Speech To Sign Language Conversion using Convolutional Neural Networks

1st Sreeraksha M R
Dept.Information Science and Engg
JSS Science and Technological University Mysuru

2nd Vani H Y
Dept.Information Science and Engg
JSS Science and Technological University Mysuru

3rd Phani Bhushan
Dept.Information Science and Engg
JSS Science and Technological University Mysuru

4th D K Shivkumar
Dept.Information Science and Engg
JSS Science and Technological University Mysuru

***Abstract*—Speech to Sign Language Recognition targets on interpreting the speech to text and sign language, so as to facilitate the communication between deaf-mute people and ordinary people. This task has broad social impact, but is still very challenging due to the complexity and large variations in hand actions. The objective of the project is to convert Speech to sign language using convolution neural network to enhance the communication capabilities of people with hearing disabilities or speaking disabilities. In the proposed method we have achieved 93% accuracy using MFCC as a feature extraction method and CNN as a Classifier.**

Index Terms:- Convolution Neural NetworkCNN), Mel Fre-quency Co-efficient (MFCC).

I. INTRODUCTION

The main aim of speech recognition is the transcription of human speech to sign language. It is a very challenging task because human speech signals are highly variable due to various speaker attributes, different speaking styles, uncertain environmental noises, and so on. In a first step, the data is transformed into features, usually composed of a dimension- ality reduction phase and an information selection phase, based on the task-specific knowledge of the phenomena. These two phases have been carefully handcrafted, leading to state-of- the-art features such as mel frequency cepstral coefficients (MFCCs). In a second step, the likelihood of sub-word units such as, phonemes is estimated using generative models or discriminative models. In a final step, dynamic programming techniques are used to recognize the word sequence given the lexical and syntactical constraints.

Recent advances in machine learning have made possible systems that can be trained in an end-to-end manner, i.e. systems where every step is learned simultaneously, taking into account all the other steps and the final task of the whole system. It is typically referred to as deep learning, mainly be- cause such architectures are usually composed of many layers (supposed to provide an increasing level of abstraction), com- pared to classical “shallow” systems. As opposed to “divide and conquer” approaches presented previously (where each step is independently optimized) deep learning approaches are often claimed to lead to more optimal systems, as they alleviate the need of finding the right features by instead training a stack of features in a end-to-end manner, for a given task of interest. While there is a good success record of such approaches in the computer vision or text processing fields, deep learning approaches for speech recognition still rely on spectral-based features such as MFCC [4]. Some systems have proposed to learn features from “intermediate” representation of speech, like mel filter bank energies and their temporal derivatives.

II. LITERATURE SURVEY

The author Dimitri Palaz, et. al. [1] proposed the automatic speech recognition systems model the relationship between acoustic speech signal and phone classes in two stages, namely, extraction of spectral-based features based on prior knowledge followed by training of acoustic model, typically an artificial neural network (ANN). It was shown that Convolu- tional Neural Networks (CNNs) can model phone classes from raw acoustic speech signal, reaching performance on par with other existing feature-based approaches. The paper extends the CNN-based approach to large vocabulary speech recognition task. More precisely, the proposed method compares the CNN- based approach against the conventional ANN-based approach on Wall Street Journal corpus. The studies show that the CNN-based approach achieves better performance than the conventional ANN-based approach with as many parameters. We also show that the features learned from raw speech by the CNN-based approach could generalize across different databases.

Identify applicable funding agency here. If none, delete this.

The author Ossama Abdel-Hamid, et. al. [2] proposed method, the error rate reduction can be obtained by using convolutional neural networks (CNNs). first present a concise description of the basic CNN and explain how it can be used for speech recognition. further propose a limited-weight- sharing scheme that can better model speech features. The special structure such as local connectivity, weight sharing, and pooling in CNNs exhibits some degree of invariance to small shifts of speech features along the frequency axis, which is important to deal with speaker and environment variations. Experimental results show that CNNs reduce the error rate by 6%-10% compared with DNNs on the TIMIT phone recogni- tion and the voice search large vocabulary speech recognition tasks.

The author Jui-Ting Huang, et. al. [3] the proposed method aims to provide some detailed analysis of CNNs by visualizing the localized filters learned in the convolutional layer, shows that edge detectors in varying directions can be automatically learned. Then identify four domains we think CNNs can consistently provide advantages over fully-connected deep neural networks (DNNs): channel-mismatched training-test conditions, noise robustness, distant speech recognition, and low-footprint models. For distant speech recognition, a CNN trained on 1000 hours of Kinect distant speech data obtains relative 4%-word error rate reduction (WERR) over a DNN of a similar size. This is the largest corpus so far reported in the literature for CNNs to show its effectiveness. Lastly, establish the CNN structure combined with maxout units is the most effective model under small-sizing constraints for the purpose of deploying small-footprint models to devices. This setup gives relative 9.3% WERR from DNNs with sigmoid units.

