Indian Institute of Technology

CS 753 Project Report

Speech to Sign-Language(with emotions) for the Hearing-Impaired

Reported By

Soham Naha	Mohit Agarwala	Abhinav Goud Bingi
193079003	19307R004	180050002

May 14, 2021

1 Problem Statement and Motivation

A speech and/or hearing impaired person communicates with others using Sign-Language (SL), that is different from the speech modes of communication. There are a lot of literatures that deal with the problem of conversion from Sign-Language (predominantly American Sign Language or ASL) letters to speech, but the reverse domain is not much explored.

1.1 Problem Statement

We choose to explore the domain from speech to ASL. So, our primary task was, given a speech utterence from a speaker convey a message to a person who is hearing-impaired and/or voiceless, then speech has to be converted in a form that the other person can understand, i.e. in a sign language.

1.2 Motivation

Most of the literature that is present regarding the field of SL and Speech are related to the field of SL-to-Speech conversion with limited vocabulary datasets.

But research related to the other direction, i.e. from Speech-to-SL is not much explored. In this project, we thus aimed to explore the domain from Speech to Sign-Language with the help of Automatic Speech Recognition. This could be used as a conversation model for the speech and/or hearing impaired to interact freely with people who don't have knowledge of SL.

It is also noticeable that conversion of speech-to-SL does not necessarily contain the emotion conveyed through the utterence. So, we also tried to add this emotion part along with the other decodings.

2 Task at Hand

In order to convert from Speech-to-SL along with emotions, we tried to follow the approach as stated below:

- Convert Speech to English Text
- Guess the Emotion conveyed within the utterence in parallel to text estimation
- Use the predicted text to estimate the sign-language patterns.

So, we modularized the complete pipeline into different blocks that can be trained and modelled independently and integrated again. The modules that we have are thus:

- Speech2text
- Emotion Detection from Audio
- Text2ASL

Now we look into each of these individual modules sequentially.

2.1 Speech2Text

A large part of Automatic Speech Recognition (ASR) is dedicated to the task of conversion of Speech to Text. This conversion can predict as output entire word (or grapheme) sequences, or phoneme sequences (which can be later converted to words using deterministic mappings). It began as a statistical modelling task using Gaussian Mixture Models (GMMs), and then gradually as the complexity of the task unfolded, researchers switched to Hidden Markov Models (HMMs) that could use context-based information (using Weighted Finite State Transducers and Dynamic Programming) and GMMs in their hidden states to produce mapping from speech to phones. But these architectures dependent on many assumptions of independence and statistical properties of the input speech. Also there were different models for different tasks, like the Acoustic Model(to convert from Speech to Phones),the Pronunciation Model (that maps the phonemes to words) and the Language Model (to find the correct words), and the overall accuracy was dependent on the product of each of the individual model accuracies. So, there was a need to model the complexities in a more accurate manner for better precision.

On the advent of the field of Deep Neural Networks (DNNs), researchers started to focus on Hybrid and Tandem models that used both HMMs and DNNs to perform the task of conversion from speech to text, with well-defined objective functions to train such DNNs. But still the models were not able to achieve human-level accuracy.

As the complexity of the DNN architectures increased, the accuracy of their predictions also increased. In the recent years, after the idea of processing this task in an end-to-end manner, CNN-LSTM based models got

much attention in the field of ASR, with Mozilla's DeepSpeech model trained of Common Voice Dataset, earning very high accuracy. This is due to the fact that conversion of speech to text can be thought of a sequence to sequence task, where the speech signal is a sequential input data (temporal sequence) and the output of the model (words or phonemens) is also sequential in nature (temporal sequence again). This allowed to train the complete model at once, removing the dependence on separate models for the different tasks.

