CS 753 Project

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

Soham Naha (193079003)

Mohit Agarwala (19307R004)

Abhinav Goud Bingi (180050002)

Indian Institute of Technology, Bombay

May 13, 2021

Table of Contents

Problem Statement

Given a speech utterence from a speaker who is trying to 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.

Table of Contents

- Most of the literature in the field of ASL and Speech are based on the conversion of sign-to-speech.
- But the converse model that completes the cycle of sign-to-speech, i.e. speech-to-sign, is mostly unexplored.
- In this project, we explore the speech-to-sign paradigm Deep Learning ASR models.
- This project could be used as a conversation model for the speech and/or hearing-impaired to interact with people who dont have knowledge of sign-language.

Table of Contents

Subtasks

- Speech to text conversion (ASR)
- Speech to emotion recognition
- Text to Sign Language Conversion (Future Work)

Table of Contents

Task1: Speech to Text

- The speech2text problem is one of the most classical problems in ASR.
- It started from a statistical modelling problem with separate model components like the acoustic model, language model and the pronunciation model.
- On the advent of the era of Deep Learning, this changed to an End-to-End modelling paradigm.
- It started with Tandem and Hybrid networks, with the present state-of-the-art(SOTA) model being the Conformer-based model.

Task1: Speech to Text

- Dataset: LIBRISPEECH[1]
- We initially wanted to train our own Speech2Text network using a CTC-Beam Search based CNN-LSTM model.
- But due to constraint in resources could not train the model.(code is present but only able to run 30 epochs over 4days)
- So, we reverted to the ESPNet Toolkit[2].
- From this toolkit we used the model here
- It uses a conformer based architecture for the acoustic model and a transformer based architecture for the language model.
- We have used the pre-trained models for both.
- The metrics claimed by the model on `test-clean` are:

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

[1] Panayotov et al. 2015.
[2] Guo et al. 2020.

Table of Contents

Short Description of ESPNet Architecture

The Conformer-based model architecture[3] is as follows:

- The input speech signal is converted into a sequence of 80 dimensional log-mel filterbank features with/without 3-dimensional pitch features.
- Then, passed through a Conformer-based encoder.
- The output of the above block is passed through a Transformer-based decoder.
- The encoder-decoder model was trained using joint-CTC-attention training and decoding.
- Followed by a token/word-level language model (transformer based) via shallow fusion.

[3] Guo et al. 2020.

ESPNet Architecture (Contd.)

