Enhancing the Technique of Speech Emotion recognition using Feature Learning

Mrs. T. Sunitha #1, *K. Poojitha #2, S. Venkata Rakesh #3, G. Bhanu Swetha #4, B. Kasi Priyanka #5*

#1 Associate. Professor, #2,3,4,5 B.Tech., Scholars
Department of Computer Science and Engineering,
QIS College of Engineering & Technology

Abstract-The human voice can be characterized by several attributes such as pitch, timbre, loudness, and vocal tone. It has often been observed that humans express their emotions by varying different vocal attributes during speech generation. This paper presents an algorithmic approach for detection of human emotions with the help speech .The prime objective of this paper isto recognize emotionsin speech and classify them in 6 emotion output classes namely angry, fear, disgust, happy, sad and neutral. The proposed approach is based upon the Mel Frequency Cepstral coefficients (MFCC) uses Crema-D database of emotional speech. Data Augmention is perfomed on input data audio file,such as Noise, High Speed, Low Speed etc. are added, thus more the varied data is available to the model better the model understands. Feature extraction is done using MFCC and then the extracted features are Normalized(for Independent Variable), Label Encoding(for Dependent Variable(for SVM,RF)),One Hot Encoding(for Dependent Variable(for CNN))is done. After thisthe datasetis divided into Train, Test and given to different models such as Convolutional Neural Network(CNN),Support Vector Machine(SVM), Random Forest(RF) for Emotion prediction. We report accuracy, f-score, precision and recall for the different experiment settings we evaluated our models in. Convolutional Neural Network(CNN) was found to have the highest accuracy and predicted correct emotion 88.21%ofthe time. Hence, deduction of human emotionsthrough speech analysis has a practical plausibility and could potentially be beneficial for improving human conversational and persuasion skills.

I. INTRODUCTION

The human voice is extremely adaptable and conveys a huge number of feelings. Feeling in discourse conveys additional understanding about human activities. Human discourse passes on data and setting through discourse, tone, pitch and some such qualities of the human vocal framework. As humanmachine cooperations advance, there is a need to brace the results of such communications by preparing the PC also machine communicates with the capacity to perceive the feeling of the speaker. Feelings assume an essential part in human correspondence. To broaden its job towards the human-machine cooperation, itis attractive for the PCs to have a few inherent capacities for perceiving the unique passionate conditions of the client [2,5].Today, a lot of assets and endeavors are being placed into the improvement of man-made reasoning, and savvy machines, allfor theprimary reason for improving on human existence. Research studies have given proof that human feelings impact the dynamic interaction partially [1-4]. On the off chance that the machine can perceive the hidden feeling in human discourse, it will bring about both valuable reaction also correspondence. To convey effectivelywith people,the frameworks need to comprehend the feelings in speech. Therefore, there is a need to foster machines that can perceive the paralinguistic data like feeling to have powerful clear correspondence like people. One significant information in paralinguistic data is Emotion,which is conveyed along with discourse. A great deal of AI calculations have been created and tried to group these feelings conveyed by discourse. The mean to foster machines to decipher paralinguistic information,

similar to feeling, helps in human-machine cooperation and it assists with making the connection more clear and normal. In this concentrate on various characterization models, for example, CNN,SVM,RF are utilized to predictin discourse sample.The MFCC is utilized for the component extraction .To prepare the model CREMA - D dataset was utilized alongside Data Augmention.

II. RELATEDWORKS

The task of speaking recognition is split into numerous subtasks in traditional ASR systems (Fig.1-conventional approach), each of them optimized individually. In [12, 9], an end-to-end strategy for acoustic modeling was presented, which includes both the characteristics and the classifier. A feature-learning phase consisting of multiple layers of convolution and a classification phase consisting of fully connected (FC) layers (also referred to as the multi layer perceptron (MLP)) and an output layer is the basis of the CNN-based end-to-end acoustic modeling approach as shown in the figure 1.

