Audio-Visual Speech Recognition Using Visual and Thermal Images

by

Zhaniya Koishybayeva

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Data Science

at the

NAZARBAYEV UNIVERSITY

July 2021

Author Zhaniya Koishybayeva ....
Department of Computer Science
July 19, 2021

Certified by ....
Michael Lewis
Associate Professor
Thesis Supervisor

Accepted by ....
Vassilios D. Tourassis
Dean, School of Science and Technology

Audio-Visual Speech Recognition Using Visual and Thermal Images

by

Zhaniya Koishybayeva

Submitted to the Department of Computer Science
on July 19, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Data Science

Abstract

In this thesis I examine the hypothesis that the performance of lipreading systems can be improved by including thermal image data in combination with the usual visual image streams. I test the hypothesis by constructing a system based on the Lip2Wav model for lipreading using deep learning methods. The system takes silent video as an input and generates synthesized audio as an output. System performance is evaluated using standard metrics such as the Word Recognition Rate (WRR), to assess the contribution of the thermal input to the accuracy of the lipreading system, and qualitative assessments of the synthesized audio such as Short-Term Objective Intelligibility (STOI) and Extended STOI (ESTOI), and Perceptual Evaluation of Speech Quality (PESQ). The model is trained using three variations of input channels: visual images only, thermal images only, and a synthesis of the visual and thermal images. The model uses a novel dataset, SpeakingFaces LipReading (SFLR), comprised of aligned streams of visual and thermal images of a person reading short imperative commands that are representative of typical human-computer interaction with devices such as personal digital assistants. The results as shown in Table 5.2 suggest that with the inclusion of aligned thermal data I was able to approximate the system performance from the previously published results. However the addition of thermal image stream did not show improvement in the performance.

Thesis Supervisor: Michael Lewis
Title: Associate Professor

Acknowledgments

I extend my gratitude to the team at the Institute for Smart Systems and Artificial Intelligence (ISSAI), under the leadership of Dr. Atakan Varol. The time I spent there as a research assistant was invaluable to the design and implementation of the project. The team was responsible for the collection of the original SpeakingFaces dataset, and the extended form SpeakingFaces LipReading (SFLR) that was specifically collected for the purposes of this project. To these ends, I thank Madina Abdrakhmanova, ISSAI Data Scientist and the SpeakingFaces project leader who designed the setup and supervised data collection, and ISSAI team members Askat Kuzdeuov, who wrote the code for data collection and supervised the data collection setup, Almas Mirzakhmetov, who performed the data acquisition and verification, and especially Zhanat Makhataeva for her kind participation as the main subject of the SFLR dataset.

Contents

List of Figures

List of Tables

Chapter 1

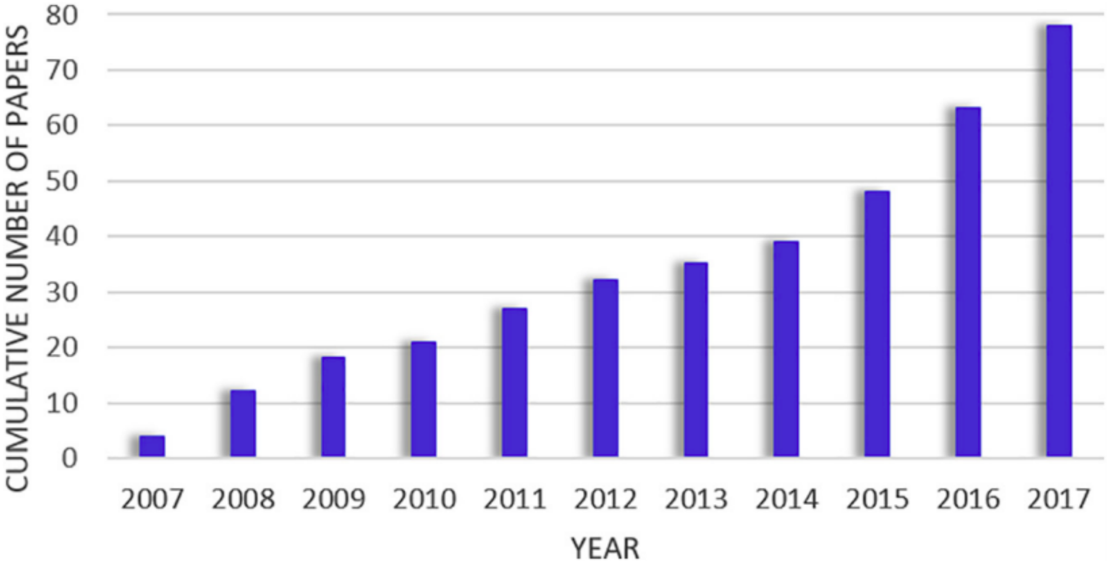
Introduction

The number of lipreading systems has increased sharply in recent years (Fig. 1-1) due to the emergence of high-performance deep learning architectures and the availability of relevant large-scale databases. Lipreading systems are generally distinguished by their output type: those that generate textual output are known as "lip2text", while those that produce synthesized audio are described as "lip2speech". State-of-the-art lip2text models have progressed significantly, achieving 85% accuracy [21] in lipreading from silent video of subjects speaking "in the wild", that is, not in controlled laboratory settings.

While the achieved recognition rate is quite an improvement over prior results, it nevertheless leaves substantial room for improvement. A promising development is the potential augmentation of the visual video stream with data acquired from high-resolution thermal cameras. Given the recent trending of thermal sensors towards higher resolution at lower cost, it is feasible that such sensors will become more commonplace as a component of popular commercial digital devices, such as smartphones [1].

This trend, if realized, would increase the utility of research into the effective utilization of thermal data. However, thus far, there is apparently only one published conference paper which considers the use of thermal data aligned with visual video in the performance of a speech recognition system [32]. While innovative, the authors used a very short and phonetically limited dataset, and the research is now outdated.

Figure 1-1: Cumulative number of lipreading papers (2007-2017) [12]



The objective for this thesis is to test the hypothesis by replicating and upgrading recent work on lipreading utilizing the usual visual image stream paired with the corresponding thermal image stream, and producing an output which can be analyzed in terms of speech recognition as demonstrated by the standard metrics of the field.

For this purpose I chose to use a lip2speech model called Lip2Wav, which has become known as a state-of-the-art system for lipreading in the lip2speech domain [29]. The architecture of the system includes subtasks such as the identification and cropping of the region-of-interest (ROI) in the image streams, where in this case the ROI consists of a tight bound of subject's face. Then the ROI data is submitted to an encoder module (a convolutional neural network) that serves to extract the observable features from a given n-gram, and then finally a decoder (a recurrent neural network) to map the visual features into speech.

The choice of the Lip2Wav model has several advantages. First, the authors have shown that the model is suitable for datasets collected from both controlled environments and speaking in the wild settings. Second, the Lip2Wav model uses the entire face as the ROI, rather than just the region of the lips and mouth; consideration of the larger ROI may contribute additional information and potentially improve the accuracy rate. Third, Lip2Wav is a speaker-specific model, which is convenient for data collection: it requires only one subject to accept the hypothesis, and allows to make some improvements in the setup compared to the existing thermal dataset (discussed in a later chapter). Fourth, the architecture of the model uses up-to-date deep learning methods for both feature extraction and classification parts of the system. Finally, as the fifth reason, the Lip2Wav model accommodates the consideration of other datasets.

Based on the model and the noted advantages, the thesis presents the first publicly available lip2speech study (as compared to lip2text) with the addition of thermal image data. I tested the model using a new dataset SpeakingFaces LipReading (SFLR), which was designed and collected for the purposes of the project by a team of researchers at ISSAI.

The SFLR dataset's transcript was built from phrases which are typical for users' voice commands for smart devices, thus, these findings may be applied to improve the speech recognition field of human-computer interaction in adverse environments, such as transport hubs or industrial settings, where audio speech recognition is effectively impeded.

Chapter 2

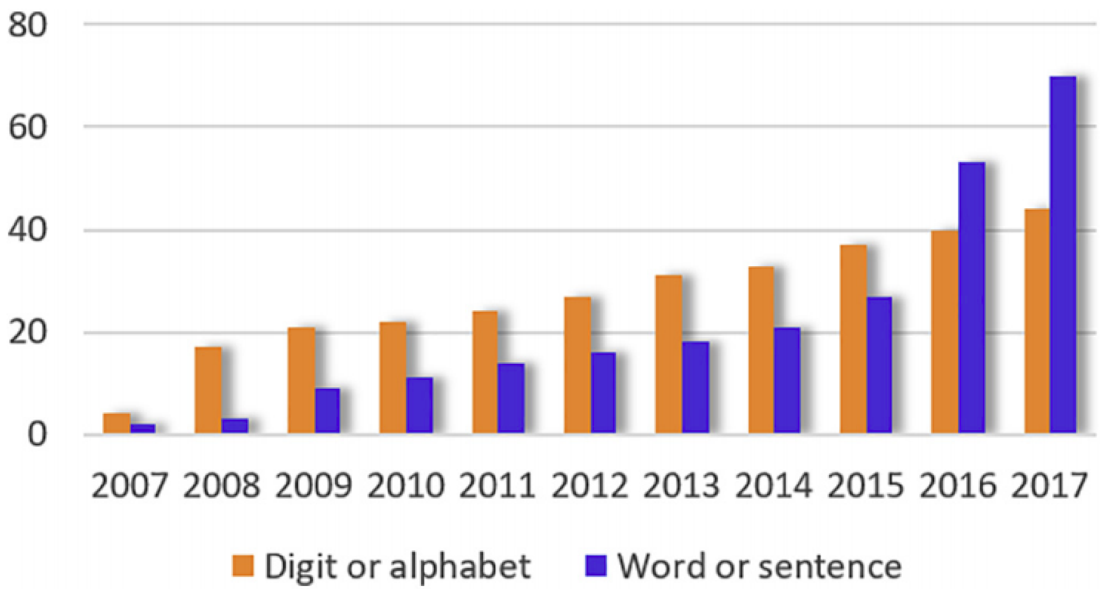
Literature Review

Researchers in the domain of lipreading typically differentiate amongst five types of audio-visual databases, depending on type of utterances that the subjects are tasked to pronounce: alphabetical characters, digits, words, phrases and sentences.

