scared. What is there inside the dark room. I think I will fall sick.
Angry	মোক সি গালি পাৰিছে । কি বনাইচ্ এইবোৰ তয়ে খাঁ । সি কিয় চিঞৰিব মোৰ উপৰত ।	He scolded me What rubbish have you cooked, you eat yourself. How dare he shout at me.
Surprise	সেইটো চুন ধুনীয়া পখিলা । কি কোৱা হে, মই নাজানো । তুমি কেতিয়া আহিলা ।	Such a beautiful flower What are you saying! I don't know! When did you arrive?
Disgust	ইছ্ ইমান লেতেৰা মানুহজন । চেহ, তেণ্ডুলকাৰ আকৌ শূণ্যত আউট হল। ধেই নোৱাৰি খাব এইবোৰ ।	The person is very dirty. Ohho, Tendulkar again got out at 0. Yuck, this is not edible.

Table 1. Assamese Speech Samples from the dataset along with their corresponding English representations.

3.1. ANALYSIS OF GLOTTAL SOURCE WAVEFORM

Affective as well as paralinguistic attribute like valence, activations and emotion should not be dependent on the speaker or the phonetics and should not have any sensitivity towards the identity of the speaker or the content of utterance. Investigation would also be done for improving the performance of classification by filtering out the various factors responsible for variations like the identity of the speaker or the information related to phonetics from the signal before considering the training of the denoising encoder along with the BLSTM-RNN. The glottal source waveform generally possesses this property and is achieved through glottal inverse filtering of speech samples that uses Iterative Adaptive Inverse Filtering algorithm [19]. While this method might be an approximation of the actual glottal waveform and might result in a biased form of outcome, The

IAIF algorithm is used because of its wide applications in literature. Moreover the motivation also comes from the performance of the features based on glottal flow like Normalized Open Quotient and Quasi-open Quotient and their success related to assessment of depression[20] and classification of voice quality[21].

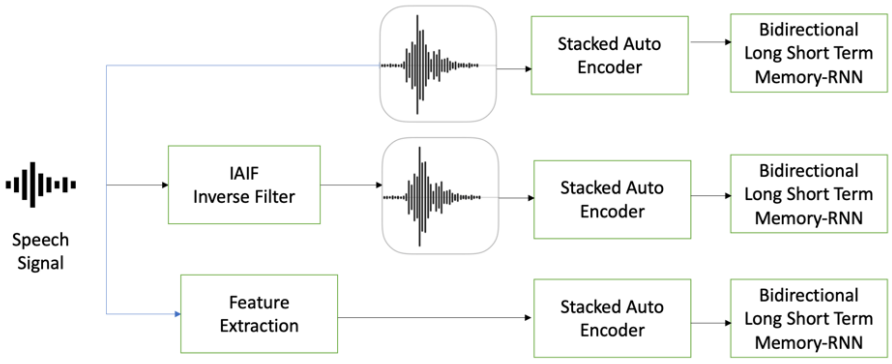

Figure 1: Experimental setup for BLSTM-RNN

3.2. REPRESENTATION OF THE SPECTROGRAMS AND BASELINE FEATURE EXTRACTION

The spectrograms are generated using (a) the waveforms of the speech and the (b) waveforms generated from the glottal flow. The width of the frame used is 15ms along with a frame overlap of 5ms for carrying out the extraction. These spectrograms comprises of 128 Fast Fourier Transform bins corresponding to each and every frame and Subsequently a context window is stacked using 4 adjacent Frames together that results in 512-dimensional vector in-order to capture the context based information in a better way. This 512 -D feature vectors are used as an input to the stacked autoencoder that is represented using the architecture 512-256-128-52 motivated from [22][23] and reports that performance of the emotion classification. A frequency domain log scale is used as more emphasis on the low frequencies proves to be more significant in terms of auditory perception. In the baseline feature extraction common features like Mel Frequency Cepstral Coefficients(MFCC), Fundamental Frequency, Normalized Amplitude Quotient, Quasi Open Quotient, LPCC and pitch is extracted. This is followed by a generation of a vector of size 160 by stacking 25 feature vectors using 4 adjacent frames which is then provided as input to the stacked autoencoder having an architecture of 160-64.