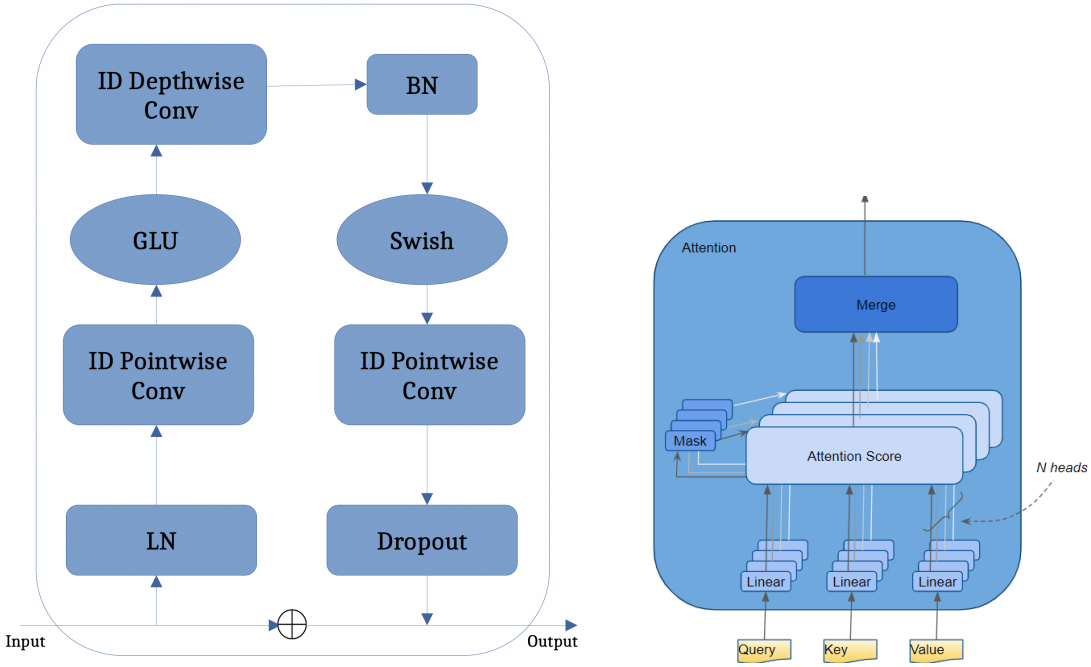
The current state-of-the-art (SOTA) model architecture for ASR are the **Conformer** [3] based models. They use ConvNets and Transformers do decode the text from the speech input. These have gained much popularity as on March, 2021.

We tried to train a CNN-LSTM based DNN for the ASR task, but due to constrained resources, we haven't yet completed the training on LIBRISPEECH [1] dataset (around 1000hrs data of audio data). The code for the same can be found here. So, we had to switch to using a pretrained model, and the toolkit available that had the present SOTA model support was the ESPNet toolkit [4].

From the ESPNet Model Zoo, we choose the Conformer-based model [5]. The Conformer model, used a encoder-decoder network to produce words as output that were again rescored using a Language-Model(LM) using shallow fusion.

The encoder network was a Conformer based network which takes as input speech data and then converts the data into 80-dimensional Mel-Spectrogram features with/without 3 pitch features, and then uses this to create an encoded form of the input signal. This encoded output passes through Attention Networks and is fed to the decoder network which is a Transformer based model. The output of the transformer is the word sequence related to the speech input. This is then rescored using a RNN-Transformer based LM, which produces the final sequence of words. The encoder-decoder network was jointly trained using joint-CTC loss. The metrics claimed by this model is as follows:

Metric	Sub	Del	Ins	Err
WER	2.1	0.2	0.3	2.6
CER	1.2	0.8	0.7	2.7



(a) Conv Architecture (b) Multi-Head Soft Attention

Figure 1: Conformer Building Blocks

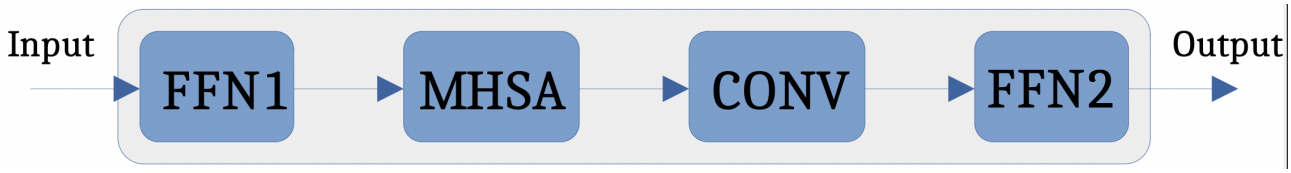


Figure 2: Conformer Architecture

2.2 Emotion to Speech

Motivation:

Often the emotion carried out by an utterence cannot be properly conveyed using sign-language.

Dataset:

We used RAVDESS [2] dataset to train our model. Since the whole dataset is about 25 GB consisting of about 7356 files. Because of our limited resources, we downloaded about 3000 files and trained our model on it. As a result, the model has accuracy in the range 47-89% for different emotions.

Model description:

The model consists of Conv1D layer, ReLU activation layer, Dropout layer, Maxpooling 1D layer, Conv1D layer, Dropout layer, Flatten layer, Dense layer, Softmax Activation layer.
The optimizer used was an RMSprop optimizer with learning rate = 0.00005, $\rho = 0.9$

Layer (type)	Output Shape	Param #
conv1d_10 (Conv1D)	(None, 40, 128)	768
activation_15 (Activation)	(None, 40, 128)	0
dropout_10 (Dropout)	(None, 40, 128)	0
max_pooling1d_5 (MaxPooling1	(None, 5, 128)	0
conv1d_11 (Conv1D)	(None, 5, 128)	82048
activation_16 (Activation)	(None, 5, 128)	0
dropout_11 (Dropout)	(None, 5, 128)	0
flatten_5 (Flatten)	(None, 640)	0
dense_5 (Dense)	(None, 10)	6410
activation_17 (Activation)	(None, 10)	0

