Audio Visual Speech Recognition

Submitted by
Keerthana Krishnamurthy Burly - 7023672
Prashant Pombala - 7039928
`Speech-based Adaptation of Personalized User Interfaces`

Abstract

Audio-visual speech recognition is one of the major technological innovations in the automotive industry, especially for self-driving cars where the vehicle needs to adapt to a user and surroundings. With this goal, we aim to implement a personalized speech recognition system for audio and video modalities that adapts the machine learning algorithm based on features in the input. The Temporal convolution Network model is used for training and the dataset used is the LRW dataset which includes different words from the English language. To incorporate adaptation we train and evaluate the models for different syllables. The results show that training the different syllable words separately and having dedicated models for each syllable showed better speech recognition results

1 Motivation

Audiovisual speech recognition is a growing technology that is finding various applications in an expanding number of domains. It is predicted that Audiovisual speech recognition could play a pivotal role in the automobile industry. Audiovisual speech recognition could be used in a wide range of applications in the automobile industry that range from giving commands to the automobile audio system to giving driving commands to self-driving cars. In the automobile industry speech recognition typically involves recognizing speech in a dynamic and noisy environment. While working in a noisy environment, trying to recognize the speech using only the noisy audio modality may lead to a drastic decrease in the recognition accuracy. Recognizing the users' speech and taking the required action can be pivotal and at times life-saving in the context of the automobile industry. As an example, a speech recognition system used to recognize the driving commands in a self-driving car could play a vital role in saving the user's life in emergency situations. Hence it is very essential to use both audio and visual modalities in speech recognition with the aim of increasing accuracy. Adaptation is an important aspect that could be integrated into such Audiovisual speech recognition systems. Its application range from personalizing the system for the user to improving the recognition accuracy. There are many kinds of adaptations that could be performed on the audiovisual system. Some of the common types of adaptations are modality adaptation, user adaptation, and learning model adaptation. This project focused on learning model adaptation that adapted the machine learning algorithm based on some features in the input to increase the speech recognition accuracy.

2 Goals

The goal of this project can be divided into two sections

1. Implement a speech recognition system for both audio and video modalities.
2. Perform a model adaptation that changes the speech recognition machine learning model based on some input parameter.

The input parameter that was chosen to do the adaptation of the machine learning algorithm was the number of syllables in a word.

3 Related Work

There are multiple algorithms to perform speech recognition and Audiovisual speech recognition. Some of the commonly used algorithms are end-to-end audiovisual speech recognition and Automatic audiovisual speech recognition. In the context of automobiles, speech recognition must be performed in a robust manner with very good speed. Algorithms which identify the features from the input data in the first step and use these features to perform speech recognition are good, but generally very slow. Hence end-to-end audio-visual speech recognition was found to be suitable for this project. In end-to-end, audio-visual speech recognition all the appropriate features are automatically extracted by the machine learning model, and speech recognition is performed in a single step. This acts like a black box system to us, where we only need to worry about giving the input in the right format and the machine learning algorithms directly give us the decoded output. This performs speech recognition in a very efficient and swift manner.

The algorithm mentioned in the paper[1] performs end-to-end audio-visual speech recognition. It uses ResNet which is a convolutional neural network to first extract useful features and then uses BGRUs which are a simplified form of LSTMs to model temporal dynamics and perform speech recognition. This is done for both the audio and video modalities. A further improvement work[2] focused on using MS TCNs in place of BGRUs to further improve performance.