4. METHODOLOGY

The entire samples in the dataset are split into 5 different sessions where every session comprises of a certain scripted sentences pertaining to the emotional categories between 3 different speakers. The experiments in this paper are done in a strategy which similar to leave out one[14]. As there are 15 speakers in the dataset. As there are 15 speakers in the dataset and every session consisted of 3 speakers. Therefore in ever fold speech samples from 12 speakers i.e. 4 sessions would correspond to training set and from the session that was not included 1 speaker was chosen for hyperparameter validation while the testing is performed on the other 2 and this is done in a circular manner. Results based on weighted and non-weighted classification accuracy for the complete testing dataset(scripted and improvised) as well as a subset which comprises of improvised samples only have been reported. The accuracy which considers the accuracy taking into consideration all the testing speech samples of the dataset is known as weighted accuracy whereas the average accuracy considering every emotional category pertains to the non or unweighted accuracy. For every session, pretraining using stacked encoder is done on all the training set speech samples. The emotion classification work is done for the signals that are labelled with the primary considered emotions by using BLSTM-RNN. Figure 1 shows a snapshot of the entire approach. The work that has been done is compared with the following standards:

i. Deep Neural Network-Extreme Learning Machine method mentioned in [13] where experiments are done on training a Deep Neural Network with an Extreme Learning Machine.
ii. Recurrent Neural Network-Extreme Learning Machine [14] where experiments are done on training a Recurrent Neural Network with an Extreme Learning Machine.
iii. Work done by Jin et. Al [15] where higher level representations are using a fusion of lexical and acoustic features. Here some standard features and techniques like Bag of words Modelling (BoW) is also used. However, for the acoustic features results are reported to avoid any biasness.

iv. The experimented model where the features are extracted and are stacked in order to generate context descriptors consisting of 4 frames instead of generating spectrograms.

In this paper the classification of emotions is categorized into 7 different categories-Neutral, Happy, Sad, Disgust, Angry, Surprise and fear. There are in total 1786 speech utterances which are labelled with valence and activation intensities. Sun et. al [24] presented a detailed view on the corelation of the dimensional affect and the categories of emotion are extensively researched upon. Further research has been carried out in the current work in terms of investigating if these features derived from representation learning for dimension regression is also considering the selection of emotion classification. For each of the speech signals in the dataset, an aggregate score related to the activation dimensions and valence is calculated by making an average of all the labelled ratings followed by training of a BLSTM-RNN network corresponding to a regression model. Similar to the classification of emotion, replication of at each step is done for the targets where each of them is a 2D vector that consists of the activation score and valence for the speech sample. The network architecture resembles the BLSTM classifier having an exception of a 2D feedforward linear unit in top layer. The network is trained using the Sum of Squared Errors and eventually the topmost layer is replaced with softmax layer. In the adaptation stage the training is again resumed and here instead of activation and valence, the target results in an emotional annotation which would belong to one of the 7 different emotional categories. Transfer Learning provides a perspective to the relation between the emotions and dimensional effectiveness along with the enabling of the RNN that makes use of additional unannotated samples.

5. RESULTS AND DISCUSSIONS

Classification experiments are conducted and the performance of the experiments are compared based on the representations derived from the speech and spectrograms based on glottal flow. Table 1 represents leave-one-session-out setup for various emotions and features. According to the results, Happy and Angry categories have near similar accuracy and are often confused whereas the sad category has the most accurate performance.

The authors in [14] has reported the highest unweighted and weighted accuracies which are resulted by the Deep Neural Network-Extreme Learning Machines and states to be 52.13% and 57.91%. In the work the accuracies without the ELM are 56% and 57.91% respectively which are near comparable with the accuracy reported in this paper.

Speech Category	**Feature Set**	**Weighted**	**Unweighted**	**Neutral**	**Happy**	**Sad**	**Fear**	**Angry**	**Disgust**	**Surprise**
Overall	Traditional Features	48.22	49.10	49.56	42.50	65.60	39.90	43.40	38.70	44.50
	Spectrogram (speech)	**52.80**	**54.20**	**53.20**	**53.20**	**72.50**	**53.70**	**53.70**	**51.10**	**52.60**
	Spectrogram (Glottal)	**57.20**	**58.40**	**57.10**	**57.30**	**74.50**	**57.30**	**56.40**	**55.20**	**55.70**
Improvised	Traditional Features	49.82	50.10	50.50	50.20	59.20	52.20	53.40	42.50	45.60
	Spectrogram (speech)	**54.60**	**55.50**	**54.70**	**54.70**	**67.80**	**54.70**	**55.10**	**49.10**	**47.30**
	Spectrogram (Glottal)	**59.10**	**60.22**	**59.60**	**59.60**	**62.40**	**59.30**	**60.20**	**50.50**	**51.60**