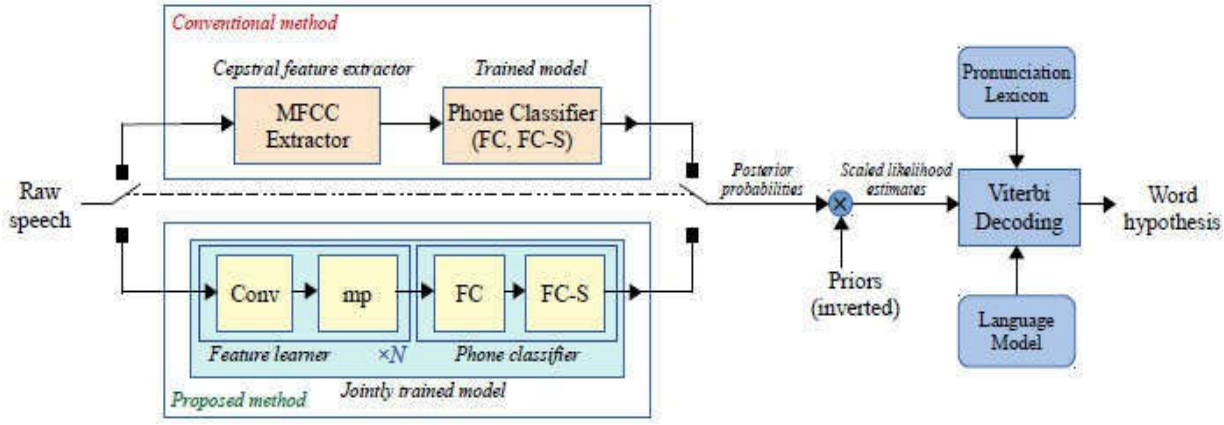

Fig. 1: ASR system flow illustrating the conventional and proposed methods.

The hyper parameters of the system include: (i) the window sizeof the speech input (w_{seq}), (ii) the number of convolution layers N,(iii) for each convolution layer i ∈ f{1..N}, kernel width kW_i,kernel shift dW_i, number of filters nf_i and maxpooling size $mp_i$and (iv) the number of hidden layers in the MLP.All these hyperparameters were identified via cross validation in the original paper. This technique also influences how quickly the input talk is processed. In particular, the first kernel layer width of the convolution layer (i.e. kW_1) and the kernel shift (i.e. dW_1) are respectively the frame size and frame shift that work on the signal. Figure 2 shows the processing of the first layer of convolution. Note that the frame rate of the system is determined by the shiftof input speech window of size w_{seq}, which was fixed to 10 ms, asdone conventionally.

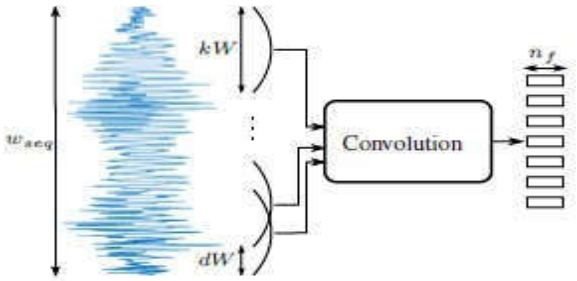

Fig. 2. Illustration of first convolution layer processing.

In [9] the first convolution modelling of the "sub-segment," i.e. a 2ms signal that is smaller than one pitch, was identified. After analysing the filters with two distinct approaches, spectral dictionary based analysis[12] and back propogation-based analysis[13], the CNN learned to model forming frequency information for post-probability assessment of the phone.In addition, this method has been proven to provide equivalent performance or better than the traditional cepstral functional system with fewer parameters.This study will take use of these two characteristics, namely the automated functional learning and the less parameters of systems, to enhance children's ASR systems' performance.

III. CONFIGURATION SETTING

This section discusses the databases and protocols first and then the created systems.