The author Sunchan Park, et. al. [4] The Convolutional neural network (CNN) acoustic models showed lower word error rate (WER) in distant speech recognition than fully- connected DNN acoustic models. To improve the performance of reverberant speech recognition using CNN acoustic models, the proposed method uses the multiresolution CNN that has two separate streams: one is the wideband feature with wide- context window and the other is the narrowband feature with narrow-context window. The experiments on the ASR task of the REVERB challenge 2014 showed that the proposed mul- tiresolution CNN based approach reduced the WER by 8.79% and 8.83% for the simulated test data and the real-condition test data, respectively, compared with the conventional CNN based method.

III. PROPOSED METHODOLOGY

A. Dataset

The RAVDESS is an emotive and song-validated multina- tional database. The database is gender balanced and consists of 24 professional actors, vocalised in a North American ac- cent.The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files. . 24 professional actors are in the database (12 female, 12 male). Speaking includes expressions of calm, happy, sad, angry, fearful, surprising and disgust. Songs contain emotions of peaceful, joyful, sad and furious. The dataset contains:-(01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful,
07 = disgusting, 08 = surprised).

B. Speech to Sign Language Recognition

The figure 4.1. depicts the architecture of speech to sign language recognition system using MFCC as feature extraction method and CNN as classifier. The speech to sign language recognition system is carried out in three main steps namely., speech signal processing, feature extraction and classification. The architecture of the speech to sign language recognition is illustrated in figure 4.1. The first step is to extract the features from speech signal uttered by the speaker. The features will become the basic unit for classifying. The signals from the same users are tested and verified with CNN for the required output. At the end the speech signal which is given as input is mapped to respective sign language letters.

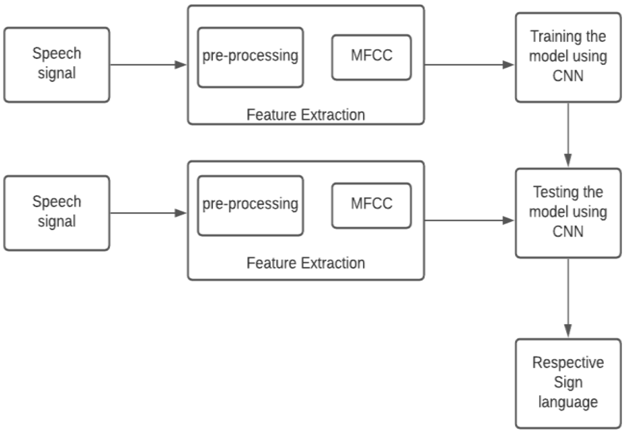


Fig. 1. The architecture of speech to sign language conversion

C. Feature Extraction – MFCC

The figure 4.2 depicts the how MFCC extract the features from speech signal. The main purpose of extraction and processing of speech signal is to extract important features from raw audio. The proposed method uses MFCC approach to extract the features of the cepstral frequency. The most straightforward technique involves determining the average energy of the signal. This metric, along with total energy in the signal, indicates the volume of the speaker. The signal in the frequency domain through the (Fast) Fourier Transform is processed. The windowed samples is used to get accurate representations of the frequency content of the signal at different points in time. By taking the square value of the signal at each window sample, power spectrum can be derived. then the values of the power spectrum as features. The three largest frequency peaks for each window is obtained and add those to the feature vector.

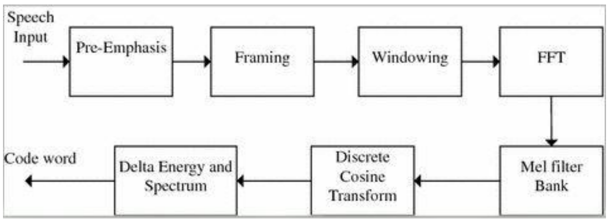


Fig. 2. Mel-frequency Cepstrum Coefficients (MFCC)

D. CNN

The figure 4.3. depicts the architecture diagram of Convo-lutional Neural Network. There are three basic layers in CNN,