Table 2: Test based accuracies for various feature sets on overall and improved speech utterances. The results in bold represent the proposed experiment accuracies

In [14] it was not clearly mentioned as to which speaker was utilized for testing or validation. In the experiments sone in this paper 1 speaker is allotted for validation while the other 2 were allotted for testing. The experiments were repeated by switching the speaker responsibilities and the average of the performance of all the test sets were considered for evaluation. In [14] the approach also performs experiments on the utterances that are improvised whereas the current work focusses both on the improvised as well as the utterances that are scripted in nature. The authors of [15] have reported an acoustic approach where they made use of 10 fold leave one speaker out validation accuracies on the scripted as well as improvised utterances. However they do not specifically do the evaluation on a testing set. The reported accuracy which is 49% for weighted accuracy to a range of 55.4% for the feature fusion and is comparable to the accuracy

done in the validation phase i.e. 57%. These results represents that the classifications could be done even form lower level spectrograms. The results generated from the glottal flow signals performs better than the results generated from the speech signals for the weighted category 4.2 % which proves that it is more advantageous to eliminate the identity of the speaker and information related to phonetics before classification. For more detailed evaluations the confusion matrix is presented in Figure 2 and Figure 3. This represents the confusion related to the labelled emotions with other emotions for the prediction. Since Happy and Angry has relatable acoustic characteristics they have major performance related to the accuracy. It has been observed that the glottal flow representations reduces the confusion with the Happy and Angry classes to a great extent. The improvisation is 2.3% for the Happy category and 14.44% for the Angry category and is influenced from the results reported in [25]. Moreover they share a similar activation level when their location is considered in correspondence to valence activation. It is believed that the representations using glottal flow are not affective in confusion to Angry and Happy since they extract the differences on the dimension of valences in a better way. Based on these findings it can be concluded that the speech based representation learning could be improved by filtering out the factors like speaker identity and phonetics before the classification process.

	Happy	**Neutral**	**Sad**	**Fear**	**Angry**	**Disgust**	**Surprise**
Happy	38.2	26.7	15.5	18.2	32.4	28.7	29.6
Neutral	37.4	55.6	21.3	19.2	33.7	31.2	28.9
Sad	19.2	43.2	69.3	22.5	22.5	21.4	32.2
Fear	28.7	34.5	29.2	55.2	33.4	42.1	47.4
Angry	25.6	41.7	32.5	44.1	50.5	23.7	31.2
Disgust	31.2	35.5	37.1	32.1	28.7	47.2	22.6
Surprise	29.4	27.9	33.4	38.1	23.5	37.2	49.7