Datasets used for the recognition of digit and alphabetical utterances were popular in the early stages of audio-visual speech recognition due to the constrained vocabulary and the likelihood of having large numbers of instances per class [12]. However, the recognition scope was very limited, and the results difficult to extrapolate to more complex and practical tasks such as the recognition of full words and sentences. While the research on the simpler models is still ongoing, over the past decade the interest of researchers has shifted towards the more complicated structures (Fig. 2-1). The transition has accelerated with the emergence of high-performance Deep Learning (DL) architectures and the availability of large-scale databases which support modern machine-learning methods.

Cognizant of these trends, and with access to the computational and technical resources of the Institute for Smart Systems and Artificial Intelligence (ISSAI), I chose to focus on the more complex databases consisting of words, phrases and sentences. Note that the majority of the examples described are used as source material for lipreading in general. As will be described later, there are very few datasets that include the thermal data stream in the process of recording the utterances of research participants.

Figure 2-1: Cumulative number of lipreading papers (2007-2017) targeting different utterance types [12]



2.1 Datasets

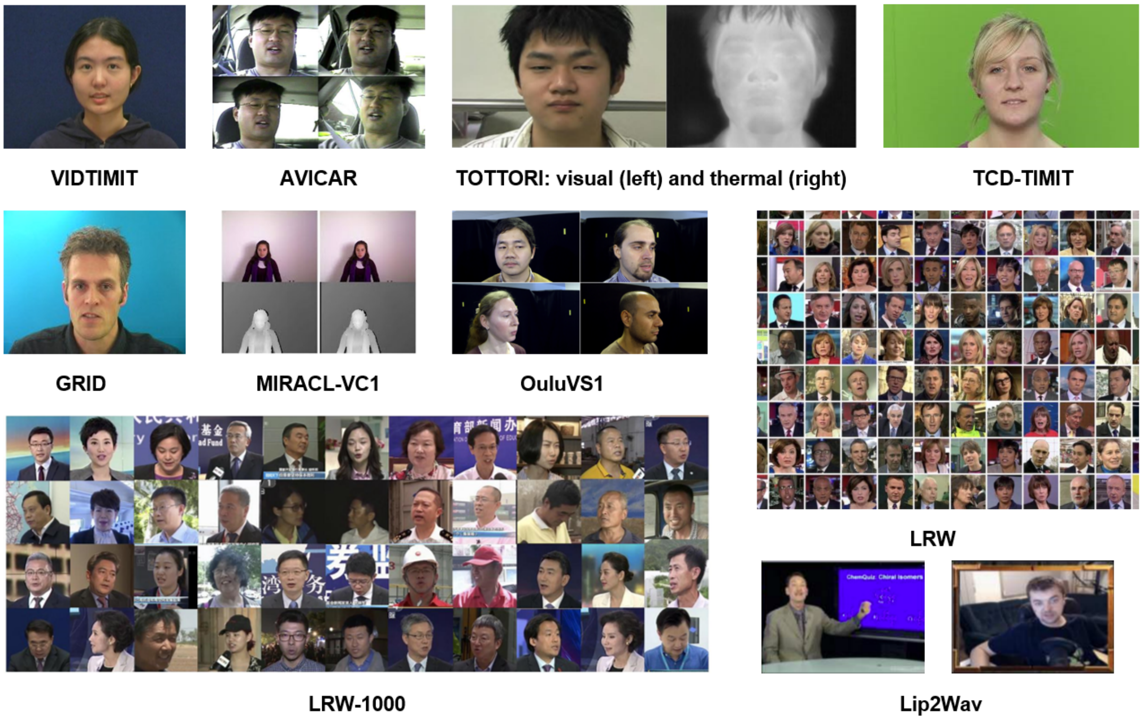
Table A.1 summarizes a representative set of commonly-referenced databases. The table includes information on the year of release, types of speech units, number of unique classes (such as vocabulary size, unique phrases or sentences, depending on the class type), the number of unique speakers, image resolution, frame rate and the approximate total duration recordings in the database. There are numerous similar databases in the field; the table provides an overview of a sample of the largest and most commonly cited databases.

The IBM corporation was an early entrant to the field, and generated numerous proprietary databases of 30-50 hours each for the purpose of audio-visual speech recognition; unfortunately none of them are publicly available. IBMViaVoice is one of these datasets, it was collected in 2000 and included 290 speakers pronouncing sentences with vocabulary of approximately 10,500 words.

I note here the well-known and widely referenced TIMIT database. TIMIT was published in 1989 by a team comprised of members from Texas Instruments (TI) the Massachusets Institute of Technology (MIT), and the Stanford Research Institute (SRI) International. The project was officially known as the DARPA-TIMIT Acoustic-Phonetic Continuous Speech Corpus. The researchers recorded 10 spoken sentences, of duration 30 seconds, from 630 speakers. The sentences and speakers were chosen to maximize acoustic-phonetic coverage of the English language, with an explicit objective to balance regional dialects and gender [17]. It was stated that the sentences were designed to provide a sufficient balance among both phoneme instances and phoneme pairs. The TIMIT data (in whole or part) are widely used to benchmark new methods and generate customized datasets.

One of the earliest publicly available audio-visual datasets was VIDTIMIT [39]. 43 participants were uttering 10 sentences each, out of 346 different TIMIT sentences. In addition to the sentences, each person performed a head rotation sequence in each recording session.

Figure 2-2: Example images from different audio-visual speech datasets



AVICAR is another multi-speaker and multi-view dataset [3]. It was developed in a car cabin, where four cameras were distributed on the dashboard to record a subject's face from four different view angles. The audio was collected by eight microphones placed on the front passenger seat. 86 speakers were asked to pronounce 4 types of utterances: digits, letters, phone numbers and TIMIT sentences. At the time, this database solved the problem of a small number of speakers.

The next relevant database was completed by Japanese researchers. The database is herein referenced as "Tottori", in recognition of the university where the researchers were based. They gathered five sets of five Japanese words from three subjects. It is a very small database: the total duration of the video is approximately 4 minutes. It is noteworthy as it seems to be the first published conference paper that combines the thermal image data with the typical visual video data.

The GRID corpus was published in 2006 and its popularity has increased over time [38]. The corpus consists of 34 subjects uttering 1000 constrained phrases, produced as all combinations of "color", "digit" and "letter" with additional words, for example "Place red in C 3 please" or "Lay green by B 2 now".

The OuluVS1 [24] and OuluVS2 [25] databases are two of the most commonly used datasets for evaluating visual speech recognition systems. OuluVS2 in particular has come to be regarded as a standard public benchmark multi-view dataset. The data was collected with high resolution from 52 subjects generating nearly 1600 utterances. The data collection process was divided to three stages for each speaker. During the first stage, a speaker was uttering continuously ten randomly generated sequences of digits. In the second stage the subject was asked to pronounce 10 short English phrases of daily use, such as "Thank you" and "You are welcome", and in the third stage the subject read ten random TIMIT sentences, which are considered to be phonetically rich.

MIRACL-VC1 is a dataset which consists of both visual and depth streams of images [22]. 15 speakers pronounced a set of ten words and ten phrases ten times each.

TCD-TIMIT is a multi-angle high-quality audio-visual database, which includes 62 subjects reading a total of 6913 TIMIT sentences [37]. Video recording was done from two different angles for each speaker: straight on, and at 30 degrees. Notably, TCD-TIMIT was shot with a greenscreen panel surface behind speakers' backs "for possible speaker segmentation applications as in CUAVE" [23].

While all preceding datasets were collected in a laboratory setting, most of the following were collected in the wild. Speaking in the wild datasets have become more popular over time, as the field had reached a kind of plateau in terms of laboratory results, and the more interesting applications and challenging research problems were associated with lipreading of video collected in public and therefore noisy spaces such as transit hubs (airports and train stations).