- The Convolutional layer
- The Pooling layer
- The output layer Along with these three layers

there twoother layers which helps in classification. They are,

- Activation function
- Dropout layer

The input audio is passed to the first convolution layer and the convoluted output is obtained as an activation map. The filters applied in the convolution layer extract relevant features from the input image to pass further. Each filters gives a different feature to aid the correct class prediction. In case if need to retain the size of the image, use same padding, otherwise valid padding is used since it helps to reduce the number of features. The Pooling layers are then added to further reduce the number of parameters Several convolution and pooling layers are added before the prediction is made. Convolutional layer help in extracting features. As we go deeper in the network more specific features are extracted as compared to a shallow network where the features extracted are more generic. The output layer in a CNN as mentioned previously is a fully connected layer, where the input from the other layers is attended and sent so as the transform the output into the number of classes as desired by the network. The output is then generated through the output layer and is compared to the output layer for error generation. A loss function is defined in the fully connected output layer to compute the mean square loss. The gradient of error is then calculated. The error is then backpropagated to update the filter and bias values. One training cycle is completed in a single forward and backward pass.

Fig. 5. Training and Validation accuracy for CNN

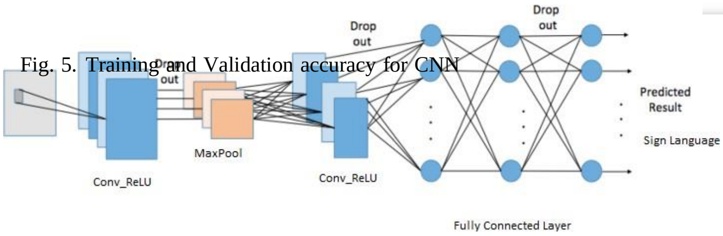


Fig. 3. Convolution Neural Network (CNN)

IV. Result

In the RAVDESS dataset there are 4909 audio files avail- able, among them 3923 file is used as training, 490 is used for validation and 491 is used as testing. Each model is trained for 20 epochs to guarantee fair comparison between all models.

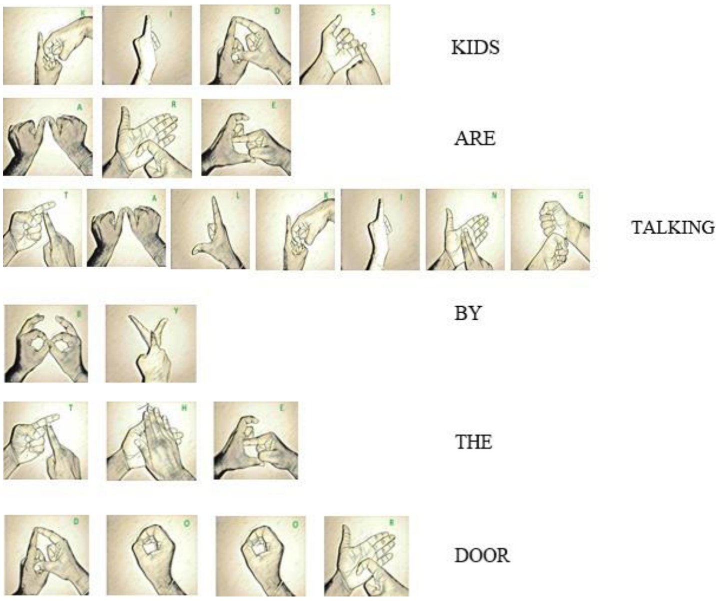
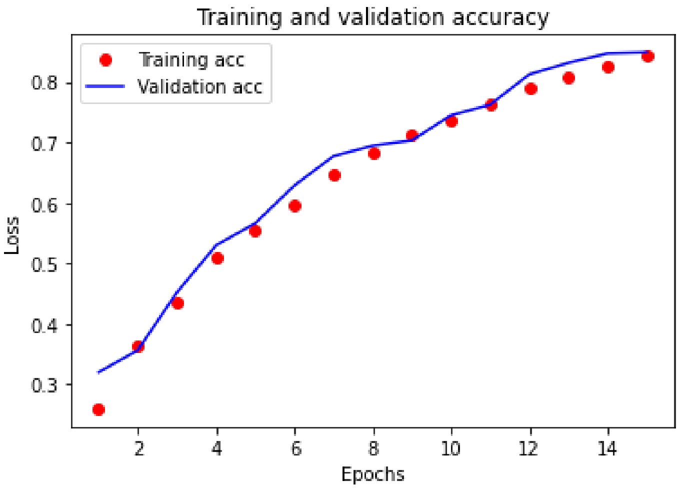


Fig. 4. The speech signal which is given as input is: KIDS ARE TALKINGBY THE DOOR



V. Conclusion

The field of machine learning is sufficiently new to still be rapidly expanding, often from innovation in new formalization of machine learning problems driven by practical applications. In the proposed method we have investigated the scalability of a Speech to sign language recognition based on CNN, which takes as input the raw speech signal, to large vocabulary task.