Table 3: Confusion Matrix for Speech representation

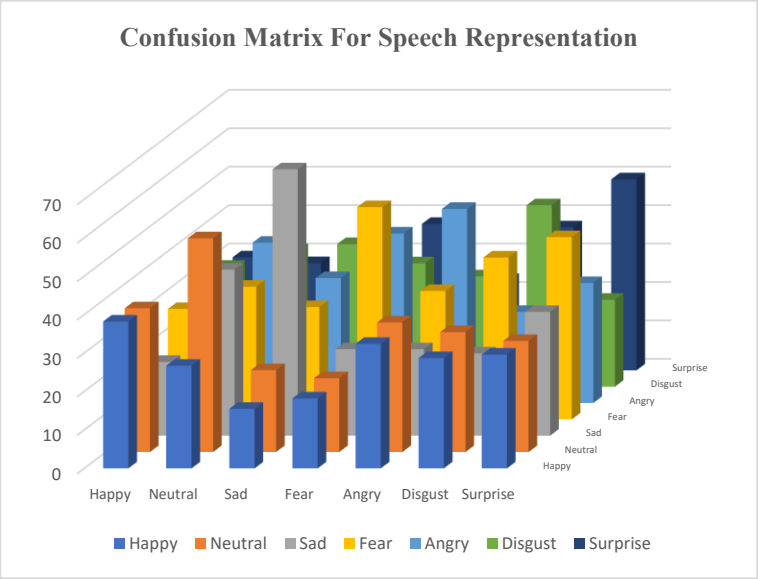

Figure 2: Representation of Confusion Matrix for Spee

	Neutral	Happy	Sad	Fear	Angry	Disgust	Surprise
Neutral	39.5	28.8	17.5	19.2	35.5	29.9	31.7
Happy	39.5	65.8	27.7	26.8	35.8	33.5	30.6
Sad	22.5	44.7	72.3	25.6	28.6	26.5	34.5
Fear	31.6	37.5	31.2	58.4	35.4	44.3	49.7
Angry	29.8	45.3	33.5	45.3	54.3	29.7	33.5
Disgust	34.7	36.4	38.5	38.2	31.2	50.5	29.7
Surprise	30.5	29.5	35.2	40.4	24.5	38.1	52.3

Table 3: Confusion Matrix for Glottal representation

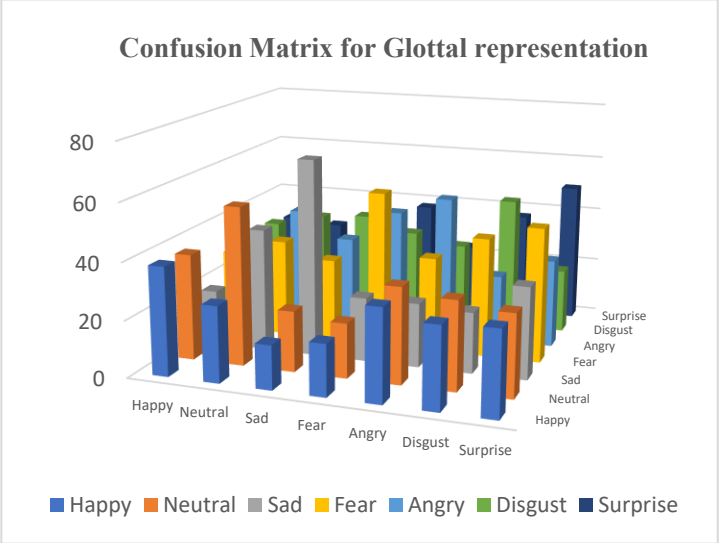

Figure 3: Representation of Confusion Matrix for glottal representation

In order to address the question of whether data insufficiency could be addressed through transfer learning using some affective and significant attributes Like valence and activation regarding emotions the mentioned BLSTM-RNN is pretrained to be a regression model for the same regarding the entire training set and then fine tune it for the 7 category speech emotion recognition. It has been observed that it is necessary that the BLSTM weights in context to the pre trained network could improve the performance on the adaptation task. In order to finetune it, the validation of the hyper parameters over a grid needs to be done. As the best performance among the competing features is shown by the glottal flow spectrogram representation the transfer learning is done using only those representations. This results in a weighted accuracy of 54.65% and an unweighted accuracy of 58.92% considering the test set having category based emotions as 40.20% (Happy), 57.20% (Angry), 68.45% (Sad), 48.50% (Surprise), 42.55% (Fear), 44.50% (Disgust) and 51.20% (Neutral). The representations using transfer learning improves the accuracy by 2.55% for weighted and 0.52 % for unweighted accuracy respectively. However it does not show a significant impact in comparison with the direct learning of emotions from the glottal flow representations.

6. CONCLUSIONS

In this paper extensive research has been done on the investigation of representation learning using the spectrograms of speech and the glottal flow signals. The experiments done in this paper demonstrates that the features extracted from representation learning are discriminative of the classification of emotions and can be compared with the state of art approaches. It has also been found that inverse filtering which results in filtering out the speaker and phonetic information reduces the confusion between the overlapping categories like Happy and Angry emotions. Also minimal improvement has been noticed in the performance related to the transfer learning from valence and activation regarding the emotion categories. Overall the findings are encouraging in context to improvement of performance in the system involving multiclassifiers due to the diversity in the resulting errors. In future more elaborate and extensive transfer learning

experiments could be done with much larger datasets.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Netw. 61 (2015), 85–117.

[2] G. Hinton, L. Deng, Y. Dong, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, Signal Proc. Mag. 29 (2012), 82– 97.