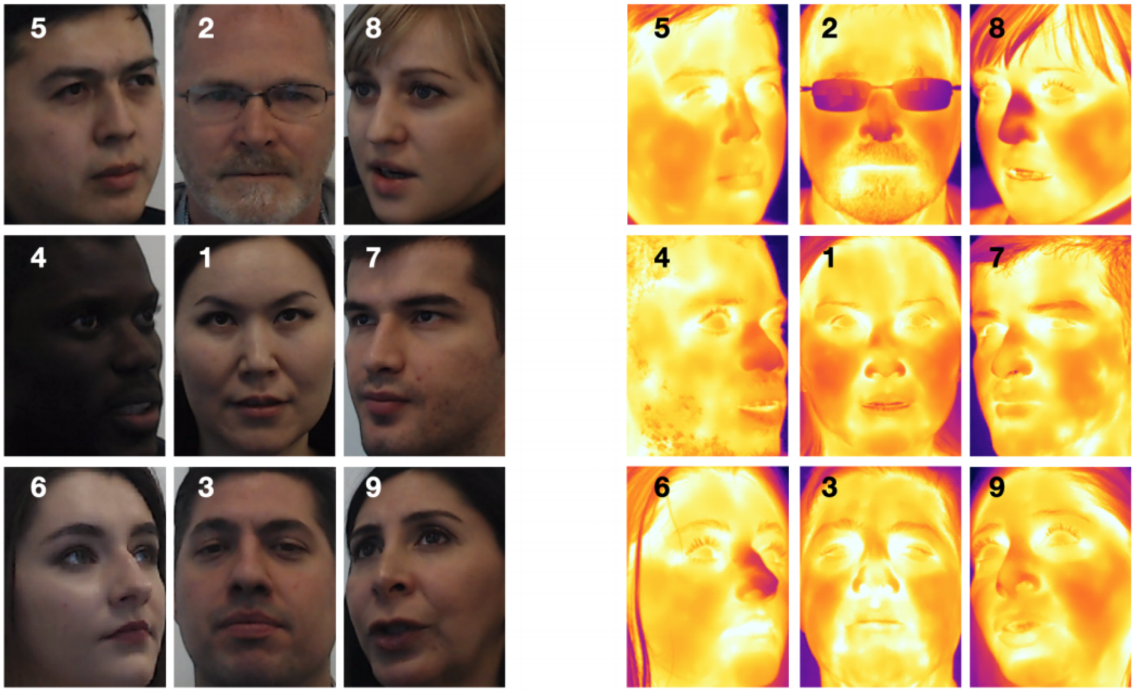
The next three databases were collected by one research team; the corresponding papers are mentioned further below. These three examples are very large-scale visual speech recognition datasets, as the team extracted thousands of hours of spoken text from BBC TV broadcasts covering a wide vocabulary size of thousands of different words, from over one thousand different speakers [19].

The lexicon for the first one, Lip Reading in the Wild (LRW), was collected by picking out the 500 most frequently occurring words with length ranging from 5 to 10 characters. Its duration is approximately 111 hours in total. The Lip Reading Sentences (LRS) database consists of individual sentences/phrases which were separated using the punctuation in the transcript. The sentences were constrained to 100 characters or 10 seconds in length. The dataset contains thousands of different speakers and lasts about 328 hours. The Multi-View Lip Reading Sentences (MV-LRS) database is is an extension of LRS, in that it includes a wider range of subject's profile. The database was fitted to the purposes of the paper, which was modeling the lip reading in profile. For this reason, the MV-LRS authors took the faces angled from 0 to 90 degrees, where 0 is front view, and 90 is profile (the previous databases were constrained to include face angles of no more than 30 degrees).

One of the most recent databases is LRW-1000, collected in 2019 [20]. The authors claim that it is currently the largest word-level audio-visual dataset and also the only public large-scale lipreading dataset for the Mandarin dialect of the Chinese language. It was extracted from Chinese national TV stations, in a manner similar to that of the three previous datasets. It contains one thousand classes with 718018 samples from more than two thousand individual speakers. Each class corresponds to the syllables of a Mandarin word.

The Lip2Wav dataset was collected by the authors of the paper [29] in 2020. It consists of lectures downloaded from YouTube and presented by 5 speakers on the topics of chemistry, chess, deep learning, hardware security and ethical hacking. The overall duration of the dataset is 120 hours, it was gathered for the purpose of exploring individual speaker specifics, with approximately 20 hours of talking for each speaker.

Figure 2-3: Examples of corresponding visual and thermal image pairs of SpeakingFaces dataset [1]



Finally, Abdrakhmanova et al., of the Institute for Smart Systems and Artificial Intelligence (ISSAI) of Nazarbayev University collected the SpeakingFaces [1] dataset in 2020. It is a publicly available multiview large-scale dataset that includes aligned streams of visual, thermal and audio recordings. SpeakingFaces was collected from a balanced sample of 142 subjects. Each person was asked to utter approximately 100 phrases out of a pool of 1800 unique human-computer interaction phrases (such as ‘play Despacito’), with an average total duration per speaker of approx. 20 minutes.

The video was captured from a total of nine different positions (angles) of a subject's face (Fig. 2-3).

While there are additional thermal imaging datasets that include the facial region, they are not relevant to this work as the subjects are not speaking, thus, lipreading is not an achievable task.

2.2 Related Works

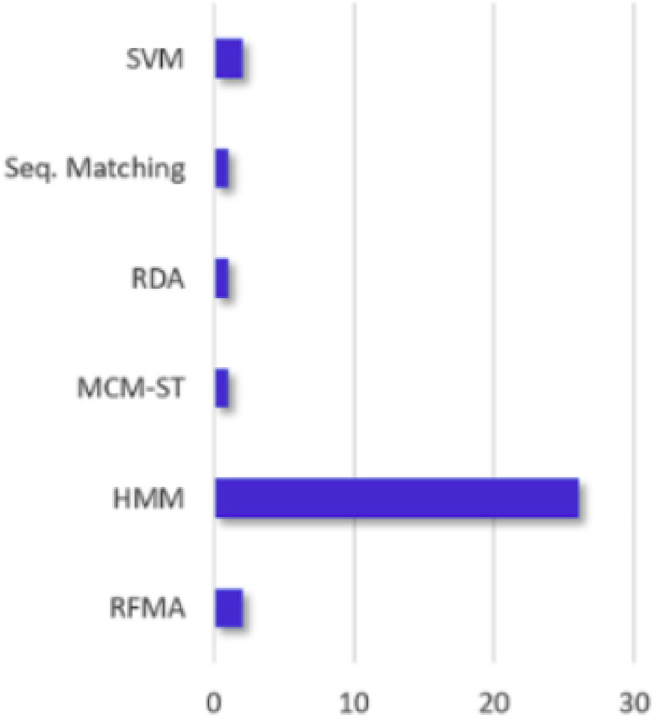
Table A.2 lists the main works on the topic, which have influenced developments in the field. This section is divided into two part depending on what kind of output the studied model was expected to produce: lip2text models to generate text output, and lip2speech models to generate audio speech.

2.2.1 Lip2text Generation

In this section I have prioritized the papers published since 2016, based on the observation that the major improvements in results have only occurred in recent years. I have recorded the databases that were used, described the models and summarized the final results. The exception here is the Saitoh and Konishi paper [32], which is noteworthy due to the innovation in the use of thermal image data streams. It is an older paper, from 2006, with a very small database. They reported modest success of their thesis, citing a word recognition rate (WRR) of 76% while using visual image only, 44% WRR using thermal image only and the modest improvement to an 80% WRR using both channels. The authors gave an explanation to this increase that the thermal images register the changes in the temperature in the mouth area when a speaker breathes out, and therefore gives additional data for the model for distinguishing different visemes (an image of a face/mouth which depicts some sound, adapted from "phoneme").

Regarding the methods, the year 2016 also marks the emergence of the now-common practice used to solve the speech recognition problem by first extracting facial features using Convolutional Neural Networks (CNNs) and then classifying the

Figure 2-4: Classification methods and the number of times they were utilized in the lipreading papers (2007-2017) [12]



phonemes and visemes using some kind of Recurrent Neural Network. Prior to 2016 the most popular method of classification was Hidden Markov Models (HMMs) (Fig. 2-4), which at the time were achieving only 14 to 70% of accuracy [12].

In 2016 Wand et al. [40] showed that the neural network based lipreading system performs significantly better in terms of WRR than a system based on a conventional processing pipeline at that time (such as HOG for feature extraction and SVM for classification). This outcome can be seen in the Table A.2 where the Capital "V" in the "Accuracy" column indicates the WRR on video-only speech recognition, while "A" and "AV" refer to audio-only and audio-visual experiments, respectively.

Assael et al. [2] used the GRID database to establish new state-of-the-art performance on the dataset at the time - WRR of 95.20%, using a system with CNN encoders and bi-GRU as a classifier. The authors claimed that their architecture was the first end-to-end model which was using sentences for automatic lipreading.

There are two teams in particular which are well-known for their dedication to the lip-reading problem. The first one, consisting of Chung and Zisserman et al, released a number of papers [5, 6, 7, 8] on the topic and introduced a new benchmark for performance – their LRW database [6]. In 2017 Chung and Zisserman created a new database MV-LRS [7], which included faces in profile unlike the previous one, and proposed a model, which gave an accuracy of 88.9% on OuluVS2 database and 37.20% on the MV-LRS database for profile speech recognition. In the same year they created one more database based on BBC TV extractions, this time consisting of sentences, and included an audio channel to the research. The accuracy of video-only speech recognition was 49.8%, audio-only – 37.1% and audio-visual – 58% [5]. In 2018 they returned to LRW and OuluVS2 and improved the performance on these datasets to 66% and 94.1% [8].

The second prominent team, Petridis, Pantic, et al., started with OuluVS1 and trained a deep autoencoder with a bottleneck layer, gaining 81.8% WRR [26]. In 2017 they moved to OuluVS2 and using bi-LSTM achieved maximum accuracy 96.9% [28]. In 2018 and 2020 they started experimenting on audio and audio-visual inputs. In 2018 they achieved the best performance on LRW [27], beating the previous record of Stafylakis and Tzimiropoulos from 2017 [35]. In 2020 they improved the performance on LRW, and to the best of my knowledge have the current state-of-the-art performance on this dataset with word recognition rate of 85.3% for video-only, 98.46% for audio-only and 98.96% for audio-visual speech recognition [21].

Yang et al. [41] presented a benchmark for lipreading in the wild in Chinese, named LRW-1000 [20]. Their results showed no improvement in Word Recognition Rates (WRRs), showing WRR of 34.76%, but the database is new, so there is some space for improvement. Martinez et al. [21] for now have the best performance to date on this dataset as well with 41.1% accuracy on lipreading.