Datasets

For children's speech experiments, we utilised PF-STAR [14] and for adult speech WSJCAM [15]. Both data sets have utterances captured using two microphones in British English. PFSTAR is a big vocabulary dataset containing 140% of the speakers. It comprises 158 children aged between the ages of 4 and 14 years old. WSJCAM0. For PF-STAR ASR we utilised BEEP [16] lexicon. We have utilised the standard BEEP lexicon protocol for WSJCAM0, supplemented with CMU dictionary pronunciations for invisible words.

Data from both the recorded channels - head mounted microphones (denoted channel A) and microphones from far-field (denoted channel B) - were utilised to train models for tests with PF-STAR, since this was partly possible for overcoming data shortages. The evaluation/adaptation data of PF-STAR is used as a cross-validation set for neural network training. We provide results independently on the A and B channels of test data.

For experimentation, standard WSJCAM0 training (train), development (dev) and test sets were employed. In decoding of WSJCAM0 utterances, Standard 20k trigram LMs of WSJ corpus have been utilised.The PF-STAR language model has been developed as follows: one LM is from the Witten-Bell-Smoothing training set and another one from Witten-Bell Smoothing with standard MGB-3 text [17]. In order to remove the lower probabilities with 10^{-8} as a threshold, the LMs of the two have been interpolated linearly by weights chosen based on their concerns in the cross-validation set PF-STAR (explained above).

GMM-HMM systems

To train all GMM-HMM systems was Kaldi's toolkit[18] for usage.We have monophonous, triphonic and LDA+MLLT, as well as LDA+MLLT+fMLLR+SAT, using conventional training system procedures. The sheet nodes were limited to up to 2,500 nodes and 15,000 Gaussians for context-dependent clustering in all systems. Then, SGMM systems with 2500 leaf nodes, 9,000 substates and 400 mixes per state were trained.

Table 1. CNN architectures. N_f : number of filters, kW: kernelwidth, dW: kernel shift, mp: max-pooling.

Model	Layer	Conv			mp
		n_f	kW	dW	
CNN3	1	80	30	10	3
	2,3	60	7	1	3
CNN4	1	200	30	5	4
	2,3,4	100	7	1	2
CNN5	1	200	30	5	4
	2	100	9	1	2
	3	100	8	1	2
	4	100	7	1	2
	5	100	6	1	2

DNN-HMM systems

Keras[19] was used to train all neural networks using Tensorflow [20] backend. The feature utilised was 429-size MFCC 13-size CMVN with 11-screen splicing and associated coefficients. The DNNs, referred to as DNN1 and DNN3, consisted of 1 or 3 hidden layers, each with 1024 nodes, followed by a softmax output layer with activation with a rectified linear unit (ReLU). Monophone DNNs were intended for single-phone states, whereas SGMM clusters were intended for the triphone systems.The systems were trained using the alignments from the relevant systems. The Glorot uniform distribution technique was used to initialise the DNN parameters, default in Keras. Training took place on a stochastic gradient descent with a cross-entropical loss, where everything except the last layer dropped by 20 percent. When cross-validation loss stopped decreasing, the learning rate was half in the 10^{-1} to 10^{-6} range.Scaled up by priors (computed from goals used for training) and used to decode or forcibly align neural networks in Kaldi.The HMM state transition probability was derived from the GMM-HMM system they were learned from during decoding. During decoding The DNN training was followed by an alignment procedure utilising the DNN-HMM system, as monophone system alignments were poor. Then the DNNs were randomly re-exercised. It's been repeated twice.

CNN-HMM systems

Keras-Tensorflow was used to train the CNNs. Raw voice signals were shown in 250ms chunks with a 10ms shift. Each segment was removed mean (by its scalar average) and normalised before the CNN was fed. Table 1 shows the architectures of the CNN. Each CNN contains a single completely connected hidden 1024 node layer, followed by a softmax output FC layer. A 20% dropout was applied to the concealed FC layer. For the training of CNNs were utilised the central labels of the segment, based on the training alignments. The training processes for the DNNs were identical.