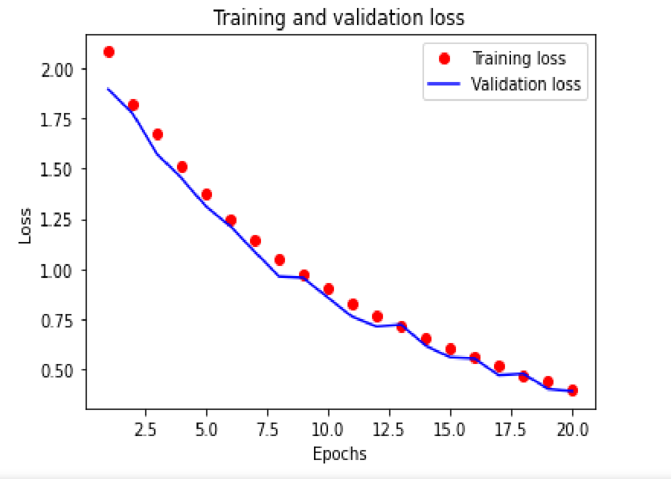
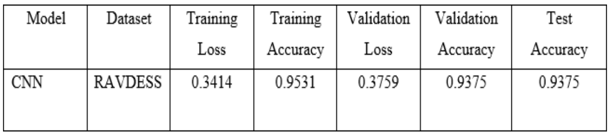


Fig. 6. Training and Validation loss for CNN



Model	Dataset	Training Loss	Training Accuracy	Validation Loss	Validation Accuracy	Test Accuracy
CNN	RAVDESS	0.3414	0.9531	0.3759	0.9375	0.9375

Fig. 7. Table of Speech to Sign Recognition using CNN

In the proposed method we have achieved 93% accuracy usingCNN.

REFERENCES

[1] Palaz, Dimitri, Mathew Magimai Doss, and Ronan Collobert. ”Convo- lutional neural networks-based continuous speech recognition using raw speech signal.” In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4295-4299. IEEE, 2015.

[2] Abdel-Hamid, Ossama, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu. ”Convolutional neural networks for speech recognition.” IEEE/ACM Transactions on audio, speech, and language processing 22, no. 10 (2014): 1533-1545.

[3] Noda, Kuniaki, Yuki Yamaguchi, Kazuhiro Nakadai, Hiroshi G. Okuno, and Tetsuya Ogata. ”Audio-visual speech recognition using deep learn- ing.” Applied Intelligence 42, no. 4 (2015): 722-737.

[4] Park, Sunchan, Yongwon Jeong, and Hyung Soon Kim. ”Multiresolution CNN for reverberant speech recognition.” In 2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O- COCOSDA), pp. 1-4. IEEE, 2017.

[5] HY, Vani, and Anusuya MA. ”A Neuro Fuzzy Classifier with Linguistic Hedges for Speech Recognition.” EAI Endorsed Transactions on Internet of Things 5, no. 20 (2020).

[6] Damodar, Navya, H. Y. Vani, and M. A. Anusuya. ”Voice emotion recognition using CNN and decision tree.” Int J Innov Technol ExplEng 8, no. 12 (2019): 4245-4249.

[7] https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio.

[8] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. ”New types of deep neural network learning for speech recognition and related applications: An overview.” In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 8599-8603. IEEE, 2013.

[9] Rownicka, Joanna, Peter Bell, and Steve Renals. ”Multi-scale octave convolutions for robust speech recognition.” In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Pro- cessing (ICASSP), pp. 7019-7023. IEEE, 2020.

[10] Koller, Oscar, Sepehr Zargaran, Hermann Ney, and Richard Bowden. ”Deep sign: Enabling robust statistical continuous sign language recogni- tion via hybrid CNN-HMMs.” International Journal of Computer Vision 126, no. 12 (2018): 1311-1325.

[11] Wang, Weizhe, Xiaodong Yang, and Hongwu Yang. ”End-to-end low- resource speech recognition with a deep cnn-lstm encoder.” In 2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP), pp. 158-162. IEEE, 2020.

[12] Deng, Li, Geoffrey Hinton, and Brian Kingsbury. ”New types of deep neural network learning for speech recognition and related applications: An overview.” In 2013 IEEE international conference on acoustics, speech and signal processing, pp. 8599-8603. IEEE, 2013.