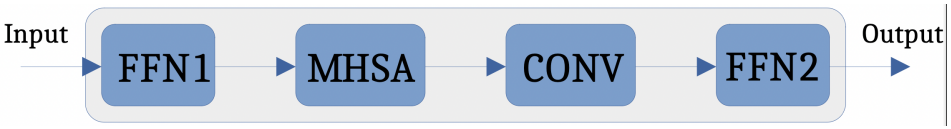


Figure: Conformer Module Architecture

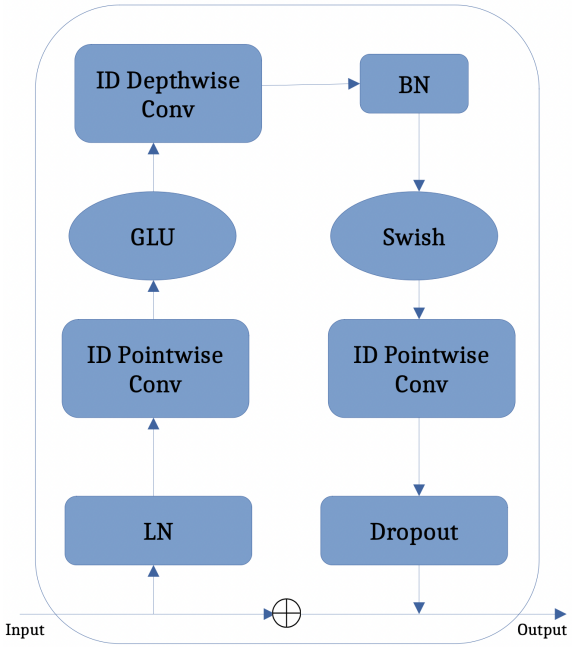


Figure: Conv Module in the Conformer Model

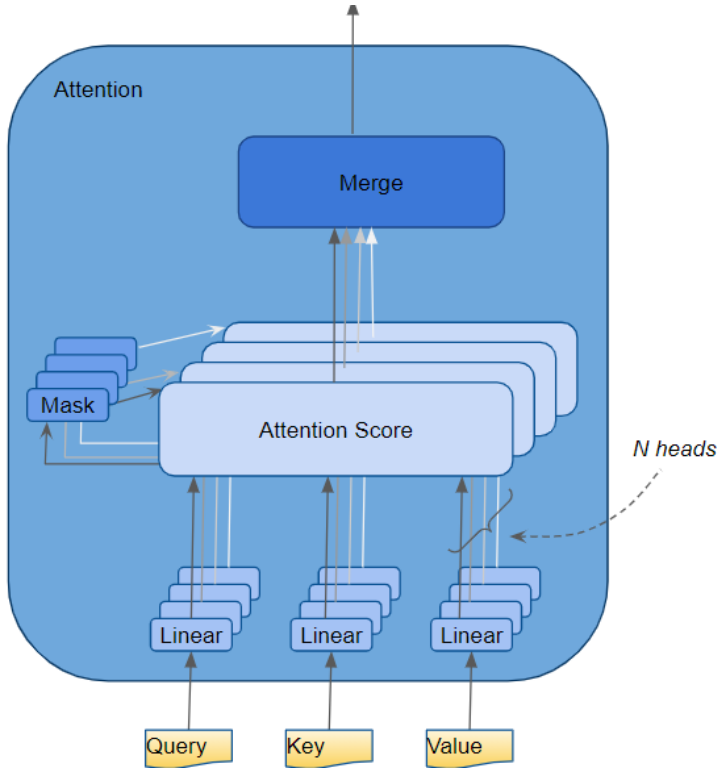


Figure: Multi-Head Attention Module

Table of Contents

Emotion Recogition

- **Motivation**:
  Often the emotion carried out by an utterence cannot be properly conveyed using sign-language.
- **Dataset**: RAVDESS[4]
- **Data Pre-processing**:
  - Convert the input speech signal into 40-dimensional MFCC features.
- **Model**:
  - The model used is a CNN-based model with a softmax layer, trained on Sparse Categorical Cross-Entropy Loss.
  - Model Accuracy: 72%

[4] Livingstone and Russo 2018.

Results from the model trained

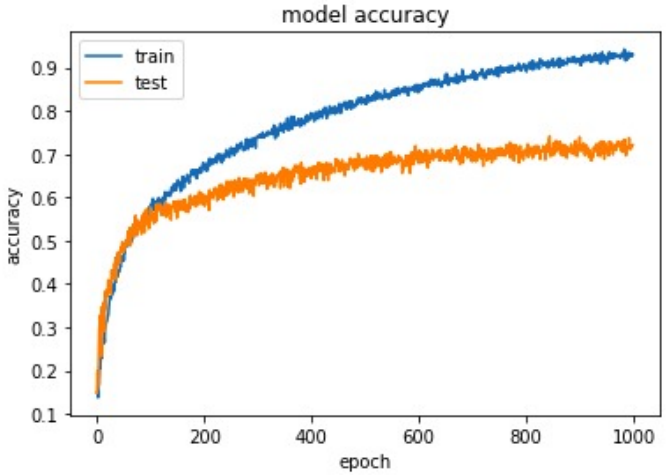
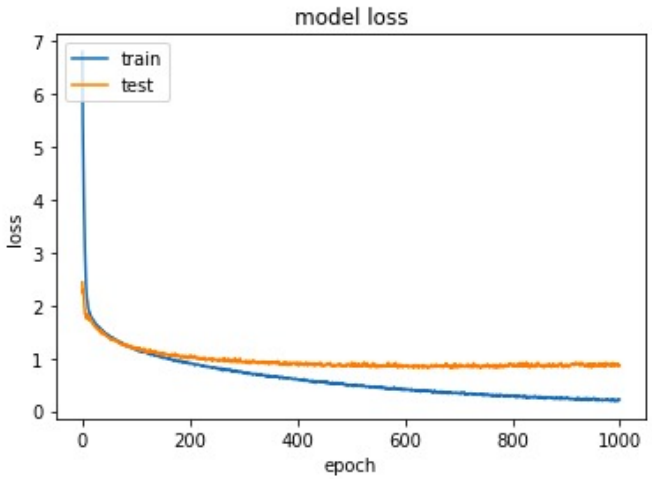