IV. RESULTS AND DISCUSSION

On kid speech test set (Channels A and B), Table 2 displays word error rates (WER) by using models trained in speech and additional speech by the adult. We note that CNN systems regularly perform best or better than their counterparts GMM/HMM and DNN/HMM. The SGMM systems also profit from data shortages and decoding of multipasses to produce competent results. It must be noted that 11.99% WER is the best-reported PF-STAR corpus [21, 22], to the best of our knowledge.The effect of integrating kid data into adult ASR on WER is seen in Table 3. We see that it lowers performance by adding child voice data.

Table 2. Comparison of WER on children test data with childrenmodels and children+adult models.

Model trained on →		Children data		Added adult data	
Children test set →		A	B	A	B
mono	GMM	17.84	19.27	18.43	20.63
	DNN1	15.67	16.63	15.88	17.69
	DNN3	15.84	17.21	15.62	17.60
	CNN3	**15.09**	**15.63**	15.12	16.72
	CNN4	16.21	16.13	15.68	16.90
	CNN5	17.35	17.00	15.82	17.37
tri	SGMM	13.18	14.64	12.38	14.54
	DNN1	14.65	15.52	14.77	16.28
	DNN3	15.54	16.34	14.37	16.41
	CNN3	13.25	13.87	**11.99**	14.42
	CNN4	14.09	14.40	12.49	14.40
	CNN5	13.43	14.21	12.24	**13.77**

Table 3. Comparison of WER on adult test data with adult models and adult+children models, showing the effect of adding children data on adult speech recognition.

Model trained on →		Adult data		Added children data	
Adult test set →		dev	test	dev	test
mono	GMM	28.28	28.27	28.84	29.04
	DNN1	15.60	15.69	18.27	18.01
	DNN3	13.12	13.18	14.63	14.37
	CNN3	14.96	14.12	16.91	16.18
	CNN4	13.99	13.68	15.74	15.04
	CNN5	14.32	13.80	16.14	15.43
tri	SGMM	9.10	9.44	9.32	9.56
	DNN1	10.98	10.64	11.53	11.80
	DNN3	9.66	9.29	10.30	10.44
	CNN3	10.83	10.24	12.09	11.44
	CNN4	10.31	9.70	11.51	11.08
	CNN5	9.93	9.53	10.85	10.55

In [12], it was proposed to comprehend the information represented on the first convolution layer of the spectral dictionary. The technique was used for understanding the spectral information modelled on the CNNs in other research, such as [23] and [24]. The spectrum reaction of the filters to the input language is determined in this way:

(1) $s_t^c$was taken as the input speech segment. For the sake of simplicity,a window size of 30 ms similar to the one used in standardshort term processing is used in our analysis.

(2) Successive windows of kW samples (30 samples for all models)interspaced by dW samples (10 samples for CNN3, 5 samplesfor CNN4 and CNN5 models) are taken froms_t^c

(3) For each of these successive window signals (s_t), the outputs ofthe filters y_t to the input speech signal $S_t = S_t\text{-}(kW\text{-}1)/2 \ldots S_t + (kW\text{-}1)/2$ are estimated as

$$y_t[m] = \sum_{l=-(kW-1)/2}^{l=+(kW-1)/2} f_m[l].s_{t+l} \qquad (1)$$

where f_m denotes the m^{th} filter in first convolution layer and$y_t[m]$ denotes the output of the m^{th} filter at time frame t.
The frequency response S_t of the input signal s_t is estimated as

$$S_t = |\sum_{m=1}^{M} y_t[m].\mathcal{F}_m|, \qquad (2)$$

Based on the confusion matrix in [25], the subset of telephones and speakers were chosen. The 30-ms-Frame from the steady-state area of/and/of the boy speaker displays spectral response in Figure 3 (b23). The formant values are often consistent with the range in the data set. In various vowels and speakers we saw comparable tendencies.