2.2.2 Lip2speech Generation

All of the afore-mentioned works were focusing on lip2text generation. Starting from 2017 there was an increase in the studies interested in synthesizing audio speech from silent lip movements. For example, Le Cornu and Milner [18] reported 85% word accuracy on GRID corpus using regression and classification methods for feature extraction and RNN for generating the audio. Table A.2 presents lip2speech papers highlighted with gray color, distinguishing them from the lip2text papers. These entries include complementary metrics (STOI, ESTOI and PESQ), which are specific to the audio processing field and indicate the intelligibility level of the synthesized audio.

The same year year Ephrat and Peleg [11] published an end-to-end architecture called Vid2Speech. Using the GRID corpus, they showed how CNN can extract visual features in order to reconstruct the audio based on silent video. Later that year they improved the model and tested it additionally on the TCD-TIMIT dataset [10].

Kumar et al. [16] used all 53 speakers of OuluVS2 database and claimed that multi-view videos got better results compared to single-view using an architecture with first was classifying a frame by its angle and then processing it with a deep neural network to obtain audio or text results.

Finally, Prajwal et al. [29] collected their own dataset consisting of Youtube lectures (on subjects such as chemistry, chess, and deep learning), and built a lip2speech model, but could not report on resulting word recognition rate due to the lack of text transcripts in the videos. They also tested their architecture on other benchmark datasets apart from their own: applied to the GRID corpus they achieved a recognition rate of 85.92%, TCD-TIMIT 68.74% and Lip reading in the wild dataset – 65.8%.

Overall, the field of lipreading through visual speech recognition has achieved significant progress in the past several years, but their performance still lags behind the audio-based systems. Additionally, even though researchers achieved remarkable results in building deep learning architectures for lip-reading in basic databases like OuluVS2 and GRID, their performance has not yet been extended to more complicated examples like LRW or LRS. Continuous natural speech recognition is also expected to be a developing field at least in the next decade. The addition of thermal images to the lipreading problem seems to be an unexplored area for further work, with only one paper currently published on the topic.

Chapter 3

Research Methodology

The Lip2Wav model [29] was selected to serve as the baseline system to test the hypothesis of this thesis work; the rationale is described below.

The Lip2Wav model accommodates experiments on the full range of tested dataset types: those focused on uttering words, phrases and sentences, recorded both in a fixed environment and in the wild. The model is speaker-specific, such that it is trained for each individual person speaking; the authors state that they were inspired by the fact that it is easier for professionals to lip read people with whom they interact frequently.

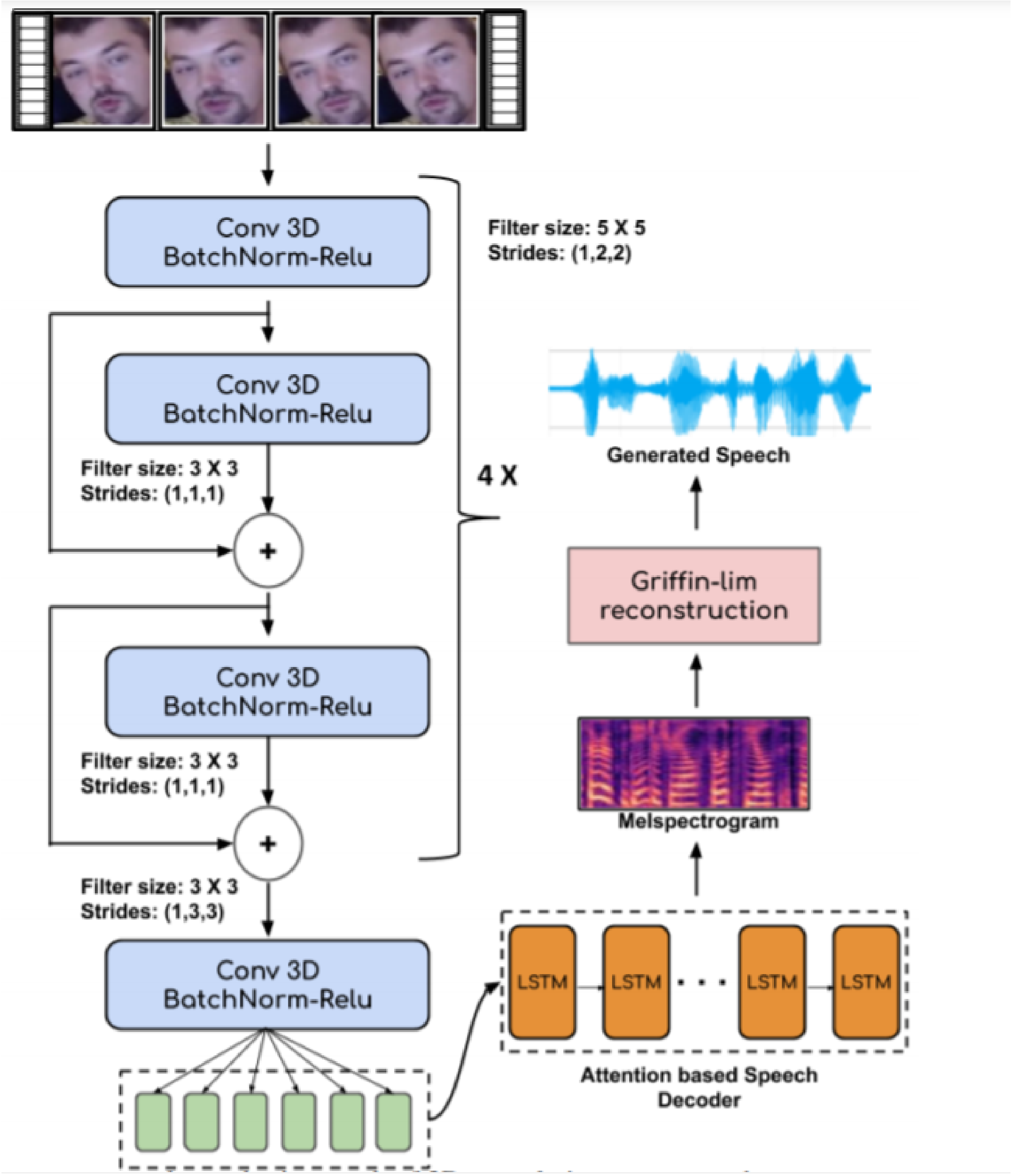
The Lip2Wav uses more up-to-date methods as compared to the Tottori architecture [32]. Tottori used LDA as a feature extractor from the image streams and eigenimage waveforms as a decoder of the ROI embeddings, while Lip2Wav uses CNN layers for encoding the images and LSTM + attention layers for decoding them into the output.

This chapter will describe how the original model processed the videos and how I modified the system to add thermal images to the input.

3.1 Original Architecture

Fig. 3-1 illustrates the architecture of the model used in this study. During the preprocessing step the model splits the video into frames, identifies the face as the region-of-interest (ROI) and then extracts the facial ROI from each frame. For this purpose the authors use the pre-trained s3fd face detector [42]. The data is further divided into training, validation and testing sets, 90%-5%-5%, respectively.

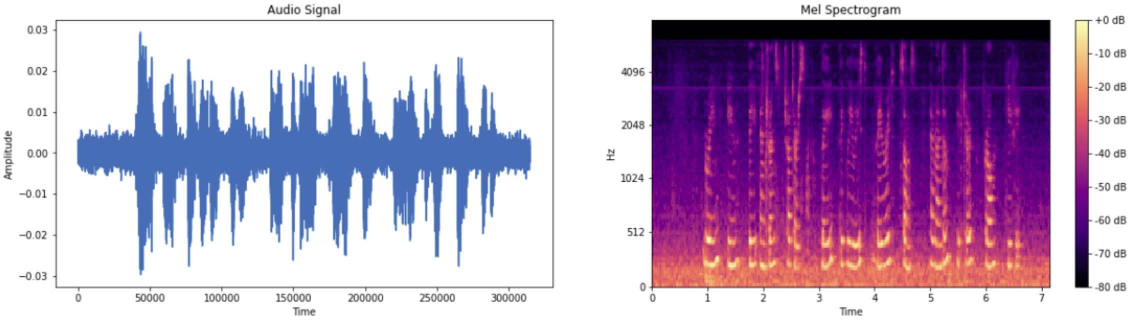
Figure 3-1: The architecture of the model [29]



The sequence of faces is fed to a face encoder, which consists of a stack of 3D convolutions with residual skip connections and batch normalizations, and outputs a vector for each image. For the decoder the authors chose to adapt Tacotron 2 [33], which was originally created to synthesize audio speech from text input, conditioning it on the encoded visual stream. The melspectrogram transformation of an audio extracted from the corresponding video is used as the ground truth for training.

Most of the lip2speech models use melspectrograms as an encoding for audios, as it is not so straightforward for the neural networks to process the raw audio signal; the melspectrogram can be treated like a visualization of an audio data (Fig. 3-2). The fast Fourier transform algorithm is utilized here to convert the signal into spectrum of frequencies, additionally mapped according to the mel scale, which gives the resulting melspectrogram.

