**AUDIO – VISUAL SPEECH RECOGNITION**

**Team members:**

2010030119 – Harika Nethi

2010030088 – Tejaswi Kottakki

2010030155 – Shreya Chintawar

2010030009 – Amrutha Varshini

**Introduction**

Communication with people is a basic need for everyone, but some people, such as deaf people, are not able to communicate well. Deaf people communicate visually and physically rather than audibly. Therefore, they have some problems in their relationship with people.

Usually, deaf people learn and use sign-language to communicate with each other, but people have no desire to learn it. For this reason, many people feel awkward or become frustrated trying to communicate with deaf people, especially when no interpreter is available.

So, our aim is to create a new system where AR and AVSR technologies are combined. This is a new system with multiple features, but helping deaf people to communicate with people is its main goal. This system uses the narrator’s audio, video, and facial expressions to make the narrator’s speech visible to deaf people on AR display.

**Literature Survey**

**Article-1:** LRS3-TED: a large-scale dataset for visual speech recognition

This paper introduces a new multi-modal dataset for visual and audio-visual speech recognition. It includes face tracks from over 400 hours of TED and TEDx videos, along with the corresponding subtitles and word alignment boundaries. The new dataset is substantially larger in scale.

The dataset consists of .mp4 files for videos and also plain test files for subtitles. The dataset is organized into three sets: pre-train, train-val and test. The first two overlap in terms of content but the last is completely independent.

These TED and TEDX videos were selected for multiple reasons:  
(1) a wide range of speakers appears in the videos, unlike movies or dramas with a fixed cast;   
(2) shot changes are less frequent, therefore there are more full sentences with continuous face tracks; (3) the speakers usually talk without interruption, allowing us to obtain longer face tracks. TED videos have previously been used for audio-visual datasets for these reasons.

The models/techniques mentioned in this article are CNN face detector based on the Single Shot MultiBox Detector (SSD) to detect face appearances in the individual frames, the subtitles in the YouTube videos are broadcast in sync with the audio only at sentence-level, therefore the Penn Phonetics Lab Forced Aligner (P2FA) [12] is used to obtain a word-level alignment between the subtitle and the audio signal. The alignment is double-checked against an off-the-shelf Kaldi-based ASR model, a two-stream network (SyncNet) to synchronize the two streams. The same network is also used to determine which face’s lip movements match the audio, and if none matches, the clip is rejected as being a voice-over.

**Article-2:** Deep Audio-Visual Speech Recognition

The goal is to recognize phrases and sentences being spoken by a talking face, with or without the audio. Unlike previous works that have focused on recognizing a limited number of words or phrases, they tackled lip reading as an open-world problem – unconstrained natural language sentences, and in the wild videos.

The key contributions are:

(1) we compare two models for lip reading, one using a CTC loss, and the other using a sequence-to-sequence loss. Both models are built on top of the transformer self-attention architecture;

(2) we investigate to what extent lip reading is complementary to audio speech recognition, especially when the audio signal is noisy;

(3) we introduce and publicly release a new dataset for audio-visual speech recognition, LRS2-BBC, consisting of thousands of natural sentences from British television.

The models that we train surpass the performance of all previous work on a lip-reading benchmark dataset by a significant margin. They finally demonstrate that visual information helps to improve speech recognition performance even when the clean audio signal is available. Especially in the presence of noise in the audio, combining the two modalities leads to a significant improvement.

**Article-3:** Audio-visual speech recognition techniques in augmented reality environments

This article says that many recent studies show that Augmented Reality (AR) and Automatic Speech Recognition (ASR) technologies can be used to help people with disabilities. Many of these studies have been performed only in their specialized field. Audio-Visual Speech Recognition (AVSR) is one of the advances in ASR technology that combines audio, video, and facial expressions to capture a narrator’s voice.

In this paper, they combined AR and AVSR technologies to make a new system to help deaf and hard-of-hearing people. Their proposed system can take a narrator’s speech instantly and convert it into a readable text and show the text directly on an AR display. Therefore, in this system, deaf people can read the narrator’s speech easily.

In addition, people do not need to learn sign-language to communicate with deaf people. The evaluation results show that this system has lower word error rate compared to ASR and VSR in different noisy conditions. Furthermore, the results of using AVSR techniques show that the recognition accuracy of the system has been improved in noisy places. Also, the results of a survey that was conducted with 100 deaf people show that more than 80 % of deaf people are very interested in using their system as an assistant in portable devices to communicate with people.

The technologies used in creating this new system are AR engine, the AVSR engine, the Joiner Algorithm, Audio-Visual contents, Automated Process Scripts, and Face Detection techniques.

**Article-4:** VoxCeleb2: Deep Speaker Recognition

The objective of this paper is speaker recognition under noisy and unconstrained conditions.

They made two key contributions.

1.) They introduced a very large-scale audio-visual speaker recognition dataset collected from open-source media. Using a fully automated pipeline, they curate VoxCeleb2 which contains over a million utterances from over 6,000 speakers. This is several times larger than any publicly available speaker recognition dataset.

2.) They developed and compared Convolutional Neural Network (CNN) models and training strategies that can effectively recognize identities from voice under various conditions. The models trained on the VoxCeleb2 dataset surpass the performance of previous works on a benchmark dataset by a significant margin.

VoxCeleb2 consists of over a million utterances from over 6k speakers. Since the dataset is collected ‘in the wild’, the speech segments are corrupted with real world noise including laughter, cross-talk, channel effects, music and other sounds. The dataset is also multilingual, with speech from speakers of 145 different nationalities, covering a wide range of accents, ages, ethnicities and languages.