[3] A. Graves, N. Jaitly, Towards end-to-end speech recognition with recurrent neural networks, In: Proceedings of the 31st International Conference on Machine Learning, PMLR 32(2) (2014), 1764-1772.

[4] F. Yu, E. Chang, Y.-Q. Xu, H.-Y. Shum, Emotion Detection from Speech to Enrich Multimedia Content, in: H.-Y. Shum, M. Liao, S.-F. Chang (Eds.), Advances in Multimedia Information Processing — PCM 2001, Springer Berlin Heidelberg, Berlin, Heidelberg, 2001: pp. 550–557.

[5] G.-B. Huang, Q.-Y. Zhu, C.-K. Siew, Extreme learning machine: Theory and applications, Neurocomputing. 70 (2006), 489–501.

[6] B. Schuller, S. Steidl, A. Batliner, The Interspeech 2009 emotion challenge. In: Proc. INTERSPEECH (2009), pp 312–315.

[7] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013), 1798–1828.

[8] R. Raina, A. Battle, H. Lee, B. Packer, A.Y. Ng, Self-taught learning: transfer learning from unlabeled data, in: Proceedings of the 24th International Conference on Machine Learning - ICML ’07, ACM Press, Corvalis, Oregon, 2007: pp. 759–766.

[9] M.F.A. Hady, F. Schwenker, Semi-supervised Learning, in: M. Bianchini, M. Maggini, L.C. Jain (Eds.), Handbook on Neural Information Processing, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013: pp. 215–239.

[10] N.Jaitly, G.E. Hinton, http://www.cs.toronto.edu/~fritz/absps/capsules_speech.pdf

[11] A. Graves, A. Mohamed, G. Hinton, Speech recognition with deep recurrent neural networks, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Vancouver, BC, Canada, 2013: pp. 6645–6649.

[12] Y. Kim, H. Lee, E.M. Provost, Deep learning for robust feature generation in audiovisual emotion recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Vancouver, BC, Canada, 2013: pp. 3687–3691.

[13] K. Han, D. Yu, I. Tashev, Speech emotion recognition using deep neural network and extreme learning machine, Proc. INTERSPEECH (2014), pp. 223–227.

[14] J. Lee, I. Tashev, High-level feature representation using recurrent neural network for speech emotion recognition. In: Proc. INTERSPEECH (2015), pp. 1537–1540.

[15] Q. Jin, C. Li, S. Chen, H. Wu, Speech emotion recognition with acoustic and lexical features, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, South Brisbane, Queensland, Australia, 2015: pp. 4749–4753.

[16] P. Baldi, Autoencoders, unsupervised learning, and deep architectures. In: Proceedings of ICML workshop on unsupervised and transfer learning, (2012), pp 37–49.

[17] P. Vincent, H. Larochelle, Y. Bengio, P.-A. Manzagol, Extracting and composing robust features with denoising autoencoders, in: Proceedings of the 25th International Conference on Machine Learning - ICML ’08, ACM Press, Helsinki, Finland, 2008: pp. 1096–1103.

[18] N. Choudhury, U Sharma, Emotion recognition in standard spoken Assamese language using support vector machine and ensemble model, Indian J. Computer Sci. Eng. 12 (2021), 148-158

[19] P. Alku, Glottal wave analysis with Pitch Synchronous Iterative Adaptive Inverse Filtering, Speech Commun. 11 (1992), 109–118.

[20] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T.F. Quatieri, A review of depression and suicide risk assessment using speech analysis, Speech Commun. 71 (2015), 10–49.

[21] S. Scherer, J. Kane, C. Gobl, F. Schwenker, Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification, Computer Speech Lang. 27 (2013), 263–287.

[22] Y. Kim, E.M. Provost, Emotion classification via utterance-level dynamics: A pattern-based approach to

characterizing affective expressions, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, Vancouver, BC, Canada, 2013: pp. 3677–3681.

[23] D. Le, E.M. Provost, Emotion recognition from spontaneous speech using Hidden Markov models with deep belief networks, in: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, IEEE, Olomouc, Czech Republic, 2013: pp. 216–221.

[24] R. Sun, E.I. Moore, Empirical study of dimensional and categorical emotion descriptors in emotional speech perception, in Twenty-Fifth International FLAIRS Conference, 2012.

[25] R.J. Davidson, K.R. Scherer, H. Goldsmith, Handbook of affective sciences, Oxford University Press, Oxford, 2003.