Figure 3-2: An example of audio signal (left) and its melspectogram (right)



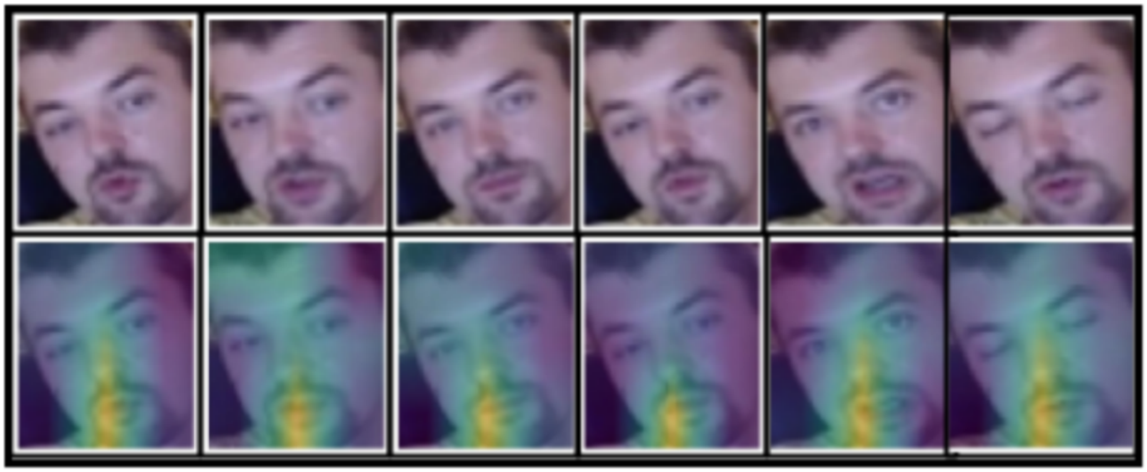
The decoder is trained to output a melspectrogram, which is then further converted into audio using the Griffin-Lim reconstruction algorithm, which is a kind of reverse function mapping of a spectogram back to the time domain (audio signal).

The authors suggest to train the system until such time as the loss value plateaus for more than 30,000 epochs.

After training, the authors plotted the activations of the penultimate layer of the encoder and the attention alignment from the decoder and concluded that the system focuses not only on the mouth area, but also a slightly wider region of interest, including features such as the nose, the brows and the forehead (Fig. 3-3). Therefore the region of interest encompasses much of the the whole face, not only the lips (as previously expected).

There is a relatively standard set of metrics used to assess the performance of lipreading systems; I have adopted these metrics for this study for purposes of comparability. The Word Recognition Rate (WRR) is used to assess the accuracy of the

Figure 3-3: Visualization of the attention of the decoder [29]



synthesized text, while the Short-Term Objective Intelligibility (STOI), the Extended STOI (ESTOI), and the Perceptual Evaluation of Speech Quality (PESQ) are applied for estimating the quality of the audio output.

The most significant metric is the WRR, which is a calculation of words recognized correctly over the total number of words; it is this measure that is used to determine the potential improvement of my approach over prior methods. In this instance, it is an indication of the relative contribution of the thermal data to the results. It can be referred as word accuracy (WAcc), and it is similar to the widely familiar accuracy measure which is commonly used for classification problems: the WRR ranges between 0 and 1 (or 0% and 100%) and indicates the closeness of a synthesized output to its true value in automatic speech recognition problems. The WRR can be computed as follows:

$$WRR = \frac{N - S - D - I}{N},$$

where N is the overall number of words in the reference text, S is the number of words that were substituted with another ones, D is the number of deletions which occur when a whole word was skipped in the resulting text, and I is the number of insertions such as when a word has been replaced by a consonant phrase. Many studies in the field also like to use the word error rate (WER), which is associated with the WRR:

$$WER = 1 - WRR.$$

For the purpose of this study, I emphasize WRR as an estimate of the relative accuracy, rather than WER as an estimate of the incorrectness of the fit. WRR and WER are most typically calculated for lip2text methods, and omitted while working with lip2speech architecture, only assessing the quality of the resulting audio. The authors of Lip2Wav [29] obtained the WER by using out-of-the-box Google speech to text API to compare the accuracy of their model to existing studies.

In many prior papers, the STOI, the ESTOI, and the PESQ are also used to assess the overall quality of the synthesized audio output. These are standard speech quality metrics used in the lip2speech papers to provide different estimations of the intelligibility of an audio file. The PESQ metric is now considered slightly dated, as it has been superseded by a new ITU-T standard, the Perceptual Objective Listening Quality Assessment (POLQA) [4], but for purposes of direct comparability and due to the restrictive licensing requirements of the ITU, I utilize the PESQ metric for this study.

In brief, the STOI [36] and ESTOI [14] metrics measure the correlation value between processed or distorted audio speech and clear sound. STOI uses the average of linear correlations between short temporal envelops of original and noisy audio, while ESTOI follows the same procedure, but uses spectral correlation coefficients. Therefore, both STOI and ESTOI range from 0 to 1, from least to most intelligible audio prediction, respectively. Lastly, PESQ [31] is an algorithm built to predict mean opinion score (MOS), a "true" value of audio quality, calculated by averaging the scores given by subjects ranging between 1 (bad) and 5 (excellent).

These metrics are calculated using standard code libraries, thus, they did not require re-implementation, and also enabled better direct comparisons with earlier results.

Table 3.1 lists the four datasets that were tested for the Lip2Wav model and shows

Table 3.1: Lip2Wav results

Dataset	Hours per speaker	STOI	ESTOI	PESQ	WRR
GRID	0.8	0.731	0.535	1.772	85.92%
TCD-TIMIT	0.5	0.558	0.365	1.350	68.74%
LRW	0.03-0.08	0.543	0.344	1.197	65.80%
Lip2Wav (chem)	20	0.416	0.284	1.300	-

how their attributes such as durations per speaker, STOI, ESTOI, PESQ, and WRR measures affect each other. All four estimates are roughly proportional, for example if we compare the results on GRID and LRW as the datasets with the largest and the smallest WRR, respectively, we can notice that the system performed best using the GRID corpus, as shown by the STOI, ESTOI and PESQ metrics, while the LRW dataset yielded the lowest values of audio intelligibility measures among the datasets with defined WRR. This could mean that the better the quality of the synthesized audio is, the better speech2text system performs in terms of accuracy.

Additionally, there is a pattern in the dependence of the model performance from the type of environment in which the dataset was collected, and the average duration of speech. The first two datasets – GRID and TCD-TIMIT were collected in fixed environments, while LRW and Lip2Wav are in-the-wild video datasets, therefore the performance of systems using those datasets are slightly degraded. Even 20 hours per speaker of Lip2Wav dataset do not increase the intelligibility metrics. Furthermore, the performance of systems using the TCD-TIMIT dataset is worse than those using the GRID dataset, probably because TCD-TIMIT is a richer dataset with larger vocabulary, while GRID consists of combinations of “color”, “digit” and “letter” with additional words (though these datasets have comparable numbers of hours per speaker).

3.2 Changes in the Model

I first replicated the original model by training with the visual images of a new dataset (described in the next chapter). I decided to leave the thermal image as it is (3 channels) in order to match the input dimension of the original model for visual stream, therefore there was no need to adapt the system in terms of data dimensionality.

For the purpose of encoding the combined streams, a concatenated array of the visual and thermal images was fed to the system. This method of concatenating two different streams of images was used in a number of similar research. For instance, Pujar et al. [30] combined visual and depth images into four channels (RGBD) and fed it as an input to CNN encoder for indoor scene classification. Another example was published by Shopovska et al. [34], where they concatenated visual and thermal images into 6-channeled array, and trained their model in order to increase pedestrians' visibility using deep neural network.

Similarly, the new input for the Lip2Wav model's combined input represents not only visual information but also indicates thermal differences between the pixels. In this case the thermal image was converted to grayscale to preserve an equivalent number of input channels: 3 channels for the original model and 3+1 for the combined model so as to include the additional thermal image channel.

In order to include a thermal data stream matching the visual stream it was necessary to identify the facial region on the thermal image. This was achieved by aligning the visual and thermal images (described in the next chapter) and cropping the face based on mapping the bounding box from the corresponding visual image.

The main structure of the CNN encoder did not require any changes due to the similar nature of the images (comparable size of the images). The only modification in the encoder was a new option for input shape: when the system received visual or thermal input, it would be conditioned for 3-channeled input, and for combined input it would expect a 4-channeled array. Additionally, for combined image the shape of the initial kernel was changed to 4 channels to match the input's dimensions.

The RNN decoder needed some adjustments, as it was fit for training big data files of Lip2Wav dataset. When the authors trained the system on the GRID and TCD-TIMIT datasets they chose to halve the number of layers of the LSTM network in order to avoid overfitting. The relative size of the dataset I used for training is comparable to both GRID and TCD-TIMIT, therefore I followed their "halving" approach and set the decoder units to 512 instead of 1024.

Chapter 4

Datasets Description

The working hypothesis for this thesis is that the inclusion of thermal video data aligned with the corresponding visual data stream could improve lipreading recognition rates. Thus, it is imperative that the dataset used in the experiment include synchronized thermal data, such that the model as described can take the two video streams as inputs and then accurately generate the audio as output.

As noted, there are few existing datasets that include the thermal data stream. In this section I review the work done to identify and evaluate the existing datasets, and describe in some detail SpeakingFaces, the largest such dataset collected to date, and then describe why it was necessary to design, collect, and prepare a new dataset for the purpose of the project.

4.1 Tottori

As described previously, the "Tottori" dataset was the first publicly referenced dataset that recorded the thermal data stream and used the data in the analytics for the purpose of visual speech recognition; in their case, the inclusion of thermal data marginally improved the lipreading recognition rate. However, the dataset is quite small, and thus insufficient for the purpose of this thesis. Moreover, it is not publicly available.