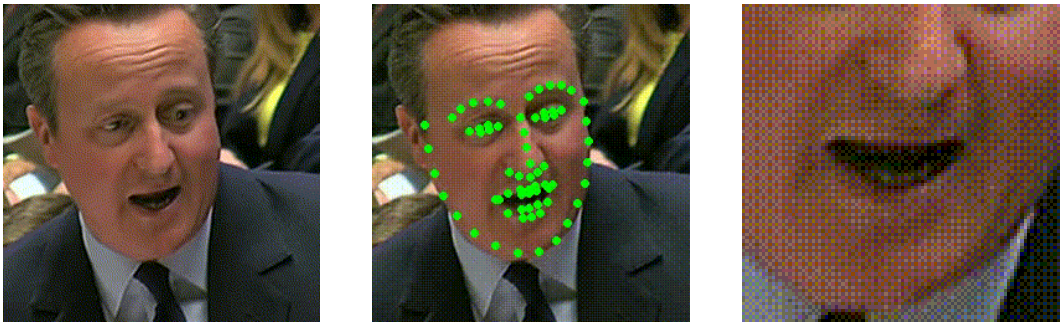
After performing end-to-end Audio Visual speech recognition we need to perform model adaptation. Based on the work described in the research paper[3] context-aware machine learning was performed. In the research paper[3] described the authors wanted to perform training and testing machine learning models for predicting benzene concentration. The authors found that the accuracy of the prediction was dependent on the climate and more specifically the month of the year. Initially, the authors decided to train the entire data on a general machine learning algorithm and noted the accuracy of prediction, this served as a baseline to compare the accuracy obtained after adaptation. Next, the authors decided to separate the input data space based on the parameter month. Then, they trained the data obtained for different months separately using separate machine learning algorithms and used these models to test the data belonging to those months. Example: Only input data for November was used to train a specific model. When certain testing data belonged to November the data was tested using the model specifically trained using November data. The authors compared the accuracy of prediction done using specifically trained models with the baseline and found significant improvements in the accuracy. This research successfully achieves the goal of adapting the machine learning algorithm based on some parameters in the input space and achieving better prediction accuracy. Similar to the paper mentioned above we aim to train our Audiovisual Speech recognition machine learning model separately for words of different syllables and verify if separating the input data space based on syllables and using specific machine learning models to train words of different syllables separately provides better accuracy than training all the data with a generic algorithm.

4 Data

For our project, the Lip Reading in the Wild (LRW) dataset was used[5]. The dataset consists of up to 1000 examples of 500 different words. The audio and video for the data were taken from BBC News. The dataset contained the video files and the metadata. The data contained about 800 - 1000 training videos for each word and about 50 test videos. This dataset was chosen at it contained a comprehensible list of audio and video data of words from real-world scenarios. The video had speakers from different angles and accents, this aspect resembles typical data that could be obtained from an automobile. The dataset also had videos of different illuminations which is similar to the data obtained by driving an automobile at different times of the day.

5 Preprocessing

The data that is used is 'in the wild' data which means it is more natural and does not have any fixed format. To standardize the data for training we need to pre-process it. The videos contain different



(a) Original image (b) Landmarks detection (c) Cropped ROI

Figure 1: Data Preprocessing

Table 1: Accuracy for generic model

Words	Audio	Video
Dataset(72 words)	94.0%	50.0%
1 syllable(5 words)	84.4%	45.0%
3 syllable(5 words)	88.8%	72.8%

backgrounds, angles, and genders. The pre-processing as shown in Figure 1 is started with detecting the face regions and alignment using face landmarks as shown in Figure 1(b). Each face frame is aligned to a reference mean face shape. Later the image is cropped to a fixed 96 x 96 pixels wide ROI(Region of Interest) from the aligned image so that the mouth region is always roughly centered on the image crop as shown in Figure 1(c). To avoid color or RGB conflicts the image is scaled to a gray level. Similarly in the case of audio, Liborsa - python package is used to extract audio from video. Both audio and video extracted files are stored in npz file formats.

6 Training and Syllabification

The architecture used is the Temporal Convolution network which uses the standard ResNet-18 network for classification. The JSON configuration file of ResNet-18 uses the pre-processed data for training. The pre-processed files from audio and video are saved in npz format, which is one of the arguments passed for training along with the modality i.e., audio or video, dataset, and model configuration. The training is done for 72 words which are around 70000 video and audio files. The training time is longer as it is done on CPU machines. It takes around 24 hours for audio files training and around 26 hours for video files. The testing was done with around 3000 video and audio files.

Further to bring in the adaptation of syllables, we further segregate the dataset into 1 and 3-syllable words. The selection excluded 2-syllable words as the lip movement(video) and audio indentation(audio) difference between 1 and 2-syllable words are considerably low. This gives how the model adapts to different contexts and is hence trained on seven 1-syllable and 3-syllable words. The testing was done on five words with their respective syllable words.