Total params: 89,226
Trainable params: 89,226
Non-trainable params: 0

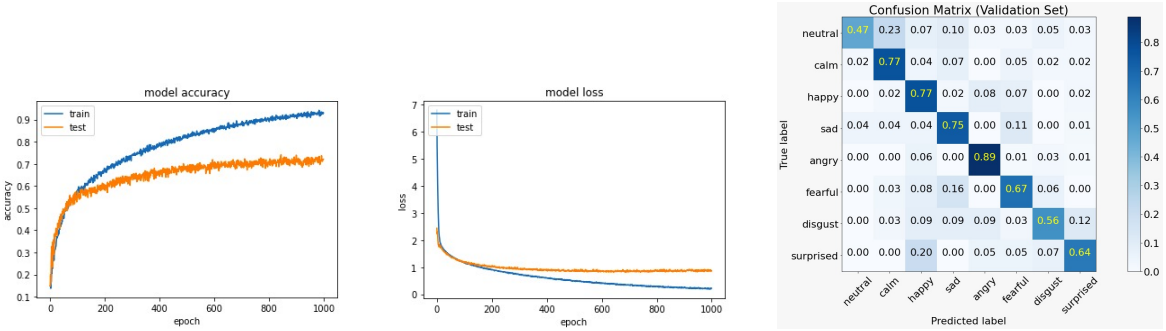
Figure 3: Emotion Recognition Model Architecture

Training:

The input speech signal was converted into 40-dimensional MFCC features and trained on a CNN-based model with a softmax output, trained on Sparse Categorical Cross-Entropy Loss.

Results:

The model achieved a test accuracy of 72% on RAVDESS audio test data consisting of all emotions.



(a) Model accuracy vs epochs (b) Train and Validation Loss (c) Confusion Matrix

Figure 4: Emotion Recognition Training Stats

2.3 Text to ASL

So, in order to complete the pipeline, the final step is to convert the decoded text into ASL. But this conversion is not straight forward, because the grammar of ASL is different from Spoken grammar. We tried to find resources that would enable us to complete the pipeline. Most of the resources that we gathered converted each letter of a text to some ASL mapping, but that was not what we intended as it would result in a spelling type output. We also went through some implementations that used Unity or Blender based Avatar Models, to convert text to ASL, but they were limited by their dataset. We came across this repository, which explored this idea in some depth.

It proposed a GAN based interpolation network that used a ASL glossary derived from the decoded English text. The workflow in this repository is as follows:

- Translate text to ASL glossary using Transformer model.
- Align the ASL Glossary to poses using OpenPose.
- Interpolate the poses generated using a Fully-Connected neural network (FCN).
- Generate avatar images for each pose using pix2pix GAN and compile as a video.

But given the constrained resources, we could not attempt to explore this path. Hence, the final step of the project remains to be completed in the near future.

3 Web-App for Integration

We created a Web-App using python's `streamlit` module, that integrated the different subtasks into one window. THe Web-App contains the facility to record one's own audio, visualize and hear-out the recorded stream and then look into the decoded output from the Speech2Text model as well as the Emotion Recognition model. A demo video for the same is made available here.

Here are some snippets that were captured while we were experimenting with the app.

(a) Overview of the App (b) Play the recording (c) Decoded Output

Figure 5: WebApp Demo Snippets

4 Work Load Division

- **Soham Naha**: Speech2Text, WebApp intergration, Presentation
- **Mohit Agarwala**: Presentation, Speech to ASL research
- **Abhinav Goud Bingi**: Speech to Emotion

References

[1] Vassil Panayotov et al. "Librispeech: An ASR corpus based on public domain audio books". In: *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[2] Steven R. Livingstone and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English". In: *PLOS ONE* 13.5 (May 2018), pp. 1–35. DOI: 10.1371/journal.pone.0196391. URL: https://doi.org/10.1371/journal.pone.0196391.

[3] Anmol Gulati et al. *Conformer: Convolution-augmented Transformer for Speech Recognition*. 2020. arXiv: 2005.08100 [eess.AS].

[4] Pengcheng Guo et al. *Recent Developments on ESPnet Toolkit Boosted by Conformer*. 2020. arXiv: 2010.13956 [eess.AS].

[5] kamo naoyuki. *ESPnet2 pretrained model: kamo-naoyuki/librispeech _asr_train_asr_conformer6_n_fft512_hop_length256_raw_en_bpe5000_scheduler_confwarmup_steps40000_opti m_conflr0.0025_sp_valid.acc.ave, fs=16k, lang=en.*