4.2 SpeakingFaces (SF)

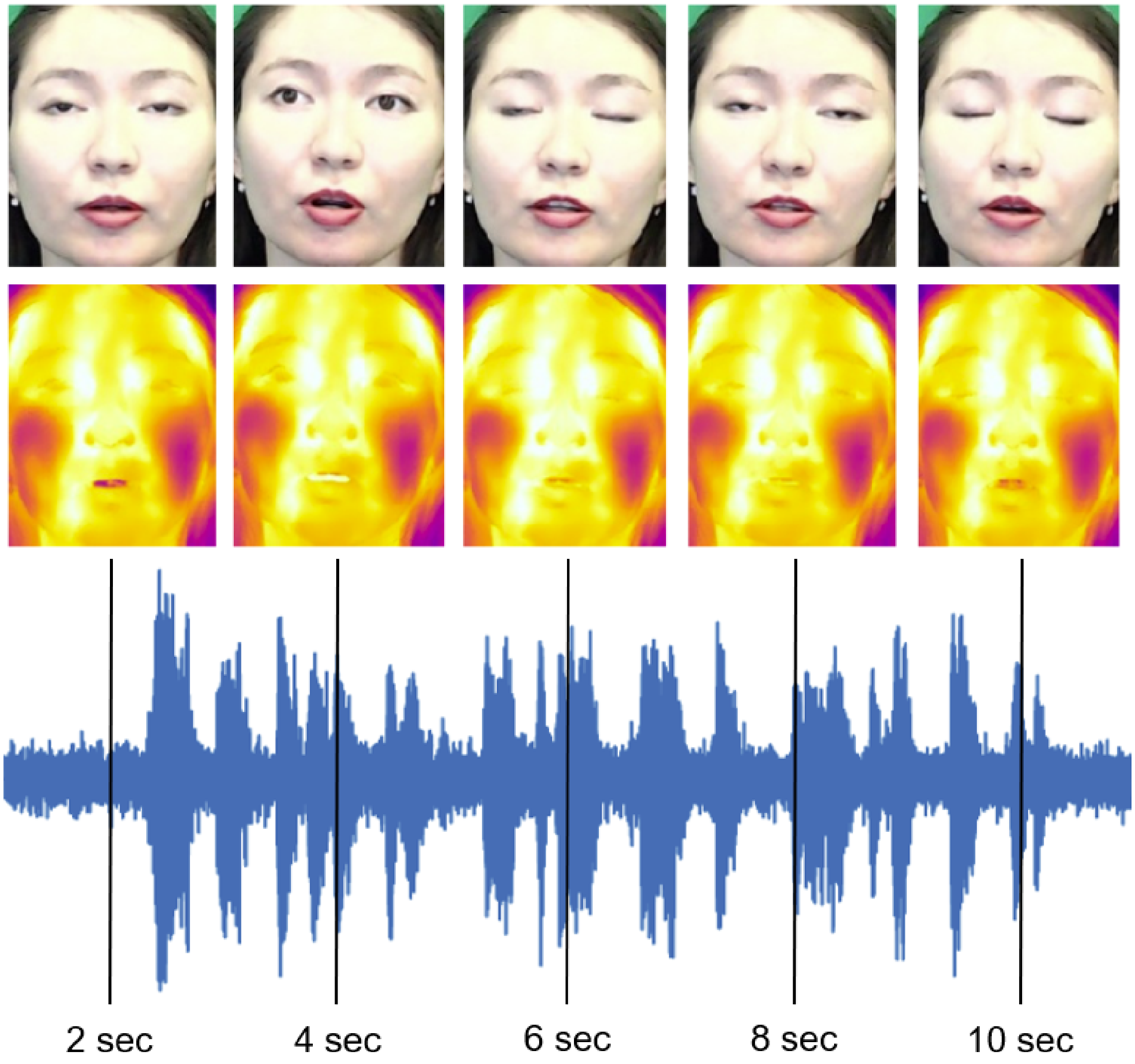
As noted in the Literature Review, SpeakingFaces is a large dataset, collected from numerous subjects, well-balanced and publicly available for purposes of research. At the time of publication, SpeakingFaces was the most extensive dataset of its kind, and was the first option considered for the purposes of the project. However, after initial experimentation with the system architecture using a subset of the SpeakingFaces data, it became apparent that it was not the best fit for the project. Firstly, the dataset is too large to consider in full, for purposes of training, due to the computational requirements and time constraints. Secondly, and of greater significance, the duration of recordings per subject is relatively short, ranging from 10-20 minutes per subject. By comparison, the Lip2Wav dataset has on average 20 hours per speaker. Based upon the preliminary investigations using SF, it was determined that for the purposes of the project a speaker-specific dataset comprised of longer duration recordings of individual speakers would be more effective; the optimal duration was set at approx. 2 hours, as explained below.

4.3 SpeakingFaces LipReading (SFLR)

Based on the prior review of existing datasets, and the detailed examination of the full SpeakingFaces dataset, I determined to use instead a speaker-specific model of extended recording duration. Given the relative lack of such data sources, the ISSAI team agreed to design and collect an extension of the SF dataset, designated as SpeakingFaces LipReading (SFLR), consisting of the two main streams of visual and thermal video (Fig. 4-1), but enhanced with features more convenient to the purposes of the project.

This section will describe the similarities and differences between SF and SFLR datasets, changes in the setup and how the changes affected the process of data cleaning and preparation.

Figure 4-1: Snapshots of the visual and thermal image streams with 2-second intervals



4.3.1 Data Collection

The SFLR dataset was collected by the ISSAI team, on site at the research labs in Nur-Sultan, Kazakhstan. For the purpose of proof-of-concept, it was sufficient to capture the data of a single subject; if successful, the dataset can be readily expanded, and the model scaled to consider the additional subjects.

The subject was asked to utter the phrases from the same pool of 1800 phrases used for SF dataset. The list of these utterances was gathered from several sources that specialize in the style of imperative commands typical of human-computer inter-

Table 4.1: SF and SFLR datasets

Name	Year	Type	Classes	Speakers	Resolution	Duration
SpeakingFaces	2020	Phrases	1800	142	768 × 512 (visual), 464 × 348 (thermal), 28 fps	45 hours
SpeakingFaces LipReading	2021	Phrases	1298	1	768 × 512 (visual), 464 × 348 (thermal), 28 fps	2 hours

action; sources include the Stanford University open source digital assistant command database, and a selection of common commands used with popular digital assistants such as Siri and Alexa.

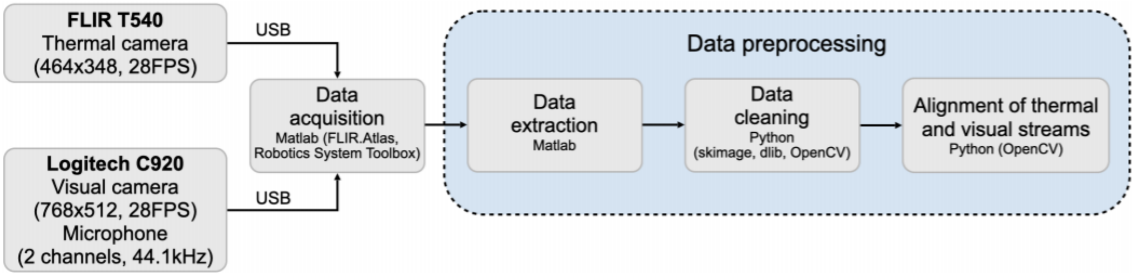
The participant spoke for a longer period of time compared to the subjects in SpeakingFaces. The duration was selected as a middle ground between the examples of the datasets used in Lip2Wav paper. Lip2Wav dataset has 20 hours per speaker, but the videos can be classified as speaking in the wild, and hence this amount of data was necessary for training the model. The GRID and TCD-TIMIT datasets with 0.8 and 0.5 hours per speaker, respectively, have shown a good performance under the Lip2Wav model, but the vocabulary per speaker was much more simple, shot in fixed environments and therefore did not require long recordings.

SFLR has also been recorded in controlled laboratory environment, hence there was no need for 20 hours of video, but the utterances are more complex and with wider vocabulary, so 0.5 hours was considered insufficient. Thus, after due consideration, the requisite recording interval of the subject was estimated at approximately two hours.

The subject pronounced 1298 phrases taken from the SF utterances, which summed up to 1.9 hours of speaking. The utterances were divided to sessions, two utterances per video, for easier batching during the training. The main features of the dataset and their comparison to the SF dataset are listed in Table 4.1.

The setup was slightly modified to one more convenient for recording and to reduce the need for further preprocessing. There is only one position for the speaker - straight face, therefore it was possible to fix the cameras on a tripod, such that the image would be more stable. The visual camera was attached on the top of thermal camera in order to more closely align the frames. The cameras and their specifications were the same as the ones used in SF data collection (Fig. 4-2). For SFLR the team added two sources of light, so as to avoid the facial shadows caused by the single-source overhead illumination that had afflicted the original SF dataset. The lights were adjusted to prevent shadows on the background plane as well. A greenscreen was placed behind speaker's back following the example of TCD-TIMIT corpus, for easier face recognition in visual images, as occasional errors in face detection were encountered while preprocessing the original SF dataset.