Table 2: Accuracy for adapted model

Words	Audio	Video
1 syllable(5 words)	96.8%	51.6%
3 syllable(5 words)	99.4%	93.2%

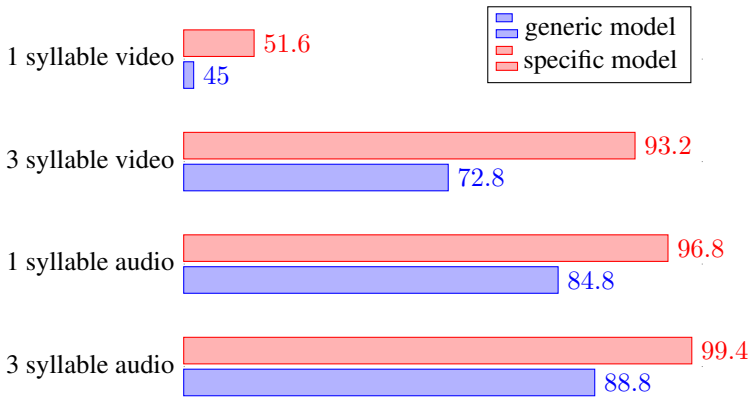
7 Results

The results for generic model test data are shown in Table 1. The testing on the entire test data gives an accuracy of 94% for audio but only 50% for video. The second and third row shows the accuracy for testing only 1 and 3 syllable words using the generic trained model. The test accuracy for audio on 1-syllable is 84.4% but for video, it is relatively very low at only 45%. But in the case of 3-syllable audio accuracy remains almost consistent at 88.8% but there is a large change for the video which increases to 72.8%

In Table 2 we see the results for adapted trained models. Here the models are trained specially for 1 and 3 syllables separately. The 1-syllable trained model when tested gives an accuracy of 96.8% but 51.6% for video. Similarly, for 3-syllable words, audio has a very good accuracy of 99.5% and also we can see a huge increase in the video data with an accuracy of 93.2%. We shall further interpret these results in the next section.

8 Interpretation

Accuracy of specific and generic algorithms for audio and video modalities for 1 syllable and 3 syllable words



It can be observed from the graph above that training 1-syllable and 3-syllable words separately and testing the respective words with the specially trained model gives a higher speech recognition accuracy for both audio and video modalities. Training the models separately for words of different syllables optimally tunes the model weights, and this in turn enables the model to perform better speech recognition.

9 Challenges and future work

The training of the data comes with its challenges. The data set is not easily available as it includes the image of faces and voices of individuals which causes privacy concerns. The training was done on a CPU machine and the time taken was longer when compared to a GPU system training. Since the original data set was huge containing 500 words it could not be entirely used and was limited to only 72 words for our work. There were also trials of the concatenation of audio and visual models which lead to errors as there was a mismatch in the frames of images and audio npz files.

Even though there are several challenges there is a lot of scope for future work. As mentioned the audio-visual concatenation would lead to end-to-end Audio-Visual speech recognition. Since the focus on adaptation is on syllable training, it can also be extended to syllabification using other machine learning techniques. The model can also adapt to different desired modalities based on the noise level in the data set. This could further help in adapting to audio in absence of video and vice-versa.

10 Summary and Conclusion

The entire project phase can be summarized into the main stages of preparing the dataset, training the generic model, training the adapted model, and testing and evaluation. At the end of all the phases, we try to evaluate the adaptation of models based on different contexts which give different results.

Hence it can be concluded that testing the model with specialized models provides better speech recognition accuracy than testing on a generically trained algorithm for 1 and 3 syllable words.

References

[1] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropou- los, and M. Pantic, *End-to-end audiovisual speech recognition*

[2] Brais Martinez, Pingchuan Ma, Stavros Petridis, Maja Pantic *Lipreading using temporal convolutional networks*

[3] Nathalia Nascimento, Paulo Alencar, Carlos Lucena, Donald Cowan, *A Context-Aware Machine Learning-based Approach*

[4] Gerald Schwiebert, Cornelius Weber, Leyuan Qu, Henrique Siqueira, Stefan Wermter, *A Multimodal German Dataset for Automatic Lip Reading Systems and Transfer Learning*

[5] J. S. Chung, A. Zisserman, *Lip Reading in the Wild*