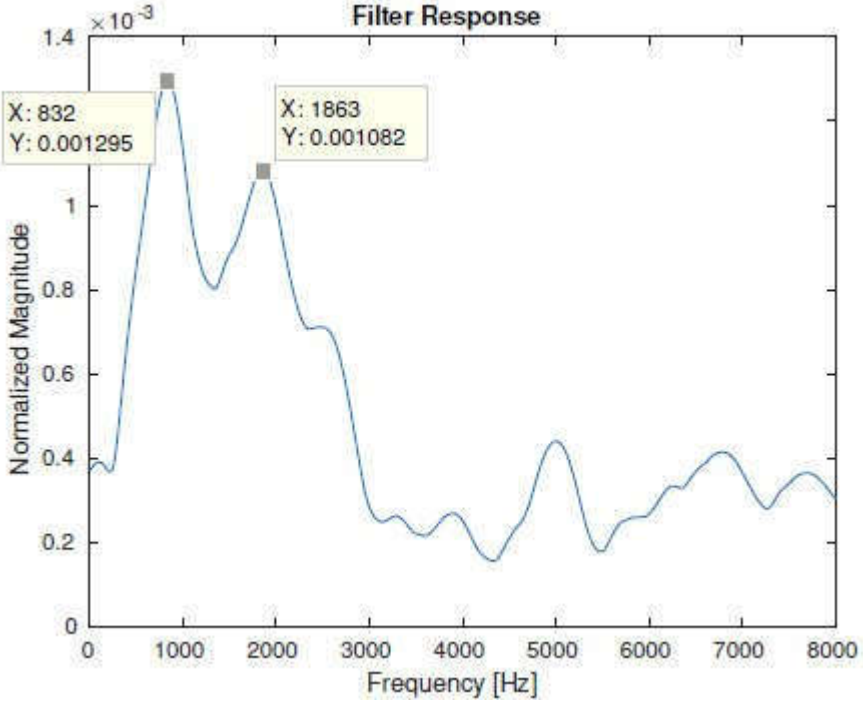

Fig. 3. Average filter response for a speech segment /er/ from CNN3trained on children speech

V. FUTURE SCOPE AND CONCLUSION

This article compares the traditional cepstral ASR methodology with a CNN-based end-to-end acoustic modelling technique to learn the key characteristics simultaneously and the raw language telephone classification for children to learn the language. Our PF-STAR corpus investigations have shown that CNN end-to-end acoustic modelling produces superior systems than those that have conventional characteristics such as MFCCs.Our tests have shown shown the system may be further improved by increasing child data with adult voice. An examination of the trained CNNs has shown that CNNs have learnt to represent formational information invariant in children's and adult speech acoustic differences.

REFERENCES

[1] A. Potamianos, S. Narayanan, and S. Lee, “Automatic speechrecognition for children.” in Proceedings of Eurospeech, 1997.