Figure 4-2: Data pipeline for SpeakingFaces [1]

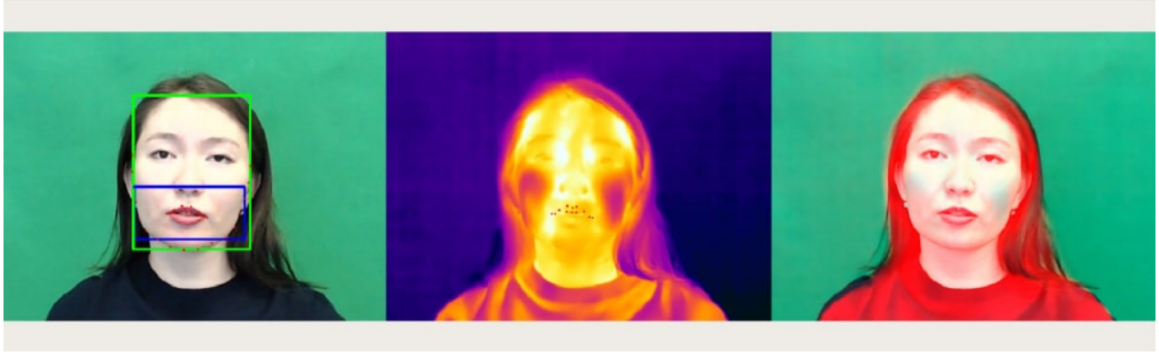


4.3.2 Data Preparation

The raw data required some cleaning and adjusting. For example, it was necessary to align the visual and thermal recordings, even though there was only one position for the speaker and the cameras were attached; they still had slightly different viewing angles. Additionally, the thermal camera had an autofocus property, which would occasionally change the shift of the frame. It was necessary to align the corresponding visual and thermal images, matching them manually, as part of the data preprocessing stage. As the recordings were collected over an interval of nearly two months, it became apparent that for each day of the recording the thermal camera was setting up the autofocus differently, hence more than one aligning process was needed.

Unfortunately, the data collection and occasional re-shoots introduced minor complications in session identification and alignment. The issue was resolved by manual classification of each session. As a result, I distinguished 12 classes of aligning.

Figure 4-3: An example of the process of aligning SFLR dataset's subject



The aligning was done by detecting the lip landmarks of a visual frame and matching them with the lips on a corresponding thermal image, such that while cropping the region of interest (ROI) on a visual image, the program would be able to use the same coordinates to crop the ROI on the thermal image as well (Fig. 4-3). After ensuring that the alignment was correct for several random frames of a particular session, the vertical and horizontal shift values were recorded and applied for all frames in that session.

The artifacts common for the SF dataset, such as "freezing" of the thermal camera's stream, frame flickerings, and image blur were detected during data collection, corresponding sections were deleted and re-shot on-the-go, so there was no need to search for them.

Additionally, after preliminary experiments with the dataset, it was apparent that the audio volume was uneven from one video to another, as the volume of synthesized audio was noticeably sometimes different from the original. Therefore, all audios were normalized using ffmpeg-normalize program in accordance with the EBU R128 loudness normalization standard.

Chapter 5

Results and Analysis

The goal of this chapter is to present and discuss the results of my thesis work using the metrics described above: STOI, ESTOI, PESQ and WRR. Based upon the literature review, I identified the current state-of-the-art model for lip2speech systems, Lip2Wav, and was able to download and install the system locally, using the DGX computational environment of the Institute for Smart Systems and Artificial Intelligence. I was able to configure and conduct test-runs of that system, and then adapt it for the inclusion of the thermal data stream. I ran the system for training purposes using the SFLR dataset.

Table 5.1 presents the results of the trained model on visual image from SpeakingFaces LipReading data, and compares them to the ones of the original Lip2Wav paper. As shown in the table, the current performance metrics for the visual stream are lower when I run the system on the SPLR dataset than they are when run on the Lip2Wav dataset, but comparable to the overall results in the field.

Table 5.2 shows the results of training of Lip2Wav model on SLFR dataset's

Table 5.1: The results of training Lip2Wav on the original dataset and on SFLR visual image

Dataset	STOI	ESTOI	PESQ
Lip2Wav [29]	0.282	0.183	1.671
SpeakingFaces LipReading	0.134	0.041	1.395

Table 5.2: The results of training Lip2Wav on different inputs from SFLR dataset

Channels	STOI	ESTOI	PESQ	WRR
Visual	0.134	0.041	1.395	14.2%
Thermal	0.045	0.002	1.141	0.00%
Both	0.125	0.031	1.372	14.3%

different types of data: visual image only, thermal image only and both streams simultaneously. The metrics of the thermal-only model are significantly worse than those of visual only model. This can be attributed to the fact that a thermal image contains less amount of information on facial features compared to the corresponding visual image. Taking into the account the small numbers in the metrics for thermal image training and the insignificant change in them from visual image only and combined input models, I conclude that the thermal image in its current resolution does not contribute any significant information for lipreading in this model.

Additionally, it should be pointed out that the word accuracy scores do not meet the state-of-the-art standards. As the metrics were derived by using speech recognition model on synthesized audio, the intelligibility of the resulting audio is the reason for these decreased numbers. The suggestions for their improvements are enumerated in the conclusion.

The training was performed on a DGX-2 server. All of the preprocessing and training procedures were conducted using a set of Python programs adapted from the authors' original source code (available at https://github.com/Rudrabha/Lip2Wav). The environment was set up in accordance with the directions of the authors of Lip2Wav [29].

As shown in the literature review section, lip2text models have higher performance levels than lip2speech models, as measured by the Word Recognition Rate. Taking this into account, if we compare directly the achieved results against the published results of other lip2speech models, we can see that the implemented system was comparable to the others, but did not improve on those results. In short, I was able to configure and adapt a state-of-the-art system, and replicate comparable results, but not yet further improve them.

Chapter 6

Conclusion

The level of interest in lipreading systems has increased in recent years, due to rapid improvements in system performance and the potential utility of lipreading in applications ranging from human-computer interaction to the use of speech2text systems for the hearing-impaired people. However, the challenge of lipreading has not yet been fully met: the results obtained from silent video rarely exceed a Word Recognition Rate (WRR) of 85%, thus leaving substantial room for improvement.

This thesis examines the conjecture that the recognition rate could be improved by augmenting visual image data with aligned thermal image data. The recent improvements in the resolution of thermal cameras provides an increased level of facial feature granularity that could contribute additional information to the machine learning process and thereby potentially improve lipreading accuracy.

Upon reviewing the recent literature, and assessing the current state-of-the-art, I chose to base my work on the Lip2Wav model, as described in the Methodology, and adapt the system to incorporate the thermal data.

There are few existing datasets that include aligned thermal data, as noted in the Literature Review. One of the largest such datasets is known as SpeakingFaces, with which I began my initial investigations by conducting data preprocessing and preliminary analytics. However, I determined that for the purpose of this study it was necessary to have extended data collected from individual speakers, beyond the approximately 20 minutes of utterances per speaker available in the SF dataset.

To these ends, the ISSAI team designed an extended version known as SpeakingFaces LipReading (SFLR), consisting of approximately two hours of recordings of a single speaker, collected under the conditions of the original SF dataset.

I obtained the code for the open-source Lip2Wav system, and configured the code for local execution, then adapted the system to take into account the thermal data as provided in the novel SFLR dataset. I conducted experiments on three variations of the data streams consisting of the visual image stream alone, the thermal image stream alone, and the two combined. As shown in the Results, I was able to replicate the system, generate comparable results for the PESQ measure on visual streams, but PESQ results were lower on the thermal and combined streams.

Upon reflection, the system can be further enhanced by enlarging and improving the dataset. First, as the SFLR dataset's transcripts consist of rather complex phrases, the collection of additional data could potentially increase the training results. Second, the thermal and visual camera images were not matched pixel-by-pixel, i.e. there is still some minor shift in the view angle, which affects the precision of the alignment. Refinements of the recording setup could have a positive impact on the performance of the model. Additionally, the extended dataset could include variations of head postures, as Kumar et al. [16] pointed out that multi-view data gives better results compared to single-view data.

Changes in the model may also cause a positive dynamics in the results. Apart from further fine-tuning of the Lip2Wav system, one can try to implement alternative fusion approaches for the combined model, such as first encoding each image separately, and then concatenating them [15], and more complex architectures [9, 13]. Another option is to try adapting other lipreading models to test the hypothesis, not necessarily lip2speech and speaker-specific one. Furthermore, it is recommended to use POLQA metrics for assessing synthesized audio intelligibility, as an improved successor of PESQ, once its implementation is publicly available.

As future work, the Lip2Wav system as implemented produces synthesized audio tracks; it is feasible that such output can be used as input in a similar lip2text system so as to facilitate the association of the ROI with specific audio outputs in the deep learning process, following Kaldi or ESPNET-based model recipes. Possible findings include the detection of patterns unique to the movements on thermal images, gaining higher lipreading performance through adding thermal video on top of visual image, increasing the robustness of audio-visual speech recognition in adverse environments, and the investigation of results on how the inclusion of audio input affects each of these methods.

Appendix A

Tables

Table A.1: Literature Review: Datasets

Name	Year	Type	Classes	Speakers	Resolution	Duration
TIMIT	1989	Sentences	6300	360	-	30 hours
IBMViaVoice	2000	Sentences	10,500*	290	704 × 480, 30 fps	50 hours
VIDTIMIT [39]	2002	Sentences	346	43	512 × 384, 25 fps	30 minutes
AVICAR [3]	2004	Sentences	1317	86	720 × 480, 30 fps	33 hours
Tottori [32]	2006	Words	5	3	720 x 480, 30 fps	4 minutes
GRID [38]	2006	Phrases	1000	34	720 × 576, 25 fps	28 hours
OuluVS1 [24]	2009	Phrases	10	20	720 × 576, 25 fps	16 minutes
MIRACL-VC1 [22]	2014	Words Phrases	10	15	640 × 480, 15 fps	3 hours
OuluVS2 [25]	2015	Phrases Sentences	10	52	1920 × 1080, 30 fps	2 hours
TCD-TIMIT [37]	2015	Sentences	6913	62	1920 × 1080, 30 fps	6 hours
LRW [19]	2016	Words	500	1000+	256 × 256, 25 fps	111 hours
LRS [19]	2017	Sentences	17428*	1000+	160 × 160, 25 fps	328 hours
MV-LRS [19]	2017	Sentences	14960	1000+	160 × 160, 25 fps	207 hours
LRW-1000 [20]	2019	Syllables	1000	2000+	1024 × 576, 25 fps	57 hours
Lip2Wav [29]	2020	Sentences	5000	5	various, 25-30 fps	120 hours
SpeakingFaces [1]	2020	Phrases	1800	142	768 × 512 (visual), 464 × 348 (thermal), 28 fps	45 hours