**Article-5:** Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization

This article says that Audio-visual speech recognition (AVSR) is thought to be one of the potential solutions for robust speech recognition, especially in noisy environments. Compared to audio only speech recognition, the major issues of AVSR include the lack of publicly available audio-visual corpora and the need of robust knowledge fusion of both speech and vision.

In this work, based on the recently released NTCD-TIMIT audio-visual corpus, they addressed the challenges of AVSR through three aspects:

1) optimal integration of acoustic and visual information;

2) robust performance with multi-condition training;

3) robust modeling against missing visual information during decoding.

They proposed a bimodal-DFSMN to jointly learn feature fusion and acoustic modeling, and utilize a per-frame dropout approach to enhance the robustness of AVSR system against the missing of visual modality. In the experiments, they constructed two setups based on the NTCD-TIMIT corpus that consists of 5 hours clean training data and 150 hours multi-condition training data, respectively. As a result, they achieved a phone error rate of 12.6% on clean test set and an average phone error rate of 26.2% on all test sets (clean, various SNRs, various noise types), which both dramatically improve the baseline performance in NTCD-TIMIT task.

**Article-6:** Large – scale visual speech recognition

This work presents a scalable solution to open-vocabulary visual speech recognition. To achieve this, they constructed the largest existing visual speech recognition dataset, consisting of pairs of text and video clips of faces speaking (3,886 hours of video).

They designed and trained an integrated lipreading system, consisting of a video processing pipeline that maps raw video to stable videos of lips and sequences of phonemes, a scalable deep neural network that maps the lip videos to sequences of phoneme distributions, and a production-level speech decoder that outputs sequences of words.

The proposed system achieves a word error rate (WER) of 40.9% as measured on a held-out set. In comparison, professional lipreaders achieve either 86.4% or 92.9% WER on the same dataset when having access to additional types of contextual information. Their approach significantly improves on other lipreading approaches, including variants of LipNet and of Watch, Attend, and Spell (WAS), which are only capable of 89.8% and 76.8% WER respectively.

The combination of methods in this work represents a significant improvement in lipreading performance, a technology which can enhance automatic speech recognition systems, and which has enormous potential to improve the lives of speech impaired patients worldwide.

**Article-7:** Visual speech recognition

Machine Learning techniques gives computers the capability to learn using sample inputs and their outputs which creates a model to test against test cases instead of being explicitly programmed. Visual speech recognition is a process of conversion of speech to text in the absence of audio where the lip features of the person are extracted to track the pattern formed.

This paper also contains the overview of different Machine Learning algorithms and image processing procedures to effectively extract and track the lip movements. Nowadays image processing procedure is turning into a key technique for extracting the key features and considering various other environmental features to enhance the output.

This paper predominantly centers prediction of text based on the lip movements. This paper mainly focuses on reviewing various algorithms used for VSR. There are so many classification techniques such as LSTMs, CNNs, Decision Tree and Neural networks.

**Article-8:** A Survey on Different Visual Speech Recognition Techniques

In automatic speech recognition (ASR) visual speech information plays a pivotal role especially in the presence of acoustic noise. This paper provides a short review of the different methods for visual speech recognition systems (VSR). Here, they discussed the different stages of VSR including the face and lip localization techniques and different visual feature extraction techniques. They also provided the details of audio-visual database related to this study.

They have made a short review on the face and lip detection methods, visual feature extraction techniques and databases related to the visual speech recognition (VSR). The quality of VSR mainly depends upon the database and visual features selected.

The review reveals that only the two databases AVCAR and XM2VTS considered about the speaker number above 70 and most of them considered the acoustic noise only. For a better database design for VSR, they have to include large number of speakers, natural conversations and facial expressions, both acoustic and visual noise.

Finally, the comparative analysis of visual feature extraction techniques shows that the transform-based approaches include more visual information than the geometric feature-based approaches. But the combination of both transform and geometric based technique will lead to a better visual feature extraction approach.

**Article-9:** Audio-Visual Speech Recognition for People with Speech Disorders

Speech recognition of disorder people is a difficult task due to the lack of motor-control of the speech articulators. Multimodal speech recognition can be used to enhance the robustness of disordered speech.

This paper introduced an automatic speech recognition system for people with dysarthria speech disorder based on both speech and visual components. The Mel-Frequency Cepestral Coefficients (MFCC) is used as features representing the acoustic speech signal. For the visual counterpart, the Discrete Cosine Transform (DCT) Coefficients are extracted from the speaker's mouth region. Face and mouth regions are detected using the Viola-Jones algorithm. The acoustic and visual input features are then concatenated on one feature vector. Then, the Hidden Markov Model (HMM) classifier is applied on the combined feature vector of acoustic and visual components.

The system is tested on isolated English words spoken by disorder speakers from UA-Speech data. Results of the proposed system indicate that visual features are highly effective and can improve the accuracy to reach 7.91% for speaker dependent experiments and 3% for speaker independent experiments.

**Article-10:** Audio-Visual Speech Recognition is worth 32 X 32 X 8 voxels

In this work they designed a transformer visual front-end for the AV-ASR task. To the best of their knowledge, the use of a purely transformer-based visual front-end in combination with a transformer encoder makes this the first fully transformer end-to-end architecture for AV-ASR.

Their proposed model outperforms a strong baseline employing convolutions for the visual frontend on the lip-reading scenario, and matches its performance on the audio-visual ASR scenario.

Finally, they fine-tuned their model on the publicly available LRS3-TED dataset and were able to achieve a new state of-the-art word error rate on the lip-reading scenario, while matching the performance of the best published model when both acoustic and visual signals are employed.