[2] S. Lee, A. Potamianos, and S. Narayan, "Acoustics of children'sspeech: Developmental changes of temporal and spectralparameters," The Journal of the Acoustical Society ofAmerica, vol. 105, no. 3, pp. 1455–1468, 1999.
[3] J. Hillenbrand, L. Getty, M. Clark, and K. Wheeler, "Acousticcharacteristics of American English vowels," The Journal ofthe Acoustical Society of America, vol. 97, pp. 3099–111, 061995.
[4] S. Palethorpe, R. Wales, J. Clark, and T. Senserrick, "Vowelclassification in children," The Journal of the Acoustical Societyof America, vol. 100, no. 6, pp. 3843–3851, 1996.
[5] R. Serizel and D. Giuliani, "Deep neural network adaptationfor children's and adult's speech recognition," in Proceedingsof Italian Computational Linguistics Conference, 2014.
[6] P. Shivakumar, A. Potamianos, S. Lee, and S. Narayan, "Improvingchildren's speech recognition using acoustic adaptationand pronunciation modeling," in Proceedings of the Workshopon Child Computer Interaction, 2014.
[7] H. Liao, G. Pundak, O. Siohan, M. Carroll, N. Coccaro,Q. Jiang, T. Sainath, A. Senior, F. Beaufays, and M. Bacchiani,"Large vocabulary automatic speech recognition for children,"in Proceedings of Interspeech, 2015.
[8] J. Fainberg, P. Bell, M. Lincoln, and S. Renals, "Improvingchildren's speech recognition through out-of-domain data augmentation,"in Proceedings of Interspeech, 2016.
[9] D. Palaz, R. Collobert, and M. Magimai.-Doss, "Estimatingphoneme class conditional probabilities from raw speech signalusing convolutional neural networks," in Proceedings of Interspeech,2013.
[10] T. N. Sainath, R. J. Weiss, A. Senior, K. W. Wilson, andO. Vinyals, "Learning the Speech Front-end With Raw WaveformCLDNNs," in Proceedings of Interspeech, 2015.
[11] P. Golik, Z. T¨uske, R. Schl¨uter, and H. Ney, "ConvolutionalNeural Networks for Acoustic Modeling of Raw Time Signalin LVCSR," in Proceedings of Interspeech, 2015.
[12] D. Palaz, M. Magimai-Doss, and R. Collobert, "Endto-End Acoustic Modeling using Convolutional NeuralNetworks for HMM-based Automatic Speech Recognition,"Speech Communication, 2019. [Online]. Available: https://doi.org/10.1016/j.specom.2019.01.004
[13] H. Muckenhirn, V. Abrol, M. Magimai.-Doss, andS. Marcel, "Gradient-based spectral visualization ofCNNs using raw waveforms," Idiap Research Institute,Tech. Rep. Idiap-RR-11-2018, Jul 2018. [Online].
[14] A. Batliner, M. Blomberg, S. D'Arcy, D. Elenius, D. Giuliani,M. Gerosa, C. Hacker, M. Russell, S. Steidl, and M. Wong,"The PF STAR children's speech corpus," in Proceedings ofNinth European Conf. Speech Communication and Technology,2005.
[15] T. Robinson, J. Fransen, D. Pye, J. Foote, and S. Renals, "WSJCAMO:a british english speech corpus for large vocabularycontinuous speech recognition," in Proceedings of ICASSP,1995.
[16] "BEEP dictionary," http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Lexical/beep.html, accessed: 01-07-2018.
[17] "MGB challenge lexicon," http://data.cstr.ed.ac.uk/asru/MGB3/data/lm/mgb.normalized.lm, accessed: 01-07-2018.
[18] D. Povey et al., "The Kaldi speech recognition toolkit," inIEEE workshop on Automatic Speech Recognition and Understanding,2011.
[19] F. Chollet et al., "Keras," https://github.com/fchollet/keras,2015.
[20] M. Abadi et al., "TensorFlow: Large-scale machine learningon heterogeneous systems," http://tensorflow.org/, 2015.

[21] M. J. Russell, S. D'Arcy, and L. P.Wong, "Recognition of readand spontaneous children's speech using two new corpora," inProceedings of Interspeech, 2004.
[22] M. J. Russell and S. D'Arcy, "Challenges for computer recognitionof children's speech," in Proceedings of the Workshopon Speech and Language Technology in Education (SLaTE),2007, pp. 108–111.
[23] H. Muckenhirn, M. Magimai.-Doss, and S. Marcel, "Towardsdirectly modeling raw speech signal for speaker verificationusing CNNs," in Proceedings of ICASSP, 2018.
[24] S. Kabil, H. Muckenhirn, and M. Magimai.-Doss, "On learningto identify genders from raw speech signal using cnns," inProceedings of Interspeech, 2018.
[25] "American vowels database," https://homepages.wmich.edu/_hillenbr/voweldata.html, accessed: 15-07-2018.