Table A.2: Literature Review: Papers

Year	**Reference**	**Database**	**Extractor**	**Classifier**	**Accuracy**
2006	Saitoh and Konishi [32]	Tottori	LDA	Eigenimage waveform + DP matching	RGB: 76.00% Thr: 44.00% Both: 80.00%
2016	Wand et al. [40]	GRID	Eigenlips	SVM	V: 70.60%
			HOG	SVM	V: 71.30%
			Feed-forward	LSTM	V: 79.60%
2016	Assael et al. [2]	GRID	CNN	Bi-GRU	V: 95.20%
2016	Chung and Zisserman [6]	LRW	CNN	CNN	V: 61.10%
		OuluVS1	CNN	CNN	V: 91.40%
		OuluVS2	CNN	CNN	V: 93.20%
2016	Petridis and Pantic [26]	OuluVS1	DBNF + DCT	LSTM	V: 81.80%
2017	Chung and Zisserman [7]	OuluVS2	CNN	LSTM+attention	V: 88.90%
		MV-LRS	CNN	LSTM+attention	V: 37.20%
2017	Chung et al. [5]	LRS	CNN	LSTM+attention	V: 49.80% A: 37.10% AV: 58.00%
2017	Petridis et al. [28]	OuluVS2	Autoencoder	Bi-LSTM	V: 96.90%
2017	Stafylakis and Tzimiropoulos [35]	LRW	3D-CNN + ResNet	Bi-LSTM	V: 83.00% A: 97.72%
2017	Le Cornu and Milner [18]	GRID	AAM	RNN	V: 33% ESTOI: 0.434 PESQ: 1.686
2017	Ephrat and Peleg [11]	GRID	CNN	CNN	STOI: 0.584 PESQ: 1.190
2017	Ephrat et al. [10]	GRID	CNN	CNN	STOI: 0.7 ESTOI: 0.462 PESQ: 1.922
		TCD-TIMIT	CNN	CNN	STOI: 0.63 ESTOI: 0.447 PESQ: 1.612
2018	Chung and Zisserman [6]	LRW	CNN	LSTM	V: 66.00%
		OuluVS1	CNN	LSTM	V: 94.10%
2018	Petridis et al. [27]	LRW	CNN	ResNet + Bi-GRU	V: 83.39% A: 97.72%

Continued on next page

Table A.2 – continued from previous page

Year	Reference	Database	Extractor	Classifier	Accuracy
					AV: 98.38%
2019	Kumar et al. [16]	OuluVS2	VGG-16 + STCNN	Bi-GRU	V: 97.00% PESQ: 2.002
2019	Yang et al. [41]	LRW	CNN	3D-DenseNet	V: 78.00%
		LRW-1000	CNN	3D-DenseNet	V: 34.76%
2020	Martinez et al. [21]	LRW	CNN	ResNet+MS-TCN	V: 85.30% A: 98.46% AV: 98.96%
		LRW-1000	CNN	ResNet+MS-TCN	V: 41.10%
2020	Prajwal et al. [29]	GRID	3D-CNN	LSTM + attention (Tacotron 2)	V: 85.92% STOI: 0.731 ESTOI: 0.535 ESTOI: 1.772
		TCD-TIMIT	3D-CNN	LSTM + attention (Tacotron 2)	V: 68.74% STOI: 0.558 ESTOI: 0.365 PESQ: 1.350
		LRW	3D-CNN	LSTM + attention (Tacotron 2)	V: 65.80% STOI: 0.543 ESTOI: 0.344 PESQ: 1.197
		Lip2Wav	3D-CNN	LSTM + attention (Tacotron 2)	STOI: 0.416 ESTOI: 0.284 ESTOI: 1.300

Bibliography

[1] M. Abdrakhmanova, A. Kuzdeuov, S. Jarju, Y. Khassanov, M. Lewis, and H.A. Varol. Speakingfaces: A large-scale multimodal dataset of voice commands with visual and thermal video streams. *Sensors*, 21(10):3465, January 2021.

[2] Y.M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599, year = 2016,*.

[3] AVICAR corpus. Available at: http://www.isle.illinois.edu/sst/AVICAR/.

[4] Schmidmer C. Berger J. Obermann M. Ullmann R. Pomy J. Beerends, J.G. and M. Keyhl. Perceptual objective listening quality assessment (polqa), the third generation itu-t standard for end-to-end speech quality measurement part i—temporal alignment.

[5] J.S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3453. IEEE, July 2017.

[6] J.S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, Cham, November 2016.

[7] J.S. Chung and A.P. Zisserman. Lip reading in profile. British Machine Vision Association and Society for Pattern Recognition, 2017.

[8] J.S. Chung and A.P Zisserman. Learning to lip read words by watching videos. *Computer Vision and Image Understanding Journal*, 173:76–85, 2018.

[9] L. Ding, Y. Wang, R. Laganiere, D. Huang, and S. Fu. Convolutional neural networks for multispectral pedestrian detection.

[10] A. Ephrat, T. Halperin, and S. Peleg. Improved speech reconstruction from silent video. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 455–462. IEEE, 2017.

[11] A. Ephrat and S. Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017.

[12] A. Fernandez-Lopez and F.M. Sukno. Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing Journal*, 78:53–72, 2018.

[13] C. Hangil, S. Kim, P. Kihong, and K. Sohn. Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, page 621, 2016.

[14] J. Jensen and C.H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers.

[15] B. Khalid, A. M. Khan, M. U. Akram, and S. Batool. Person detection by fusion of visible and thermal images using convolutional neural network. In *2019 2nd International Conference on Communication, Computing and Digital systems (C-CODE)*, page 143, 2019.

[16] Y. Kumar, R. Jain, K.M. Salik, R.R. Shah, Y. Yin, and R. Zimmermann. Lipper: Synthesizing thy speech using multi-view lipreading. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2588–2595. IEEE, 2019.

[17] L.F. Lamel, R.H. Kassel, and S. Seneff. Speech database development: Design and analysis of the acoustic-phonetic corpus. January 1989.

[18] T. Le Cornu and B. Milner. Generating intelligible audio speech from visual speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(9):1751–1761, June 2017.

[19] Lip Reading in the Wild and Lip Reading Sentences in the Wild Datasets. Available at: https://www.bbc.co.uk/rd/projects/lip-reading-datasets.

[20] LRW-1000: Lip Reading database. Available at: http://vipl.ict.ac.cn/en/view$_d$$atabase.php?id = 13$.

[21] B. Martinez, P. Ma, S. Petridis, and M. Pantic. Lipreading using temporal convolutional networks. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6319–6323. IEEE, May 2020.

[22] MIRACL-VC1 dataset. Available at: https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1.

[23] Harte N. and Gillen E. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, February 2015.

[24] OuluVS database. Available at: https://www.oulu.fi/cmvs/node/41315.

[25] OuluVS2 database. Available at: http://www.ee.oulu.fi/research/imag/OuluVS2/.

[26] S. Petridis and M. Pantic. Deep complementary bottleneck features for visual speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2304–2308. IEEE, March 2016.

[27] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic. End-to-end audiovisual speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6548–6552. IEEE, April 2018.

[28] S. Petridis, Y. Wang, Z. Li, and M. Pantic. End-to-end multi-view lipreading. *arXiv preprint arXiv:1709.00443, year = 2017,*.

[29] K.R. Prajwal, R. Mukhopadhyay, V.P. Namboodiri, and C.V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13796–13805. IEEE, 2020.

[30] K. Pujar, S. Chickerur, and M.S. Patil. Combining rgb and depth images for indoor scene classification using deep learning. In *2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, pages 1–8. IEEE, December 2017.

[31] A.W. Rix, J.G. Beerends, M.P. Hollier, and A.P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 749–752. IEEE, 2001.

[32] T. Saitoh and R. Konishi. Lip reading using video and thermal images. In *2006 SICE-ICASE International Joint Conference*, pages 5011–5015. IEEE, October 2006.

[33] J. Shen, R. Pang, R. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. Saurous, Y. Agiomvrgiannakis, and Y. Wu. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783, 2018.

[34] I. Shopovska, L. Jovanov, and W. Philips. Deep visible and thermal image fusion for enhanced pedestrian visibility.

[35] T. Stafylakis and G. Tzimiropoulos. Combining residual networks with lstms for lipreading. *arXiv preprint arXiv:1703.04105, year = 2017,*.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE international conference on acoustics, speech and signal processing*, pages 4214–4217. IEEE, 2010.

[37] TCD-TIMIT corpus. Available at: https://sigmedia.tcd.ie/TCDTIMIT/.

[38] The GRID audiovisual sentence corpus . Available at: http://spandh.dcs.shef.ac.uk/gridcorpus/.

[39] VidTIMIT Audio-Video Dataset. Available at: http://conradsanderson.id.au/vidtimit/.

[40] M. Wand, J. Koutník, and J. Schmidhuber. Lipreading with long short-term memory. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6115–6119. IEEE, March 2016.

[41] S. Yang, Y. Zhang, D. Feng, M. Yang, C. Wang, J. Xiao, K. Long, S. Shan, and X. Chen. Lrw-1000: A naturally-distributed large-scale benchmark for lip reading in the wild. In *2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019)*, pages 1–8. IEEE, May 2019.

[42] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S.Z. Li. S3fd: Single shot scale-invariant face detector. In *Proceedings of the IEEE international conference on computer vision*, pages 192–201. IEEE, 2017.