Figure: Train and Validation Accuracy Plot



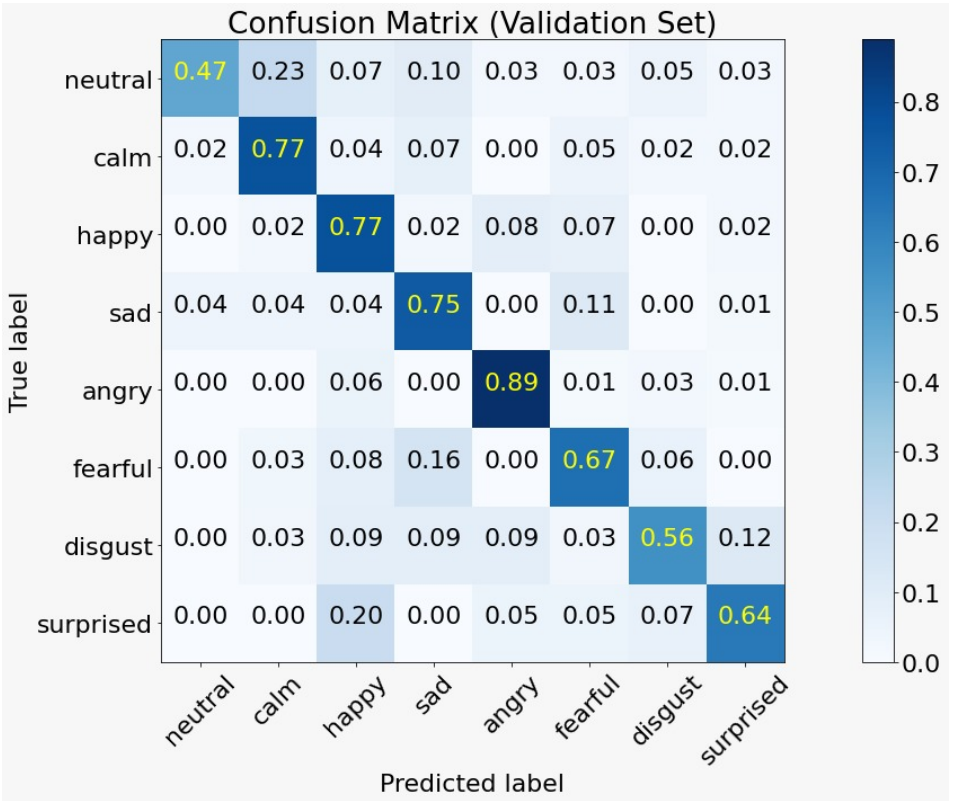


Figure: Confusion Matrix

Table of Contents

Text to Speech Conversion

- We thought of building a speech to sign-language tool, by first converting the speech to text and for every word map to american sign-language (ASL) representation.
- **Disclaimer: We could not implement the text ot sign language part.
- We came across several datasets that converted the letters to ASL hand-images, but that was not what we intended.
- There were other Unity and Blender based Avatar models, but that were not truly capable of direct text to ASL conversion, because of limited datasets.
- We got an architecture that treated this problem as a GAN-based model, but we could not implement it due constraints.
- The Workflow was 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.

Table of Contents

Web Tool for the pipeline

- We created a toolkit using the `streamlit` module of python where we can record a 10sec audio and can detect text and emotion.

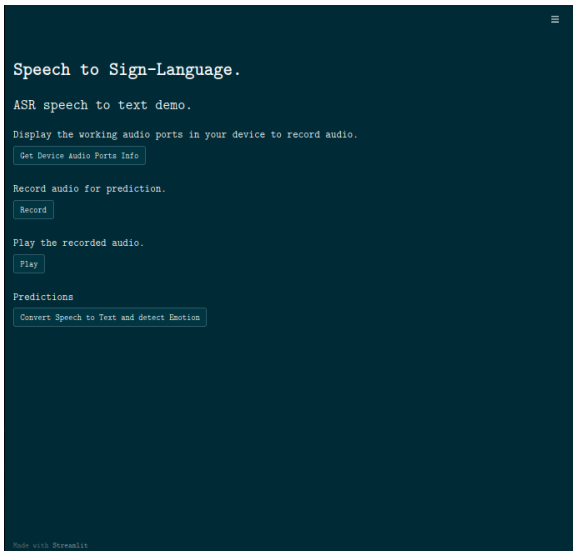


Figure: The Tool Overview

Figure: Play the Audio

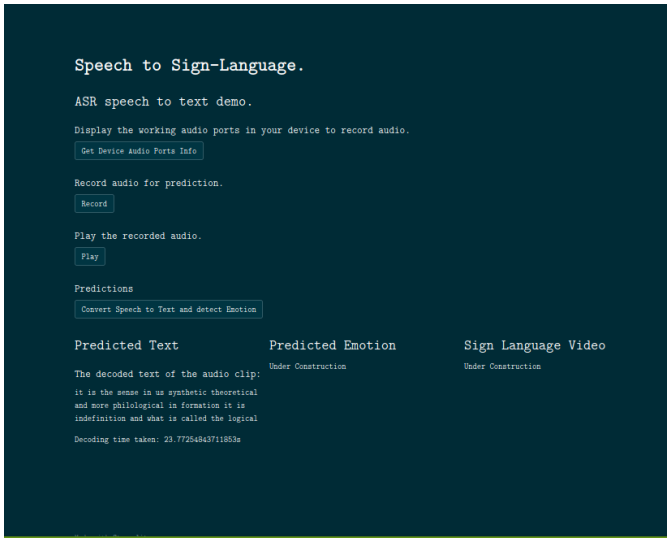


Figure: Audio Decoding

The accuracy of the toolkit depends on the accuracy of the models in general and we hope to integrate the text to ASL feature in future.

Table of Contents

Work Load Division

- **Soham Naha**: Speech2text, Webapp Integration using `streamlit`, presentation
- **Mohit Agarwala**: Speech2text, presentation, Speech to ASL research
- **Abhinav Goud Bingi**: Speech to Emotion

Table of Contents

References

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

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.

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.